Foundations in
Signal Processing, Communications and Networking 16
Series Editors: Holger Boche · Rudolf Mathar · Wolfgang Utschick

Rudolf Ahlswede
Lectures on Information Theory 6

Identification and Other Probabilistic Models

Alexander Ahlswede · Ingo Althöfer
Christian Deppe · Ulrich Tamm *Editors*

Springer

Foundations in Signal Processing, Communications and Networking

Volume 16

This book series presents monographs about fundamental topics and trends in signal processing, communications and networking in the field of information technology. The main focus of the series is to contribute on mathematical foundations and methodologies for the understanding, modeling and optimization of technical systems driven by information technology. Besides classical topics of signal processing, communications and networking the scope of this series includes many topics which are comparably related to information technology, network theory, and control. All monographs will share a rigorous mathematical approach to the addressed topics and an information technology related context.

** Indexing: The books of this series are indexed in Scopus and zbMATH **

More information about this series at http://www.springer.com/series/7603

Rudolf Ahlswede

Identification and Other Probabilistic Models

Rudolf Ahlswede's Lectures on Information Theory 6

Alexander Ahlswede • Ingo Althöfer • Christian Deppe • Ulrich Tamm
Editors

Springer

Author
Rudolf Ahlswede (deceased)
Bielefeld, Germany

Editors
Alexander Ahlswede
Bielefeld, Germany

Ingo Althöfer
Faculty Mathematics and Computer Science
Friedrich-Schiller-University Jena
Jena, Germany

Christian Deppe
Institute for Communications Engineering
Technical University of Munich
München, Germany

Ulrich Tamm
Fachbereich Wirtschaft und Gesundheit
Fachhochschule Bielefeld
Bielefeld, Germany

ISSN 1863-8538 ISSN 1863-8546 (electronic)
Foundations in Signal Processing, Communications and Networking
ISBN 978-3-030-65070-4 ISBN 978-3-030-65072-8 (eBook)
https://doi.org/10.1007/978-3-030-65072-8

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Words and Introduction of the Editors

Rudolf Ahlswede was one of the worldwide accepted experts on information theory. Many main developments in this area are due to him. Especially, he made big progress in multi-user theory. Furthermore, with identification theory and network coding, he introduced new research directions. Ahlswede died in December 2010.

Several highlights of Ahlswede's research are the Ahlswede-Daykin inequality and the Ahlswede-Khachatrian complete intersection theorem, which even include his name. He also described the capacity region of the multiple-access channel (many senders, one receiver). Together with Tom Cover's corresponding result for the broadcast channel (one sender, several receivers), this is the theoretical backbone for many algorithms in mobile communication, for instance, in the 5G standard. In 1990 jointly with his student Gunter Dueck, he initiated a whole new area of research—the theory of identification. Gunter Dueck in the supplement of this volume describes how things got started in this direction. Their paper found immediate interest. Shortly after its appearance, Ahlswede and Dueck received the Best Paper Award of the IEEE Information Theory Society. This is very much remarkable, especially, when taking into account that Ahlswede had received this award only two years before for his joint work with Imre Csiszar—it is quite extraordinary that the same author is honored with such an important award twice in such a short time.

This whole volume is devoted to identification and related concepts. In classical information theory, a sender transmits a message to a receiver over a noisy channel. The question that has to be answered at the receiving end hence is "Which message was sent?". Claude Shannon derived the famous channel capacity C: approximately 2^{nC} messages can be sent over the channel, such that the receiver can still reliably answer this question, where the message length n tends to infinity. Ahlswede and Dueck considered a new scenario, in which the receiver now has to answer the question: "Is this the message the one I am interested in?" This might be illustrated with the following example where a car owner (the sender) presses the button of his key and his car (the receiver) opens the door automatically. To obtain this result, a code is transmitted and the receiver is not really interested in the question which car should be opened but only if the car itself should be opened or not. In order

to reliably answer this question, two conditions have to be guaranteed: (1) The car of interest should open (with very high probability) and (2) no other car should open its door when receiving the transmitted signal. Ahlswede and Dueck found that also for this problem a capacity theorem exists. Approximately, $2^{2^{nC}}$ messages can be identified over the same noisy channel. Surprisingly, the number C, i.e., the identification capacity, is the same as Shannon's capacity for transmission. However, now, the expression is doubly exponential.

In the sequel, Ahlswede was working intensively on a general theory of information transfer that should include transmission and identification of information as special cases. To this aim, he was awarded a prestigious project in the center for interdisciplinary research (ZiF) in Bielefeld. Actually, this work occupied him for the rest of his life and was also the main reason for the delay of these lecture notes. He wanted to publish them when the general theory of information transfer was mature to some degree. For instance, his research led to the conjecture that the non-secure identification capacity (C_{ID}) might be the same as the common randomness capacity (C_{CR}) for channels without extra resources (like feedback). His student Christian Kleinewächter found a counterexample in which $C_{\mathrm{CR}} > C_{\mathrm{ID}}$. Ahlswede himself also showed that $C_{\mathrm{ID}} > C_{\mathrm{CR}}$ can hold (see 6). In his Shannon Lecture 2006 at the IEEE Symposium on Information Theory in Seattle, Ahlswede mentioned that this conjecture had helped him in the derivation of further capacity results.

As this example shows, the analysis of information identification led to many new concepts and problems. Source coding and data compression for identification are different from the corresponding concepts in information transmission. New probabilistic algorithms and the underlying randomness had to be studied. Further, there is a strong relation to hypothesis testing, when hypotheses have to be discriminated. All these directions are presented and studied in the corresponding chapters of this volume.

Chapter "Testing of Hypotheses and Identification" are lecture notes that were prepared by Marat Burnashev for a lecture he gave in Bielefeld in 2001. Ahlswede later used his notes for his lecture. We thank Marat Burnashev for allowing us to add his text in this book. Furthermore, we add Part VI to the book, which is a survey by Holger Boche, Christian Deppe, and Wafa Labidi of results in the theory of identification in the last 10 years.

Special thanks go to Wafa Labidi for the sixth volume. She has put a lot of work into creating index directories, proofreading, and rewriting. We also thank Gerhard Kramer for his support by financing Wafa Labidi. Finally, our thanks go to Bernhard Balkenhol who combines the first approximately 2000 pages of lecture scripts in different styles (AMS-TeX, LaTeX, etc.) to one big lecture script. He can be seen as one of the pioneers of Ahlswede's lecture notes.

Alexander Ahlswede, Ingo Althöfer, Christian Deppe, Ulrich Tamm

Preface

After an introduction to classical information theory, we present now primarily own methods and models, which go considerably beyond it. They were also sketched in our Shannon Lecture 2006. There are two main components: our combinatorial approach to information theory in the late seventies, where probabilistic source and channel models enter via the skeleton, a hypergraph based on typical sequences, and our theory of identification, which is now generalized to a general theory of information transfer (GTIT) incorporating also as ingredient a theory of common randomness, the main issue in cryptology. We begin with methods, at first with collections of basic covering, coloring, and packing lemmata with their proofs, which are based on counting or the probabilistic method of random choice.

Of course, these two methods are also closely related: the counting method can be viewed as the method of random choice for uniform probability distributions. It must be emphasized that there are cases where the probabilistic method fails, but the greedy algorithm (maximal coding) does not or both methods have to be used in combination. A striking example, Gallager's source coding problem, is discussed. Particularly useful is a special case of the covering lemma, called the link. It was used by Körner for zero-error problems, which are packing problems, in his solution of Rényi's problem. Very useful are also two methods, the elimination technique and the robustification technique, with applications for arbitrarily varying channel and unidirectional memories.

Coloring and covering lemmata find also applications in many lectures on combinatorial models of information processing: communication complexity, interactive communication, write-efficient memories, ALOHA. They are central in the theory of identification, especially in the quantum setting, in the theory of common randomness, and in the analysis of a complexity measure by Ahlswede, Khachatrian, Mauduit, and Sárkozy for number theoretical crypto-systems.

Bielefeld, Germany — Rudolf Ahlswede[1]

[1]This is the original preface written by Rudolf Ahlswede for the second 1.000 pages of his lectures. This volume consists of the last third of these pages.

Preamble

As long as algebra and geometry proceed along separate paths, their advance was slow and their applications limited. But when these sciences joined company, they drew from each other fresh vitality and hence forward marched on at a rapid pace towards perfection.

Joseph Louis Lagrange

Contents

Notation and Abbreviations

$< x^n \mid x >$	Number of occurrences of letter x in sequence x^n
$\| \cdot \|_1$	Statistical distance
1_A	Characteristic function of set A
A^c	Complement of set A
A-code	Average-list-size code
AVC	Arbitrarily-varying channel
AVWC	Arbitrarily-varying wiretap channel
C	Channel capacity
CC	Compound channel
CR	Common randomness
CWC	Compound Wiretap Channel
$D(X\|\|Y)$	I-divergence between X and Y
$D(X\|\|Y\|P)$	Conditional I-divergence between X and Y given P
DMC	Discrete memoryless channel
ED	Empirical distribution
$\mathbb{E}(X)$	Expectation of X
$H(X)$	Entropy of X
$H(X\|Y)$	Conditional entropy of X given Y
ID	Identification
IDF	Identification with feedback
IDf	Identification with passive feedback
$I(X \wedge Y)$	Mutual information between X and Y
$I(P, W)$	Mutual information between P and PW
$M(n, \lambda)$	Max. codesize for transmission codes
$\mu(A)$	Lebesgue measure of set A
NRA	Non-randomized (deterministic) average-list-size
NRI	Non-randomized (deterministic) identification
NRS	Non-randomized (deterministic) separation
$\mathbb{N}$	Natural numbers
$N(n, \lambda)$	Max. codesize for ID codes
PD	Probability distribution

$\mathcal{P}(A)$	The set of all probability distributions on the set A
$\mathbb{R}$, $\mathbb{R}^+$	Real and positive real numbers
RV	Random variable
SAVC	Strongly-arbitrarily-varying channel
SP	Separation
UCR	Uniform Common Randomness
V, W	Stochastic matrices
$W(\cdot\|i)$	$i-th$ row of W
WIDSI	Watermarking IDentification with side information at transmitter and receiver
WIDK	Watermarking IDentification with secure key
wtf	wiretap with feedback

Part I
Identification via Channels

Identification via Channels

Contrasting to Shannon's classical coding scheme for the transmission of a message over a noisy channel in the theory of identification the decoder is not really interested in what the received message is, but he only wants to decide whether a message, which is of special interest to him, had been sent or not. If the sender knows this certain message, this is a trivial problem. He just transmits one bit over the channel, namely "Yes, my message is the same as your message" or "No, it is not". However, if he does not know this message or if there are several receivers, each one interested in different messages, this is not possible. So there is need for a different method. There are also algorithmic problems where it is not necessary to calculate the solution, but only to check whether a certain given answer is correct. Depending on the problem, this answer might be much easier to give than finding the solution. "Easier" in this context means using less resources like channel usage, computing time or storage space.

The main result of Ahlswede and Dueck in [5] was, that in contrast to transmission problems, where the possible code sizes grow exponentially fast with blocklength, the size of identification codes will grow doubly exponentially fast. To become more specific, let $M(n, \lambda)$ and $N(n, \lambda)$ denote the maximal code sizes for transmission respectively randomised identification codes for a given discrete memoryless channel (abbreviated as DMC) W with blocklength n and error probability λ. It was proved by Shannon that

$$\lim_{n\to\infty} \frac{\log M(n, \lambda)}{n} = C \qquad \text{for all } \lambda \in (0, 1)\,,$$

where C is a constant which we denote as channel capacity. In the first chapter we will show for identification codes that

$$\lim_{n\to\infty} \frac{\log\log N(n, \lambda)}{n} = C \qquad \text{for all } \lambda \in (0, 1/2)\,.$$

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_1

The previous statements are valid for maximal error bounds. For average error bounds, the problem is much easier, one can use for example conventional checksum schemes. For transmission codes on a DMC, there is no difference between maximal and average error bounds, but for identification codes, the results are quite different, and the maximal error case needs quite sophisticated methods to be analysed.

There are even further qualitative differences between transmission and identification: Although feedback has no effect on the channel transmission capacity for the DMC, it was shown in [6], that it can increase the identification capacity. There are even cases where a noisier channel is the better one.

1 Results and Preliminaries

To put the coding theorems for identification into proper perspective, we describe first the analogous classical situation for transmission. This was discussed in detail in the first volume [7].

1.1 Notation and Known Facts

We use essentially the notations of [4]. Script capitals $\mathcal{X}, \mathcal{Y}, \ldots$ denote finite sets. The cardinality of a set $\mathcal{A}$ is denoted by $|\mathcal{A}|$. The letters P, Q always stand for probability distributions on finite sets. $X, Y, \ldots$ denote random variable (RV's). The functions "log" and "exp" are understood to be to the base 2. Let $x \in \mathbb{R}$ then we define the following functions $[x]^+ = \max\{x, 0\}$, $\lfloor x \rfloor = \max\{z \in \mathbb{Z} : z \leq x\}$ (floor function), and $\lceil x \rceil = \min\{z \in \mathbb{Z} : z \geq x\}$ (ceiling function).

1.1.1 Entropy and Information Quantities

Definition 1 For an arbitrary set $\mathcal{S}$ we will denote by $\mathcal{P}(\mathcal{S})$ the set of all probability distributions (PD's) on $\mathcal{S}$.

Let X be a RV with values in $\mathcal{X}$ and distribution $P \in \mathcal{P}(X)$. Let Y be a RV with values in $\mathcal{Y}$ such that the joint distribution of (X, Y) on $\mathcal{X} \times \mathcal{Y}$ is given by

$$\Pr(X = x, Y = y) = P(x) \cdot V(y|x), \qquad V \in \mathcal{W},$$

where $\mathcal{W}$ is the set of stochastic matrices, namely $\mathcal{W}$ denotes the set of all channels V with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$.

Definition 2 (Entropy) Let P be a probability distribution on a finite set $\mathcal{X}$. The entropy of P is defined as

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

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

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

Definition 3 (Conditional Entropy) Let X, Y be RVs on finite sets $\mathcal{X}, \mathcal{Y}$ with distributions P and Q respectively. The conditional entropy of Y given X is defined by

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

Definition 4 (Mutual Information) Let X and Y be random variables on finite sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Then we define the mutual information between X and Y by

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

Definition 5 (Discrete Memoryless Channel) A discrete memoryless channel (DMC) is a triple $(\mathcal{X}, \mathcal{Y}, W)$, where $\mathcal{X}$ and $\mathcal{Y}$ are finite sets denoted as input- respectively output alphabet, and $W = \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ is a stochastic matrix. The probability for a sequence $y^n = (y_1, \ldots, y_n) \in \mathcal{Y}^n$ to be received if $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$ was sent is defined by

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

If it is clear which alphabets are to be used, we will omit them if we are talking about the channel. If P is a probability distribution on $\mathcal{X}$ and $W = \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ a stochastic matrix, we define

$$I(P, W) \triangleq I(X\wedge Y),$$

where X is a RV with distribution P and Y has conditional distribution $W(\cdot|x)$ given $X = x$. Informally, one could also write

$$I(P, W) \triangleq I(P\wedge PW),$$

where for $P \in \mathcal{P}$, $V \in \mathcal{W}$ we write PV for the PD on $\mathcal{Y}$ given by

$$PV(y) = \sum_x P(x)V(y|x), \qquad y \in \mathcal{Y}.$$

For $P, \tilde{P} \in \mathcal{P}$

$$D(\tilde{P}||P) = \sum_x \tilde{P}(x) \log \frac{\tilde{P}(x)}{P(x)}$$

denotes the relative entropy and for $V, \tilde{V} \in \mathcal{W}$ the quantity

$$D(\tilde{V}||V|P) = \sum_x P(x) D(\tilde{V}(\cdot|x)||V(\cdot|x))$$

stands for the conditional relative entropy.

1.1.2 Channels, Empirical Distributions, Generated Sequences

For a stochastic $|\mathcal{X}| \times |\mathcal{Y}|$-matrix W we have already defined the transmission probabilities W^n of a DMC, and we have also introduced $\mathcal{P}(\mathcal{X}^n)$ as the set of PD's on $\mathcal{X}^n$. We abbreviate $\mathcal{P}(\mathcal{X})$ as $\mathcal{P}$.

For $x^n \in \mathcal{X}^n$ one can count the relative frequency of letters in x^n and gets an n-type P on $\mathcal{X}$:

Definition 6 Define for $x \in \mathcal{X}$ and $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$

$$< x^n | x >= |\{i \in \{1, \ldots, n\} | x_i = x\}|$$

Then the PD P on $\mathcal{X}$ defined by

$$P_{x^n}(x) = \frac{< x^n | x >}{n}$$

is an n-type and we will call x^n to be P-typical, while P is called the type of x^n.

Definition 7 For positive integers n we set

$$\mathcal{P}_n = \{P \in \mathcal{P} | P(x) \in \{0, 1/n, 2/n, \ldots, 1\} \text{ for all } x \in \mathcal{X}\}.$$

For any $P \in \mathcal{P}_n$, called empirical distribution (ED) or n-type, we define the set

$$\mathcal{W}_n(P) = \left\{ V \in \mathcal{W} \middle| V(y|x) \in \left\{ 0, \frac{1}{nP(x)}, \frac{2}{nP(x)}, \ldots, 1 \right\}, x \in \mathcal{X}, y \in \mathcal{Y} \right\}.$$

P_{x^n} is a member of $\mathcal{P}_n$ by definition. P_{x^n} is called ED of x^n.

Similarly, we define the ED $P_{x^n y^n}$ for pairs $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$.

Definition 8 For $P \in \mathcal{P}$ the set $\mathcal{T}_P^n$ of all P-typical sequences in $\mathcal{X}^n$ is given by

$$\mathcal{T}_P^n = \{x^n | P_{x^n} = P\}.$$

For $V \in \mathcal{W}$, a sequence $y^n \in \mathcal{Y}^n$ is said to be V-generated by x^n if, for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$,

$$P_{x^n y^n}(x, y) = P_{x^n}(x) \cdot V(y|x).$$

The set of those sequences is denoted by $\mathcal{T}_V^n(x^n)$.

Notice that $\mathcal{T}_P^n \neq \emptyset$ if and only if $P \in \mathcal{P}_n$ and $\mathcal{T}_V^n(x^n) \neq \emptyset$ if and only if $V \in \mathcal{W}_n(P_{x^n})$. $\mathcal{T}_{PV}^n$ is the set of PV-typical sequences in $\mathcal{Y}^n$.

1.1.3 Elementary Properties of Typical Sequences and Generated Sequences

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|} \tag{1}$$

$$|\mathcal{W}_n(P)| \leq (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|} \qquad \forall\ P \in \mathcal{P}_n \tag{2}$$

$$|\mathcal{T}_P^n| \leq \exp\{nH(P)\}$$

$$|\mathcal{T}_P^n| \geq (n+1)^{-|\mathcal{X}|} \cdot \exp\{nH(P)\} \quad \forall\ P \in \mathcal{P}_n \tag{3}$$

$$|\mathcal{T}_V^n(x^n)| \leq \exp\{nH(V|P)\}$$

$$|\mathcal{T}_V^n(x^n)| \geq (n+1)^{-|\mathcal{X}|\cdot|\mathcal{Y}|} \cdot \exp\{nH(V|P)\}$$
$$\forall\ P \in \mathcal{P}_n, V \in \mathcal{W}_n(P), x^n \in \mathcal{T}_P^n \tag{4}$$

$$W^n(y^n|x^n) = \exp\{-n(D(V||W|P) + H(V|P))\}$$
$$\forall\ P \in \mathcal{P}_n, V \in \mathcal{W}_n(P), x^n \in \mathcal{T}_P^n, y^n \in \mathcal{T}_V^n(x^n) \tag{5}$$

$$W^n(\mathcal{T}_V^n(x^n)|x^n) \leq \exp\{-nD(V||W|P)\}$$

$$W^n(\mathcal{T}_V^n(x^n)|x^n) \geq (n+1)^{-|\mathcal{X}|\cdot|\mathcal{Y}|} \cdot \exp\{-nD(V||W|P)\}$$
$$\forall\ P \in \mathcal{P}_n, V \in \mathcal{W}_n(P), x^n \in \mathcal{T}_P^n. \tag{6}$$

1.1.4 Formulation of the Classical Transmission Problem

Definition 9 An (n, M, λ) code for W is a set of pairs $\{(u_i, \mathcal{D}_i) : i = 1, \ldots, M\}$ with the properties

$$u_i \in \mathcal{X}^n, \quad \mathcal{D}_i \subset \mathcal{Y}^n, \qquad \text{for all } i \in \{1, \ldots, M\} \tag{7}$$

$$\mathcal{D}_i \cap \mathcal{D}_j = \emptyset, \qquad \text{for all } i, j \in \{1, \ldots, M\} \text{ with } i \neq j \tag{8}$$

$$W^n(\mathcal{D}_i|u_i) \geq 1 - \lambda, \qquad \text{for all } i \in \{1, \ldots, M\}. \tag{9}$$

Let $M(n, \lambda)$ be the maximal integer M for which an (n, M, λ) code exists.

Theorem 10 (Shannon [14])

$$\lim_{n\to\infty} \frac{1}{n} \log M(n, \lambda) = C \qquad \textit{for all } \lambda \in (0, 1)$$

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

Actually, Shannon proved in [14] only the direct part of the theorem. The so-called strong converse was proved by Wolfowitz (see [15]).

1.2 Formulation of the Identification Problem

Definition 11 A (randomized) identification (ID) code $(n, N, \lambda_1, \lambda_2)$ is a family

$$\{(Q(\cdot|i), \mathcal{D}_i)|i = 1, \ldots, N\}$$

of pairs with

$$Q(\cdot|i) \in \mathcal{P}(\mathcal{X}^n), \quad \mathcal{D}_i \subset \mathcal{Y}^n, \qquad \text{for all } i = 1, \ldots, N \tag{10}$$

and with errors of the first (resp. second) kind satisfying

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|i) W^n(\mathcal{D}_i^c|x^n) \leq \lambda_1 \tag{11}$$

and

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|j) W^n(\mathcal{D}_i|x^n) \leq \lambda_2 \tag{12}$$

for all $i = 1, \ldots, N$, $j = 1, \ldots, N$ with $j \neq i$.

Of course, we also could have defined deterministic ID codes where the $Q(\cdot|i)$ denote point masses on points $u_i \in \mathcal{X}^n$. However, the study of deterministic ID codes leads only to very poor results (see 4). Therefore, we consider only the much more powerful randomized ID codes.

The essential difference between ID codes and classical transmission codes is that no disjointness condition is imposed on the decoding sets $\mathcal{D}_i$. In an ID code, the decoding sets have to be only pairwise significantly different in the sense defined in (11) and (12).

We now explain how ID codes arise naturally as the appropriate code concept in an identification problem. Assume there is a set $\mathcal{E} = \{e_1, \ldots, e_N\}$ of events (or objects), any one of which may occur. The event is known to the sender of the channel, but unknown to the receiver. On the receiver's side is a set of persons (or devices) $\mathcal{F} = \{F_1, \ldots, F_N\}$ observing the output of the channel. *Person F_i wants to know whether or not event e_i occurred*. The sender can transmit his knowledge of the event over the channel. For this transmission procedure, randomization is allowed, that is, an encoding rule for an event e_i is formally described by a probability distribution $Q(\cdot|i)$ out of $\mathcal{P}(\mathcal{X}^n)$. Clearly, F_i can choose a decision rule specifying sequences y^n for which s/he assumes that e_i has occurred. This rule is represented by the decoding set $\mathcal{D}_i \subset \mathcal{Y}^n$. Thus one is led to the notion of an ID code as described above. Randomized decision rules on the receiver's side are not considered because they yield only minor improvements in the present coding problem.

The identification problem can also be stated in the following way. Instead of N persons, we can assume that the receiver wants to know whether or not e_j occurred. The parameter j is not known to the sender; that is, the sender does not know what the receiver wants to identify.

At the end of the lecture we give examples for which the present model is suitable and discuss its relation to identification problems found in the literature [12, 16].

1.2.1 The Double Exponent Coding Theorem

Let $N(n, \lambda)$ be the maximal number N such that an $(n, N, \lambda_1, \lambda_2)$ ID code with $\lambda_1, \lambda_2 \leq \lambda$ exists, and let C be Shannon's *transmission* capacity of the DMC W.

Theorem 12 (Coding Theorem and Strong Converse)

(i) $\liminf_{n\to\infty} \frac{1}{n} \log\log N(n, \lambda) \geq C, \quad$ *for all* $\lambda \in (0, 1]$.

(ii) $\limsup_{n\to\infty} \frac{1}{n} \log\log N(n, \lambda) \leq C, \quad$ *for all* $\lambda \in (0, \frac{1}{2})$.

Note that (ii) is not true for $\lambda > 1/2$.

Originally, instead of (ii) it was proved only that

$$\limsup_{n\to\infty} \frac{1}{n} \log\log N(n, 2^{-n\epsilon}) \leq C, \qquad \text{for all } \epsilon > 0.$$

This statement was called "soft" converse, because the error probability on the left side is exponentially small. In the usual terminology a "weak" converse would mean

$$\inf_{\lambda\in(0,1)} \limsup_{n\to\infty} \frac{1}{n} \log\log N(n,\lambda) \le C.$$

We derive better estimates than those stated in Theorem 12. For their description we use the notions of mutual information and relative entropy, and some notation from [4] and [9]. In those sharper estimates, we are concerned with *error exponents*, which can be achieved with a certain (second-order) rate.

The triple (R, E_1, E_2) is called achievable if, for all $\delta > 0$ and $n \ge n(\delta, |\mathcal{X}|, |\mathcal{Y}|)$, and ID code exists for N messages and error probabilities $\lambda_1(n)$, $\lambda_2(n)$ such that

$$\frac{1}{n}\log\log N \ge R - \delta, \quad \lambda_i \le \exp\{-n(E_i - \delta)\}, \qquad i = 1, 2. \tag{13}$$

For achievable triples we have the following result.

Theorem 13

(i) If $P \in \mathcal{P}(\mathcal{X})$ satisfies $I(P, W) > R + 2E_2$, then

$$\left(R, \min_{I(P,V)\le R+2E_2} D(V||W|P), E_2\right)$$

is achievable.

(ii) If $E_1 > 0$ and $R + 2E_2 > C$, then (R, E_1, E_2) is not achievable.

Some remarks are due:

1. Theorem 13(i) clearly implies Theorem 12(i).
2. Theorem 13(ii) implies formally, that

$$\liminf_{n\to\infty} \frac{1}{n} \log\log N(n, 2^{-n\epsilon}) \le C.$$

   From the proof of Theorem 13(ii), however, it will become clear, that the same is true for the limit superior.
3. Since $D(V||W|P)$ is a continuous function in (V, W) with the property $D(V||W|P) = 0$ if and only if $V = W$ almost everywhere we see that the condition $I(P, W) > R + 2E_2$ in Theorem 13(i) implies that

$$\min_{I(P,V)\le R+E_2} D(V||W|P) > 0.$$

4. Theorem 13 completely characterizes the set of achievable pairs (R, E_2) in the limit $E_1 \to 0$. More precisely,

$$\lim_{E_1 \to 0} \{(R, E_2) : (R, E_1, E_2) \text{ is achievable}\} = \{(R, E_2) : R \leq C - 2E_2\}.$$

In the remainder of this subsection we prepare the reader for the results of Theorem 12 and its proof. The fact that the maximal code length grows doubly exponentially can more easily be understood for the very special case of a noiseless binary channel. We include a complete proof. We then comment on our proof for the direct part of Theorem 12 for the general DMC, and, finally, on the proof of the converse part.

We start with the construction of n-block ID codes for the binary channel W given by the input alphabet $\mathcal{X} := \{0, 1\}$, $\mathcal{Y} := \{0, 1\}$, and $W(1|1) = W(0, 0) = 1$. We use the standard maximal coding argument.

Let n be the block length, and let $\lambda \in (0, 1/2)$ be given. Let 2^l be the smallest power of 2, such that

$$\lambda \cdot \log(2^l - 1) > 1 \qquad \text{and } 2^l > 6.$$

Suppose that n is large compared with 2^l. Set

$$M := 2^{n-l}.$$

We define an n-block ID code

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

such that $\log \log N$ is close to $n \log 2$. We restrict our attention to distributions $Q(\cdot|i)$ which are equidistributions on sets $\mathcal{A}_i \subset \mathcal{X}^n$ with cardinality M. Since M equals 2^{n-l}, we therefore consider only equidistributions on relatively large subsets of $\mathcal{X}^n$. Suppose now we have found subsets $\mathcal{A}_1, \ldots, \mathcal{A}_N \subset \mathcal{X}^n$, all of which have cardinality M and such that

$$|\mathcal{A}_i \cap \mathcal{A}_j| < \lambda \cdot M, \qquad \text{for all } i, j \in \{1, \ldots, N\}, \quad i \neq j. \tag{14}$$

Then we define

$$Q(\cdot|i) := \text{ equidistribution on } \mathcal{A}_i$$

$$\mathcal{D}_i := \mathcal{A}_i, \qquad \text{for all } i = 1, \ldots, N.$$

Consider $\{(Q(\cdot|i), \mathcal{D}_i) : i = 1, \ldots, N\}$. We claim that this system is an $(n, N, 0, \lambda)$ ID code. This is true because

$$\sum_{x^n} Q(x^n|i) W^n(\mathcal{D}_i|x^n) = 1$$

$$\sum_{x^n} Q(x^n|j) W^n(\mathcal{D}_i|x^n) = M^{-1}|\mathcal{A}_i \cap \mathcal{A}_j| < \lambda$$

for $i, j \in \{1, \ldots, N\}$, $i \neq j$. Here we used the special nature of W and assumption (14).

We have seen now that it suffices to show the existence of a large family $\mathcal{A}_1, \ldots, \mathcal{A}_N$ of sets of cardinality M which satisfies (14).

Proposition 14 *Let $\mathcal{Z}$ be a finite set and let $\lambda \in (0, 1/2)$ be given. If ϵ is so small that*

$$\lambda \log\left(\frac{1}{\epsilon} - 1\right) > 2 \qquad \frac{1}{\epsilon} > 6.$$

then a family $\mathcal{A}_1, \ldots, \mathcal{A}_N$ of subsets of $\mathcal{Z}$ exists satisfying

$$|\mathcal{A}_i| = \lfloor \epsilon |Z| \rfloor, \qquad \text{for all } i \in \{1, \ldots, N\},$$

$$|\mathcal{A}_i \cap \mathcal{A}_j| < \lambda \lfloor \epsilon \cdot |\mathcal{Z}| \rfloor, \qquad \text{for all } i, j \in \{1, \ldots, N\}, \quad i \neq j$$

and

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

Proof Choose as a starting point an arbitrary $\mathcal{A}_1 \subset \mathcal{Z}$, $|\mathcal{A}_1| = \lfloor \epsilon \cdot |\mathcal{Z}| \rfloor$. We count how many sets $\mathcal{A} \subset \mathcal{Z}$ exist with cardinality $\lfloor \epsilon \cdot |\mathcal{Z}| \rfloor$ and

$$|\mathcal{A}_1 \cap \mathcal{A}| \geq \lambda \lfloor \epsilon |\mathcal{Z}| \rfloor.$$

We define $M' := \lfloor \epsilon |\mathcal{Z}| \rfloor$. The number of those sets $\mathcal{A}$ in question is then

$$\sum_{i=\lceil \lambda \cdot M' \rceil}^{M'} \binom{|\mathcal{Z}| - M'}{M' - i} \binom{M'}{i}. \tag{15}$$

For $\lambda < 1/2$ and $1/\epsilon > 6$ the first summand in the sum is the maximal one. This is easy to establish. Therefore, the sum in (15) can be upper-bounded by

$$M' \cdot \binom{|\mathcal{Z}| - M'}{M' - \lceil \lambda M \rceil} \binom{M'}{\lceil \lambda M' \rceil} \leq M' \cdot \binom{|\mathcal{Z}|}{M' - \lceil \lambda M' \rceil} \cdot 2^{M'} =: T. \tag{16}$$

Hence at most T sets $\mathcal{A}$ of cardinality M' exist such that

$$|\mathcal{A}_1 \cap \mathcal{A}| \geq \lambda \cdot M'.$$

There are $\binom{|\mathcal{Z}|}{M'}$ sets of cardinality M'. If $T < \binom{|\mathcal{Z}|}{M'}$, then an $\mathcal{A}_2 \subset \mathcal{Z}$ exists with $|\mathcal{A}_2| = M'$ and $|\mathcal{A}_1 \cap \mathcal{A}_2| < \lambda \cdot M'$. Furthermore, if $2T < \binom{|\mathcal{Z}|}{M'}$, then $\mathcal{A}_3 \subset \mathcal{Z}$ exists with $|\mathcal{A}_3| = M'$, $|\mathcal{A}_3 \cap \mathcal{A}_1| < \lambda \cdot M'$, and $|\mathcal{A}_3 \cap \mathcal{A}_2| < \lambda M'$. By repeatedly using this argument, we get the following result.

There are N sets $\mathcal{A}_1, \ldots, \mathcal{A}_N \subset \mathcal{Z}$ of cardinality M' such that $|\mathcal{A}_i \cap \mathcal{A}_j| < \lambda \cdot M'$ for every $i \neq j$, if

$$N \cdot T < \binom{|\mathcal{Z}|}{M'}.$$

Hence a family of sets $\mathcal{A}_1, \ldots, \mathcal{A}_N$ exists with

$$N := \left\lfloor \binom{|\mathcal{Z}|}{M'} \cdot T^{-1} \right\rfloor - 1. \tag{17}$$

Recall that T was defined in (16). It is now easy to lower-bound N. By (16) and (17),

$$N \geq 2^{-M'} \cdot M'^{-1} \cdot \prod_{i=1}^{\lceil \lambda M' \rceil} \frac{|\mathcal{Z}| - M' + i}{M' - \lceil \lambda M' \rceil + i} - 1.$$

Since $M' = \lfloor \epsilon |\mathcal{Z}| \rfloor$ and $\lambda \leq 1/2$, for $i \in \{1, \ldots, \lceil \lambda M' \rceil\}$

$$\frac{|\mathcal{Z}| - M' + i}{M' - \lceil \lambda M' \rceil + i} \geq \frac{1}{\epsilon} - 1.$$

Hence

$$\begin{aligned} N + 1 &\geq 2^{-M'} \cdot M'^{-1} \cdot \left(\frac{1}{\epsilon} - 1 \right)^{\lceil \lambda \cdot M' \rceil} \\ &\geq 2^{-M'} \cdot \left(\frac{1}{\epsilon} - 1 \right)^{\lambda \cdot M'} \cdot M'^{-1} \\ &= 2^{M'(\lambda \log((1/\epsilon)-1)-1)} \cdot M'^{-1} \\ &\geq 2^{M'} \cdot |\mathcal{Z}|^{-1} = 2^{\lfloor \epsilon \cdot |\mathcal{Z}| \rfloor} \cdot |\mathcal{Z}|^{-1}, \end{aligned}$$

where we have used the assumption in the proposition. The proof is complete. □

We return to the binary noiseless channel. We apply the result of Proposition 14 to $\{0, 1\}^n$ instead of $\mathcal{Z}$ and with 2^{-l} instead of ϵ. We conclude that there are at least

$$N := 2^{-n} \cdot 2^{2^{n-l}} - 1$$

sets $\mathcal{A}_1, \ldots, \mathcal{A}_N$ of $\{0, 1\}^n$ with cardinality 2^{n-l} such that

$$|\mathcal{A}_i \cap \mathcal{A}_j| < \lambda \cdot 2^{n-l} \qquad \text{for all } i \neq j.$$

In other words, we have found an $(n, N, 0, \lambda)$ identification code. Clearly, $(1/n) \log \log N$ is arbitrarily close to $\log 2$, the capacity of the binary noiseless channel, if n grows to infinity.

Thus Theorem 12(i) is proved for the noiseless binary channel. The validity of Theorem 12(ii) is easy to see for this channel. Obviously, the number of messages in an identification code cannot exceed the number of possible decoding sets, because all the decoding sets have to be different.

Since all the decoding sets of n-block length are subsets of $\{0, 1\}^n$, there are at most 2^{2^n} decoding sets in an identification code. We shall see that the construction of Proposition 14 can be used to construct good ID codes with the help of an underlying classical transmission code with rate close to capacity.

The original converse part, however, is highly complicated. One has to show that there is no advantage in considering equidistributions on subsets with a cardinality larger than $\exp\{nC\}$.

We originally wanted to show that, starting with a given code with equidistributions on large sets, we can find "smaller" equidistributions on sets with cardinality smaller than $\exp\{nC\}$ such that the resulting decoding sets in the new code are nearly the same as in the originally given code. Then one could conclude that any ID code could have at most as many messages as subsets of $\mathcal{X}^n$ with cardinality smaller than $\exp\{nC\}$, and the proof would be complete.

Unfortunately, we were not able to prove the converse part in this elegant version; we were, however, able to prove it following this basic idea.

2 The Direct Parts of the Coding Theorems

Proof of the Direct Part: Theorem 12(i)
We simply apply Proposition 14 for a classical transmission code for W. Let $\lambda \in (0, 1)$ be given, and let $\epsilon > 0$ be so small that

$$\lambda \log \left(\frac{1}{\epsilon} - 1 \right) > 2, \qquad \frac{1}{\epsilon} > 6.$$

By Theorem 10 there exists for any large n an n-length block code,

$$\mathcal{C} = \{(u_i, \mathcal{E}_i) : i = 1, \ldots, M\}$$

with maximal error bounded by λ and

$$M \geq 2^{n(C-\epsilon)}.$$

Let $\mathcal{Z} := \{u_i, \ldots, u_n\}$. By Proposition 14 a family of subsets $\mathcal{A}_1, \ldots, \mathcal{A}_N$ of $\mathcal{Z}$ exists satisfying

$$|\mathcal{A}_i| = \lfloor \epsilon |\mathcal{Z}| \rfloor = \lfloor \epsilon \cdot M \rfloor \qquad \text{for all } i \in \{1, \ldots, N\} \tag{18}$$

$$|\mathcal{A}_i \cap \mathcal{A}_j| < \lambda \lfloor \epsilon \cdot |\mathcal{Z}| \rfloor \qquad \text{for all } i, j \in \{1, \ldots, N\}, i \neq j, \tag{19}$$

$$N \geq |\mathcal{Z}|^{-1} \cdot 2^{\lfloor \epsilon |\mathcal{Z}| \rfloor}. \tag{20}$$

From the sets $\mathcal{A}_1, \ldots, \mathcal{A}_N$ we construct an ID code in the following simple manner. Define for $i \in \{1, \ldots, N\}$

$$Q(\cdot|i) := \text{ equidistribution on } \mathcal{A}_i$$

and

$$\mathcal{D}_i := \bigcup_{u_k \in \mathcal{A}_i} \mathcal{E}_k.$$

Define the ID code

$$\{(Q(\cdot|i), \mathcal{D}_i) : i = 1, \ldots, N\}.$$

We look at the errors of first and second kind of this ID code (recall (11) and (12)). Let $i \in \{1, \ldots, N\}$, and let $u_k \in \mathcal{A}_i$. Then

$$W^n(\mathcal{D}_i^c|u_k) \leq W^n(\mathcal{E}_k^c|u_k) \leq \lambda$$

because of $\mathcal{E}_k \subset \mathcal{D}_i$ and because $\mathcal{E}_k$ is the decoding set for u_k in the transmission code $\mathcal{C}$. Hence

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|i) W^n(\mathcal{D}_i^c|x^n) = \sum_{u_k \in \mathcal{A}_k} \frac{1}{|\mathcal{A}_k|} \cdot W^n(\mathcal{D}_i^c|u_k) \leq \lambda.$$

On the other hand, for each $j \in \{1, \ldots, N\}$, $j \neq i$,

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|j) W^n(\mathcal{D}_i|x^n) = \sum_{u_l \in \mathcal{A}_j} \frac{1}{|\mathcal{A}_j|} \cdot W^n(\mathcal{D}_i|u_l)$$

$$= \frac{1}{|\mathcal{A}_j|} \left(\sum_{u_l \in \mathcal{A}_j \cap \mathcal{A}_i} W^n(\mathcal{D}_i|u_l) + \sum_{u_l \notin \mathcal{A}_j \cap \mathcal{A}_i} W^n(\mathcal{D}_i|u_l) \right)$$

$$\leq \frac{1}{|\mathcal{A}_j|} \cdot \left(|\mathcal{A}_j \cap \mathcal{A}_i| + \sum_{u_l \notin \mathcal{A}_j \cap \mathcal{A}_i} W^n(\mathcal{D}_i|u_l) \right).$$

If $u_l \in \mathcal{A}_i^c$, then $\mathcal{E}_l \cap \mathcal{D}_i = \emptyset$. Hence for such u_l the relation $\mathcal{D}_i \subset \mathcal{E}_l^c$ holds. This observation together with (18) and (19) yields

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|j) W^n(\mathcal{D}_i|x^n) \leq 2\lambda.$$

Equation (20) finally gives

$$N \geq M^{-1} \cdot 2^{\lfloor \epsilon |\mathcal{Z}| \rfloor} \geq |\mathcal{X}|^{-n} \cdot 2^{\lfloor \epsilon 2^{n(C-\epsilon)} \rfloor}.$$

In summary, $\{(Q(\cdot|i), \mathcal{D}_i) : i = 1, \ldots, N\}$ is an $(n, N, \lambda, 2\lambda)$ ID code with $(1/n) \log \log N$ close to $C - \epsilon$. Since λ and ϵ could be chosen arbitrarily small, Theorem 12 is proved.

Remark Observe that for the proof we needed only Proposition 14 (which is just Gilbert's bound for constant weight sequences) and a given code for the channel W. Thus we can conclude that Theorem 12(i) holds in fact for all channels having a capacity. It is not necessary to assume that W is discrete or memoryless.

Proof of the Direct Part: Theorem 13(i)
Of course, one could easily derive exponential error bounds with the construction in the preceding subsection. The difference would be instead of a code with maximal error λ one would start with a code having exponentially small error probability. Furthermore, one would choose $2^{-n\xi}$ instead of ϵ.

However, Theorem 13(i) gives a stronger result than the one obtainable by this simple method. Theorem 13(i) gives, in the sense expressed in the fourth remark following Theorem 13, a best possible error exponent. The principal idea is random selection of ID codes rather than a maximal coding idea which led to Proposition 14. The key step is the application of Proposition 15 which we will present soon. Its proof is rather technical, so we give here only the short proof of Theorem 13(i) assuming Proposition 15 below holds. The proof of Proposition 15 can be found in the following subsection.

Let $P \in \mathcal{P}_n$. We consider here only ID codes of a special structure. Every message i is encoded by the uniform distribution on a family $\mathcal{U}_i$ of members of $\mathcal{T}_P^n$ satisfying $|\mathcal{U}_i| = M$ for all $i = 1, \ldots, N = \lceil 2^{2^{nR}} \rceil$.

Let (R, E_2) be given. We assign to $\mathcal{U}_i$ a decoding set $\mathcal{D}_i = \mathcal{D}(\mathcal{U}_i)$ defined by

$$\mathcal{D}(\mathcal{U}_i) = \bigcup_{u \in \mathcal{U}_i} \mathcal{F}_u \tag{21}$$

where

$$\mathcal{F}_u = \bigcup_{V: I(P,V) > R+2E_2} \mathcal{T}_V^n(u). \tag{22}$$

First notice that

$$\begin{aligned} W^n(\mathcal{F}_u^c|u) &\leq \sum_{V: I(P,V) \leq R+2E_2} W^n(\mathcal{T}_V^n(u)|u) \\ &\leq \sum_{V: I(P,V) \leq R+2E_2} \exp\{-n D(V||W|P)\} \\ &\leq (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|} \cdot \exp\left\{-n \min_{V: I(P,V) \leq R+2E_2} D(V||W|P)\right\} \end{aligned}$$

by (6) and (2) and thus

$$\frac{1}{M} \sum_{u \in \mathcal{U}_i} W^n(\mathcal{D}_i^c|u) \leq (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|} \cdot \exp\left\{-n \min_{V: I(P,V) \leq R+2E_2} D(V||W|P)\right\}.$$

This means that regardless of the choice of $\mathcal{U}_i$ the error exponent of the first kind

$$E_1 := \min_{V: I(P,V) \leq R+2E_2} D(V||W|P)$$

is achievable. We now specify the sets $\mathcal{U}_i$ to achieve (R, E_2). We choose

$$M = \lfloor \exp\{n(R + E_2)\} \rfloor \tag{23}$$

and define the $\mathcal{U}_i$ by random selection as follows.

Let U_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, M$ be independent RV's, all uniformly distributed over $\mathcal{T}_P^n$. Define the random families

$$\overline{\mathcal{U}}_i = \{U_{i1}, \ldots, U_{iM}\}, \qquad i = 1, \ldots, N.$$

Every realization of the U_{ij} gives rise to an ID code as described before.

We want to show that a large fraction of these randomly selected ID codes has an error exponent of the second kind at most $E_2 - \delta$ ($\delta > 0$ arbitrarily small, n

sufficiently large for δ). In fact, we can get this result (and therefore Theorem 13(i)) by the following result for two messages.

Let $\mathcal{U}_1$ be any subset of $\mathcal{T}_P^n$ of cardinality M and let $\overline{\mathcal{U}}_2$ be as described before. We consider the probability P^* that for $\gamma > 0$:

$$\frac{1}{M} \sum_{u \in \mathcal{U}_1} W^n(\mathcal{D}(\overline{\mathcal{U}}_2)|u) \leq \exp\{-n(E_2 - 3\gamma)\} \tag{24}$$

and

$$\frac{1}{M} \sum_{u \in \overline{\mathcal{U}}_2} W^n(\mathcal{D}(\mathcal{U}_1)|u) \leq \exp\{-n(E_2 - 3\gamma)\}. \tag{25}$$

Then we have the following proposition.

Proposition 15 *For any $\gamma > 0$ and $n \geq n(\gamma)$,*

$$P^* \geq 1 - (n+1)\exp\{-(n\gamma - 2) \cdot \exp\{nR\}\}.$$

A proof is given below. We show here that Theorem 13(i) follows from Proposition 15. Imagine that the random selection is performed iteratively and that the realizations $\mathcal{U}_1, \ldots, \mathcal{U}_t$ have the desired error performances.

Then, by Proposition 15, with the choice $\gamma = \delta/3$, $(\mathcal{U}_1, \ldots, \mathcal{U}_t, \overline{\mathcal{U}}_{t+1})$ has the desired error performances with probability exceeding $1 - t(1 - P^*)$. Therefore, there exists a sufficiently good realization $\mathcal{U}_{t+1}$ of $\overline{\mathcal{U}}_{t+1}$ if

$$1 - t(1 - P^*) > 0.$$

Since by Proposition 15 even $1 - N(1 - P^*) > 0$ for large n, it is possible to find a realization $(\mathcal{U}_1, \ldots, \mathcal{U}_N)$ of $(\overline{\mathcal{U}}_1, \ldots, \overline{\mathcal{U}}_N)$ with the desired error performances.

The proof of Proposition 15 follows the large deviations approach of [2] in the improved form of [4]. Specifically, we shall need a very useful lemma on large deviations which we state now.

Lemma 16 (Generalized Chebyshev Inequality, Bernstein's Trick) *Let $\Psi_1, \ldots, \Psi_M$ be i.i.d. RV's with values in $\{0, 1\}$. Suppose that the expectation $\mathbb{E}\Psi_1$ of Ψ_1 satisfies $\mathbb{E}\Psi_1 \leq \mu < \lambda \leq 1$; then*

$$\Pr\left(\sum_{j=1}^{M} \Psi_j > M \cdot \lambda\right) \leq \exp\{-M \cdot D(\lambda||\mu)\},$$

where $D(\lambda||\mu)$ denotes the relative entropy between $(\lambda, 1-\lambda)$ and $(\mu, 1-\mu)$.

Proof of Proposition 15
The following lemma is Lemma 1, p.433, in [4].

Lemma 17 *Let U be uniformly distributed on $\mathcal{T}_P^n$, $P \in \mathcal{P}_n$. Let $\mathcal{U}$ be a subset of $\mathcal{T}_P^n$, $|\mathcal{U}| = \lfloor \exp\{nR\} \rfloor$. Define for any $V, V' \in \mathcal{W}$ and $u^* \in \mathcal{X}^n$:*

$$g_{V,V'}(u^*) = \left| \bigcup_{u \in \mathcal{U}} \mathcal{T}_{V'}^n(u) \cap \mathcal{T}_V^n(u^*) \right|. \tag{26}$$

Then

(i) $\mathbb{E} g_{V,V'}(U) \le (n+1)^{|\mathcal{X}|} \cdot \exp\{n(H(V|P) - [I(P,V') - R]^+)\}$;
(ii) for all $\eta > 0$, $\xi \ge 0$, and $n \ge n(\eta, |\mathcal{X}|, |\mathcal{Y}|)$,

$$\Pr\left(g_{V,V'}(U) \ge \exp\left\{n\left(H(V|P) - \left[I(P,V') - R\right]^+ + \xi + 2\eta\right)\right\} \right.$$
$$\left. \text{for all } V, V' \in \mathcal{W}\right) \le \exp\{-n(\eta + \xi)\}.$$

The significance of the functions $g_{V,V'}$ lies in the fact that they can be used in deriving upper bounds on the error probabilities of the second kind. Indeed we have the following.

Lemma 18 *Suppose that, for every $V, V' \in \mathcal{W}$, $u^* \in \mathcal{T}_P^n$ satisfies*

$$g_{V,V'}(u^*) \le \exp\left\{n\left(H(V|P) - \left[I(P,V') - R - E_2\right]^+ + 2\eta + \xi\right)\right\}; \tag{27}$$

then

$$W^n(\mathcal{D}(\mathcal{U}) \cap \mathcal{F}_{u^*}|u^*) \le (n+1)^{2|\mathcal{X}|\cdot|\mathcal{Y}|} \exp\{-n(E_2 - 2\eta - \xi)\}, \tag{28}$$

$$\sum_{u \in \mathcal{U}} W^n(\mathcal{F}_{u^*}|u) \le (n+1)^{2|\mathcal{X}|\cdot|\mathcal{Y}|} \exp\{-n(E_2 - 2\eta - \xi)\}. \tag{29}$$

Proof Notice that

$$W^n(\mathcal{D}(U) \cap \mathcal{F}_{u^*}|u^*)$$

$$\le \sum_{\substack{V: I(P,V) \ge R + 2E_2 \\ V': I(P,V') \ge R + 2E_2}} W^n\left(\mathcal{T}_V^n(u^*) \cap \bigcup_{u \in \mathcal{U}} \mathcal{T}_{V'}^n(u) | u^*\right)$$

$$\le (n+1)^{2|\mathcal{X}|\cdot|\mathcal{Y}|} \max_{\substack{V: I(P,V) \ge R + 2E_2 \\ V': I(P,V') \ge R + 2E_2}} g_{V,V'}(u^*) \cdot \exp\{-n(D(V||W|P) + H(V|P))\}$$

by (2), (5), and (26). Furthermore, by (27) and $D(V||W|P) \geq 0$

$$\begin{aligned} & g_{V,V'}(u^*) \exp\{-n(D(V||W|P) + H(V|P))\} \\ & \leq \exp\left\{-n\left([I(P,V') - R - E_2]^+ - 2\eta - \xi\right)\right\} \\ & \leq \exp\{-n(E_2 - 2\eta - \xi)\}, \qquad \text{if } I(P,V') \geq R + 2E_2. \end{aligned}$$

Substitution of this bound in the previous bound gives (28). We show now (29). Clearly

$$\sum_{u\in\mathcal{U}} W^n(\mathcal{F}_{u^*}|u) \leq \sum_{u\in\mathcal{U}} \sum_{V} \sum_{V':I(P,V')\geq R+2E_2} W^n(\mathcal{T}^n_{V'}(u^*) \cap \mathcal{T}^n_V(u)|u).$$

Obviously, $\mathcal{T}^n_{V'}(u^*) \cap \mathcal{T}^n_V(u) \neq \emptyset$ implies $PV = PV'$. Now use (2), (26), (27), and $D(V||W|P) \geq 0$ to obtain the upper bound

$$\sum_{\substack{V':I(P,V')\geq R+2E_2 \\ V:PV=PV'}} \exp\left\{-n\left(H(V|P) - H(V'|P) + [I(P,V) - R - E_2]^+ - 2\eta - \epsilon\right)\right\}.$$

It remains to be shown that

$$H(V|P) - H(V'|P) + [I(P,V) - R - E_2]^+ \geq E_2.$$

If $I(P,V) \leq R + E_2$, then because of $PV = PV'$ and $I(P,V') \geq R + 2E_2$

$$I(P,V) \leq I(P,V') - E_2 \qquad \Rightarrow \qquad H(V|P) - H(V'|P) \geq E_2$$

Since $I(P,V') \geq R + 2E_2$,

$$H(V|P) - H(V'|P) + [I(P,V) - R - E_2]^+ \geq R + 2E_2 - R - E_2 = E_2.$$

On the other hand, if $I(P,V) \geq R + E_2$

$$\begin{aligned} H(V|P) - H(V'|P) + I(P,V) - R - E_2 &= H(PV) - H(V'|P) - R - E_2 \\ &= I(P,V') - R - E_2 \\ &\geq R + 2E_2 - R - E_2 = E_2. \end{aligned}$$

□

Lemma 18 says something about the error contribution of a u^* satisfying (27) if taken as a member of a $\mathcal{U}_2$, say, and if $\mathcal{U}_1 = \mathcal{U}$ is already specified. Our last auxiliary result concerns large deviations. We keep $\eta > 0$ fixed in (27) and prove

Lemma 19 *Let $U_{21}, U_{22}, \ldots, U_{2M}$ be defined as above. For any $\xi \geq 0$ we define*

$$S_{j\xi}(U_{2j}) = \begin{cases} 0, & \text{if } U_{2j} \text{ equals a } u^* \text{ satisfying (27)} \\ 1, & \text{otherwise} \end{cases}$$

for $j = 1, \ldots, M$. Then for every $\xi \in [0, E_2]$ and $n \geq n(\eta, E_2)$:

$$\Pr\left(\sum_{j=1}^{M} S_{j\xi} > \exp\{-n\xi\} \cdot M\right) \leq \exp\{-\exp\{nR\} \cdot (n\eta - 2)\}. \tag{30}$$

Proof For fixed ξ the $S_{1\xi}, \ldots, S_{M\xi}$ are i.i.d. RV's with values in $\{0, 1\}$. Lemma 17(ii) gives

$$\mathbb{E}S_{1\xi} \leq \exp\{-n(\eta + \xi)\}.$$

We apply Lemma 16 and get

$$\Pr\left(\sum_{j=1}^{M} S_{j\xi} > M \cdot \exp\{-n\xi\}\right) \leq \exp\{-M \cdot D(2^{-n\xi} || 2^{-n(\xi+\eta)})\}. \tag{31}$$

We have to estimate the relative entropy:

$$\begin{aligned} D(2^{-n\xi} || 2^{-n(\xi+\eta)}) &= 2^{-n\xi} \log(2^{-n\xi} 2^{n(\xi+\eta)}) + (1 - 2^{-n\xi}) \log \frac{1 - 2^{-n\xi}}{1 - 2^{-n(\xi+\eta)}} \\ &\geq 2^{-n\xi} \cdot \eta n + (1 - 2^{-n\xi}) \log(1 - 2^{-n\xi}) \\ &\geq 2^{-n\xi} \cdot \eta n + \log(1 - 2^{-n\xi}), \end{aligned}$$

because $t \log t \leq 0$ for $t \in [0, 1]$.

For small $t > 0$ one can estimate

$$\log(1 - t) \geq -2t.$$

Therefore,

$$D(2^{-n\xi} || 2^{-n(\xi+\eta)}) \geq 2^{-n\xi} \cdot \eta \cdot n - 2 \cdot 2^{-n\xi} = (n\eta - 2) \cdot 2^{-n\xi}, \tag{32}$$

if n is large enough. Since $M = \lfloor \exp\{n(R + E_2)\} \rfloor$ (see (23)) and $\xi \in [0, E_2]$, (31) and (32) imply (30). □

From Lemmas 17–19 to Proposition 15. We apply Lemmas 17–19 with $\eta = \gamma$. Choose in Lemma 19

$$\xi_k = \frac{E_2 \cdot k}{n}, \qquad k = 0, \ldots, n$$

to obtain

$$\Pr\left(\sum_{j=1}^{M} S_{j\xi_k} > M \cdot \exp\{-n\xi_k\} \quad \text{for some } k \in \{0, \ldots, n\}\right) \tag{33}$$

$$\leq (n+1)\exp\{-\exp\{nR\} \cdot (n\gamma - 2)\}.$$

It remains to be shown that (24) and (25) hold if

$$\sum_{j=1}^{M} S_{j\xi_k} \leq M \cdot \exp\{-n\xi_k\} \qquad \text{for all } k \in \{0, \ldots, n\}. \tag{34}$$

Suppose now that (34) holds. Choose $j \in \{0, \ldots, M\}$. Note that if $S_{j\xi} = 0$ for a $\xi \geq 0$, then also $S_{j\xi'} = 0$ for every $\xi' \geq \xi$. Similarly, if $S_{j\xi'} = 1$ for a $\xi \geq 0$, then also $S_{j\xi'} = 0$ for every $0 \leq \xi' \leq \xi$. By the definition of S there is a minimal ξ such that $S_{j\xi} = 0$.

Choose $k \in \{0, \ldots, n\}$. The number of $j \in \{1, \ldots, M\}$ such that the corresponding minimal ξ is contained in the interval $(k/n, (k+1)/n)$ is upper-bounded by $\sum_{j=0}^{M} S_{j\xi_k}$ (because of the monotonicity property discussed above). For these j, of course, $S_{j\xi_{k+1}} = 0$ holds, and *a fortiori*, (29) holds with ξ_{k+1}. On the other hand, the number of $j \in \{0, \ldots, M\}$ such that the corresponding minimal ξ is larger than E_2 $(= \xi_n)$ is upper-bounded by $\sum_{j=1}^{M} S_{j\xi_n}$.

With these arguments we get the following estimate:

$$\frac{1}{M}\sum_{u \in \mathcal{U}_1} W^n(\mathcal{D}(\mathcal{U}_2)|u) \leq \frac{1}{M}\sum_{j=1}^{M}\left(\sum_{u \in \mathcal{U}_1} W^n(\mathcal{F}_{U_{2j}}|u)\right)$$

$$\leq \frac{1}{M}\sum_{k=0}^{n}\left(\sum_{j=1}^{M} S_{j\xi_k}\right) \cdot (n+1)^{2|\mathcal{X}|\cdot|\mathcal{Y}|} \cdot \exp\{-n(E_2 - 2\gamma - \xi_{k+1})\} + \frac{1}{M}\sum_{j=1}^{M} S_{j\xi_n}.$$

Since $\xi_k - \xi_{k+1} = -1/n$ we can continue with (34)

$$\frac{1}{M}\sum_{u\in\mathcal{U}_1} W^n(\mathcal{D}(\mathcal{U}_2)|u)$$

$$\leq \left(\sum_{k=0}^{n}(n+1)^{2|\mathcal{X}|\cdot|\mathcal{Y}|}\cdot\exp\{-n(E_2-2\gamma-\xi_{k+1}+\xi_k)\}\right)+\exp\{-nE_2\}$$

$$\leq \exp\{-n(E_2-3\gamma)\}$$

for $n \geq n(\gamma, E_2)$. Using (28) instead of (29) in the foregoing derivation one obtains by the very same arguments that for all $n \geq n(\gamma, E_2)$

$$\frac{1}{M}\sum_{u\in\overline{\mathcal{U}}_2} W^n(\mathcal{D}(\mathcal{U}_1)|u) \leq \frac{1}{M}\sum_{j=1}^{M} W^n\left(\bigcup_{u\in\mathcal{U}_1}\mathcal{F}_u\cap\mathcal{F}_{U_{2j}}|U_{2j}\right)$$

$$\leq \exp\{-n(E_2-3\gamma)\}.$$

3 The Strong Converses

The strong converse to the coding theorem for identification via a DMC was conjectured in [5] (In case of complete feedback the strong converse was established already in [6]) and proved by Han and Verdu [10] and in a simpler way in [11]. We will present this proof in the next subsection. However, the authors used and developed analytical methods and take the position that combinatorial techniques for instance of [1, 3] find their limitations on this kind of problem (see also Newsletter on Moscow workshop in 1994). We demonstrate in Sect. 3.2 that this is not the case.

3.1 Analytic Proof of the Strong Converse

$$\lim_{n\to\infty}\frac{\log\log N(n,\lambda)}{n} \leq C \text{ for all } \lambda\in(0,1/2)$$

This proof follows the one given in [10]. We will need a few properties of *types*:

Recall that a PD Q on a finite set Ω is called n-type, if

$$Q(\omega)\in\left\{\frac{0}{n},\frac{1}{n},\ldots,\frac{n}{n}\right\} \text{ for all } \omega\in\Omega$$

Lemma 20 (Type Counting) *The number of different n-types on Ω is upper bounded by $|\Omega|^n$ and $(n+1)^{|\Omega|}$. The exact number is $\binom{|\Omega|+n-1}{|\Omega|-1}$.*

Proof The upper bounds are obvious: For the first bound notice that there are n "probability units" that can be distributed onto $|\Omega|$ different positions and for the second there are $n+1$ different masses for each $\omega \in \Omega$. For the exact number (which will not be needed in the following), consider the following encoding of an n-type $Q = (q_1/n, q_2/n, \ldots, q_{|\Omega|}/n)$ as 0-1 sequences of length $|\Omega| + n - 1$: First we take q_1 zeroes, then a one, q_2 zeroes, a one and so on until we take $q_{|\Omega|}$ zeroes. Obviously this procedure yields a bijective mapping between all n-types on Ω and 0-1 sequences of length $n + |\Omega| - 1$ which contain exactly $|\Omega| - 1$ ones. □

In this section n-types on $\mathcal{X}$ will be referred to as types, the set of all types is denoted by $\mathcal{P}_n$. Recall for $x \in \mathcal{X}$ and $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$ we set

$$< x^n | x >= |\{i \in \{1, \ldots, n\} : x_i = x\}|$$

and the PD P on $\mathcal{X}$ defined by

$$P(x) = P_x^n(x) := \frac{< x^n | x >}{n}$$

is an n-type and we will call x^n to be P-typical, while P is called the type of x^n. The set of all P-typical n-sequences is denoted as $\mathcal{T}_P^n$.

Definition 21 For a given PD Q and a type P we denote the restriction of Q on $\mathcal{T}_P^n$ by

$$Q^P(x^n) = Q(x^n | \mathcal{T}_P^n) = \begin{cases} \frac{Q(x^n)}{Q(\mathcal{T}_P^n)}; & x^n \in \mathcal{T}_P^n \\ 0; & x^n \notin \mathcal{T}_P^n \end{cases}$$

Definition 22 (Homogeneous ID Code) An ID Code

$$\{(Q(\cdot|i), \mathcal{D}_i) | i = 1, \ldots, N\}$$

which satisfies for all $P \in \mathcal{P}_n$

$$Q_1(\mathcal{T}_P^n) = \cdots = Q_N(\mathcal{T}_P^n) \tag{35}$$

is called homogeneous.

Lemma 23 *For every $(n, N, \lambda_1, \lambda_2)$ ID code, $\delta > 0$, $\lambda_1' > \lambda_1$, $\lambda_2' > \lambda_2$, n large enough there exists a homogeneous $(n, N \exp(-\delta n (n+1)^{|\mathcal{X}|}), \lambda_1', \lambda_2')$ ID code.*

Proof Let $\{(Q_i, \mathcal{D}_i) \,|i = 1, \ldots, N\}$ be the original code. We define an equivalence relation on the Q_i in the following way:

$$Q_i \leftrightarrow Q_j \Longleftrightarrow \forall\, P \in \mathcal{P}_n,\ \exists\, i_P \in \{0, \ldots, \lfloor \exp(n\delta/2) \rfloor\} :$$
$$i_P \exp(-n\delta/2) \leq Q_i(\mathcal{T}_P^n),\, Q_j(\mathcal{T}_P^n) < (i_P + 1)\exp(-n\delta/2).$$

If $\exp(n\delta/2) \in \mathbb{N}$, we will allow equality on the right side for $i_P = \exp(-n\delta/2) - 1$ and omit the last interval. Let $\mathcal{E}$ be the largest equivalence class and take an arbitrary $a \in \{1, \ldots, N\}$ with $Q_a \in \mathcal{E}$. Define for all $i \in \{1, \ldots, N\}$

$$\hat{Q}_i(x^n) = Q_a(\mathcal{T}_P^n) Q_i^P(x^n) \text{ for } x^n \in \mathcal{T}_P^n.$$

Then for all i and for all $P \in \mathcal{P}_n$ we have $\hat{Q}_i(\mathcal{T}_P^n) = Q_a(\mathcal{T}_P^n)$, so the ID code

$$\left\{ \left(\hat{Q}_i, \mathcal{D}_i \right) |\, Q_i \in \mathcal{E} \right\}$$

is homogeneous. To lowerbound its size we notice that there are at most $\lceil \exp(n\delta/2) \rceil^{|\mathcal{P}_n|}$ (possibly empty) equivalence classes, so

$$\begin{aligned} |\mathcal{E}| &\geq N \lceil \exp(n\delta/2) \rceil^{-|\mathcal{P}_n|} \\ &\geq N \exp(-n\delta |\mathcal{P}_n|) \\ &\geq N \exp\left(-n\delta (n+1)^{|\mathcal{X}|} \right) \end{aligned}$$

Also we have for arbitrary $\mathcal{D} \subset \mathcal{Y}^n$ and for all $i \in \{1, \ldots, N\}$

$$\begin{aligned} \hat{Q}_i W^n(D) &= \sum_{P \in \mathcal{P}_n} \sum_{x^n \in \mathcal{T}_P^n} \hat{Q}_i(x^n) W^n(\mathcal{D}|x^n) \\ &= \sum_{P \in \mathcal{P}_n} \sum_{x^n \in \mathcal{T}_P^n} Q_a(\mathcal{T}_P^n) Q_i^P(x^n) W^n(\mathcal{D}|x^n) \\ &= \sum_{P \in \mathcal{P}_n} \sum_{x^n \in \mathcal{T}_P^n} Q_i(\mathcal{T}_P^n) Q_i^P(x^n) W^n(\mathcal{D}|x^n) \pm \sum_{P \in \mathcal{P}_n} \exp(-n\delta/2) \\ &= Q_i W^n(D) \pm (n+1)^{|\mathcal{X}|} \exp(-n\delta/2) \end{aligned}$$

But $(n+1)^{|\mathcal{X}|} \exp(-n\delta/2)$ converges to zero for $n \to \infty$, so for sufficiently large n our code has the desired error probabilities. □

Homogeneous codes have a restriction concerning the weight of each type, but we also need some control of codistributions induced by the types:

Definition 24 (M-Regular ID Code) Let $M \in \mathbb{N}$. An ID code

$$\{(Q_i, \mathcal{D}_i)|i = 1, \cdots, N\}, \quad Q(\cdot|i) \in \mathcal{P}\left(\mathcal{X}^n\right), \quad \mathcal{D}_i \subset \mathcal{Y}^n \quad \text{for all} i = 1, \cdots, N$$

is called M-regular, if for all $P \in \mathcal{P}_n$ and $i = 1, \ldots, N$, Q_i^P is M-type.

We will now state a lemma about approximation of channel output statistics by distributions, which are t-typical for certain t.

Lemma 25 *There exists $\epsilon_0, \delta_0 > 0$ such that for all $P \in \mathcal{P}_n$, $\epsilon \in [0, \epsilon_0]$, $\delta \in [0, \delta_0]$, PD Q on $\mathcal{T}_P^n$, n large enough there exists an $\lceil \exp(nC + n\gamma) \rceil$-type distribution Q', such that for all $Y \subset \mathcal{Y}^n$*

$$Q'W^n(Y) \leq \frac{(1+\epsilon)}{1-\exp(-n\delta)} QW^n(Y) + \exp(-n\delta)$$

$$Q'W^n(Y) \geq (1-\epsilon)(1-\exp(-n\delta))QW^n(Y) - \exp(-n\delta)$$

where $\gamma = \rho(\delta)$ and $\rho : [0, \delta_0] \to \mathbb{R}^+$ is a continuous strictly increasing function with $\rho(0) = 0$.

A proof for this lemma will be given later below, while more general results about approximation of output statistics can be found in [11]. We apply this lemma to convert homogeneous codes to M-regular codes, where $M \sim \exp(nC)$.

Lemma 26 *For every homogeneous $(n, N, \lambda_1, \lambda_2)$ ID code*

$$\{(Q(\cdot|i), \mathcal{D}_i) | i = 1, \ldots, N\}$$

with error probabilities λ_1 and λ_2, for all $\lambda_1' > \lambda_1$, $\lambda_2' > \lambda_2$, $\gamma > 0$, n large enough, there exists a homogenous $\lceil \exp(nC + n\gamma) \rceil$-regular $(n, N, \lambda_1', \lambda_2')$ ID code.

Proof Modify the original code the following way: Define new codistributions

$$Q_i'(x^n) = Q_i(\mathcal{T}_P^n) Q_i^{'P}(x^n), \text{ if } x^n \in \mathcal{T}_P^n$$

where $Q_i^{'P}$ is an $\lceil \exp(nC + n\gamma) \rceil$-type obtained from Q_i^P using Lemma 25 with $\delta = \rho^{-1}(\gamma)$ and

$$\epsilon < \min\left\{\frac{\lambda_2'}{\lambda_2} - 1, \frac{1-\lambda_1'}{1-\lambda_1}\right\}.$$

The decoding sets remain unchanged. This is a homogeneous $\lceil \exp(nC + n\gamma) \rceil$-regular code with the same size as the code we started with. We will now upper bound its error probabilities. Let $a \neq b$. Then for large n

$$\begin{aligned} Q_a' W^n(\mathcal{D}_b) &= \sum_{P \in \mathcal{P}_n} Q_a(\mathcal{T}_P^n) Q_a'^P W^n(\mathcal{D}_b) \\ &\leq \exp(-n\delta) + \sum_{P \in \mathcal{P}_n} Q_a(\mathcal{T}_P^n)(1+\epsilon)(1-\exp(-n\delta))^{-1} Q_a^P W^n(\mathcal{D}_b) \\ &= (1+\epsilon)(1-\exp(-n\delta))^{-1} Q_a W^n(\mathcal{D}_b) + \exp(-n\delta) \\ &\leq (1+\epsilon)(1-\exp(-n\delta))^{-1} \lambda_2 + \exp(-n\delta) \\ &\leq \lambda_2'. \end{aligned}$$

Also

$$\begin{aligned} Q_a' W^n(\mathcal{D}_a) &= \sum_{P \in \mathcal{P}_n} Q_a(\mathcal{T}_P^n) Q_a'^P W^n(\mathcal{D}_a) \\ &\geq (1-\epsilon)(1-\exp(-n\delta)) Q_a W^n(\mathcal{D}_a) - \exp(-n\delta) \\ &\geq 1 - \lambda_1'. \end{aligned}$$

Thus for sufficiently large n, the new code has the desired error probabilities. □

Now we show that homogenous M-regular codes cannot be too big:

Lemma 27 *Given a homogeneous M-regular $(n, N, \lambda_1, \lambda_2)$ ID code with $\lambda_1 + \lambda_2 < 1$. Then*

$$\log N \leq n(n+1)^{|\mathcal{X}|} M \log |\mathcal{X}|.$$

Proof If $\lambda_1 + \lambda_2 < 1$, then all Q_i have to be different. But there are at most $|\mathcal{X}|^{nM}$ different M-types on $\mathcal{X}^n$, $|\mathcal{P}_n|$ different types, we can upperbound the number of different M-regular codistributions by

$$N \leq |\mathcal{X}|^{nM|\mathcal{P}_n|}.$$

This holds because of (35), every type $\mathcal{T}_P^n$ has constant weight under all codistributions. Therefore, if

$$Q_i \neq Q_j \Rightarrow \exists P : Q_i^P \neq Q_j^P.$$

(If $Q_i(x^n) \neq Q_j(x^n)$ choose the type of x as P). Also we have $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$.

Finally, we will put all the results together: Given arbitrary λ_1, λ_2 with $\lambda_1 + \lambda_2 < 1$ and an arbitrary sequence of $(n, N, \lambda_1, \lambda_2)$ ID codes. Choose $\delta > 0$, $\gamma > 0$,

$\lambda_1'' > \lambda_1' > \lambda_1$ and $\lambda_2'' > \lambda_2' > \lambda_2$, such that $\lambda_1'' + \lambda_2'' < 1$. Using the preceding results, via a homogeneous $(n, N', \lambda_1', \lambda_2')$ ID code we can construct a homogeneous $\lceil \exp(nC + n\gamma) \rceil$-regular $(n, N', \lambda_1'', \lambda_2'')$ ID code, where $N' = N \exp\left(-n\delta(n+1)^{|\mathcal{X}|}\right)$ (for sufficiently large n). From Lemma 27 we get that

$$\log N' = \log N - \delta n(n+1)^{|\mathcal{X}|} \leq n(n+1)^{|\mathcal{X}|} M \log |\mathcal{X}|.$$

Therefore for large n

$$\begin{aligned} \frac{\log \log N}{n} &\leq \frac{\log n(n+1)^{|\mathcal{X}|} + \log(\exp(nC + n\gamma) \log |\mathcal{X}| + \delta)}{n} \\ &\leq \gamma + \frac{\log \log |\mathcal{X}|}{n} + C + 2\gamma \\ &\leq C + 4\gamma \end{aligned}$$

which completes the proof as $\gamma > 0$ could be chosen arbitrarily. □

3.1.1 Proof of Lemma 25

Proof This proof is the second part of [10]. It is separated form the rest of the proof of the ID-coding theorem for two reasons: On the one hand, it is quite a long series of arguments. On the other hand, this result may be useful also in other contexts: Since n-type distributions can be modelled as empirical distributions of code words of length n, it estimates the number of input bits to be given into a channel to approximate certain output distributions.

First we will decompose our given DMC W into channels, which have a certain structure and are therefore easier to analyse:

Definition 28 (Conditional Type) Let $V = \{V(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ be a stochastic matrix. We say that $y^n \in \mathcal{Y}$ has conditional type V given $x^n \in \mathcal{X}^n$, if for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$

$$< (x^n, y^n) | (x, y) > = < x^n | x > V(y|x).$$

If $x^n \in \mathcal{X}^n$, we will denote the set of all sequences in $\mathcal{Y}^n$ with conditional type V given x^n by $\mathcal{T}_V^n(x^n)$ (this set was defined in Definition 8).

The conditional type of y^n given x^n can be viewed as an estimate of the channel that produces y^n for an input sequence x^n, it yields the empirical distribution of pairs of input/output letters. The conditional type is uniquely determined except for the rows, whose corresponding letter does not occur in x^n.

Since the cardinality of $\mathcal{T}_V^n(x^n)$ depends only on V and the type of x^n (if a permutation of indices changes x^n to x'^n, then the same permutation changes $\mathcal{T}_V^n(x^n)$ to $\mathcal{T}_V^n(x'^n)$) we can define

$$L_V^P = |\mathcal{T}_V^n(x^n)|$$

where x^n is a sequence of type P. Now we will define a channel, which maps input sequences of the same type to output sequences of the corresponding conditional type: Let P be a type and V be given. Then define

$$W_V^P(y^n|x^n) = \begin{cases} \frac{1}{L_V^P} & x^n \in \mathcal{T}_P^n \text{ and } y^n \in \mathcal{T}_V^n(x^n) \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

We will call such a channel an equitype channel. Also we define the set of conditional types congruent with P by

$$\Lambda^P = \{V : (L_V^P > 0) \text{ and } (\forall x : P(x) = 0 \implies V(\cdot|x) = W(\cdot|x))\}.$$

The value of $W^n(y^n|x^n)$ is uniquely determined by the type of x^n and the conditional type of y^n given x^n, so we can define

$$c_V^P = W^n(\mathcal{T}_V^n(x^n)|x^n)$$

where x^n can be arbitrarily chosen from $\mathcal{T}_P^n$. Then for all $x^n \in \mathcal{T}_P^n$ and $y^n \in \mathcal{T}_V^n(x^n)$

$$W^n(y^n|x^n) = \frac{c_V^P}{L_V^P} = c_V^P W_V^P(y^n|x^n).$$

For every (x^n, y^n) there is a unique pair (P, V), $P \in \mathcal{P}_n$, $V \in \Lambda^P$ with $x^n \in \mathcal{T}_P^n$ and $y^n \in \mathcal{T}_V^n(x^n)$, because we defined Λ^P in such a way, that we just copied the original values of W where the conditional type is not uniquely determined. So we have

$$W^n(y^n|x^n) = \sum_{P \in \mathcal{P}_n} \sum_{V \in \Lambda^P} c_V^P W_V^P(y^n|x^n).$$

Thus we can write our channel as a sum of equitype channels. Now we will consider the set of input sequences that can with positive probability produce a certain output on a given equitype channel: Let $y^n \in \mathcal{Y}^n$ and W_V^P be an equitype channel. Then

$$H_V^P(y^n) = \{x^n \in \mathcal{T}_P^n : W_V^P(y^n|x^n) > 0\}$$

which we will denote as inverse image of y^n. From Eq. (36) we get that

$$QW_V^P(y^n) = \frac{Q\left(H_V^P(y^n)\right)}{L_V^P}. \tag{37}$$

Lemma 29 *We will define for $P \in \mathcal{P}_n$, $V \in \Lambda^P$ and $\delta > 0$ the set*

$$G_V^P = \left\{ y^n \in \mathcal{Y}^n : Q\left(H_V^P(y^n)\right) \geq \exp(-nI(P,V) - n\delta) \right\}.$$

Then for all n it holds, that

$$QW_V^P(G_V^P) \geq 1 - \exp(-n\delta)(n+1)^{|\mathcal{X}||\mathcal{Y}|}.$$

Proof Define the set

$$F_V^P = \left\{ y^n \in \mathcal{T}_{PV}^n : \frac{Q\left(H_V^P(y^n)\right)}{L_V^P} > \frac{\exp(-n\delta)(n+1)^{|\mathcal{X}||\mathcal{Y}|}}{|\mathcal{T}_{PV}^n|} \right\}.$$

Here $\mathcal{T}_{PV}^n$ denotes the set of all output sequences of unconditional type PV. Obviously,

$$QW_V^P(F_V^P) \geq 1 - \exp(-n\delta)(n+1)^{|\mathcal{X}||\mathcal{Y}|}.$$

Also $F_V^P \subset G_V^P$, because for $x^n \in \mathcal{T}_P^n$ [9, Problem 3(b), Chapter 1 Section 2], yields

$$\frac{\exp(-nI(P,V))}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} \leq \frac{|\mathcal{T}_V^n(x^n)|}{|\mathcal{T}_{PV}^n|} = \frac{|L_V^P|}{|\mathcal{T}_{PV}^n|}. \qquad \square$$

Now we will concentrate on those equitype channels, that are "close" to the original channel in the sense of the conditional relative entropy.

Let $V = \{V(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ and $W = \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ be stochastic matrices and let P be a PD on $\mathcal{X}$. Recall that the conditional relative entroy between W and V with respect to P was defined as $D(V||W|P) \triangleq \sum_{x\in\mathcal{X}} P(x)D(V(\cdot|x)||W(\cdot|x))$.

Now define

$$\Lambda_\delta^P = \{V \in \Lambda^P : D(V||W|P) \leq \delta\}$$

and a new channel matrix $W_\delta^* = \{W(y^n|x^n) : x^n \in \mathcal{X}^n, y \in \mathcal{Y}^n\}$ by

$$W_\delta^*(y^n|x^n) = \sum_{P\in\mathcal{P}_n} \sum_{V\in\Lambda^P} \bar{c}_V^P W_V^P(y^n|x^n),$$

where

$$\bar{c}_V^P = \begin{cases} c_V^P / \sum_{U \in \Lambda_\delta^P} c_U^P & V \in \Lambda_\delta^P \\ 0 & \text{otherwise.} \end{cases}$$

We will now show that this modification does not change our channel too much:

Lemma 30 *For all n, $x^n \in \mathcal{X}^n$, $Y \subset \mathcal{Y}^n$ and $\delta > 0$ we have*

$$W^n(Y|x^n) \geq (1 - \exp(-n\delta))(n+1)^{|\mathcal{X}||\mathcal{Y}|} W_\delta^*(Y|x^n), \tag{38}$$

$$W^n(Y|x^n) \leq W_\delta^*(Y|x^n) + \exp(-n\delta)(n+1)^{|\mathcal{X}||\mathcal{Y}|}. \tag{39}$$

Proof Let $x^n \in \mathcal{X}^n$, let P be the type of x^n and define

$$\alpha_\delta^P = \sum_{U \in \Lambda_\delta^P} c_U^P.$$

Then we can write

$$W^n(Y|x^n) = \sum_{V \in \Lambda^P} c_V^P W_V^P(Y|x^n),$$

$$W_\delta^*(Y|x^n) = \sum_{V \in \Lambda^P} \bar{c}_V^P W_V^P(Y|x^n),$$

which implies on the one hand

$$\alpha_\delta^P W_\delta^*(Y|x^n) \leq W^n(Y|x^n)$$

and on the other

$$\begin{aligned} W^n(Y|x^n) &= \sum_{V \in \Lambda_\delta^P} c_V^P W_V^P(Y|x^n) + \sum_{V \in \Lambda^P \setminus \Lambda_\delta^P} c_V^P W_V^P(Y|x^n) \\ &\leq \sum_{V \in \Lambda_\delta^P} \bar{c}_V^P W_V^P(Y|x^n) + \sum_{V \in \Lambda^P \setminus \Lambda_\delta^P} c_V^P W_V^P(Y|x^n) \\ &\leq W_\delta^*(Y|x^n) + 1 - \alpha_\delta^P. \end{aligned}$$

To estimate α_δ^P note, that for all $P \in \mathcal{P}_n$

$$\begin{aligned} 1 - \alpha_\delta^P &= \sum_{V \in \Lambda^P \setminus \Lambda_\delta^P} c_V^P \\ &= \sum_{V \in \Lambda^P \setminus \Lambda_\delta^P} W^n(\mathcal{T}_V^n(x^n)|x^n) \\ &\leq \sum_{V \in \Lambda^P \setminus \Lambda_\delta^P} \exp(-nD(V||W|P)) \\ &\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-n\delta), \end{aligned}$$

where we used Lemma 1.2.6 from [9] and the fact, that $|\Lambda^P| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}$, since its elements are n-types on $\mathcal{X} \times \mathcal{Y}$. This together with the two previous inequalities yields the desired result. □

Using this lemma we will show the existence of an $\lceil \exp(nC + n\gamma) \rceil$-type distribution Q', with

$$\begin{aligned} Q'W^n(Y) &\leq (1+\epsilon) QW_\delta^*(Y) + \exp(-n\delta) \\ Q'W^n(Y) &\geq (1-\epsilon) QW_\delta^*(Y) - \exp(-n\delta) \end{aligned}$$

for sufficiently large n. Finally we will show that this is sufficient to complete the proof.

To prove the existence, we have to upperbound $I(P, V)$:

Lemma 31 *Let* $\sqrt{D(V||W|P)} < \min\{\frac{1}{8}\log e, 1\}$. *Then*

$$|I(P, V) - I(P, W)| \leq 2g(D(V||W|P)) + \sqrt{D(V||W|P)} \log |Y|,$$

where

$$g(x) = \begin{cases} -\sqrt{\frac{2x}{\log e}} \log \sqrt{\frac{2x}{\log e}} & x > 0 \\ 0 & x = 0. \end{cases}$$

Proof Let Q and R be arbitrary PDs with $D(Q||R) < \frac{1}{8}\log e$. Then [9, Chapter 1, Lemma 2.7 and Problem 17 Section 3] yield

$$|H(Q) - H(R)| \leq g(D(Q||R)).$$

Now define the set $L = \left\{x : D(V(\cdot|x)||W(\cdot|x)) \leq \sqrt{D(V||W|P)}\right\}$ and apply Markov's Lemma (cf. for example [13, Satz 3.15]) on L^c to obtain $P(L^c) \leq \sqrt{D(V||W|P)}$. Thus

$$\begin{aligned}
&|H(V|P) - H(W|P)| \\
&= \left| \sum_{x \in \mathcal{X}} P(x)(H(V(\cdot|x)) - H(W(\cdot|x))) \right| \\
&\leq \sum_{x \in \mathcal{X}} P(x)|H(V(\cdot|x)) - H(W(\cdot|x)|) \\
&\leq \sum_{x \in L} P(x) g(D(V(\cdot|x)||W(\cdot|x))) + \sum_{x \in L^c} P(x) \log |\mathcal{Y}| \\
&\leq \sum_{x \in L} P(x) g(D(V(\cdot|x)||W(\cdot|x))) + \sqrt{D(V||W|P)} \log |\mathcal{Y}| \\
&\leq g(D(V||W|P)) + \sqrt{D(V||W|P)} \log |\mathcal{Y}|
\end{aligned}$$

where we used for the second inequality the fact, that the entropy of a random variable can be upperbounded by the logarithm of the size of its image, and for the last we used the fact, that g is concave. Also we have

$$D(PV||PW) \leq D(V||W|P) \leq \sqrt{D(V||W|P)} \leq \frac{1}{8} \log e$$

since we also assumed $\sqrt{D(V||W|P)}$ to be smaller than 1. Therefore

$$|H(PV) - H(PW)| \leq g(D(V||W|P))$$

and

$$\begin{aligned}
&|I(P, V) - I(P, W)| \\
&\qquad = |H(PV) - H(V|P) - H(PW) - H(W|P)| \\
&\qquad \leq |H(PV) - H(PW)| + |H(V|P) - H(W|P)| \\
&\qquad \leq g(D(V||W|P)) + g(D(V||W|P)) + \sqrt{D(V||W|P)} \log |\mathcal{Y}|
\end{aligned}$$

□

Now we will define our function ρ from the lemma: Set

$$\rho(\delta) = 2\delta + 2g(\delta)\sqrt{\delta}|\mathcal{Y}|.$$

It is obvious to see that there is a $\delta_0 > 0$ (which depends on $|\mathcal{Y}|$) such, that ρ (and also g) is continuous and strictly increasing on $(0, \delta_0)$. If now $V \in \Lambda_\delta^P$ (which means by definition that $D(V||W|P) \leq \delta$), then by the previous lemma

$$\begin{aligned} I(P,V) + \delta &\leq I(P,W) + 2g(D(V||W|P)) + \sqrt{D(V||W|P)}\log|Y| + \delta \\ &\leq I(P,W) + 2g(\delta) + \sqrt{\delta}\log|Y| + \delta \\ &= I(P,W) + \rho(\delta). \end{aligned}$$

This with $C = \sup_{P\in\mathcal{P}(\mathcal{X})} I(P,W)$ implies

$$\sup_{P\in\mathcal{P}_n}\sup_{V\in\Lambda_V^P} I(P,V) + \delta \leq C + \rho(\delta). \tag{40}$$

We will set for abbreviation $\gamma = \rho(\delta)$ and $M = \lceil \exp(nC + n\gamma)\rceil$. If $y \in G_V^P$ and $V \in \Lambda_\delta^P$, then

$$\begin{aligned} Q(H_V^P(y^n)) &\geq \exp(-n(I(P,V) + \delta)) \\ &\geq \exp(-n(C + \rho(\delta))) \\ &\geq \exp(-n(C + \rho(\delta) + \delta)) \\ &= \frac{1}{M}\exp(-n\delta). \end{aligned}$$

Now we will show the existence of our M-type distribution Q' by considering realisations of independent identically distributed random variables $(U_1, \ldots, U_M)$, where all U_i have distribution Q. The empirical distribution of a realisation of this RVs is of course an M-type, and with positive probability it will approximate our original PD Q as needed:

Lemma 32 *There exists $(u'_1, \ldots, u'_M)$, $u'_i \in \mathcal{T}_P^n$ such, that for all $V \in \Lambda_\delta^P$, $y^n \in G_V^P$ the following inequalities hold:*

$$\frac{1}{M}\sum_{i=1}^{M} 1_{\{u'_i \in H_V^P(y^n)\}} \leq (1+\epsilon)Q(H_V^P(y^n))$$

$$\frac{1}{M}\sum_{i=1}^{M} 1_{\{u'_i \in H_V^P(y^n)\}} \geq (1-\epsilon)Q(H_V^P(y^n))$$

$$\frac{1}{M}\sum_{i=1}^{M} W_V^P((G_V^P)^c|u'_i) \leq \exp\left(-\frac{n\delta}{3}\right).$$

Proof Let $0 < \delta' < \delta$. Then

$$\Pr\left(\frac{1}{M}\sum_{i=1}^{M} U_i W_V^P((G_V^P)^c) > \exp\left(-\frac{n\delta'}{2}\right)\right) < \exp\left(-\frac{n\delta'}{2}\right).$$

Therefore

$$\begin{aligned}\Pr&\left(\frac{1}{M}\sum_{i=1}^{M} U_i W_V^P((G_V^P)^c) > \exp\left(-\frac{n\delta'}{2}\right) \text{ for some } V \in \Lambda_\delta^P\right)\\ &< (n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\left(-\frac{n\delta'}{2}\right)\\ &\leq \exp\left(-\frac{n\delta}{3}\right)\end{aligned}$$

for large n.

Using a similar argument as for Lemma 16 or using the following

Theorem 33 (Sanov's Theorem) *Let A be a set of probability distributions over an alphabet $\mathcal{X}$, and let Q be an arbitrary distribution over $\mathcal{X}$ (where Q may or may not be in A). Suppose we draw n i.i.d. samples from X, represented by the vector x^n. Further, let P_{x^n} be the empirical distribution, of the samples falling within the set A. Then,*

$$Q^n(x^n) \leq (n+1)^{|\mathcal{X}|}2^{-nD(P^*||Q)},$$

where P^* is the information projection of Q onto A.

Furthermore, if A is a closed set,

$$\lim_{n\to\infty}\frac{1}{n}\log Q^n(x^n) = -D(P^*||Q).$$

One gets for $\Psi_1, \cdots, \Psi_M$ a Bernoulli chain with values 0 and 1 and $\Pr(\Psi_i = 1) = \mathbb{E}\Psi_i \geq \mu > \lambda$ that

$$\Pr\left(\sum_{j=1}^{M}\Psi_j < M\cdot\lambda\right) \leq \exp\left(-M\cdot D\left(\lambda||\mu\right)\right). \tag{41}$$

Together with Lemma 16 and the property of the relative entropy, that there is an $\epsilon_0 > 0$ with

$$D((1+t)\mu||\mu) \geq \frac{t^2\mu\log e}{2} \quad \text{for all } t \in [-\epsilon_0, \epsilon_0]$$

we get

$$\Pr\left(\frac{1}{M}\sum_{i=1}^{M} 1_{\{u_i'\in H_V^P(y^n)\}} > (1+\epsilon)Q(H_V^P(y^n))\right) \tag{42}$$
$$\leq \exp_e\left(-\frac{M\epsilon^2\mu}{3}\right) \leq \exp_e\left(-\frac{\epsilon^2\exp(n\delta)}{2}\right)$$

and also

$$\Pr\left(\frac{1}{M}\sum_{i=1}^{M} 1_{\{u_i'\in H_V^P(y^n)\}} < (1-\epsilon)Q(H_V^P(y^n))\right) \tag{43}$$
$$\leq \exp_e\left(-\frac{M\epsilon^2\mu}{3}\right) \leq \exp_e\left(-\frac{\epsilon^2\exp(n\delta)}{2}\right)$$

for all $\epsilon \in (0,\epsilon_0)$. To see this, we set $\Psi_i = 1_{\{u_i'\in H_V^P(y^n)\}}$ (which leads to $\mu = Q(H_V^P(y^n)) \geq \frac{n\delta}{M}$) and apply Lemma 19 respectively inequality (41). Now using the union bound one easily gets

$$\Pr\left(\frac{1}{M}\sum_{i=1}^{M} 1_{\{u_i'\in H_V^P(y^n)\}} > (1+\epsilon)Q(H_V^P(y^n))\right) \leq \exp_e\left(-\frac{\epsilon^2\exp(n\delta)}{3}\right)$$
$$\Pr\left(\frac{1}{M}\sum_{i=1}^{M} 1_{\{u_i'\in H_V^P(y^n)\}} < (1-\epsilon)Q(H_V^P(y^n))\right) \leq \exp_e\left(-\frac{\epsilon^2\exp(n\delta)}{3}\right)$$

for large n. Therefore, the sum of the probabilities, that the realisation fails at least one of the conditions, becomes smaller than 1 for n sufficiently large, i.e. a realisation with the desired properties exists which concludes the proof of Lemma 32. □

To complete the proof of Lemma 25 we will finally show the properties of the $(u_1', \ldots, u_M')$ (and therefore the empirical distribution Q). we show that we obtain from Lemma 32 the required quality of approximation:

Let $y^n \in G_V^P$. Then

$$\begin{aligned} Q'W_V^P(y^n) &= \frac{Q'(H_V^P(y^n))}{L_V^P} \\ &\leq (1+\epsilon)\frac{Q(H_V^P(y^n))}{L_V^P} \\ &= (1+\epsilon)QW_V^P(y^n) \end{aligned}$$

Thus for $Y \subset \mathcal{Y}^n$, $V \in \Lambda_\delta^P$ we get

$$Q'W_V^P(Y) = Q'W_V^P(Y \cap G_V^P) + Q'W_V^P(Y \cap (G_V^P)^C)$$
$$\leq (1+\epsilon)Q(W_V^P(Y)) + \exp\left(\frac{-n\delta}{3}\right)$$

and analogously

$$Q'W_V^P(Y) \geq (1-\epsilon)QW_V^P(Y) - \exp\left(\frac{-n\delta}{3}\right).$$

Since this inequalities hold for all equitype channels, they hold also for our channel W, which is a convex combination of the W_V^P, and Lemma 25 is proved. □

3.2 Combinatorial Proof of the Strong Converse

Here we come back to the very first idea from [5], essentially to replace the distributions P_i by uniform distributions on "small" subsets of $\mathcal{X}^n$, namely with cardinality slightly above $\exp(nC(W))$.

The core of the proof is a result on hypergraph coloring, which is explained in all details in [8]. We will give here the definitions and the result.

Definition 34 A hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ consists of a (finite) vertex set $\mathcal{V}$ and a set of hyper-edges $\mathcal{E}$, where each edge $E \in \mathcal{E}$ is a subset of $E \subset \mathcal{V}$.

The vertices will usually be labelled by $\mathcal{V} = (v_1, \ldots, v_I)$ with $I = |\mathcal{V}|$, and the edges by $\mathcal{E} = (E_1, \ldots, E_J)$ with $1 \leq J \leq 2^{|\mathcal{E}|}$.

We call a hypergraph uniform if in $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ all edges $E \in \mathcal{E}$ have the same cardinality D. For a uniform distribution P on $\mathcal{E}$ we can define the associated (output) distribution Q,

$$Q(v) = \sum_{E \in \mathcal{E}} D(E) 1_E(v) = \sum_{E \in \mathcal{E}} \frac{1}{|\mathcal{E}|} 1_E(v). \tag{44}$$

Our goal is to find an $\mathcal{E}^* \subset \mathcal{E}$ as small as possible such that the distribution Q^*,

$$Q^*(v) \triangleq \sum_{E \in \mathcal{E}^*} \frac{1}{|\mathcal{E}^*|} 1_E(v) \quad \text{for all } v \in \mathcal{V} \tag{45}$$

is a good approximation of Q in the following sense. For some $\mathcal{V}^* \subset \mathcal{V}$

$$\sum_{u \in \mathcal{V}^*} Q(v) \leq \delta \tag{46}$$

and

$$(1-\varepsilon)Q(v) \le Q^*(v) \le (1+\varepsilon)Q(v) \quad \text{for all } v \in \mathcal{V} \setminus \mathcal{V}^*. \tag{47}$$

Lemma 35 (Multiple Covering Lemma) *For the uniform hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, and any $\varepsilon, \delta > 0$ there is a $\mathcal{E}^* \subset \mathcal{E}$ and a $\mathcal{V}^* \subset \mathcal{V}$ such that for Q^* defined in (44), (45) holds and*

$$|\mathcal{E}^*| \le \frac{\delta|\mathcal{V}|}{\varepsilon^2|E|} \log|\mathcal{V}|.$$

Let $\{(P_i, \mathcal{D}_i) : i = 1, \ldots, N\}$ be an $(n, N, \lambda_1, \lambda_2)$ ID code, $\lambda_1 + \lambda_2 = 1 - \lambda < 1$. Our goal is to construct an $(n, N, \lambda_1 + \lambda/3, \lambda_2 + \lambda/3)$ ID code $\{(\bar{P}_i, D_i) : i = 1, \ldots, N\}$, where $\bar{P}_i \in \mathcal{P}_k$ is a k-type on $\mathcal{X}^n$. Recall that all the probabilities are rational with common denominator k (see Definition 6). We fix i for the moment.

Let T be a distribution on $\mathcal{X}$. Recall that $\mathcal{T}_T^n$, is nonempty if T is an empirical distribution (ED).

We already saw that there are less than $(n+1)^{|\mathcal{X}|}$ many empirical distributions.

For the emprirical distribution T the restricted type (see Definition 21)

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

is a probability distribution on $\mathcal{T}_T^n$.

Note:

$$P_i = \sum_{T \in \mathcal{P}_n} P_i(\mathcal{T}_T^n) P_i^T.$$

Up to now we only considered typical sequences for probability distributions arising from relative frequencies, which we call ED. For general probability distributions the following definition is more appropriate.

Definition 36 $x^n \in \mathcal{X}^n$ is (n, P, α)-*typical*, if

$$|< x^n | x > - n \cdot P(x)| \le \alpha \cdot n \quad \text{for all } x \in \mathcal{X}.$$

So for $\alpha = 0$ we again have the typical sequences. As pointed out above this notion can be applied to any probability distribution. Another advantage is that we are free to choose the parameter α. For example α might be a constant. When Chebyshev's inequality and the weak law of large numbers are involved usually $\alpha = \frac{c}{\sqrt{n}}$ for a constant c is chosen. The set of (n, P, α)-typical sequences is denoted as $\mathcal{T}_{P,\alpha}^n$ (or $\mathcal{T}_{X,\alpha}^n$).

Definition 37 Let $\mathcal{X}$ and $\mathcal{Y}$ be the input and output alphabet of a DMC specified by the transmission matrix $W = \big(W(y|x)\big)_{x\in\mathcal{X},\, y\in\mathcal{Y}}$. A sequence $y^\ell \in \mathcal{Y}^\ell$ is called $(\mathbf{x}^\ell, \alpha)$-*conditionally typical*, if for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$

$$\Big| < x^\ell, y^\ell | x, y > - < x^\ell | x > \cdot W(y|x) \Big| \le \alpha \cdot \ell.$$

The set of all $(\mathbf{x}^\ell, \alpha)$-*conditional typical* sequences is denoted as

$$\mathcal{T}^\ell_{Y|X,\alpha}(x^\ell) = \mathcal{T}^\ell_{W,\alpha}(x^\ell).$$

Let us give a short interpretation of the previous definitions. Assume that a sequence $x^\ell \in \mathcal{X}^\ell$ of absolute type $(< x^\ell | x >)_{x\in\mathcal{X}}$ is transmitted over the channel W. Then the received sequence y^ℓ will "typically" contain $< x^\ell | x > \cdot w(y|x)$ y's in those positions where the original sequence x^ℓ had the letter x. So with high probability a (x^ℓ, α)-generated sequence will be received, when x^ℓ is transmitted over the channel.

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

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

we consider the set of *conditional typical sequences* $\mathcal{T}^n_{W,\alpha}(x^n)$.

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

Define now the measures Q_{x^n} by

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

By the properties of typical sequences and choice of α we have

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

where $\| \cdot \|_1$ denotes the statistical distance.

Now with $\varepsilon = \tau = \lambda/36$ apply Lemma 35 to the hypergraph with vertex set $\mathcal{T}^n_{TW,\alpha}$ and edges $\mathcal{T}^n_{W,\alpha}(x^n)$, $x^n \in \mathcal{T}^n_T$, carrying measure $W(\cdot|x^n)$, and the probability distribution P_i^T on the edge set: we get an L-type $\bar{P}_i^T$ with

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

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

where the constants depend explicitly on α, δ, τ. By construction we get

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

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

Now choose a K-type R on the set of all empirical distributions such that

$$\sum_{T \in \mathcal{P}_n} |P_i(\mathcal{T}_T^n) - R(T)| \leq \frac{\lambda}{3},$$

which is possible for

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

Defining

$$\bar{P}_i = \sum_{T \in \mathcal{P}_n} R(T) \bar{P}_i^T$$

we can summarize

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

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

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

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

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

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

where the last inequality holds only if n is large enough.

4 Discussion

In all coding problems previously studied in information theory, the maximal code lengths grow only exponentially in block length. Therefore, our double exponent coding theorem is the first of its kind. The identification problem solved seems to be a natural one. In our judgement it enlarges the basis of information theory, which in Shannon's foundation was restricted to the transmission problem. The success of Shannon's theory relies on the fact that the semantic aspect of information was excluded, but the identification problem also has its place in a presemantic theory. Therefore it is satisfying to see that this meaningful question finds an answer in a smooth mathematical theory. Moreover, the result is quite sophisticated from the mathematical point of view.

A few historical remarks seem in order. In 1970 the author presented a manuscript entitled "A New Information Theory: Information Transfer at Rates Above Shannon's Capacity" to the late Jack Wolfowitz. Within 24 h Wolfowitz responded with a letter entitled "New Information Theory for Those who Don't Know the Old". He was absolutely right, because the calculation of the error probability for a random encoding procedure used only two-codeword error probabilities and had completely ignored the union bound. Nonetheless, somehow information was conveyed, and in another letter 2 days later, Wolfowitz wrote "The result is perhaps completely useless, but I like it!"

At the Information Theory Workshop at Gränna, Sweden, during a discussion on Yao's two-way communication complexity (see [16]), Ephremides drew attention to a recent unpublished work of Ja'Ja' (see [12]). Immediately, the bell rang. The ancient result had a proper interpretation in the context of identification. The observation of Ja'Ja' is that, for the binary symmetric channel with crossover probability $\epsilon \neq 1/2$, one can identifiy at a rate arbitrarily close to 1. This is immediately clear, if one uses Gilbert's bound for the Hamming distance $d = \delta \cdot n$, $\delta \to 0$ and Hamming spheres of radius equal to $(\epsilon + \eta)n$, $\epsilon < 1/n$, $\eta \ll \epsilon$ as decoding sets.

One can apply the same idea to the general DMC to get a (non-randomized) identification capacity equal to $\log_2$ of the number of distinct row vectors in W. The unsatisfactory aspect of this result is that the actual values of the positive entries in W do not matter.

Our idea to use randomization in the encoding therefore is fruitful in two respects: it leads to much better performance and also eliminates the shortcoming mentioned. Since $\sum_{x^n} Q(x^n|i)W(\mathcal{D}_i|x^n) \geq 1-\lambda$ implies the existence of a u_i with $W(\mathcal{D}_i|u_i) \geq 1-\lambda$, the effect of randomization is on the error of the second kind. For the transmission problem on the DMC, it does not help at all!

It must also be emphasized that even for noiseless channels our result is of interest. Suppose that one out of N possible events occurred. Shannon was concerned with the question, "Which event occurred?" The question asked in identification is "Did event i occur?" Here i could be any member of $\{1, 2, \ldots, N\}$. There are many situations in which the answer to this question is of interest.

Example Let $S_1, \ldots, S_N$ be sailors on a ship, and let sailor S_i be associated with lady L_i. In a stormy night one sailor, say S_j, drowns in the ocean. One could now broadcast his name to the radio stations of the country from which all sailors are known to come, hoping that the lady L_j listens to the news, so that she hears about the tragic event. However, this takes $\lceil \log_2 N \rceil$ bits and the news is (primarily) of interest to only one lady. If we now permit a certain error probability, which is not much of a price in an imperfect (as the tragedy shows) world, then by our result $O(\log \log N)$ bits suffice! ▲

Example In many countries the winning m-digit state lottery number is made public on radio and television. Again, by tolerating a certain error probability, this number could be replaced by a properly produced random number of $O(\log m)$ digits and still every winner and every loser would be informed correctly with probability close to 1. Also one could modify the lottery so that the chance errors become part of the lottery. ▲

These examples show that there is a need for explicit *constructions* of ID codes. If such codes achieve positive second-order rates, then they are already much better than the naive error-free identification codes.

There is a multitude of other problems which can now be studied. Almost every known coding theorem concerning the transmission problem can be reconsidered in the context of identification. Also, new phenomena arise. Chapters "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" and "On Identification via Multi-Way Channels with Feedback: Mystery Numbers" expand the discussion.

References

1. R. Ahlswede, Coloring hypergraphs: a new approach to multi-user source coding, part I. J. Combin. Inform. Syst. Sci. **4**(1), 76–115 (1979)
2. R. Ahlswede, A method of coding and an application to arbitrarily varying channels. J. Comb. Inf. Syst. Sci. **5**(1), 10–35 (1980)
3. R. Ahlswede, Coloring hypergraphs: a new approach to multi-user source coding, part II. J. Combin. Inform. System Sci. **5**(3), 220–268 (1980)
4. R. Ahlswede, G. Dueck, Good codes can be produced by a few permutations. IEEE Trans. Inform. Theory **28**(3), 430–443 (1982)
5. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inform. Theory **35**(1) (1989)
6. R. Ahlswede, G. Dueck, Identification in the presence of feedback – a discovery of new capacity formulas. IEEE Trans. Inform. Theory **35**(1) (1989)
7. A. Alexander, I. Althöfer, C. Deppe, U. Tamm (eds.), *Storing and Transmitting Data Rudolf Ahlswede's Lectures on Information Theory 1*, vol. 10, 1st edn. (Springer, Berlin, 2014). Series: Foundations in Signal Processing, Communications and Networking
8. A. Alexander, I. Althöfer, C. Deppe, U. Tamm (eds.), *Combinatorial Methods and Models, Rudolf Ahlswede's Lectures on Information Theory 4*, vol. 13, 1st edn. (Springer, Berlin, 2018). Series: Foundations in Signal Processing, Communications and Networking
9. I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Channels* (Academic Press, New York, 1981)

10. T.S. Han, S. Verdú, New results in the theory of identification via channels. IEEE Trans. Inform. Theory **38**(1), 14–25 (1992)
11. T.S. Han, S. Verdú, Approximation theory of output statistics. IEEE Trans. Inform. Theory **39**(3), 752–772 (1993)
12. J. Ja'Ja', Identification is easier than decoding. Annu. Symp. Found. Comput. Sci. **26**, 43–50 (1985)
13. U. Krengel, Einführung in die Wahrscheinlichkeitstheorie und Statistik, Vieweg (1991)
14. C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948)
15. J. Wolfowitz, *Coding Theorems of Information Theory*, 3rd ed. (Springer, New York, 1978)
16. A.C. Yao, Some complexity questions related to distributive computing. Proceedings of the 11th Annual ACM Symposium on Theory of Computing, Atlanta, pp. 209–213 (1979)

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

The main contribution of our earlier work, "Identification via Channels", was that $N = \exp\{\exp\{nR\}\}$ objects can be *identified* in block length n with arbitrarily small error probability via a discrete memoryless channel (DMC), if *randomization* can be used for the encoding procedure and if $R < C(W)$. Moreover, in this case the second-order identification capacity equals Shannon's transmission capacity $C(W)$, where W is the transmission matrix of the DMC. Here we study the identification problem in the presence of a *noiseless feedback channel* and determine the second-order capacity C_f (resp. C_F) for deterministic (resp. randomized) encoding strategies. We encounter several important phenomena.

1. Although feedback does not increase the transmission capacity of a DMC, it does increase the (second-order) identification capacity. We actually prove that

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

and

$$C_F(W) = \max_{P} H(P \cdot W), \qquad \text{if } C(W) > 0.$$

2. Notice that $C_f = 0$ if W is a matrix with 0 and 1 as entries only. Thus *noise increases* C_f!
3. The structure of the new capacity formulas is apparently much simpler than Shannon's familiar formula. This has the effect that proofs of converses become easier than in our previous work.

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_2

1 The Results

In the beginning of chapter "Identification via Channels", we discussed the notions of classical transmission codes and (randomized) identification codes. We refer the reader to chapter "Identification via Channels" for definitions, and start right away with the analogous concepts for discrete memoryless channels with feedback.

Definition 38 A transmission feedback code $\left\{\{f_j, \mathcal{D}_j\} : j = 1, \ldots, M\right\}$ is described as follows. There is given a finite set of messages $\mathcal{M} = \{1, \ldots, M\}$. One of these messages is to be sent over the channel. Message $j \in \mathcal{M}$ is encoded by a (vector-valued) function

$$f_j = \left[f_j^1, f_j^2, \ldots, f_j^n\right]$$

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

The distribution of the RV's Y_t, $t = 1, 2, \ldots, n$ is determined by f_j and W. We denote the probability of receiving $y^n = (y_1, \ldots, y_n) \in \mathcal{Y}^n$, if j has been encoded, by $W^n(y^n|f_j) = W(y_1|f_j^1) \cdot W(y_2|f_j^2(y_1)) \cdots W(y_n|f_j^n(y_1 \cdots y_n))$. Again the $\mathcal{D}_j \subset \mathcal{Y}^n$, $j = 1, \ldots, M$ are disjoint decoding sets and we require that

$$W^n(\mathcal{D}_j|f_j) \geq 1 - \lambda, \qquad \text{for all } j = 1, \ldots, M.$$

Now let $M_f(n, \lambda)$ be the maximal integer M for which an (n, M, λ) feedback code exists.

Theorem 39 (Shannon–Kemperman–Kesten (SKK))

$$\lim_{n\to\infty} \frac{1}{n} \log M_f(n, \lambda) = C \qquad \textit{for all } \lambda \in (0, 1).$$

The proof and the apportionment of the credit for it can be found in [5] and [3].

Remark It is also known that randomization in the encoding or/and decoding does not increase the capacity.

Now let us turn again to the identification problem. We consider two concepts, deterministic and randomized identification-feedback (IDF) codes, and make the following important observations:

1. Even in the deterministic case, feedback causes the maximal code length to grow doubly exponentially in block length.
2. If, in addition, we allow randomization in the encoding, this results in a further improvement to the extent that the aforementioned double exponent increases.

3. In both cases the capacities are characterized in terms of entropy measures. Mutual information, however, plays no role!
4. The formulas for the capacities show that "noise" typically increases capacity!

We now formulate the exact results. Let $\mathcal{F}_n$ be the set of all possible encoding functions of the kind defined in Definition 38. A (deterministic) (n, N, λ) IDF code for W is a system

$$\{(f_i, \mathcal{D}_i) : i = 1, \ldots, N\} \qquad \text{with } f_j \in \mathcal{F}_n, \mathcal{D}_i \subset \mathcal{Y}^n, \quad \text{for all } i \in \{1, \ldots, N\}$$

and

$$W^n(\mathcal{D}_i^c | f_i) \leq \lambda, \qquad W^n(\mathcal{D}_j | f_i) \leq \lambda \tag{1}$$

for all $i, j \in \{1, \ldots, N\}$ with $i \neq j$. A randomized (n, N, λ) IDF code for W is a system

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

with $Q_F(\cdot|i) \in \mathcal{P}(\mathcal{F}_n), \mathcal{D}_i \subset \mathcal{Y}^n$, and

$$\sum_{g \in \mathcal{F}_n} Q_F(g|i) W^n(D_i^c | g) \leq \lambda, \tag{2}$$

$$\sum_{g \in \mathcal{F}_n} Q_F(g|i) W^n(\mathcal{D}_i | g) \geq \lambda \tag{3}$$

for all $i, j \in \{1, \ldots, N\}$ with $i \neq j$.

Let $N_f(n, \lambda)$ (resp. $N_F(n, \lambda)$) be the maximal integer N for which a deterministic (resp. randomized) (n, N, λ) IDF code exists. (We add f (resp. F) to the notation to indicate the model with which we are working.)

Theorem 40 (Coding Theorem and Strong Converse) *If the transmission capacity C of W is positive, then we have for all $\lambda \in (0, 1/2)$:*

(i) $\liminf_{n\to\infty} \frac{1}{n} \log\log N_f(n, \lambda) \geq \max_{x \in \mathcal{X}} H(W(\cdot|x))$

(ii) $\limsup_{n\to\infty} \frac{1}{n} \log\log N_f(n, \lambda) \leq \max_{x \in \mathcal{X}} H(W(\cdot|x))$.

In particular, for deterministic feedback strategies the second-order identification capacity $C_f(W)$ equals $\max H(W(\cdot|x))$ provided that $C(W) > 0$. $C_f(W) = 0$ if and only if $C(W) = 0$ or W is a noiseless channel, i.e. $W(y|x) \in \{0, 1\}$ for all x, y. This result says that $C_f(W)$ depends solely on the maximal per letter "output entropy" $H(W(\cdot|x^*)) = \max_{x \in \mathcal{X}} H(W(\cdot|x))$.

Also, C_f increases if $H(W(\cdot|x^*))$, "the measure of noise caused by x^*", increases. Indeed, for noiseless channels, C_f is zero.

This behavior is in surprising contrast to the familiar properties of the transmission capacity. The reader will gain a complete understanding in the course of the proof of part (i) of Theorem 40; here we give some of the underlying ideas.

In chapter "Identification via Channels" we showed that a large amount of randomization in the encoding is necessary to achieve a positive doubly exponential rate. In case of feedback, the sender has another way of performing a random experiment, namely, to send (possibly repeatedly) a letter x with $H(W(\cdot|x)) > 0$. Its outcome is known to the sender via the feedback link. The maximal amount of randomness is achieved if one uses a letter $x^* \in \mathcal{X}$ with

$$H(W(\cdot|x^*)) = \max_x H(W(\cdot|x)).^1$$

The proof of Theorem 40 shows that all good deterministic encoding strategies use such letters x^* most of the time. The situation here is quite different from what we are used to in classical coding problems. As a consequence there is almost no connection between the capacities C_f and C.

However, if we allow randomized feedback strategies, then by Theorem 12 we know that $C_F \geq C$. Actually, strict inequality holds here except for those cases which are specified in the remark below.

Theorem 41 (Coding Theorem and Strong Converse) *If the transmission capacity C of W is positive, then, for all $\lambda \in (0, 1/2)$,*

(i) $$\liminf_{n\to\infty} \frac{1}{n} \log\log N_F(n, \lambda) \geq \max_{P\in\mathcal{P}(\mathcal{X})} H(PW)$$

(ii) $$\limsup_{n\to\infty} \frac{1}{n} \log\log N_F(n, \lambda) \leq \max_{P\in\mathcal{P}(\mathcal{X})} H(PW)$$

Remark We call W essentially noiseless if there exist subsets $\mathcal{X}^* \subset \mathcal{X}, \mathcal{Y}^* \subset \mathcal{Y}$ and a bijection $g : \mathcal{X}^* \to \mathcal{Y}^*$ such that

$$W(g(x)|x) = 1, \qquad \text{for all } x \in \mathcal{X}^* \tag{4}$$

$$W(\cdot|x') \in \text{conv}\{W(\cdot|x) : x \in \mathcal{X}^*\}, \qquad \text{for all } x' \in \mathcal{X}, \tag{5}$$

where conv denotes the convex hull. We claim that $C_F = C$ if and only if $C = 0$ or W is essentially noiseless.

If W is essentially noiseless and $C > 0$ then $C = \log|\mathcal{X}^*|$ and $C_F = \log|\mathcal{Y}^*| = C$. Conversely, if $C = C_F$ and $C > 0$, then for every P' satisfying

$$I(P', W) = H(P'W) - H(W|P) = C,$$

[1] The results explain why, to identify the state of the world in a universal philosophical system, one has to proceed as follows: first choose your position and then create a lot of noise.

we have $H(W|P') = 0$. This and the optimality of P' imply that W is essentially noiseless.

Remark We make some comments concerning the proofs. In chapter "Identification via Channels" we built ID codes from large subsets of a given channel code (for transmission). We showed that Theorem 12(i) can be proved in another simple manner. The ID code is "combined" from two ordinary transmission channel codes. The first one has the sole purpose of *providing sender and receiver with the (common) knowledge of the outcome of a random experiment*. Its entropy per time unit determines the second-order rate of the ID code. This important observation also makes the role of feedback for identification transparent. Feedback makes it possible to provide sender and receiver with the knowledge of the outcome of other random experiments. In the deterministic case it is the experiment obtained by sending the letter x^* n times and in the case of randomized feedback strategies it is the experiment $(\mathcal{Y}^n, \prod_1^n PW)$, which can be performed by sending the outcome of $(\mathcal{X}^n, P^n)$ over the channel. Notice that $H(PW) = I(P, W) + H(W|P)$. Theorem 41 says that the doubly exponential rates $I(P, W)$ (achievable with randomization and no feedback) and $H(W|P)$ (achievable with feedback and no randomization) sum up to $H(PW)$ (achievable with both feedback and randomization). We choose of course a P, which maximizes $H(PW)$. The proofs of the converses essentially say that common random experiments of higher per-letter entropies do not exist under the respective circumstances.

For the second code used in the proofs of the direct part, it is only essential that its rate is positive. Thus the condition $C > 0$ enters. It can be seen by inspection of the proof that, as long as $C > 0$, an infinite identification capacity can be achieved, if sender and receiver have knowledge of the outcome of the same random experiment of an infinite entropy. It is well-known that such random experiments (also with finite entropy) can be used to increase the transmission capacity of systems of channels such as arbitrarily varying channels [1]. Their effect on the identification capacity is dramatic!

2 Notation and Known Facts

For the basic notation we again refer the reader to chapter "Identification via Channels", Sect. 1.1. We state here only two additional simple lemmas. For channels $V, V' \in \mathcal{W}$ let

$$||V - V'|| = \max_{x,y} |V(y|x) - V'(y|x)|.$$

Lemma 42 *For every $\epsilon > 0$, there is a $\delta' = \delta'(\epsilon) > 0$ such that*

$$W^n(\{y^n \in \mathcal{Y}^n : y^n \in \mathcal{T}_V^n(x^n) \text{ for a } V \text{ with } ||V - W|| \leq \epsilon\}|x^n) \leq 1 - 2^{-n\delta'}$$

for $n \geq n_0(\epsilon)$.

Lemma 43 *For every $\epsilon > 0$ there is a $c(\epsilon) > 0$ such that for $n \geq n_0(\epsilon)$*

(i) $\left| \bigcup_{V:||V-W||\leq\epsilon} \mathcal{T}_V^n(x^n) \right| \geq 2^{n(H(W|P_{x^n})-c(\epsilon))}$

(ii) $\left| \bigcup_{V:||V-W||\leq\epsilon} \mathcal{T}_V^n(x^n) \right| \leq 2^{n(H(W|P_{x^n})+c(\epsilon))}$

(iii) $|\mathcal{T}_V^n(x^n)| \geq 2^{n(H(W|P_{x^n})-c(\epsilon))}$, *if* $||V - W|| \leq \epsilon$ *and* $\mathcal{T}_V^n(x^n) \neq 0$,

and $c(\epsilon) \to 0$ *if* $\epsilon \to 0$.

3 New Proof of the Direct Part in Theorem 12

The proof in chapter "Identification via Channels", Sect. 2 [2] uses in the encoding procedure probability distributions which are uniform distributions on the sets of codewords in some classical channel codes (as defined in chapter "Identification via Channels"). There is a lot of freedom in selecting systems of such codes (see the remark at the end of the section). Here we choose a system consisting of codes, which are extensions of a single channel code. This system is designed so that, with some modifications, it can be used for the feedback case as well. It is again produced by a random selection and allows a fairly simple analysis.

We begin with two fundamental codes $\mathcal{C}'$ and $\mathcal{C}$. By Shannon's coding theorem (stated in [4]) we know that for every $\epsilon > 0, \epsilon < C$, there is a $\delta = \delta(\epsilon) > 0$ and an $n_0(\epsilon)$ such that for $n \geq n_0(\epsilon)$ an $(n, M', 2^{-n\delta})$ code

$$\mathcal{C}' = \{(u'_j, \mathcal{D}'_j) : j = 1, \ldots, M'\} \tag{6}$$

and an $(\lceil\sqrt{n}\rceil, M'', 2^{-\sqrt{n}^{\delta}})$ code

$$\mathcal{C}'' = \{(u''_k, \mathcal{D}''_k) : k = 1, \ldots, M''\} \tag{7}$$

exist with $M' = \lceil 2^{n(C-\epsilon)} \rceil$ and $M'' = \lceil 2^{\epsilon\sqrt{n}} \rceil$.

We use the abbreviation $m = n + \lceil\sqrt{n}\rceil$. Now any family $\{T_i : i = 1, \ldots, N\}$ of maps

$$T_i : \{1, \ldots, M'\} \to \{1, \ldots, M''\}$$

can be used to build an ID code $\{(Q(\cdot|i), \mathcal{D}_i) : i = 1, 2, \ldots, N\}$ from $\mathcal{C}'$ and $\mathcal{C}''$. Here $Q(\cdot|i)$ is the uniform distribution on the set of codewords

$$\mathcal{U}_i = \{u'_j \cdot u''_{T_i(j)} : j = 1, \ldots, M'\} \subset \mathcal{X}^m$$

and

$$\mathcal{D}_i = \bigcup_{j=1}^{M'} \mathcal{D}'_j \times \mathcal{D}''_{T_i(j)}.$$

We choose at random an ID code of such structure in the following way.

For $i \in \{1, 2, \ldots, N\}$ and $j \in \{1, \ldots, M'\}$ let U_{ij} be independent RV's such that U_{ij} takes the value $u'_j \cdot u''_k$ with probability $1/M''$ for $k \in \{1, \ldots, M''\}$. We consider the random sets

$$\overline{\mathcal{U}}_i = \{U_{i1}, \ldots, U_{iM'}\} \qquad \text{for all } i = 1, \ldots, N. \tag{8}$$

The uniform distributions $\overline{Q}(\cdot|i)$ on these sets become random distributions. The random decoding sets are

$$\mathcal{D}(\overline{\mathcal{U}}_i) = \bigcup_{j=1}^{M'} \mathcal{D}(U_{ij}) \tag{9}$$

where

$$\mathcal{D}(U_{ij}) = \mathcal{D}'_j \times \mathcal{D}''_k, \qquad \text{if } U_{ij} = u'_j \cdot u''_k. \tag{10}$$

We now analyze the maximal error performances of $\{(\overline{Q}(\cdot|i), \mathcal{D}(\overline{\mathcal{U}}_i)) : i = 1, \ldots, N\}$. We consider first the errors of the first kind. It is clear from the definitions (6)–(10) that for every realization $\mathcal{U}_i$ of $\overline{\mathcal{U}}_i$

$$\sum_{x^m \in \mathcal{X}^m} Q(x^m|i) W^m(\mathcal{D}(\mathcal{U}_i)^c|x^m) \tag{11}$$

$$= \frac{1}{M'} \sum_{u \in \mathcal{U}_i} W^m(\mathcal{D}(\mathcal{U}_i)^c|u) \tag{12}$$

$$\leq \frac{1}{M'} \sum_{j=1}^{M'} \left(W^n(\mathcal{D}'^c_j|u'_j) + W^{\lceil\sqrt{n}\rceil}(\mathcal{D}''^c_{T_i(j)}|u''_{T_i(j)}) \right) \tag{13}$$

$$\leq 2^{-n\delta} + 2^{-\sqrt{n}\delta}. \tag{14}$$

Thus only errors of the second kind remain to be considered. For this analysis we again use a large deviational approach to bound the probability that there does not exist a realization with a prescribed error of the second kind λ for two indices, without loss of generality say $i = 1, 2$. That bound yields the final result for all

indices i since the probability for the union of events does not exceed the sum of the probabilities of these events. Actually, it suffices to compare the random set $\overline{\mathcal{U}}_2$ with any realization $\mathcal{U}_1$ of $\overline{\mathcal{U}}_1$. Fix $\mathcal{U}_1$ and define for $j = 1, \ldots, M'$

$$\psi_j = \psi_j(\overline{\mathcal{U}}_2) = \begin{cases} 1, \text{ if } \mathcal{U}_{2j} \in \mathcal{U}_1 \\ 0, \text{ otherwise.} \end{cases} \tag{15}$$

Since the RV's U_{2j} are independent, $\psi_1, \ldots, \psi_{M'}$ are also independent. Furthermore, by our definitions

$$\mathbb{E}\psi_j = \frac{1}{M''} \qquad \text{for all } j = 1, \ldots, M'. \tag{16}$$

An elementary calculation shows that, for $M'' = \lceil 2^{\sqrt{n}\epsilon} \rceil$,

$$D(\lambda || 1/M'') \geq \lambda \cdot \sqrt{n} \cdot \epsilon - 1. \tag{17}$$

Therefore Lemma 16 implies the following.

Corollary 44 *For* $\lambda \in (0, 1)$ *and* $1/M'' < \lambda$

$$\Pr\left(\sum_{j=1}^{M'} \psi_j > M' \cdot \lambda \right) \leq 2^{-M'(\lambda\sqrt{n}\epsilon - 1)}.$$

We need one other elementary fact. Suppose that $\overline{\mathcal{U}}_2 = \mathcal{U}_2$ and $u \in \mathcal{U}_1 - \mathcal{U}_2$; then

$$W^m(\mathcal{D}(\mathcal{U}_2)|u) \leq 2^{-n\delta} + 2^{-\sqrt{n}\delta}. \tag{18}$$

To see this, let $u = u'_j \cdot u''_k$. Notice that for $u \notin \mathcal{U}_2$, $\mathcal{D}(\mathcal{U}_2) \cap (\mathcal{D}'_j \times \mathcal{D}''_k) = \emptyset$ and that therefore

$$W^m(\mathcal{D}(\mathcal{U}_2)|u) \leq W^m((\mathcal{D}'_j \times \mathcal{D}''_k)^c|u).$$

Equation (18) follows, because the Definitions 6 and 7 imply that

$$W^m((\mathcal{D}'_j \times \mathcal{D}''_k)^c|u) \leq 2^{-n\delta} + 2^{-\sqrt{n}\delta}.$$

An upper bound on the error of the second kind is now readily established:

$$\sum_{u\in\mathcal{U}_1} W^m(\mathcal{D}(\overline{\mathcal{U}}_2)|u) = \sum_{u\in\mathcal{U}_1\cap\overline{\mathcal{U}}_2} W^m(\mathcal{D}(\overline{\mathcal{U}}_2)|u) + \sum_{u\in\mathcal{U}_1-\overline{\mathcal{U}}_2} W^m(\mathcal{D}(\overline{\mathcal{U}_2})|u)$$

$$\leq |\mathcal{U}_1\cap\overline{\mathcal{U}}_2| + \sum_{u\in\mathcal{U}_1-\overline{\mathcal{U}}_2} W^m(\mathcal{D}^c(\mathcal{U}_1-\overline{\mathcal{U}}_2)|u)$$

$$\leq |\mathcal{U}_1\cap\overline{\mathcal{U}}_2| + |\mathcal{U}_1-\overline{\mathcal{U}}_2|\cdot(2^{-n\delta}+2^{-\sqrt{n}\delta})$$

$$\leq |\mathcal{U}_1\cap\overline{\mathcal{U}}_2| + M'\cdot(2^{-n\delta}+2^{-\sqrt{n}\delta})$$

$$\leq |\mathcal{U}_1\cap\overline{\mathcal{U}}_2| + M'\ \cdot 2\cdot 2^{-\sqrt{n}\delta},$$

where we have used (18). Since $|\mathcal{U}_1\cap\overline{\mathcal{U}}_2| = \sum_{j=1}^{M'}\psi_j(\overline{\mathcal{U}}_2)$,

$$\frac{1}{M'}\sum_{u\in\mathcal{U}_1} W^m(\mathcal{D}(\overline{\mathcal{U}}_2)|u) \leq \frac{1}{M'}\sum_{j=1}^{M'}\psi_j(\overline{\mathcal{U}}_2) + 2\cdot 2^{-\sqrt{n}\delta}. \tag{19}$$

Now fix $\lambda \in (0,1)$. By Corollary 44 for large n we have that with *positive* probability

$$\frac{1}{M'}\sum_{u\in\mathcal{U}_1} W^m(\mathcal{D}(\overline{\mathcal{U}}_2)|u) \leq \lambda + 2\cdot 2^{-\sqrt{n}\delta} \tag{20}$$

and similarly

$$\frac{1}{M'}\sum_{u\in\overline{\mathcal{U}}_2} W^m(\mathcal{D}(\mathcal{U}_1)|u) \leq \lambda + 2\cdot 2^{-\sqrt{n}\delta}. \tag{21}$$

Hence there is a realization $\overline{\mathcal{U}}_2 = \mathcal{U}_2$ for which (20) and (21) hold. We use this argument repeatedly for $i = 3,4,\ldots,N$ (as in chapter "Identification via Channels").

$$\Pr\{(n,N,\lambda+2\cdot 2^{-\sqrt{n}\delta}) \quad \not\ni\} = \Pr\left\{\bigcup_{\substack{l,k=1,\ldots,N\\ l\neq k}} \lambda_n^{(l,k)} \geq \lambda + 2\cdot 2^{-\sqrt{n}\delta}\right\} \tag{22}$$

$$\leq \sum_{\substack{k,l:\\ l\neq k}} \Pr\{\lambda_n^{(l,k)} \geq \lambda + 2\cdot 2^{-\sqrt{n}\delta}\} \tag{23}$$

$$\leq (N^2-1)\cdot 2^{-M'(\lambda\sqrt{n}\epsilon-1)} \tag{24}$$

Therefore, an $(n, N, \lambda + 2 \cdot 2^{-\sqrt{n}\delta})$ ID code exists, if

$$N(N-1)\Pr\left(\sum_{j=1}^{M'} \psi_j > M'\lambda\right) < 1. \tag{25}$$

From Corollary 44 and $M' = \lceil 2^{n(C-\epsilon)} \rceil$ (25) holds for every N with

$$N \leq 2^{1/2\cdot(\lambda\sqrt{n}\epsilon - 1)2^{n(C-\epsilon)}}.$$

This proves the result.

Remark Instead of extending the code $\mathcal{C}'$, one can prove the same result by making a random selection of subcodes of $\mathcal{C}'$ whose lengths are small but proportional to $|\mathcal{C}'|$.

4 Proof of the Direct Part of Theorem 40

We know already from chapter "Identification via Channels" that randomization in the encoding causes $N(n, \lambda)$ to grow doubly exponentially in n. In the preceding proof we gained additional insight. The amount of "correlated randomization", that is, the size of a random experiment, whose outcomes are known to the sender and to the receiver (with very small error probability), is the decisive quantity determining the growth of $N(n, \lambda)$.

As our random experiment we used the uniform distribution on the set of codewords of the code $\mathcal{C}'$. The outcome $u'_j \in \{u'_1, \ldots, u'_{M'}\}$ is known to the sender. Then the outcome is transmitted to the receiver with high probability. The parameter $M' = \lceil 2^{n(C-\epsilon)} \rceil$ is the size of this random experiment.

The presence of feedback allows the design of another random experiment. Feedback is used here solely for this purpose. Otherwise the coding scheme is essentially the same as previously. We now describe this random experiment and the coding scheme. Let $x^* \in \mathcal{X}$ be a letter with

$$H(W(\cdot|x^*)) = \max_{x\in\mathcal{X}} H(W(\cdot|x)). \tag{26}$$

Choose again as total block length

$$m = n + \lceil \sqrt{n} \rceil \tag{27}$$

and define $\mathcal{C}''$ as in (7). We now describe the substitute for $\mathcal{C}'$.

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

The resulting correlated random experiment $(\mathcal{Y}^n, W^n(\cdot|x^{*n}))$ needs a modification, because $W^n(\cdot|x^{*n})$ is far from being uniform on $\mathcal{Y}^n$. However, $W^n(\cdot|x^{*n})$ is essentially uniform on the set

$$\mathcal{D}^* = \bigcup_{V:||V-W||\leq\epsilon} \mathcal{T}_V^n(x^{*n}), \tag{28}$$

which carries essentially all its probability. We lump the small-probability set $\mathcal{Y}^n - \mathcal{D}^*$ together in an erasure symbol e with the understanding that

$$W^n(e|x^{*n}) = W^n(\mathcal{Y}^n - \mathcal{D}^*|x^{*n}). \tag{29}$$

We choose as our random experiment $(\mathcal{D}^* \cup \{e\}, W^n(\cdot|x^{*n}))$. The price paid for more uniformity is a small error probability, if e occurs. However, previously we still had to deal in $\mathcal{C}'$ with small error probabilities. By Lemma 43, $|\mathcal{D}^*| \sim 2^{nH(W(\cdot|x^*))}$, and this quantity now takes the role of M'.

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

In case $y^n \notin \mathcal{D}^*$ an error is declared and the sender can fill the $\lceil n^{1/2} \rceil$ positions in any way, for instance by sending $x^* \lceil n^{1/2} \rceil$ times again. Clearly, for each F_1, we have defined an encoding function $f_i \in F_m$ as introduced in Sect. 1. For the decoding we define the sets

$$\mathcal{D}(F_i) = \bigcup_{y^n \in \mathcal{D}^*} \{y^n\} \times \mathcal{D}''_{F_i(y^n)}, \qquad \text{for all } i = 1, \ldots, N. \tag{30}$$

The astute reader can avoid the following formal analysis, which is necessary only because our random experiment is not exactly uniform.

With respect to the error of the first kind notice that $W^m(\mathcal{D}(F_i)^c|f+i) \leq W^n((\mathcal{D}^*)^c|x^{*n}) + 2^{-\sqrt{n}\delta}$, and thus by Lemma 43,

$$W^m(\mathcal{D}(F_i)^c|f_i) \leq 2^{-n\delta'} + 2^{-\sqrt{n}\delta}. \tag{31}$$

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

For $i \in \{1, 2, \ldots, N\}$ and $y^n \in \mathcal{D}^*$ let $\overline{F}_i(y^n)$ be independent RV's such that $\overline{F}_i(y^n)$ takes every value $k \in \{1, \ldots, M''\}$ with probability $1/M''$. Let F_1 be any realization of $\overline{F}_1$.

In analogy to the ψ_j in Sect. 3, we define RV's $\psi_{y^n} = \psi_{y^n}(\overline{F}_2)$ for every $y^n \in \mathcal{D}^*$ by

$$\psi_{y^n} = \begin{cases} 1, \text{ if } F_1(y^n) = \overline{F}_2(y^n) \\ 0, \text{ otherwise.} \end{cases} \tag{32}$$

These RV's are independent and have expected value $1/M''$. Application of Lemma 16 in conjuncation with Lemma 43 yields the following.

Corollary 45 *For $\lambda \in (0,1)$, $1/M'' < \lambda$, and for a channel V with $||V - W|| \leq \epsilon$,*

$$\Pr\left(\sum_{\substack{y^n \in \\ \mathcal{T}_V^n(x^{*n})}} \psi_{y^n} > |\mathcal{T}_V^n(x^*n)| \cdot \lambda\right) \leq \exp\{-\exp\{nH(W(\cdot|x^{*n})) - c(\epsilon))\} \cdot (\lambda\sqrt{n}\epsilon - 1)\},$$

if $n \geq n_0(\epsilon)$. Consequently, with probability at least

$$1 - (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|} \cdot \exp\{-\exp\{nH(W(\cdot|x^{*n})) - c(\epsilon))\} \cdot (\lambda\sqrt{n}\epsilon - 1)\}$$

$\overline{F}_2$ satisfies, for all V with $||V - W|| \leq \epsilon$,

$$\sum_{y^n \in \mathcal{T}_V^n(x^{*n})} \psi_{y^n} \leq |\mathcal{T}_V^n(x^{*n})| \cdot \lambda. \tag{33}$$

We now derive an upper bound on $W^m(\mathcal{D}(\overline{F}_2)|f_1)$ for those values of $\overline{F}_2$:

$$\begin{aligned} W^m\left(\mathcal{D}(\overline{F}_2)|f_1\right) &\leq W^m\left((\mathcal{D}^* \times \mathcal{Y}^{\lceil\sqrt{n}\rceil})^c|f_1\right) + W^m\left((\mathcal{D}^* \times \mathcal{Y}^{\lceil\sqrt{n}\rceil}) \cup \mathcal{D}(\overline{F}_2)|f_1\right) \\ &\leq W^n\left((\mathcal{D}^*)^c|x^{*n}\right) \\ &\quad + \sum_{\substack{y^n \in \mathcal{D}^* \\ F_1(y^n) \neq \overline{F}_2(y^n)}} W^n(y^n|x^{*n}) \cdot 2^{-\sqrt{n}\delta} + \sum_{\substack{y^n \in \mathcal{D}^* \\ F_1(y^n) = \overline{F}_2(y^n)}} W^n(y^n|x^{*n}). \end{aligned}$$

By Lemma 42 we have $W^n(\mathcal{D}^*|x^{*n}) \geq 1 - 2^{-n\delta'}$.

The second summand is obviously not larger than $2^{-\sqrt{n}\delta}$. For an upper bound on the third summand we use (33). We get

$$\begin{aligned} W^m\left(\mathcal{D}(\overline{F}_2)|f_1\right) &\leq 2^{-n\delta'} + 2^{-\sqrt{n}\delta} + \sum_{V:||V-W||\leq\epsilon} \frac{W^n\left(\mathcal{T}_V^n(x^{*n})|x^{*n}\right)}{\left|\mathcal{T}_V^n(x^{*n})\right|} \cdot \sum_{y^n \in \mathcal{T}_V^n(x^{*n})} \psi_{y^n} \\ &\leq 2^{-n\delta'} + 2^{-\sqrt{n}\delta} + \lambda. \end{aligned}$$

The same arguments yield the same bound for

$$W^m(\mathcal{D}(F_1)|\overline{f}_2),$$

if $\overline{f}_2$ denotes the encoding function defined by the map $\overline{F}_2$. We repeatedly use this argument as in Sect. 3 and construct a code length N satisfying

$$N \geq (n+1)^{-2|\mathcal{X}|\cdot|\mathcal{Y}|} \cdot \exp\{\exp\{nH(W(\cdot|x^*)) - c(\epsilon))\} \cdot (\lambda\sqrt{n}\epsilon - 1)\}$$

and an error of the second kind less than $2^{-n\delta'} + 2^{-\sqrt{n}\delta} + \lambda$.

5 Proof of the Direct Part of Theorem 41

Since now randomization in the encoding and feedback are available, we can combine the two kinds of random experiments for the proofs of the direct parts in Theorems 12 and 40, respectively. Of course such a combination imposes restrictions to the effect that now doubly exponential capacities $\max_P I(P, W)$ and $\max_x H(W(\cdot|x)) = \max_P H(W|P)$ do not simply add. Instead, the capacity is now given by

$$\max_P (I(P, W) + H(W|P)) = \max_P H(PW). \tag{34}$$

To show this, choose a P^* such that for $Q^* = P^*W$, $H(Q^*) = \max_P H(PW)$ and define as random experiment

$$\left(\left(\bigcup_{Q:||Q-Q^*||\leq\epsilon} \mathcal{T}_Q^n\right) \cup \{e\}, Q^{*n}\right).$$

This can be realized as follows. The sender chooses a sequence x^n according to the random experiment $(\mathcal{X}^n, P^{*n})$ and sends it over the channel. $Q^{*n}(y^n)$ is the probability for receiving y^n. This sequence is also known to the sender via feedback. We can therefore substitute in the previous proof $\mathcal{D}^*$ by

$$\mathcal{D}^{**} = \bigcup_{Q:||Q-Q^*||\leq\epsilon} \mathcal{T}_Q^n \tag{35}$$

and get Theorem 41(i).

6 Proof of the Converse Part of Theorem 40

We have already mentioned that in the case of feedback the proofs of the converses become much simpler than the proofs in chapter "Identification via Channels". We need here only one auxiliary result.

Lemma 46 (Image Size for a Deterministic Feedback Strategy) *For any n-length feedback strategy f and any $\nu \in (0,1)$,*

$$\min_{\mathcal{E}\subset\mathcal{Y}^n : W^n(\mathcal{E}|f)\geq 1-\nu} |\mathcal{E}| \leq K = 2^{nH(W(\cdot|x^*))+\alpha\sqrt{n}} \tag{36}$$

where $H(W(\cdot|x^*)) = \max_{x\in\mathcal{X}} H(W(\cdot|x))$, $\alpha = \log(\beta)/\sqrt{\nu}$, and $\beta = \max(3, |\mathcal{Y}|)$.

Before we prove Lemma 46 we show that it implies Theorem 40(ii).

Let $\{(f_i, \mathcal{D}_i) : 1 \leq i \leq N\}$ be an (n, N, δ) IDF code with $\lambda \in (0, 1/2)$. We can choose ν such that $1 - \nu - \lambda > 1/2$. For f_i let $\mathcal{E}_i$ be a set for which the minimum is assumed in (36). Thus we have $W^n(\mathcal{D}_i \cap \mathcal{E}_i|f_i) > 1/2$ and the sets $\mathcal{D}_i \cap \mathcal{E}_i$, $i = 1, 2, \ldots, N$, are necessarily distinct because the errors of the second kind are smaller than $\lambda < 1/2$. Therefore by Lemma 46, $N \leq \sum_{k=0}^{K} \binom{|\mathcal{Y}|^n}{k} \leq 2^{n \log|\mathcal{Y}|\cdot K}$ and Theorem 40(ii) follows.

Proof of Lemma 46 The cardinality of the set

$$\mathcal{E}^* = \{y^n | -\log W^n(y^n|f) \leq \log K\}$$

is clearly smaller than K, and it suffices to show that $W^n(\mathcal{E}^*|f) \geq 1 - \nu$. For this we first give another description of $W^n(\mathcal{E}^*|f)$. Strategy f induces the RV's $Y^s = (Y_1, \ldots, Y_s)$, $s = 1, \ldots, n$, with distributions

$$\Pr(Y^s = y^s) = W^s(y^s|f), \qquad y^s \in \mathcal{Y}^s.$$

Defining $Z_t = -\log W(Y_t|f(Y^{t-1}))$, we can write

$$W^n(\mathcal{E}^*|f) = \Pr\left(\sum_{t=1}^{n} Z_t \leq \log K\right). \tag{37}$$

We now analyze this expression by considering the conditional expectation $\mathbb{E}(Z_t|Y^{t-1})$.

Since

$$\Pr(Y_t = y_t|Y^{t-1} = y^{t-1}) = W(y_t|f(y^{t-1})),$$

we have for $y^{t-1} \in \mathcal{Y}^{t-1}$,

$$\mathbb{E}(Z_t|y^{t-1}) = -\sum_{y_t\in\mathcal{Y}} W(y_t|f(y^{t-1})) \log W(y_t|f(y^{t-1})) \leq H(W(\cdot|x^*)),$$

and therefore

$$\mathbb{E}(Z_t|y^{t-1}) \leq H(W(\cdot|x^*)). \tag{38}$$

Finally, we introduce the RV's

$$U_t = Z_t - \mathbb{E}(Z_t|Y^{t-1}), \tag{39}$$

which obviously satisfy

$$\mathbb{E}(U_t|Y^{t-1}) = 0, \qquad \mathbb{E}U_t = 0. \tag{40}$$

Moreover, since U_s is a function of $Y_1, \ldots, Y_s$, this implies $s < t$, $\mathbb{E}(U_t|U_s) = 0$. Therefore, the RV's $U_1, \ldots, U_n$ are uncorrelated, i.e.,

$$\mathbb{E}U_sU_t = 0, \qquad \text{for all } s \neq t. \tag{41}$$

Notice that (37)–(40) and the definition of K imply

$$W^n(\mathcal{E}^*|f) \geq \Pr\left(\sum_{t=1}^{n} U_t \leq \alpha\sqrt{n}\right). \tag{42}$$

By Chebyshev's inequality,

$$\Pr\left(\sum_{t=1}^{n} U_t \leq \alpha\sqrt{n}\right) \geq 1 - \nu$$

provided that

$$\operatorname{var} U_t \leq \beta, \qquad \text{for all } t = 1, 2, \ldots, n. \tag{43}$$

Verification of (43) completes the proof.

Using (40) we can write

$$\operatorname{var} U_t = \mathbb{E}U_t^2 = \mathbb{E}(U_t - \mathbb{E}(U_t|Y^{t-1}))^2$$

$$= \sum_{y^t} \Pr(Y^{t-1} = y^{t-1}) \cdot \mathbb{E}\left((U_t - \mathbb{E}(U_t|Y^{t-1}))^2|Y^{t-1} = y^{t-1}\right)$$

and by the well-known minimality property of the expected value this can be upper-bounded by

$$\sum_{y^{t-1}} \Pr(Y^{t-1} = y^{t-1})\mathbb{E}\left((U_t - \mathbb{E}(Z_t|Y^{t-1}))^2|Y^{t-1} = y^{t-1}\right)$$

$$= \sum_{y^{t-1}} \Pr(Y^{t-1} = y^{t-1})\mathbb{E}(Z_t^2|Y^{t-1} = y^{t-1}).$$

By the definition of Z_t

$$\mathbb{E}(Z_t^2|Y^{t-1} = y^{t-1}) = \sum_{y_t=\mathcal{Y}} W(y_t|f(y^{t-1})) \cdot \log^2 W(y_t|f(y^{t-1})).$$

Since $x \log^2 x$ is bounded in $[0, 1]$, this quantity is bounded by a function of $|\mathcal{Y}|$ uniformly in t and y^{t-1}. A Lagrange multiplier argument gives the bound

$$\beta = \max(\log^2 3, \log^2 |\mathcal{Y}|).$$

Thus, $\operatorname{var} U_t \leq \beta$. □

7 Proof of the Converse Part of Theorem 41

The proof is based on the same ideas as the previous one. Here we need the following auxiliary result.

Lemma 47 (Image Size for a Randomized Feedback Strategy) *For any n-length randomized feedback strategy F and any $\nu \in (0, 1)$,*

$$\min_{\mathcal{E}' \subset \mathcal{Y}^n : W^n(\mathcal{E}'|F) \geq 1-\nu} |\mathcal{E}'| \leq K' = 2^{nH(Q')+\alpha\sqrt{n}} \tag{44}$$

where $H(Q') = \max_P H(PW)$, $\alpha = \sqrt{\beta/\nu}$, and $\beta = \max(\log^2 3, \log^2 |\mathcal{Y}|)$.

Replacing Lemma 46, $\mathcal{E}_i$, and K in the derivation of Theorem 40(ii) by Lemma 47 and the corresponding quantities $\mathcal{E}_i'$, K' we get Theorem 41(ii).

Proof of Lemma 47 The randomized strategy F can be viewed as a probability distribution Q_F on the set F_n of n-length deterministic feedback strategies. Therefore,

$$W^n(\mathcal{E}'|F) = \sum_{g \in F_n} Q_F(g) W^n(\mathcal{E}'|g). \tag{45}$$

Q_F induces the RV Y^n with distribution

$$\Pr(Y^n = y^n) = \sum_{g \in F_n} Q_F(g) W^n(y^n|g).$$

We write $Q(y^n) = \Pr(Y^n = y^n)$. The cardinality of the set

$$\mathcal{E}'^* = \{y^n | -\log Q(y^n) \leq \log K'\}$$

is clearly smaller than K', and it suffices to show now that $Q(\mathcal{E}'^*) \geq 1-\eta$. Defining $Z'_t = -\log Q(Y_t|Y^{t-1})$, we can write

$$Q(\mathcal{E}'^*) = \Pr\left(\sum_{t=1}^{n} Z'_t \leq \log K'\right). \tag{46}$$

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

Notice that

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

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

$$Q(y_t|y^{t-1}) = \sum_{g \in \mathcal{F}_n} Q_F(g) \frac{\prod_{i=1}^{t-1} W(y_i|g(y^{i-1}))}{\sum_g Q_F(g) \prod_{i=1}^{t-1} W(y_i|g(y^{i-1}))} \cdot W(y_t|g(y^{t-1})).$$

Therefore we have

$$\mathbb{E}(Z'_t|y^{t-1}) \leq H(Q'). \tag{47}$$

This is the substitute for (43). Otherwise, we continue exactly as before. We define functions

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

which again have the desired properties $\mathbb{E}U'_t = 0$, $\mathbb{E}U'_t U'_s = 0$ for $s \neq t$, and $\operatorname{var} U'_t \leq \beta$. Application of Chebyshev's inequality again establishes the result. □

Remark The method for proving the converse parts of Theorems 40 and 41 resembles the approach of Kemperman [3] for proving the strong converse of the coding theorem for memoryless channels with feedback. This "analytical" approach turns out to be better suited for coding problems involving feedback than the "typical sequences" approach. Other such instances are the coding theory for non-stationary and infinite alphabet channels. In fact, we have alternative proofs for the converse parts of Theorems 40 and 41 via typical sequences, but they are much more complicated.

References

1. R. Ahlswede, Elimination of correlation in random codes for arbitrarily varying channels. Z. Wahrscheinlichkeitstheorie u. verw. Gebiete **44**(2), 159–175 (1978)

2. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inform. Theory **35**(1) (1989)
3. J.H.B. Kemperman, Strong converses for a general memoryless channel with feedback. In: *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Function, Random Processes* (1973), pp. 375–409
4. C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948)
5. J. Wolfowitz, *Coding Theorems of Information Theory*, 3rd ed. (Springer, New York, 1978)

On Identification via Multi-Way Channels with Feedback: Mystery Numbers

"Identification for Multi-way channels" was mentioned by Ahlswede and Dueck [2] (see chapter "Identification via Channels") as a challenging direction of research. In this lecture, based on [4], we present in case of complete feedback a rather unified theory of identification. (For the classical transmission problem the dream of such a theory did not get fulfilled for more than 20 years.) Its guiding principle is the discovery of [3] (chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas), that communicators (sender and receiver) must set up a common random experiment with maximal entropy and use it as randomization for a suitable (see chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas") identification technique. Here we show how this can be done in a constructive way. The proof of optimality (weak converse) is based on a new entropy bound, which can be viewed as a substitute for Fano's Lemma (Lemma 48) in the present context. The "single-letter" characterisation of (second order) capacity regions rests now on a new "entropy characterisation problem", which often can be solved. Here this is done for the multiple-access channel with deterministic and for the broadcast channel with randomized encoding strategies.

1 Introduction

In chapter "Identification via Channels" we have introduced a new model for communication, which we call *identification* (ID), hereby contrasting Shannon's original *transmission* (TR) problem. Whereas in chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" one-way channels with feedback were analysed, we present here, as promised in chapter "Identification via Channels", contributions to the theory of multi-way channels. The discussion concentrates on cases where complete feedback links are present. We establish as an always valid principle the idea of chapter "Identification in the Presence

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_3

of Feedback: A Discovery of New Capacity Formulas", that the average maximal entropies of common random experiments among communicators determine the optimal (second order) identification rates. The achievability proof follows the method of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" to use keys selected by the common random experiment with blocklength n and short, for instance length $\sqrt{n}$, encryptions for the messages to be identified. The wide applicability of this method is due to the fact that this "$\sqrt{n}$ trick" can be applied independently for several users simply by timesharing without an essential loss in rates.

However, the converse proofs of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" use special properties of one-way channels and don't seem to be adaptable to multi-way channels. We present here a *new method* (Lemma 59 under Sect. 6), which yields weak converses for these channels. Its essence is an elementary relation in terms of entropy between the cardinalities of sets and their probabilities in arbitrary discrete probability spaces.

In our second main contribution we show how the encryption method mentioned above can be made constructive (see Sects. 7 and 8). Roughly speaking it improves a suboptimal encryption scheme of Mehlhorn and Schmidt [10] via our idea of an iterative reduction used originally for the TR problem [1, 5]. Finally we emphasize that the determination of the maximal entropies obtainable with common random experiments can be difficult for some channels (see the examples in Sect. 5).

This shows that the theory is not trivial. It cannot be expected from a general and not trivial theory that it gives detailed answers to all special questions. We remind the reader that after the foundation of mechanics there was still no explicit answer to the motion of three bodies. This hint may help to judge the state of our theory. Some examples are discussed in detail. We give now the formal statements of our concepts and results.

2 Review of Known Concepts and Results

Useful tool in many proofs of converses is the following inequality discovered by Fano [8].

Lemma 48 (Fano's Lemma) *Let* $\{(u_i, D_i) : 1 \le i \le N\}$ *be a block code with average error*

$$\lambda_Q \triangleq \sum_{i=1}^{N} Q(i) W(D_i^c | u_i).$$

Further, let U *be a random variable with* $\Pr(U = u_i) = Q(i)$ *and let* V *be a random variable induced by the channel, i.e.,* $\Pr(V = y | U = u_i) = W(y|u_i)$ *for all* $i \in \{1, \ldots, N\}$ *and* $y \in \mathcal{Y}$ *and* $\Pr(V = y) = \sum_{i=1}^{N} Q(i) \cdot W(y|u_i)$. *Then*

$$H(U|V) \le 1 + \lambda_Q \log N.$$

Fano's Lemma states that the conditional entropy is smaller (by a factor λ_Q) than $\log N$, the logarithm of the code length. In [6] the Lemma was discussed in details. The relative entropy is not a metric, it possesses several useful properties which justify to introduce $D(P||Q)$ as the statement in the following lemma.

Lemma 49 (Data Processing Lemma) *Let $\mathcal{X} = \{1, \ldots, a\}$, P and Q be two probability distributions on $\mathcal{X}$ and let $A_1, \ldots, A_t$ be a partition of $\mathcal{X}$ (hence $\mathcal{X} = \bigcup_{i=1}^{t} A_i$ and $A_i \cap A_j = \emptyset$ for $i \neq j$). Then*

$$D(P||Q) = \sum_{x \in \mathcal{X}} P(x) \cdot \log \frac{P(x)}{Q(x)} \geq \sum_{i=1}^{t} P(A_i) \cdot \log \frac{P(A_i)}{Q(A_i)}.$$

We first briefly review concepts concerning identification via one-way channels with feedback. Extensions to multi-way channels then almost suggest themselves. Unless stated otherwise we use the notation of chapters "Identification via Channels" and "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas", in particular, script capitals $\mathcal{X}, \mathcal{Y}, \ldots$ denote finite sets. $|\mathcal{A}|$ stands for the cardinality of set $\mathcal{A}$. The letters P, Q always denote PD's on finite (or countable) sets. $X, Y, \ldots$ are RV's with PD's $P_X, P_Y, \ldots$. $\mathcal{P}(\mathcal{A})$ is the set of all PD's on $\mathcal{A}$. For a stochastic $|\mathcal{X}| \times |\mathcal{Y}|$-matrix W we denote by W^n the transmission probabilities for n-length words of a DMC. Other notions such as entropies and information quantities are either standard or those from chapters "Identification via Channels" and "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas".

Let us now turn to the identification problems of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas". There are two concepts, deterministic and randomized feedback strategies, with corresponding code concepts. The vector-valued function

$$f^n = [f_1, \ldots, f_n] \tag{1}$$

is a deterministic encoding strategy of blocklength n, if $f_1 \in \mathcal{X}$ and $f_t : \mathcal{Y}^{t-1} \to \mathcal{X}$ for $t > 1$. It is understood that after the received elements $Y_1, \ldots, Y_{t-1}$ have been made known to the sender by the feedback channel, the sender transmits $f_t(Y_1, \ldots, Y_{t-1})$. At $t = 1$ the sender transmits f_1.

The distribution of the RV's $Y_t (t = 1, 2, \ldots)$ is determined by f and W. We denote the probability of receiving $y^n = (y_1, \ldots, y_n) \in \mathcal{Y}^n$ by $W^n(y^n|f) = W(y_1|f_1) \cdot W\big(y_2|f_2(y_1)\big) \ldots W\big(y_n|f_n(y_1, \ldots, y_{n-1})\big)$.

Let $\mathcal{F}_n^{\mathrm{d}}$ be the set of all possible encoding functions as defined in (1).

Definition 50 A deterministic (n, N, λ) IDF code for W is a system

$$\{(f_i^n, \mathcal{D}_i) | i = 1, 2, \ldots, N\}$$

with $f_i^n \in \mathcal{F}_n^{\mathrm{d}}$, $\mathcal{D}_i \subset \mathcal{Y}^n$ and for $i \in \{1, 2, \ldots, N\}$

$$W^n(\mathcal{D}_i^c | f_i^n) \leq \lambda$$

and for all $i, j \in \{1, 2, \ldots, N\}$ with $i \neq j$

$$W^n(\mathcal{D}_j | f_i^n) \leq \lambda.$$

Definition 51 A randomized (n, N, λ) IDF code for W is a system

$$\{(Q_F(\cdot|i), \mathcal{D}_i) | i = 1, 2, \ldots, N\}$$

with $Q_F(\cdot|i) \in \mathcal{P}(\mathcal{F}_n^{\mathrm{d}})$, $\mathcal{D}_i \subset \mathcal{Y}^n$, and with

$$\sum_{g \in \mathcal{F}_n} Q_F(g|i) W^n(\mathcal{D}_i^c | g) \leq \lambda$$

$$\sum_{g \in \mathcal{F}_n} Q_F(g|j) W^n(\mathcal{D}_i | g) \leq \lambda$$

for all $i, j \in \{1, 2, \ldots, N\}$ with $i \neq j$.

Let $N_d(n, \lambda)$ (resp. $N_r(n, \lambda)$) be the maximal N for which a deterministic (resp. randomized) (n, N, λ) IDF code exists. We summarize the results of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" as follows.

Theorem 52 (Coding Theorems and Strong Converses) *If the transmission capacity C of W is positive, then, for all $\lambda \in (0, \frac{1}{2})$, we have:*

(i) $\lim_{n \to \infty} \frac{1}{n} \log \log N_d(n, \lambda) = \max_{x \in \mathcal{X}} H\big(W(\cdot|x)\big)$

(ii) $\lim_{n \to \infty} \frac{1}{n} \log \log N_r(n, \lambda) = \max_{P \in \mathcal{P}(\mathcal{X})} H(PW).$

In identification the receiver does not necessarily want to know the message $i \in \mathcal{N} = \{1, 2, \ldots, N\}$ given to the sender, he only wants to know the answer to the question "Is it $\hat{\imath}$?". Here $\hat{\imath}$ could be any member of $\mathcal{N}$. The quantities $e_1^n = 1 - \min_{i \in \mathcal{N}} W^n(\mathcal{D}_i | f_i^n)$ and $e_2^n = \max_{i \neq \hat{\imath}} W^n(\mathcal{D}_{\hat{\imath}} | f_i^n)$ are called first kind and second kind error probabilities. (n, N, λ) IDF codes guarantee these quantities not to exceed λ. The quantity $\frac{1}{n} \log \log N$ is called (second order) rate of the code.

Clearly, analogous definitions can be given for multi-way channels. The receivers want to identify with small error probabilities of both kinds. Senders may or may not be allowed to randomize. Achievable rates are replaced by (second order) capacity regions.

Insofar we have spoken about multi-way channels without very specific definitions. We describe now a sufficiently general class, introduce then mystery numbers and use them to characterize capacity regions.

3 A General Model for Communication Systems

To describe a communication system in general, we introduce the following parameters:

(i) Ω, the set of terminals : at each terminal $\omega \in \Omega$ information can be sent and/or received;
(ii) Γ, the set of messengers : for each $\gamma \in \Gamma$, there will be a message set $\mathcal{N}_\gamma$.
The situation at each terminal $\omega \in \Omega$ is further described by
(iii) $\mathcal{A}_\omega \subset \Gamma$, the set of messengers reporting to ω;
(iv) $\mathcal{B}_\omega \subset \Gamma$, where $\gamma \in \mathcal{B}_\omega$ indicates that the messages of $\mathcal{N}_\gamma$ should be decoded at $\mathcal{B}_\omega$;
(v) $\Phi_\omega \subset \Gamma$, the set of feedback signals linked back to ω, i.e., $\omega' \in \Phi_\omega$ indicates that all symbols received at ω' are also available at ω.

Finally, the communication between terminals is governed by a discrete memoryless channel matrix , i.e., a stochastic matrix

$$W : \prod_{\omega\in\Omega} \mathcal{X}_\omega \longrightarrow \prod_{\omega\in\Omega} \mathcal{Y}_\omega,$$

with input alphabets $\{\mathcal{X}_\omega\}_{\omega\in\Omega}$ and output alphabets $\{\mathcal{Y}_\omega\}_{\omega\in\Omega}$. Notice that we assume an input and output alphabet at each terminal. However, allowing $|\mathcal{X}_\omega| = 1$ or $|\mathcal{Y}_\omega| = 1$, resp., we can effectively model also situations where ω only receives or sends signals, respectively.

The reader can convince himself that the following axioms provide plausible assumptions:

A_1 : $\mathcal{A}_\omega \cap \mathcal{B}_\omega = \emptyset$ and $\bigcup_{\omega\in\Omega} \mathcal{A}_\omega = \bigcup_{\omega\in\Omega} \mathcal{B}_\omega = \Gamma$;
A_2 : $\max\{|\mathcal{X}_\omega|, |\mathcal{Y}_\omega|\} \geq 2$;
A_3 : if $|\mathcal{X}_\omega| = 1$, then $\mathcal{A}_\omega = \emptyset$; if $|\mathcal{Y}_\omega| = 1$, then $\mathcal{B}_\omega = \emptyset$;
A_4 : if $\mathcal{A}_\omega = \emptyset$ and $|\mathcal{X}_\omega| \geq 2$ then $|\mathcal{Y}_\omega| \geq 2$; if $\mathcal{B}_\omega = \emptyset$ and $|\mathcal{Y}_\omega| \geq 2$ then $|\mathcal{X}_\omega| \geq 2$;

and as a convention to simplify notation further on, we also assume

A_5 : $\omega \in \Phi_\omega$.

These definitions and axioms define a general discrete memoryless communication system .

We will restrict our attention to the class of systems with *supervisory feedback*, i.e., where for all $\omega, \omega' \in \Omega$ it holds that

A_6 : if $\mathcal{A}_\omega \cap \mathcal{B}_{\omega'} \neq \emptyset$, then $\Phi_{\omega'} \subset \Phi_\omega$.

This assures each terminal encoding γ's messages ($\gamma \in \Gamma$) of at least all the output signals that are known at the terminals decoding these messages.

The set of decoders Δ is defined by

$$\Delta \triangleq \{\omega \in \Omega \,|\mathcal{B}_\omega \neq \emptyset\}. \tag{2}$$

We mainly consider the case of *passive decoders*, i.e., where

A_7 : for all $\omega \in \Delta$, $|\mathcal{X}_\omega| = 1$,

to avoid decoders to influence the communication.

To illustrate our model, we state the following explicit communication systems (CS):

CS 1. one-way channel: $\Omega = \{1,2\}$, $|\Gamma| \geq 1$, $\mathcal{A}_1 = \Gamma$, $\mathcal{A}_2 = \emptyset$, $\mathcal{B}_1 = \emptyset$, $\mathcal{B}_2 = \Gamma$, $\Phi_1 = \{1,2\}$, $\Phi_2 = \{2\}$, $\mathcal{X}_2 = \mathcal{Y}_1 = \{0\}$, and $W : \mathcal{X}_1 \times \{0\} \to \{0\} \times \mathcal{Y}_2$.

CS 2. multiple-access channel (MAC) : $\Omega = \{1,2,3\}$, $\Gamma = \{a,b\}$, $\mathcal{A}_1 = \{a\}$, $\mathcal{A}_2 = \{b\}$, $\mathcal{A}_3 = \emptyset$, $\mathcal{B}_1 = \emptyset = \mathcal{B}_2$, $\mathcal{B}_3 = \Gamma$, $\Phi_1 = \{1,3\}$, $\Phi_2 = \{2,3\}$, $\Phi_3 = \{3\}$, $\mathcal{X}_3 = \mathcal{Y}_1 = \mathcal{Y}_2 = \{0\}$, and $W : \mathcal{X}_1 \times \mathcal{X}_2 \times \{0\} \to \{0\} \times \{0\} \times \mathcal{Y}_3$.

CS 3. the broadcast channel (BC) : $\Omega = \{1,2,3\}$, $\Gamma = \{a,b\}$, $\mathcal{A}_1 = \Gamma$, $\mathcal{A}_2 = \mathcal{A}_3 = \emptyset$, $\mathcal{B}_1 = \emptyset$, $\mathcal{B}_2 = \{a\}$, $\mathcal{B}_3 = \{b\}$, $\Phi_1 = \{1,2,3\}$, $\Phi_2 = \{2\}$, $\Phi_3 = \{3\}$, $\mathcal{X}_2 = \mathcal{X}_3 = \mathcal{Y}_1 = \{0\}$, and $W : \mathcal{X}_1 \times \{0\} \times \{0\} \to \{0\} \times \mathcal{Y}_2 \times \mathcal{Y}_3$.

CS 4. the interference channel (IC) : $\Omega = \{1,2,3,4\}$, $\Gamma = \{a,b\}$, $\mathcal{A}_1 = \{a\}$, $\mathcal{A}_2 = \{b\}$, $\mathcal{A}_3 = \mathcal{A}_4 = \emptyset$, $\mathcal{B}_1 = \mathcal{B}_2 = \emptyset$, $\mathcal{B}_3 = \{a\}$, $\mathcal{B}_4 = \{b\}$, $\Phi_1 = \{1,3\}$, $\Phi_2 = \{2,4\}$, $\Phi_3 = \{3\}$, $\Phi_4 = \{4\}$, $\mathcal{X}_3 = \mathcal{X}_4 = \{0\} = \mathcal{Y}_1 = \mathcal{Y}_2$, and $W : \mathcal{X}_1 \times \mathcal{X}_2 \times \{0\} \times \{0\} \to \{0\} \times \{0\} \times \mathcal{Y}_3 \times \mathcal{Y}_4$.

CS 5. the relay channel (RC) : $\Omega = \{1,2,3\}$, $|\Gamma| \geq 1$, $\mathcal{A}_1 = \Gamma$, $\mathcal{A}_2 = \mathcal{A}_3 = \emptyset$, $\mathcal{B}_1 = \mathcal{B}_2 = \emptyset$, $\mathcal{B}_3 = \Gamma$, $\Phi_1 = \{1,3\}$, $\Phi_2 = \{2\}$, $\Phi_3 = \{3\}$, $\mathcal{X}_3 = \{0\} = \mathcal{Y}_1$, and $W : \mathcal{X}_1 \times \mathcal{X}_2 \times \{0\} \to \{0\} \times \mathcal{Y}_2 \times \mathcal{Y}_3$.

CS 6. the two-way channel (TWC) : $\Omega = \{1,2\}$, $\Gamma = \{a,b\}$, $\mathcal{A}_1 = \{a\}$, $\mathcal{A}_2 = \{b\}$, $\mathcal{B}_1 = \{b\}$, $\mathcal{B}_2 = \{a\}$, $\Phi_1 = \Phi_2 = \{1,2\}$, and $W : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{Y}_1 \times \mathcal{Y}_2$.

4 Classes of Feedback Strategies, Common Random Experiments and Their Mystery Numbers

In dealing with different kinds of feedback strategies it is convenient to have the following concept. Let $\mathcal{F}_n (n = 1,2,\ldots)$ be a subset of the set of all randomized feedback strategies $\mathcal{F}_n^{\mathrm{r}}$ of a DMC W with blocklength n and let it contain the set $\mathcal{F}_n^{\mathrm{d}}$

of all deterministic strategies. We call $(\mathcal{F}_n)_{n=1}^{\infty}$ a smooth class of strategies if for all $n_1, n_2 \in \mathbb{N}$ and $n = n_1 + n_2$

$$\mathcal{F}_n \supset \mathcal{F}_{n_1} \times \mathcal{F}_{n_2} \tag{3}$$

and the product means concatenation of strategies. Now for $f^n \in \mathcal{F}_n$ the channel induces an output sequence $Y^n(f^n)$. For any smooth class we define numbers

$$\mu(\mathcal{F}_n) = \max_{f^n \in \mathcal{F}_n} H(Y^n(f^n)) \,. \tag{4}$$

By (3) and the memoryless character of the channel

$$\mu(\mathcal{F}_n) \geq \mu(\mathcal{F}_{n_1}) + \mu(\mathcal{F}_{n_2}) \,, \tag{5}$$

and therefore $\mu = \mu\left((\mathcal{F}_n)_{n=1}^{\infty}\right) = \lim_{n\to\infty} \frac{1}{n}\mu(\mathcal{F}_n)$ exists. It is called *mystery number* in order to attract attention. We call $\overline{\mathcal{F}}^{\mathrm{r}} = \left(\mathcal{F}_n^{\mathrm{r}}\right)_{n=1}^{\infty}$ also the complete class of strategies. We mentioned already the class of deterministic strategies $\overline{\mathcal{F}}^{\mathrm{d}} = \left(\mathcal{F}_n^{\mathrm{d}}\right)_{n=1}^{\infty}$. Both classes are smooth. Between those classes there is a natural smooth class $\overline{\mathcal{F}}^{\mathrm{s}} = \left(\mathcal{F}_n^{\mathrm{s}}\right)_{n=1}^{\infty}$ of what may be termed stochastic strategies. For every member $F^n = \left(F_1, \ldots, F_n\right) \in \mathcal{F}_n^{\mathrm{s}}$ F_1 is a RV on $\mathcal{X}$ and $F_t : \mathcal{Y}^{t-1} \to \mathcal{X}$ for $t \geq 2$ are stochastic functions, that is, for each y^{t-1}, $F_t(y^{t-1})$ is a RV with values in $\mathcal{X}$. Stochastic functions are like channels, stochastic strategies are "stochastic versions" of deterministic strategies. One readily verifies that for a DMC

$$\mu\left(\overline{\mathcal{F}}^{\mathrm{s}}\right) = \mu\left(\overline{\mathcal{F}}^{\mathrm{r}}\right), \tag{6}$$

however for multi-way channels there are differences (see the first example below). For these channels each sender has his class of feedback strategies. If they are all smooth, then a region $\mathcal{V}$ of achievable mystery tuples is well-defined. Also, by concatenation all common random experiments are of the i.i.d. type and the asymptotic equipartition property (AEP) holds and the "$\sqrt{n}$-trick" can be applied. It yields the direct coding parts in Theorems 53 and 54 in Sect. 5.

Stochastic strategies for multi-way channels shall have the property that for $t \geq 2$ and given outputs $y_1^{t-1}, y_2^{t-1}, \ldots$ at all receivers the RV's $F_{1t}(y_1^{t-1}, y_2^{t-1}, \ldots)$, $F_{2t}(y_1^{t-1}, y_2^{t-1}, \ldots), \ldots$ in the strategies of all senders are *independent*. This condition seems reasonable, if the senders share only the knowledge of all outputs at each step.

Remark Of course the complete class $\overline{\mathcal{F}}^{\mathrm{r}}$ gives the largest rates. However, ratewise inferior classes often have other advantages such as smaller coding efforts. They therefore also should be studied.

Finally we give the formal definitions for the general communication system of Sect. 3. We assume that each terminal ω uses feedback strategies from a smooth set $\mathcal{F}_{n,\omega}$ for encoding. We will denote $\mathcal{G}_n$ for the smooth class of composite strategies,

$$\mathcal{G}_n \triangleq \left\{ g^n = \left(f_\omega^n\right)_{\omega\in\Omega} \,\middle|\, f_\omega^n \in \mathcal{F}_{n,\omega} \right\} \tag{7}$$

As before, we denote $\{\mathcal{G}_n\}_{n=1}^{\infty}$ by $\overline{\mathcal{G}}$. The channel outputs produced via the composite encoding strategy g^n can then be denoted as $Y_\omega^n(g^n)$. For every decoder $\omega \in \Delta$ (cf. (2)), we introduce

$$Z_\omega^n(g^n) = \left(Y_{\omega'}^n(g^n)\right)_{\omega'\in\Phi_\omega} \tag{8}$$

The set of *mystery vectors* for the system is then defined as

$$\mathcal{V}_\Delta(\overline{\mathcal{G}}) \triangleq \lim_{n\to\infty} \left\{ (v_\omega)_{\omega\in\Delta} \,\middle|\, \exists g^n \in \mathcal{G}_n : \forall\omega\in\Delta : 0 \le v_\omega \le \frac{H\left(Z_\omega^n(g^n)\right)}{n} \right\}, \tag{9}$$

where the convergence of sets is understood in the Hausdorff metric and follows here by the memoryless character of our channel and the smoothness assumptions for the classes of strategies.

5 Main Theorem and Consequences

Using the notation of a general (Ω, Γ) communication system in Sect. 3, we define an $\left(n, \{N_\gamma\}_{\gamma\in\Gamma}, \lambda\right)$ IDF code for a general (Ω, Γ) communication system and a smooth class of feedback strategies $\overline{\mathcal{G}}$ as a system

$$\left\{ \left(g_m^n, \left\{\mathcal{D}_{m_\omega}^{(\omega)}\right\}_{\omega\in\Delta}\right) \right\}, \tag{10}$$

with encoding strategies $g_m^n \in \mathcal{G}_n$, message vectors

$$m = (i_\gamma)_{\gamma\in\Gamma},\ m_\omega = (i_\gamma)_{\gamma\in\mathcal{B}_\omega},\ i_\gamma \in \mathcal{N}_\gamma \triangleq \{1, \ldots, N_\gamma\}, \tag{11}$$

and decoding sets

$$\mathcal{D}_{m_\omega}^{(\omega)} \subseteq \mathcal{X}_\omega \times \prod_{\omega'\in\Phi_\omega} \mathcal{Y}_{\omega'}^n, \tag{12}$$

that satisfies the upper bound λ on both kinds of error probability (which can be defined similarly to Definition 50 and 51). Achievable ID rates $(R_\gamma)_{\gamma\in\Gamma}$ are defined

as usual, and the region $\mathcal{C}(\overline{\mathcal{G}})$ of all these rates is then the (second order) ID-capacity region.

Theorem 53 (Main Theorem) *Consider an* (Ω, Γ) *communication system with passive decoders (i.e.* A_7 *holds) and supervisory feedback, and a smooth class of feedback strategies,* $\overline{\mathcal{G}}$.

(i) If all messengers $\gamma \in \Gamma$ *can transmit at positive rate, then* $(R_\gamma)_{\gamma\in\Gamma} \in \mathcal{C}(\overline{\mathcal{G}})$ *if and only if there exists some* $(v_\omega)_{\omega\in\Delta} \in \mathcal{V}_\Delta(\overline{\mathcal{G}})$ *such that*

$$0 \le R_\gamma \le v_\omega, \quad \text{for all } \omega \in \Delta \text{ and } \gamma \in \mathcal{B}_\omega. \tag{13}$$

(ii) If Γ_o *is the set of messengers which can have only transmission rate O, then* $\mathcal{C}(\overline{\mathcal{G}})$ *is obtained as a projection of the region described in (13) into the intersection of the hyperplanes* $\mathcal{R}_\gamma = O$ $(\gamma \in \Gamma_o)$.

The proof of this theorem will be given in Sect. 9.

Remark Theorem 53 of course presents a non-single-letter characterization of $\mathcal{C}(\overline{\mathcal{G}})$ in the usual language of information theory. Still, we want to state its merits:

(i) first of all, the machinery for deriving such a characterization had to be developed for the ID-situation (substitutes for Shannon's random coding argument and Fano's lemma (Lemma 48));
(ii) secondly, the characterization involves only optimization over single strategies, rather than over codebooks of strategies;
(iii) as entropy quantities, mystery numbers are easier to determine than quantities involving mutual information; this is largely responsible for the fact that we can derive a single-letter characterization from the limiting characterization in Theorem 53 directly (see Corollaries 55–58 below); we remind the reader that in the present literature on transmission there is no such direct derivation of the single-letter capacity region for the MAC from its non-single-letter characterization, and that for none of the situations studied in the corollaries below, a complete transmission result is known.

Our methods of proof for Theorem 53 also apply for communication systems not satisfying A_7, if all strategies permitted are deterministic.

Theorem 54 *For a general communication system with supervisory feedback, and for the set of deterministic strategies* $\overline{\mathcal{G}}^{\mathrm{d}}$, *the characterization of* $\mathcal{C}\big(\overline{\mathcal{G}}^{\mathrm{d}}\big)$ *is also given by the parts (i) and (ii) of Theorem 53.*

We now state some applications of Theorem 53. We restrict the discussion to the genuine part (i) since the situation in part (ii) is always obvious from there.

Corollary 55 *For the MAC as described by communication system CS2 in Sect. 3, and the set of deterministic feedback strategies $\overline{\mathcal{G}}^{\mathrm{d}}$, it holds under the condition of (i) of Theorem 53 that*

$$\mathcal{C}_{\mathrm{MAC}}\left(\overline{\mathcal{G}}^{\mathrm{d}}\right) = \left[0, \mu^{\mathrm{d}}_{\mathrm{MAC}}\right] \times \left[0, \mu^{\mathrm{d}}_{\mathrm{MAC}}\right],$$

where

$$\mu^{\mathrm{d}}_{\mathrm{MAC}} = \max_{x_1, x_2} H\left(W(\cdot|x_1, x_2)\right).$$

Notice that in the explicit entropy expressions we will discard the degenerate in- and outputs of the channel W.

Proof of Corollary 55 Since $\Delta = \{3\}$, $\mathcal{V}_{\Delta}\left(\overline{\mathcal{G}}^{\mathrm{d}}\right)$ is a one-dimensional region.

Now, for $g^n = \left(f_1^n, f_2^n\right) \in \mathcal{G}_n^{\mathrm{d}}$ and $Y^n = Y_3^n(g^n)$, since $H(Y^n) = \sum_{t=1}^n H\left(Y_t|Y^{t-1}\right)$ and

$$H\left(Y_t|Y^{t-1} = y^{t-1}\right) = H\left(W\left(\cdot|f_{1t}\left(y^{t-1}\right), f_{2t}\left(y^{t-1}\right)\right)\right) \leq \max_{x_1, x_2} H\left(W\left(\cdot|x_1, x_2\right)\right),$$

it obviously follows that $\mathcal{V}_{\Delta}\left(\overline{\mathcal{G}}^{\mathrm{d}}\right) = \left[0, \mu^{\mathrm{d}}_{\mathrm{MAC}}\right]$ and from Theorem 53 (or Theorem 54) hence also the corollary. □

Corollary 56 *For the MAC as described in CS2 and for the set of stochastic strategies $\overline{\mathcal{G}}^{\mathrm{s}}$, it holds under the condition of (i) of Theorem 53 that*

$$\mathcal{C}_{\mathrm{MAC}}\left(\overline{\mathcal{G}}^{\mathrm{s}}\right) = \left[0, \mu^{\mathrm{s}}_{\mathrm{MAC}}\right] \times \left[0, \mu^{\mathrm{s}}_{\mathrm{MAC}}\right],$$

where

$$\mu^{\mathrm{s}}_{\mathrm{MAC}} = \max_{P_1 \in \mathcal{P}(\mathcal{X}_1)} \max_{P_2 \in \mathcal{P}(\mathcal{X}_2)} H(Y),$$

and

$$P_Y(y) = \sum_{x_1} \sum_{x_2} P_1(x_1) P_2(x_2) W\left(y|x_1, x_2\right).$$

Proof Let $G^n = \left(F_1^n, F_2^n\right) \in \mathcal{G}_n^{\mathrm{s}}$, and denote $Y^n = Y_3^n(G^n)$. Since, given $Y^{t-1} = y^{t-1}$, the RV's $F_{1t}\left(y^{t-1}\right)$ and $F_{2t}\left(y^{t-1}\right)$ are independent, we have

$$H(Y^n) \leq n\, \mu^{\mathrm{s}}_{\mathrm{MAC}},$$

which proves the corollary by Theorem 53. □

Remark Using the converse methods in chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" and regarding the MAC as a one-way channel, one can obtain a strong converse to the above characterizations.

Example In [9] it was shown that the rate point $(R_1, R_2) = (0.76, 0.76)$ is achievable for transmission via the binary erasure MAC defined by

$$W\big(y|x_1, x_2\big) = 1 \text{ iff } y = x_1 + x_2,$$

and

$$\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\},\ \mathcal{Y}_3 = \{0, 1, 2\}.$$

If both senders each choose a key at random, transmit it at this rate 0.76 and decode each other's key from the feedback signal, they can each apply the "$\sqrt{n}$ trick" to the pair of keys, and achieve the ID-rate pair $(R_1 + R_2, R_1 + R_2) = (1.52, 1.52)$. As one can easily calculate that $\mu^{\mathrm{s}}_{\mathrm{MAC}} = 1.5$ for this MAC, this clearly shows that the randomized ID-capacity region $\mathcal{C}_{\mathrm{MAC}}\big(\overline{\mathcal{G}}^{\mathrm{r}}\big)$ exceeds $\mathcal{C}_{\mathrm{MAC}}\big(\overline{\mathcal{G}}^{\mathrm{s}}\big)$. ▲

Corollary 57 *For the BC (see CS3), it holds under condition (i) of Theorem 53 that*

$$\mathcal{C}_{\mathrm{BC}}\big(\overline{\mathcal{G}}^{\mathrm{r}}\big) = \mathcal{C}_{\mathrm{BC}}\big(\overline{\mathcal{G}}^{\mathrm{s}}\big) = \mathcal{R}^*_{\mathrm{BC}},$$

where

$$\mathcal{R}^*_{\mathrm{BC}} = \{(v_1, v_2)|\exists\ P \in \mathcal{P}(\mathcal{X}_1) : 0 \le v_1 \le H(PW_2),\ 0 \le v_2 \le H(PW_3)\}$$

where W_2 and W_3 are the marginal channels.

Proof Let $Q \in \mathcal{G}^{\mathrm{r}}_n = \mathcal{P}\big(\mathcal{G}^{\mathrm{d}}_n\big)$ and let Y_2^n and Y_3^n denote the corresponding channel outputs at terminals 2 and 3, respectively. Now,

$$H\big(Y_2^n\big) \le \sum_{t=1}^{n} H\big(Y_{2t}\big), \qquad H\big(Y_3^n\big) \le \sum_{t=1}^{n} H\big(Y_{3t}\big),$$

where $\Pr\big[Y_{2t} = y_2,\ Y_{3t} = y_3\big] = \sum_{x\in\mathcal{X}_1} P_t(x) W\big(y_2, y_3|x\big)$ for some $P_t \in \mathcal{P}(\mathcal{X}_1)$. Therefore,

$$\mathcal{V}_\Delta = \mathcal{R}^*_{\mathrm{BC}}.$$

Since this region is also achievable with the stochastic strategies of $\overline{\mathcal{G}}^{\mathrm{s}}$, this proves the corollary. □

Example For $\overline{\mathcal{G}} = \overline{\mathcal{G}}^{\mathrm{d}}$, a natural candidate for the single-letterization of $\mathcal{V}_{\Delta}\big(\overline{\mathcal{G}}^{\mathrm{d}}\big)$ would be the region

$$\mathcal{R} = \Big\{(R_1, R_2) | \exists\, x \in \mathcal{X}_1 : 0 \le R_1 \le H\big(W_2(\cdot|x)\big), 0 \le R_2 \le H\big(W_3(\cdot|x)\big)\Big\}.$$

However, consider the BC with $\mathcal{Y}_3 = \mathcal{X}_1$ and $W(y_2, y_3|x) = W_2(y_2|x)\ Q^*(y_3)$, where Q^* satisfies $H(Q^*W_2) = \max_{P\in\mathcal{P}(\mathcal{X}_1)}\ H(PW_2)$. If now the sender uses the deterministic strategy g, defined by

$$g_t\big(y_2^{t-1}, y_3^{t-1}\big) = y_{3,t-1},$$

he generates in this way the maximal entropy $n \cdot H(Q^*W_2)$ at terminal 2. This proves the ID achievability of $\big(H(Q^*W_2), 0\big)$, which clearly in general is not contained in the region $\mathcal{R}$. ▲

Remark In [11] the deterministic BC was treated. In that case feedback is implicitly present and therefore its non-feedback capacity region equals the feedback capacity region, in particular for randomized encoding. In [11] the direct part is proven by applying the "$\sqrt{n}$ trick" twice to $H(Y_2^n)$ and $H(Y_3^n)$ resp., and an application of the converse from chapter "Identification via Channels" gives an upper bound which coincides with this inner bound in the case of a deterministic channel.

Theorem 54 has the following consequence.

Corollary 58 *For the TWC (see CS6) it holds under the condition of (i) of Theorem 53 that*

$$\mathcal{C}_{\mathrm{TWC}}\big(\overline{\mathcal{G}}^{\mathrm{d}}\big) = \Big[0, \nu^{\mathrm{d}}_{\mathrm{TWC}}\Big] \times \Big[0, \nu^{\mathrm{d}}_{\mathrm{TWC}}\Big]$$

where

$$\nu^{\mathrm{d}}_{\mathrm{TWC}} = \max_{x_1, x_2} H\big(W(\cdot, \cdot|x_1, x_2)\big).$$

6 A Method for Proving Converses in Case of Feedback

For one-way channels the approach of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" gives sharp upper bounds (strong converse). However, it does not seem to generalize to multi-way channels with complicated interactions of feedback strategies. Settling for a weaker bound (weak converse but stronger than soft converse of chapter "Identification via Channels") we found a method (Lemma 60 below) which always works, that is, it relates rates to the number $\mu(\mathcal{F}_n)$, that is the maximal entropy of a common random

experiment of blocklength n which can be produced under the restrictions present. For example for the DMC with restriction to deterministic strategies this number equals $n \max_x H\big(W(\cdot|x)\big)$. The common random experiment with this entropy uses one coding strategy and not a whole codebook! In case of randomized strategies the number is $n \max_{P\in\mathcal{P}(\mathcal{X})} H(PW)$.

For an (n, N, λ) IDF code the encoding strategy $f_i^n \in \mathcal{F}_n$ generates a RV Y_i^n with distribution

$$\Pr\big[Y_i^n = y^n\big] = W^n(y^n|f_i^n). \tag{14}$$

Of course, for $i = 1, 2, \ldots, N$

$$H(Y_i^n) \le \mu_n \triangleq \mu(\mathcal{F}_n). \tag{15}$$

The basis of our method is now a very general entropy-setsize relation.

Lemma 59 *For $P = (P_1, P_2, \ldots) \in \mathcal{P}(\mathbb{N})$, the set of PD's on the positive integers, define*

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

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

$$\varepsilon(d) = 1 - \frac{1}{d} \text{ for all } d \ge 1.$$

Remark For discrete memoryless sources $X^n = (X_1, \ldots, X_n)$ Shannon proved that

$$\varepsilon(d, P_X^n) = \max\left\{\sum_{x^n\in J} P_X^n(x^n) : J \subset \mathcal{X}^n, |J| = \lceil 2^{dH(X^n)}\rceil\right\}$$

satisfies

$$\lim_{n\to\infty} \varepsilon(d, P_X^n) = 1, \text{ if } d > 1.$$

Lemma 59 shows what can be done for arbitrary discrete sources.

Application of Lemma 59 to the distribution of Y_i^n gives a set $\mathcal{E}_i \subset \mathcal{Y}^n$ with

$$\Pr\big[Y_i^n \in \mathcal{E}_i\big] \ge 1 - \frac{1}{d}, \tag{16}$$

$$|\mathcal{E}_i| \le \lceil 2^{dH(Y_i^n)}\rceil + 1 \le 2^{d\mu_n} + 2. \tag{17}$$

Define now $\mathcal{D}_i^* = \mathcal{D}_i \cap \mathcal{E}_i$. By Definition 51 and (16) $\Pr\left[Y_i^n \in \mathcal{D}_i^*\right] \geq 1 - \lambda - \frac{1}{d}$. Under the assumption $\lambda < 1 - \lambda - \frac{1}{d}$ by Definition 51 these $\mathcal{D}_i^*$ are necessarily *distinct*. With the abbreviation $K = 2^{d\mu_n} + 2$ we get therefore

$$N \leq \sum_{k=1}^{K} \binom{|\mathcal{Y}^n|}{k} \leq K 2^{nK \log |\mathcal{Y}|}$$

and $\log\log N \leq d\mu_n + o(n)$.

We summarize this result.

Lemma 60 *For any (n, N, λ) IDF code with coding strategies $\mathcal{F}_n^*$ and corresponding $\mu_n^* = \mu(\mathcal{F}_n^*)$*

$$\log\log N \leq d\mu_n^* + o(n),$$

provided that $d > 1$ and $\lambda < \frac{1}{2}\left(1 - \frac{1}{d}\right)$.

Remark In the case of the DMC, for deterministic strategies we have $\mu_n^* = n \max_{x \in \mathcal{X}} H\big(W(\cdot|x)\big)$ and for randomized strategies $\mu_n^* = n \max_{P \in \mathcal{P}(\mathcal{X})} H(PW)$. For λ tending to 0 in Lemma 60 we can let d tend to 1 and thus obtain weak converses.

7 A 3-Step ID Scheme for the Noiseless BSC

We begin right away with the definition of the scheme. Some heuristic understanding is provided subsequently. The proof of asymptotic optimality is given in Sect. 10. We are given a set of messages $\mathcal{M} = \{1, \ldots, M\}$. For any constant $\alpha > 1$ define

$$K = \lceil (\log M)^\alpha \rceil \tag{18}$$

and $\pi_1 < \pi_2 < \ldots \pi_K$ as the K smallest prime numbers. For $k \in \mathcal{K} = \{1, \ldots, K\}$ define a key $\varphi_k : \mathcal{M} \to \{1, \ldots, \pi_k\}$ by

$$\varphi_k(m) - 1 \equiv m \mod \pi_k. \tag{19}$$

Let $\{\varphi_k : k \in \mathcal{K}\}$ be a cipher and $\mathcal{M}' = \{\varphi_k(m) : m \in \mathcal{M}, k \in \mathcal{K}\} = \{1, 2, \ldots, \pi_K\}$ the set of all possible encipherings serving as "message set" for a further cipher $\{\varphi'_\ell : \ell \in \mathcal{K}'\}$, where $\mathcal{K}' = \{1, \ldots, K'\}$ with

$$K' = \lceil (\log \pi_K)^\alpha \rceil \tag{20}$$

and $\varphi'_\ell : \mathcal{M}' \to \{1, \ldots, \pi_\ell\}$ satisfies

$$\varphi'_\ell(m') - 1 \equiv m' \mod \pi_\ell. \tag{21}$$

Step 1. The sender chooses $k \in \mathcal{K}$ randomly according to the uniform distribution on $\mathcal{K}$ and transmits it (and therefore also φ_k) over the channel. This requires $\lceil \log K \rceil$ bits.

Step 2. Similarly the sender chooses an $\ell \in \mathcal{K}'$ at random and transmits it (and therefore also the key φ'_ℓ) over the channel. This requires $\lceil \log K' \rceil$ bits.

Step 3. $m \in \mathcal{M}$ being given to the sender he calculates $\varphi'_\ell\big(\varphi_k(m)\big)$ and sends it to the receiver. Knowing both, k and ℓ, the receiver calculates $\varphi'_\ell\big(\varphi_k(\hat{m})\big)$ and compares it with the transmitted encryption $\varphi'_\ell\big(\varphi_k(m)\big)$. In case of equality he decides $m = \hat{m}$ and otherwise he decides $m \neq \hat{m}$.

Theorem 61 (Optimality of the 3-Step Scheme) *The scheme satisfies*

$$\lim_{M\to\infty} \frac{\log\log M}{n} = \frac{1}{\alpha}$$

and thus achieves any rate below the capacity 1.

8 Extension of the 3-Step ID Scheme to the DMC With and Without Feedback

Outcomes in random experiments below must be labelled by consecutive integers $1, 2, 3 \ldots$ to make the number theoretic setting of the previous scheme possible. Otherwise the only changes on the scheme are the following: The uniform random experiments for the choice of the two keys, known to sender and receiver, are formed now under the given circumstances. We discuss three cases.

1. *Deterministic feedback strategies:* The sender sends b times a letter $x^* \in \mathcal{X}$ with $H\big(W(\cdot|x^*)\big) = \max_{x\in\mathcal{X}} H\big(W(\cdot|x)\big)$.

   As in chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" the generated sequences $\mathcal{D}^* = \bigcup_{V:\|V-W\|\le\varepsilon} \mathcal{T}^b_V(x^{*b})$ and an erasure e for the sequences in $\mathcal{Y}^b \setminus \mathcal{D}^*$ give an essentially uniform random experiment $\big(\mathcal{D}^* \cup \{e\}, W^b(\cdot|x^{*b})\big)$ used for the key selections in the first two steps with appropriate b's. A factor $\big(1 - 2\exp\{-E(\varepsilon)b'\}\big)$ enters the changes in error probabilities of the second kind. The erasure option and also a small error probability in performing step 3 add a small error probability to both kinds of errors. Since $|\mathcal{D}^*| = \exp\{b\, H\big(W(\cdot|x^*)\big) + o(b)\}$ the scheme achieves rates below $H\big(W(\cdot|x^*)\big)$, provided that W has positive transmission capacity C.
2. *Complete randomized feedback strategies:* Replace $\mathcal{D}^*$ by $\bigcup_{Q:\|Q-Q^*\|\le\varepsilon} \mathcal{T}^b_{Q^*}$ with $Q^* = P^*W$ and $H(Q^*) = \max_{P\in\mathcal{P}(\mathcal{X})} H(PW)$. Now rates below $H(Q^*)$ are achievable, if $C > 0$.

3. *Randomized encoding without feedback:* As in chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" use now standard transmission codes with uniform distribution on the set of codewords. Here sender and receiver know the outcome of the random experiments in step 1 and 2 with a small error probability only, but this can be digested. Notice that the resulting scheme is totally constructive if the transmission codes used are constructed. Here the ID capacity is C.

9 Proof of Theorems 53 and 54

Proof of Theorem 53 Since we consider supervisory feedback, the direct part of Theorem 53 follows from smoothness and memorylessness, as discussed in Sect. 4. We therefore concentrate on the converse part of Theorem 53.

Let $\varepsilon > 0$ be arbitrary and fixed. Since $\mathcal{V}_\Delta(\overline{\mathcal{G}})$ is compact, there exists a number of vectors $\left(v_\omega^{(\ell)}\right)_{\omega\in\Delta}$, $\ell = 1, \ldots, L = L(\varepsilon)$ such that

$$\mathcal{V}_\Delta(\overline{\mathcal{G}}) \subseteq \bigcup_{\ell=1}^{L} \left\{ (v_\omega)_{\omega\in\Delta} \,\middle|\, \forall \omega \in \Delta : 0 \le v_\omega < v_\omega^{(\ell)} + \varepsilon \right\}. \tag{22}$$

Let $n \ge n_0(\varepsilon)$ be sufficiently large such that for all $g^n = \left(f_\omega^n\right)_{\omega\in\Omega}$ there exists a $(v_\omega)_{\omega\in\Delta} \in \mathcal{V}_\Delta(\overline{\mathcal{G}})$ such that for all $\omega \in \Delta$

$$\frac{H\left(Z_\omega^n(g^n)\right)}{n} < v_\omega + \varepsilon. \tag{23}$$

Finally let an $\left(n, \{N_\gamma\}_{\gamma\in\Gamma}, \lambda\right)$ IDF code be given as described in (10)–(12), to which we also refer for notation. By the passive decoding Axiom A_7 and (12)

$$\mathcal{D}_{m_\omega}^{(\omega)} \subset \mathcal{Z}_\omega^n, \tag{24}$$

where we have denoted

$$\mathcal{Z}_\omega^n = \prod_{\omega'\in\Phi_\omega} \mathcal{Y}_{\omega'}^n. \tag{25}$$

For all $\ell = 1, \ldots, L$, define

$$\mathcal{M}(\ell) = \left\{ m \,\middle|\, \forall \omega \in \Delta : \frac{H\left(Z_\omega^n(g_m^n)\right)}{n} < v_\omega^{(\ell)} + 2\varepsilon \right\}. \tag{26}$$

By the definition of $\left(\nu_\omega^{(\ell)}\right)_{\omega\in\Delta}$ in (22) and the choice of n in (23), $\bigcup_{\ell=1}^{L}\mathcal{M}(\ell)$ covers all messages m.

Therefore we can choose ℓ^* such that

$$\left|\mathcal{M}(\ell^*)\right| \geq \frac{\prod_{\gamma\in\Gamma} N_\gamma}{L}. \tag{27}$$

We will now consider the marginal channels

$$W_\omega : \prod_{\omega'\in\Omega} \mathcal{X}_{\omega'} \longrightarrow \mathcal{Z}_\omega^n,$$

for all $\omega \in \Delta$, as one-way channels, and derive a marginal IDF-code for each W_ω from the above IDF-code for W. To this end, denote for $\omega \in \Delta$

$$\mathcal{N}_\omega^* = \left\{ m_\omega \,\middle|\, \exists\, m' \in \mathcal{M}(\ell^*) : \forall \gamma \in \mathcal{B}_\omega : i_\gamma = i'_\gamma \right\}. \tag{28}$$

Then it follows from (27) that for all $\gamma \in \mathcal{B}_\omega$

$$\left|\mathcal{N}_\omega^*\right| \geq \frac{N_\gamma}{L}.$$

We can assume without loss of generality that $L = L(\varepsilon) \geq 4$, and $N_\gamma \geq 4L$ for all $\gamma \in \Gamma$. Since $\log\left(\frac{a}{b}\right) \geq \frac{\log a}{\log b}$ for $b \geq 4$ and $a \geq 4b$, it then holds for all $\gamma \in \mathcal{B}_\omega$ that

$$\log\log\left|\mathcal{N}_\omega^*\right| \geq \log\log N_\gamma - \log\log L. \tag{29}$$

Let us now fix some injective mapping $\sigma : \mathcal{N}_\omega^* \to \mathcal{M}(\ell^*)$ such that for all $m_\omega = (i_\gamma)_{\gamma\in\mathcal{B}_\omega} \in \mathcal{N}_\omega^*$ it holds that

$$\sigma\left(m_\omega\right) = \left(i'_\gamma\right)_{\gamma\in\Gamma} \text{ iff } i_\gamma = i'_\gamma, \text{ for all } \gamma \in \mathcal{B}_\omega. \tag{30}$$

Let us now define the encoding strategies $\left\{h_{m_\omega}^n \,\middle|\, m_\omega \in \mathcal{N}_\omega^*\right\}$ for the one-way channel

$$W_\omega : \mathcal{X} \times \{0\} \longrightarrow \{0\} \times \mathcal{Z}_\omega^n, \tag{31}$$

with $\mathcal{X} = \prod_{\omega'\in\Omega}\mathcal{X}_{\omega'}$, by

$$h_{m_\omega}^n = g_{\sigma(m_\omega)}^n. \tag{32}$$

Then $\left\{\left(h_{m_\omega}^n, \mathcal{D}_{m_\omega}^{(\omega)}\right) : m_\omega \in \mathcal{N}_\omega^*\right\}$ forms an $\left(n, |\mathcal{N}_\omega^*|, \lambda\right)$ IDF code for W_ω in (31).

Now assume that $\lambda < \frac{1}{2}\frac{\varepsilon}{1+\varepsilon}$ and apply Lemma 60 with $d = 1 + \varepsilon > 1$. Since $\sigma(m_\omega) \in \mathcal{M}(\ell^*)$, it follows from (26) that

$$\frac{1}{n} \log\log \left|\mathcal{N}_\omega^*\right| < (1+\varepsilon)\left(\nu_\omega^{(\ell^*)} + 2\varepsilon\right) + \varepsilon\,, \tag{33}$$

if $n \geq n_1(\varepsilon)$, say.

Combination of (29) and (33) gives for all $\omega \in \Delta$ and all $\gamma \in \mathcal{B}_\omega$

$$\frac{1}{n} \log\log N_\gamma < (1+\varepsilon)(\nu_\omega^{(\ell^*)} + 2\varepsilon) + \varepsilon + \frac{1}{n} \log\log L(\varepsilon).$$

Letting $n \to \infty$ and $\varepsilon \to 0$, this proves the converse part of Theorem 53. □

Proof of Theorem 54 It can be seen from (12) that (24) does not hold in general for all communication systems (cf. two-way channel), so that $\left|\mathcal{D}_{m_\omega}^{(\omega)}\right|$ can no longer be directly related to the output-entropy $H\left(Z_\omega^n\right)$.

However, if the feedback strategies are deterministic, this functional relationship also enables us to describe $\mathcal{D}_{m_\omega}^{(\omega)}$ in such a way that (24) does hold. Thus the previous arguments apply and give the converse part of Theorem 54. The direct part goes by the "$\sqrt{n}$-trick" as usual. □

10 Proof of Theorem 61, Optimality of Our Coding Scheme

Two elementary facts from number theory are used (see e.g. [7]). The first follows from the prime factorisation theorem and the second from a weak version of the prime number theorem originally due to Chebyshev. Here are the statements.

Lemma 62

(i) The number of prime divisors of an integer m does not exceed $\log m$.
(ii) The kth prime number π_k *satisfies* $\pi_k = O(k \log k)$.

It is clear from the definition in (19) that $\varphi_k(m) = \varphi_k(\hat{m})$ exactly if $|m - \hat{m}| \equiv 0$ mod φ_k. Lemma 62(i) then implies a result basic for our analysis.

Lemma 63 *For any* $m, \hat{m} \in \mathcal{M}, m \neq \hat{m}$

$$K^{-1}\left|\left\{k \in \mathcal{K} : \varphi_k(m) = \varphi_k(\hat{m})\right\}\right| \leq K^{-1} \log M.$$

Proof of Theorem 61 *Error performance of scheme.* Since the transmission is noiseless, the error probability of the first kind is zero. The total error probability of the second kind equals

$$\begin{aligned}&\Pr\left[\varphi'_\ell\big(\varphi_k(\hat{m})\big) = \varphi'_\ell\big(\varphi_k(m)\big)|\hat{m} \neq m\right]\\ &\leq \Pr\left[\varphi_k(\hat{m}) = \varphi_k(m)|\hat{m} \neq m\right]\\ &+ \Pr\left[\varphi'_\ell\big(\varphi_k(\hat{m})\big) = \varphi'_\ell\big(\varphi_k(m)\big)|\varphi_k(\hat{m}) \neq \varphi_k(m)\right].\end{aligned}$$

By Lemma 63 we have with $K = \lceil(\log M)^\alpha\rceil$

$$\Pr\left[\varphi_k(\hat{m}) = \varphi_k(m)|\hat{m} \neq m\right] \leq \frac{1}{(\log M)^{\alpha-1}}$$

and in exact analogy

$$\Pr\left[\varphi'_\ell\big(\varphi_k(\hat{m})\big) = \varphi'_\ell\big(\varphi_k(m)\big)|\varphi_k(\hat{m}) \neq \varphi_k(m)\right] \leq \frac{1}{(\log \pi_K)^{\alpha-1}}.$$

Since with M also π_K tends to infinity and since $\alpha > 1$, the total error probability tends to zero.

Blocklength n of the scheme. Clearly, $n = \lceil\log K\rceil + \lceil\log K'\rceil + \lceil\log \pi_{K'}\rceil$. From (ii) in Lemma 62 we have $\pi_{K'} = O(K' \log K')$. Also, $\log K' = O(\log\log K)$. Therefore

$$n = [1 + o(1)]\log K = \alpha[1 + o(1)]\log\log M$$

and thus the result is proved. □

Remark

1. In the method of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" there is not our second step. However, there only the existence of appropriate keys is shown.
2. In [10] there is also not our second step. After transmission of φ_k then $\varphi_k(m)$ is *transmitted directly* (and *not only identified*, as in our scheme). Therefore $2n$ bits are used and the rate is only 1/2.
3. By further iteration one can reduce blocklength slightly at the price of a larger error probability and, vice versa.
4. In step 3, transmission could be replaced by a sub-optimal constructive ID-scheme.

References

1. R. Ahlswede, A constructive proof of the coding theorem for discrete memoryless channels with feedback, in *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Function, Random Processes* (1973), pp. 39–50
2. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inform. Theory **35**(1) (1989)
3. R. Ahlswede, G. Dueck, Identification in the presence of feedback – a discovery of new capacity formulas. IEEE Trans. Inform. Theory **35**(1) (1989)
4. R. Ahlswede, B. Verboven, On identification via multiway channels with feedback. IEEE Trans. Inform. Theory **37**(6) (1991)
5. R. Ahlswede, I. Wegener, *Search Problems*. Wiley-Interscience Series in Discrete Mathematics and Optimization (J. Wiley & Sons, Hoboken, 1987)
6. A. Alexander, I. Althöfer, C. Deppe, U. Tamm (eds.), *Storing and Transmitting Data Rudolf Ahlswede's Lectures on Information Theory 1*. Foundations in Signal Processing, Communications and Networking, vol. 10, 1st edn. (Springer, Berlin, 2014)
7. T.M. Apostol, *Introduction to Analytic Number Theory* (Springer New York, 1976)
8. R.M. Fano, *Class Notes for Transmission of Information*, Course 6.574 (MIT, Cambridge, 1952)
9. N.T. Gaarder, J.K. Wolf, The capacity region of a multiple-access discrete memoryless channel can increase with feedback. IEEE Trans. Inform. Theory **21**(1), 100–102 (1975)
10. K. Melhorn, E.M. Schmidt, Las Vegas is better than determinism in VLSI and distributed computing, in *Proceedings of the ACM 14th Symposium on the Theory of Computing* (1982), pp. 330–337
11. B. Verboven, E.C. van der Meulen, Identification via a deterministic broadcast channel. Capacity bounds for identification via broadcast channels that are optimal for the determination broadcast channel. IEEE Trans. Inform. Theory **36**(6) (1990)

Identification Without Randomization

In the theory of identification via noisy channels randomization in the encoding has a dramatic effect on the optimal code size, namely, it grows double-exponentially in the blocklength, whereas in the theory of transmission it has the familiar exponential growth.

We consider now instead of the discrete memoryless channel (DMC) more robust channels such as the familiar compound (CC) and arbitrarily varying channels (AVC). They can be viewed as models for jamming situations. We make the pessimistic assumption that the jammer knows the input sequence before he acts. This forces communicators to use the maximal error concept (see [3]) and also makes randomization in the encoding superfluous. Now, for a DMC W by a simple observation, made in [5], in the absence of randomization the identification capacity, say $C_{\mathrm{NRI}}(W)$ (where NRI stands for Non-Randomized Identification),[1] equals the logarithm of the number of different row-vectors in W. We generalize this to compound channels.

A formidable problem arises if the DMC W is replaced by the AVC $\mathcal{W}$. In fact, for 0-1-matrices only in $\mathcal{W}$ we are—exactly as for transmission—led to the equivalent zero-error-capacity of Shannon (see [1]). But for general $\mathcal{W}$ the identification capacity $C_{\mathrm{NRI}}(\mathcal{W})$ is quite different from the transmission capacity $C(\mathcal{W})$. An observation is that the separation codes of [3] are also relevant here. We present a lower bound on $C_{\mathrm{NRI}}(\mathcal{W})$. It implies for instance for the AVC

$$\mathcal{W}_\delta = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ \delta & 1-\delta \end{pmatrix} \right\}, \tag{1}$$

[1] Arguably, identification in the absence of randomization could be more straightforwardly be called deterministic identification [10]. However, here we will report the results as they appeared originally.

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_4

$\delta \in \left(0, \frac{1}{2}\right)$, that $C_{\mathrm{NRI}}(\mathcal{W}) = 1$, which is obviously tight. It exceeds $C(\mathcal{W})$, which is known [3] to exceed $1 - h(\delta)$, where h is the binary entropy function.

We observe that a separation code with worst-case average list size $\overline{L}$, which we call an NRA (Non-Randomized Average-list-size) code can be partitioned into $\overline{L}2^{n\varepsilon}$ transmission codes. This gives a non-single-letter characterization of the capacity of AVC with maximal probability of error in terms of the capacity of codes with list decoding.

We also prove that randomization in the *decoding* does not increase $C_{\mathrm{ID}}(\mathcal{W})$ and $C_{\mathrm{NRI}}(\mathcal{W})$.

Finally, we draw attention to related work on source coding [4, 6].

1 Introduction and Results

Let $\mathcal{X}, \mathcal{Y}$ be the finite input and output alphabets of the channels considered, namely, the discrete memoryless (DMC) W, the arbitrarily varying channel (AVC) $\mathcal{W}$ specified by a set of $|\mathcal{X}| \times |\mathcal{Y}|$-stochastic matrices and also written in the form $\mathcal{W} = \left\{W(\cdot|\cdot, s) : s \in \mathcal{S} \text{ finite}\right\}$, and the compound channel (CC) associated with $\mathcal{W}$.

We study here primarily identification codes without randomization (NRI codes) for $\mathcal{W}$.

Definition 64 An $(n, M, \lambda_1, \lambda_2)$ NRI code for $\mathcal{W}$ is a system of pairs $\left\{(u, \mathcal{D}_u) : u \in \mathcal{U}\right\}$ such that $\mathcal{U} \subset \mathcal{X}^n$, $\mathcal{D}_u \subset \mathcal{Y}^n$ (for $u \in \mathcal{U}$), $|\mathcal{U}| = M$, and for all $u, u' \in \mathcal{U}(u \neq u')$, $s^n \in \mathcal{S}^n$

$$W^n(\mathcal{D}_u|u, s^n) > 1 - \lambda_1 \tag{2}$$

and

$$W^n(\mathcal{D}_u|u', s^n) < \lambda_2. \tag{3}$$

Here for $s^n = (s_1, \ldots, s_n)$, $y^n = (y_1, \ldots, y_n)$, and $u = (u_1, \ldots, u_n)$

$$W^n(y^n|u, s^n) = \prod_{t=1}^{n} W(y_t|u_t, s_t).$$

(Recall that in the definition of ID codes in chapter "Identification via Channels" instead of $\mathcal{U} \subset \mathcal{X}^n$ we used more generally $\mathcal{U} \subset \mathcal{P}(\mathcal{X}^n)$, the set of all probability distributions (PD) on $\mathcal{X}^n$).

We also point out that already in [3] it had been shown that for the DMC W with distinct row vectors the capacity of NRI codes is $\log|\mathcal{X}|$ even before the concept of identification was introduced in [5].

Definition 65 A related concept, used already in [3], are (n, M, λ) non-randomized separation codes (SP-codes), which we abbreviate as NRS codes. They are defined as a system of quadruples $\{(u, u', \mathcal{D}(u, u'), \mathcal{D}(u', u)) : u, u' \in \mathcal{U}, u \neq u'\}$, where $\mathcal{U} \subset \mathcal{X}^n$, $|\mathcal{U}| = M$, $\mathcal{D}(u, u') \subset \mathcal{Y}^n$,

$$\mathcal{D}(u, u') \cap \mathcal{D}(u', u) = \emptyset \qquad \text{for all } u \neq u' \tag{4}$$

and

$$W^n(\mathcal{D}(u, u')|u, s^n) \geq 1 - \lambda \qquad \text{for all } s^n \in \mathcal{S}^n. \tag{5}$$

Notice that with the choice

$$\mathcal{D}(u, u') = \mathcal{D}_u \setminus \mathcal{D}_{u'} \tag{6}$$

we can associate with every $(n, M, \lambda_1, \lambda_2)$ ID or NRI code an (n, M, λ) SP or NRS code respectively, where by (2) and (4) $\lambda = \lambda_1 + \lambda_2$. This fact has been used in chapter "Identification via Channels" in the proof of the (soft)-converse (an exponential weak converse in the sense of [4]), because for the DMC (in case of randomization and no randomization as well) both code concepts lead to the same capacities.

Next we present a third kind of codes, called NRA codes, which were discovered in [3]. Their properties are stronger than those of NRS-codes, but weaker than those of NRI-codes.

These codes can be viewed as list codes with an additional separation property (like (4), (5)). They are essential for our analysis and described below.

Analogously we speak of A-codes (Average list size), if in the definition $\mathcal{U} \subset \mathcal{P}(\mathcal{X}^n)$, namely if the encodings are probability distributions as in the regular ID codes. For the CC only sequences $s^n = (s, \ldots, s)$ $(s \in \mathcal{S})$ are considered and the code constraints are modified accordingly.

For a system $\{(u, \mathcal{D}_u) : u \in \mathcal{U}\}$ with $\mathcal{U} \subset \mathcal{X}^n$ satisfying (2) we define the worst case average list size

$$\overline{L}_{\mathcal{U}} = \max_{u \in \mathcal{U}, s^n \in \mathcal{S}^n} \overline{L}(u, s^n), \tag{7}$$

where

$$\overline{L}(u, s^n) = \sum_{y^n \in \mathcal{D}_u} L(y^n) W^n(y^n|u, s^n) \tag{8}$$

and

$$L(y^n) = |\{u' \in \mathcal{U} : y^n \in \mathcal{D}_{u'}\}|. \tag{9}$$

Definition 66 Now we say that $\{(u, \mathcal{D}_u) : u \in \mathcal{U}\}$ is an $(n, M, \lambda_1, \lambda, \overline{L})$ Non-randomized Average-list-size code (NRA code) , if

$$\overline{L}_{\mathcal{U}} \leq \overline{L} \tag{10}$$

and for all $u, u' \in \mathcal{U}$, $u \neq u'$, there is a partition of $\mathcal{D}_u \cap \mathcal{D}_{u'}$, say $\{A(u, u'), A(u', u)\}$, such that

$$W^n\big(A(u', u)|u, s^n\big) < \lambda \text{ for all } s^n. \tag{11}$$

Obviously (11) holds for any partition of $\mathcal{D}_u \cap \mathcal{D}_{u'}$, $\lambda = \lambda_2$, if (3) is true and for $\mathcal{D}(u, u') = (\mathcal{D}_u \setminus \mathcal{D}_{u'}) \cup A(u, u')$, $\lambda = \lambda_1 + \lambda_2$, (4) holds whenever (11) holds. On the other hand, for $A(u', u) = \mathcal{D}_u \cap \mathcal{D}_{u'} \cap D(u', u)$, (5) implies (11).

Partitioning NRA-Codes into (Non-random) Transmission Codes
We start now with a first basic result.

Theorem 67 *Consider an $(n, M, \lambda_1, \lambda_2, \overline{L})$ NRA-code for the AVC $\mathcal{W}$ defined above. For every $\varepsilon > 0$, $0 < \lambda_1 < \lambda$ there exists a λ^* such that for all $\lambda_2 \leq \lambda^*$ and sufficiently large n the NRA-code can be partitioned into K transmission subcodes for $\mathcal{W}$ with maximal probability of error λ_2, if for $\overline{\ell} = \frac{1}{n} \log \overline{L}$*

$$\frac{1}{n} \log K > \overline{\ell} + \varepsilon.$$

Moreover, clearly this partition contains a subcode of size at least $\frac{M}{K}$.

A Formula for C($\mathcal{W}$)
From Theorem 67 we get a non-single letter characterisation for $C(\mathcal{W})$ involving NRS-codes for the AVC. Those codes were known (also for the DMC) already in [3] (Sects. 4 and 5, and Lemma 73, resp.). So they were known already much earlier than ID-codes and NRI-codes (see chapter "Identification via Channels"). An elegant description was used in [8]. Namely, for an integer m, we associate with $\mathcal{W}$ a graph $G^m(\mathcal{W}) = \{\mathcal{X}^m, \mathcal{E}_m\}$ such that $(x^m, x'^m) \in \mathcal{E}_m$ iff there are PD's $\pi, \pi' \in \mathcal{P}(\mathcal{S}^m)$ such that

$$\sum_{s^m} \pi'(s^m) W^m(\cdot|x^m, s^m) \equiv \sum_{s^m} \pi(s^m) W^m(\cdot|x'^m s^m) \tag{12}$$

(or $\text{conv}\left\{W^m(\cdot|x^m, s^m) : s^m \in \mathcal{S}^m\right\} \cap \text{conv}\left\{W^m(\cdot|x'^m, s^m) : s^m \in \mathcal{S}^m\right\} \neq \emptyset$ where conv denotes the convex hull).

Denote by $\mathcal{I}_m$ the family of independent sets of the graph. Then $\mathcal{U} \in \mathcal{I}_m$ is an NRS-code and we have the following auxiliary result.

Lemma 68 (Ahlswede 1980, [3]) *For any $\varepsilon > 0$, $\lambda > 0$, and sufficiently large n, one can choose $\{\mathcal{D}(u^n, u'^n) : u^n \neq u'^n, u^n, u'^n \in \tilde{\mathcal{U}}\}$ suitable to obtain an SP-*

code with probability of error λ, if the pairwise Hamming distances (with respect to alphabet $\mathcal{U}$) of codewords in $\tilde{\mathcal{U}} \subset \mathcal{U}^n$ are not smaller than $n\varepsilon$.

For a list code $\big(\mathcal{U}, (\mathcal{D}_u)_{u\in\mathcal{U}}\big)$ satisfying (2) we consider the worst case average list size $\overline{L}\big((\mathcal{D}_u)_{u\in\mathcal{U}}\big) = \overline{L}_{\mathcal{U}}$ (defined in (7)) and define

$$\overline{L}_{\mathcal{U},\lambda_1} = \min\big\{\overline{L}_{\mathcal{U}}\big((\mathcal{D}_u)_{u\in\mathcal{U}}\big) : (\mathcal{D}_u)_{u\in\mathcal{U}} \text{ satisfies (2) for } \lambda_1\big\}. \tag{13}$$

In other terms clearly,

$$\overline{L}_{\mathcal{U},\lambda_1} = \min_{(\mathcal{D}_{u''})_{u''\in\mathcal{U}} \text{ with (2)}} \max_{u\in\mathcal{U}, s^n\in\mathcal{S}^n} \sum_{u'\in\mathcal{U}\setminus\{u\}} W^m(\mathcal{D}_u \cap \mathcal{D}_{u'}|u, s^m) + 1. \tag{14}$$

Theorem 69

$$C(\mathcal{W}) = \sup_m \inf_{\lambda_1>0} \max_{\mathcal{U}\in\mathcal{I}_m} \frac{1}{m} \log \frac{|\mathcal{U}|}{|\overline{L}_{\mathcal{U},\lambda_1}|}. \tag{15}$$

On Randomization in the Decoding

We mention here the effects of randomization in the decoding on the transmission capacity $C(\mathcal{W})$ and the identification capacities $C_{\mathrm{NRI}}(\mathcal{W})$ and $C_{\mathrm{ID}}(\mathcal{W})$ for the AVC $\mathcal{W}$, that is, if the maximal probability of error criterion is used.

Theorem 70 *For every AVC $\mathcal{W}$ under the maximal error probability criterion randomization in the decoding does not lead to higher capacities than (i) $C(\mathcal{W})$, (ii) $C_{\mathrm{NRI}}(\mathcal{W})$, and (iii) $C_{\mathrm{ID}}(\mathcal{W})$, resp.*

Remark It follows immediately from the elimination technique and the positivity characterisation of [2] that also for the average error probability criterion the transmission capacity does not increase under randomized decoding.

A Lower Bound on $C_{\mathrm{NRI}}(\mathcal{W})$

Now we present a partial result for $C_{\mathrm{NRI}}(\mathcal{W})$, the quantity of our main interest in this lecture.

Theorem 71 *For $P \in \mathcal{P}(\mathcal{X})$ set*

$$\mathcal{Q}(P,\mathcal{W}) = \big\{(X,X',Y)|P_{Y|X}, P_{Y|X'} \in \overline{\overline{\mathcal{W}}},\ P_X = P_{X'} = P,$$
$$\text{and } X, X', Y \text{ form a Markov chain in this order}\big\}$$

and set

$$\hat{I}(P,\mathcal{W}) = \min_{(X,X',Y)\in\mathcal{Q}(P,\mathcal{W})} I(X' \wedge XY),$$

(where $\overline{\overline{\mathcal{W}}}$ is the row-convex hull of $\mathcal{W}$) then

$$C_{\mathrm{NRI}}(\mathcal{W}) \geq \max_{P} \hat{I}(\mathcal{P}, \mathcal{W}). \tag{16}$$

Remark $\hat{I}(P, \mathcal{W}) = H(P)$ for $P, \mathcal{W}$ such that $(X, X', Y) \in \mathcal{Q}(P, \mathcal{W})$ implies $H(X|X') = 0$.

Corollary 72 *The quantities in the inequality*

$$C_{\mathrm{NRI}}(\mathcal{W}) \geq C(\mathcal{W}) \tag{17}$$

can be different.

Example For

$$\mathcal{W}_\delta = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ \delta & 1-\delta \end{pmatrix} \right\}, \delta \in \left(0, \frac{1}{2}\right),$$

Theorem 71 (or also the lemma below) yields

$$C_{\mathrm{NRI}}(\mathcal{W}_\delta) = 1 \tag{18}$$

and

$$1 - h(\delta) < C(\mathcal{W}_\delta) < 1 \text{ (by [3])}.$$

▲

For the following class, including the example above, (17) also follows from Theorem 71.

Example Let

$$\mathcal{W} = \left\{ \begin{pmatrix} 1 & 0 \\ q(s) & 1-q(s) \end{pmatrix} : s \in \mathcal{S} \right\},$$

where $1 > q(s) > 0$ for all $s \in \mathcal{S}$ (finite).

Then for $P \in \mathcal{P}(\mathcal{X})$ of the form $P = (p, 1-p)$, $p \in (0, 1)$, $\mathcal{Q}(P, \mathcal{W}) = \emptyset$, and by (15) we get $C_{\mathrm{NRI}}(\mathcal{W}) \geq 1$. This is obviously tight. ▲

Next we give a formula for the capacities of a special class of channels, including the examples above.

Lemma 73 *Let* $\mathcal{X} = \{1, 2, \ldots, \alpha\}$, $\mathcal{Y} = \{0, 1, \ldots, \beta\}$, $|\mathcal{S}| < \infty$, *and* $\max_{x \in \mathcal{X}} \max_{s \in \mathcal{S}} W(0|x, s) < 1$.

Furthermore, consider for input alphabet $\mathcal{X} \cup \{0\}$ and output alphabet $\mathcal{Y}$ the AVC

$$\mathcal{W}^* = \big\{ W^*(y|x,s) = W(y|x,s) \text{ for all } x \in \mathcal{X},\, y \in \mathcal{Y},\, s \in \mathcal{S} \\ \text{and } W^*(0|0,s) = 1 \text{ for all } s \in \mathcal{S} \big\}.$$

Then

$$C_{\mathrm{NRI}}(\mathcal{W}^*) = \max_{0 \le p \le 1} \big[h(p) + pC_{\mathrm{NRI}}(\mathcal{W}) \big].$$

A Combinatorial Problem Related to $\mathbf{C}_{\mathrm{NRI}}$ $(\mathcal{W}_\delta)$
Finding $C_{\mathrm{NRI}}(\mathcal{W}_\delta)$ for the special case

$$\mathcal{W}_\delta = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix} \right\},$$

with $\delta \in \left(0, \frac{1}{2}\right)$, is already a formidable task.

By Theorem 71 and numerical computations of B. Balkenhol,

$$C_{\mathrm{NRI}}(\mathcal{W}_\delta) \ge C(\mathcal{W}_\delta) = 1 - h(\delta),$$

where the identity is a very special case of the capacity theorem of [3]. The heart of the matter seems to be related to the following coding problem.

We denote by $B(u^n, d) \subset \{0,1\}^n$ the Hamming ball with radius d. For numbers $1 < \beta < \delta < \frac{1}{2}$ and $\lambda \in (0,1)$ find a subset $A \subset \{0,1\}^n$ as large as possible such that for all $x^n \in A$

$$\left| B(x^n, n\delta) \cap \left[\bigcup_{y^n \in A \setminus \{x^n\}} B(y^n, n\beta) \right] \right| < \lambda |B(x^n, n\delta)|.$$

The Capacity of the Compound Channel (CC) for NRI Codes
Each member $V(\cdot|\cdot, s)$ in the compound channel with $|\mathcal{S}| < \infty$ introduces a partition $\{\mathcal{X}(1|s), \ldots, \mathcal{X}(j_s|s)\}$ of $\mathcal{X}$ such that x, x' are in the same subset exactly if $V(\cdot|x,s) = V(\cdot|x',s)$. Thus a RV X taking values in $\mathcal{X}$ induces a RV $\hat{X}(s)$ for every $s \in \mathcal{S}$ such that $\hat{X}(s) = \ell$ exactly if $X \in \mathcal{X}(\ell|s)$.

Theorem 74 *For a CC* $\mathcal{V} = \{V(\cdot|\cdot, s) : s \in \mathcal{S}\}$ *with* $|\mathcal{S}| < \infty$

$$C_{\mathrm{NRI}}(\mathcal{V}) = \max_X \min_{s \in \mathcal{S}} H\big(\hat{X}(s)\big).$$

2 Proof of Theorem 67

We color the M codewords in $\mathcal{U}$ of an $(n, M, \lambda_1, \lambda_2, L)$ NRA code $\{(u, \mathcal{D}_u) : u \in \mathcal{U}\}$ randomly and independently according to the uniform distribution with K colors and show that the probability for the existence of a coloring satisfying the conditions in the theorem is positive. To estimate the probability, we first fix $s^n \in \mathcal{S}^n$ and $u \in \mathcal{U}$ and partition $\mathcal{U} \setminus \{u\}$ into two parts $\mathcal{U}^{(i)} (i = 1, 2)$ such that $u' \in \mathcal{U}^{(1)}$ iff

$$W^n\big(A(u', u)|u, s^n\big) \le \frac{1}{n^2}. \tag{19}$$

Then

$$\begin{aligned} |\mathcal{U}^{(2)}| &< n^2 \sum_{u' \in \mathcal{U}^{(2)}} W^n\big(A(u', u)|u, s^n\big) \\ &\le n^2 \sum_{u' \in \mathcal{U}^{(2)}} W^n(\mathcal{D}_u \cap \mathcal{D}_{u'}|u, s^n) \\ &\le n^2 \sum_{u' \in \mathcal{U} \setminus \{u\}} W^n(\mathcal{D}_u \cap \mathcal{D}_{u'}|u, s^n) \\ &= n^2\big(\overline{L}(u, s^n) - 1\big) < n^2 \overline{L}, \end{aligned} \tag{20}$$

where for the equality we use the useful observation

$$\sum_{u' \in \mathcal{U} \setminus \{u\}} W^n(\mathcal{D}_u \cap \mathcal{D}_{u'}|u, s^n) = \overline{L}(u, s^n) - 1 \tag{21}$$

for all $u \in \mathcal{U}$ and $s^n \in \mathcal{S}^n$.

For the fixed u, s^n and $i = 1, 2$ let

$$Z_{u'}^{(i)} = \begin{cases} W^n\big(A(u', u)|u, s^n\big) & \text{if } u' \in \mathcal{U}^{(i)} \text{ and } u' \text{ has the same color as } u \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

If we can show that for all such u and s^n, $i = 1, 2$,

$$\Pr\left\{\sum_{u' \in \mathcal{U}^{(i)}} Z_{u'}^{(i)} > \frac{\lambda - \lambda_1}{2}\right\} \le (|\mathcal{S}^n||\mathcal{U}|)^{-1}\frac{\theta}{2} \tag{23}$$

for a $\theta \in (0, 1)$, then we can find a (coloring) partition $\{\mathcal{U}_k : k = 1, 2, \ldots, K\}$ such that for all $u \in \mathcal{U}_k (k = 1, 2, \ldots, K)$, $s^n \in \mathcal{S}^n$, $\sum_{u' \in \mathcal{U}_k \setminus \{u\}} W^n\big(A(u', u)|u, s^n\big) <$

$\lambda - \lambda_1$, and so

$$W^n\left(\mathcal{D}_u \setminus \left(\bigcup_{u' \in \mathcal{U}_k \setminus \{u\}} A(u', u)\right) | u, s^n\right)$$

$$\geq W^n(\mathcal{D}_u | i, s^n) - \sum_{u' \in \mathcal{U}_k \setminus \{u\}} W^n\big(A(u', u) | u, s^n\big)$$

$$> 1 - \lambda - (\lambda - \lambda_1).$$

Thus, if we let $\mathcal{D}'_u = \mathcal{D}_u \setminus \left\{\bigcup_{u' \in \mathcal{U}_k \setminus \{u\}} A(u, u')\right\}$, for all $u \in \mathcal{U}_k$ $(k = 1, 2, \ldots, K)$ then $\left\{(u, \mathcal{D}'_u) : u \in \mathcal{U}_k\right\}$ $(k = 1, 2, \ldots, K)$ is the family of codes required by the theorem. We first show (23) for $i = 1$. By the definition of the $Z^{(1)}_{u'}$'s

$$\Pr\left(\sum_{u' \in \mathcal{U}^{(1)}} Z^{(1)}_{u'} > \frac{\lambda - \lambda_1}{2}\right) \leq \exp_e\left\{-n^{\frac{3}{2}} \frac{\lambda - \lambda_1}{2}\right\} \prod_{u' \in \mathcal{U}^{(1)}} \exp_e\left\{n^{\frac{3}{2}} Z^{(1)}_{u'}\right\}$$

$$= \exp_e\left\{-n^{\frac{3}{2}} \frac{\lambda - \lambda_1}{2}\right\} \prod_{u' \in \mathcal{U}^{(1)}} \left[\left(1 - \frac{1}{K}\right) + \frac{1}{K} \exp_e\left\{n^{\frac{3}{2}} W^n\big(A(u, u') | u, s^n\big)\right\}\right]$$

$$= \exp_e\left\{-n^{\frac{3}{2}} \frac{\lambda - \lambda_1}{2}\right\} \prod_{u' \in \mathcal{U}^{(1)}} \left[1 + \frac{1}{K}\big(\exp_e\left\{n^{\frac{3}{2}} W^n\big(A(u', u) | u, s^n\big)\right\} - 1\big)\right]$$

$$\leq \exp_e\left\{-n^{\frac{3}{2}} \frac{\lambda - \lambda_1}{2}\right\} \prod_{u' \in \mathcal{U}^{(1)}} \left[1 + \frac{e}{K} n^{\frac{3}{2}} W^n\big(A(u, u') | u, s^n\big)\right]$$

$$\leq \exp_e\left\{-n^{\frac{3}{2}} \frac{\lambda - \lambda_1}{2}\right\} \prod_{u \in \mathcal{U} \setminus \{u\}} \left[1 + \frac{e}{K} n^{\frac{3}{2}} W^n(\mathcal{D}_u \cap \mathcal{D}_{u'} | u, s^n)\right]$$

$$\leq \exp_e\left\{-n^{\frac{3}{2}} \frac{\lambda - \lambda_1}{2} + \frac{e}{K} n^{\frac{3}{2}} \sum_{u \in \mathcal{U} \setminus \{u\}} W^n(\mathcal{D}_u \cap \mathcal{D}_{u'} | u, s^n)\right\}$$

$$= \exp_e\left\{-n^{\frac{3}{2}} \left[\frac{\lambda - \lambda_1}{2} - \frac{e}{K}\big(\overline{L}(u, s^n) - 1\big)\right]\right\} \qquad \text{(by (21))}$$

$$\leq \exp_e\left\{-n^{\frac{3}{2}} \left[\frac{\lambda - \lambda_1}{2} - \frac{e}{K}\overline{L}\right]\right\}. \tag{24}$$

Thus for any $\lambda - \lambda_1$, sufficiently large n, and fixed u and s^n, (23) holds for $i = 1$, if we choose a K satisfying

$$K \geq \frac{4e\overline{L}}{\lambda - \lambda_1}. \tag{25}$$

To show (23) for $i = 2$, it is sufficient to show that the probability of the event, that the number of $u' \in \mathcal{U}^{(2)}$ with the same color in the (random) coloring as u is larger than $\left\lfloor \lambda_2^{-1} \frac{\lambda - \lambda_1}{2} \right\rfloor = \kappa_2$, say, is not larger than the RHS of (23).

That is,

$$\sum_{j > \kappa_2} \binom{|\mathcal{U}^{(2)}|}{j} \left(\frac{1}{K}\right)^j \left(1 - \frac{1}{K}\right)^{|\mathcal{U}^{(2)}| - j} \leq [|\mathcal{S}^n||\mathcal{U}|]^{-1} \frac{\theta}{2}. \tag{26}$$

Indeed, if

$$\frac{|\mathcal{U}^{(2)}|}{K} < k_2, \tag{27}$$

then by Stirling's formula and with $\kappa_2 = \left\lfloor \lambda_2^{-1} \frac{\lambda - \lambda_1}{2} \right\rfloor$

$$\begin{aligned}
&\sum_{j > \kappa_2} \binom{|\mathcal{U}^{(2)}|}{j} \left(\frac{1}{K}\right)^j \left(1 - \frac{1}{K}\right)^{|\mathcal{U}^{(2)}| - j} \\
&\leq |\mathcal{U}^{(2)}| \frac{e}{\sqrt{2\pi}} \sqrt{\frac{|\mathcal{U}^{(2)}|}{\kappa_2 |\mathcal{U}^{(2)}| - \kappa_2)}} \left(\frac{|\mathcal{U}^{(2)}|}{K\kappa_2}\right)^{\kappa_2} \left(\frac{|\mathcal{U}^{(2)}|(K-1)}{K(|\mathcal{U}^{(2)}| - \kappa_2}\right)^{|\mathcal{U}^{(2)}| - \kappa_2} \\
&\leq |\mathcal{U}^{(2)}| \frac{e}{\sqrt{2\pi}} \left(\frac{|\mathcal{U}^{(2)}|}{K\kappa_2}\right)^{\kappa_2} \left(1 + \frac{\kappa_2}{|\mathcal{U}^{(2)}| - \kappa_2}\right)^{|\mathcal{U}^{(2)}| - \kappa_2} \\
&\leq |\mathcal{U}^{(2)}| \frac{e}{\sqrt{2\pi}} \left(\frac{|\mathcal{U}^{(2)}|e}{K\kappa_2}\right)^{\kappa_2} \leq |\mathcal{U}^{(2)}| e \left(\frac{e|\mathcal{U}^{(2)}|^2}{K}\right)^{\kappa_2}.
\end{aligned}$$

Thus (26) holds, if we choose

$$K > e|\mathcal{U}^{(2)}| \exp\left\{\frac{n}{\kappa_2} \log |\mathcal{S}||\mathcal{X}|^2 + \frac{1}{\kappa_2} \log \frac{2e}{\theta}\right\} = |\mathcal{U}^{(2)}| \exp\left\{n\varphi_\theta(\kappa_2)\right\}, \tag{28}$$

where $\varphi_\theta(z)$ is a function of z whose values are arbitrarily small when z is arbitrarily large. By (20) it is sufficient for (27), (28) to hold that

$$K > n^2 \overline{L} \exp\left\{n\varphi_\theta(\kappa_2)\right\} \text{ (for sufficiently large } n).$$

To satisfy the above inequality (25), we only need to choose

$$K > \frac{4\overline{L}}{\lambda - \lambda_1} n^2 \exp\left\{n\varphi_\theta(\kappa_2)\right\}$$

and sufficiently small λ^* (and therefore λ_2) depending on λ, λ_1, and ε.

This completes the proof.

3 Proof of Theorem 69

The converse part is absolutely trivial because an (n, M)-code $\left\{(u, \mathcal{D}_u) : u \in \mathcal{U}\right\}$ with maximal probability of error λ_1 satisfies

$$\overline{L}_{\mathcal{U},\lambda_1} = 1 \text{ and } \mathcal{U} \in \mathcal{I}_m.$$

The issue of the theorem is to show that one cannot do better by increasing the size of lists, namely, the direct part. This is an easy consequence of Theorem 67 and Lemma 68 of [3] (see second subsection of Sect. 1).

For a fixed m, $\lambda_1 > 0$ assume that $\mathcal{U} \in \mathcal{I}_m$, $\left\{(u, \mathcal{D}_u) : u \in \mathcal{U}\right\}$ achieves the maximum in (15). Then we treat $\mathcal{U}$ as an input alphabet and $\mathcal{Y}^m$ as output alphabet. Then one can find by the greedy algorithm a subset of codewords $\tilde{\mathcal{U}} \subset \mathcal{U}^\ell$ such that for all $u^\ell, u'^\ell \in \tilde{\mathcal{U}}$, $d_H(u^\ell, u'^\ell) \geq \ell\varepsilon$ (where d_H is the Hamming distance) for any fixed ε and

$$\log|\mathcal{U}| - \frac{1}{\ell}\log|\tilde{\mathcal{U}}| = o(1) \text{ (as } \varepsilon \to 0, \ell \to \infty). \tag{29}$$

Let $\tilde{\mathcal{D}}_{u^\ell}$ for $u^\ell \in \tilde{\mathcal{U}}$ to be the union of Hamming balls with radius $\ell(\lambda_1 + \varepsilon^2)$ and centers at the points in the Cartesian product $\mathcal{D}_{u_1} \times \cdots \times \mathcal{D}_{u_\ell}$, for $u^\ell = (u_1, \ldots, u_\ell)$. Then (2) holds for sufficiently large ℓ, if λ_1 is replaced by $2\lambda_1$. Moreover, for any λ and sufficiently large ℓ, by Lemma 68 of [3], (5) holds, for suitable $D(u^\ell, u'^\ell)$ and therefore (11) holds for suitable $A(u^\ell, u'^\ell)$ for all $u^\ell, u'^\ell \in \tilde{\mathcal{U}}$, $u^\ell \neq u'^\ell$.

To apply Theorem 67, we have to estimate $\overline{L}_{\tilde{\mathcal{U}}}$. Let us write $\mathcal{D}_{u_1} \times \cdots \times \mathcal{D}_{u_\ell} = \mathcal{D}_{u^\ell}$ and denote the Hamming ball with center u^ℓ and radius r in $\mathcal{U}^\ell$, by $B(u^\ell, r)$. Then for $L(\cdot)$ in (9), $u^\ell = (u_1, \ldots, u_\ell) \in \tilde{\mathcal{U}}$, $s^{m\ell} = (s_1^m, \ldots, s_\ell^m) \in \mathcal{S}^{m\ell}$, $v^\ell = (v_1, \ldots, v_\ell) \in \mathcal{D}_{u^\ell}$, and

$$\sum_{v^\ell \in B(u^\ell, \ell(\lambda_1+\varepsilon^2))} L(v^\ell) W^{m\ell}(v^\ell|u^\ell, s^{m\ell}) \leq \sum_{\substack{J \subset \{0,\ldots,\ell-1\}: \\ |J| = \ell(\lambda_1+\varepsilon^2)}} |\mathcal{U}|^{\ell(\lambda_1+\varepsilon^2)} \prod_{j \notin J} L(v_j) W^m(v_j|u_j, s_j^m).$$

Thus one can find a $\beta(\lambda_1, \varepsilon)$ such that $\beta(\lambda_1, \varepsilon) \to 0$ as $\lambda_1, \varepsilon \to 0$, and

$$\frac{1}{\ell} \log \overline{L}_{\tilde{\mathcal{U}}} \leq \frac{1}{\ell} \log \overline{L}_{\mathcal{U}} + \beta(\lambda_1, \varepsilon). \tag{30}$$

Finally, we choose $n = m\ell$ and apply Theorem 1 to $\left\{(u^\ell, \tilde{\mathcal{D}}_{u^\ell}) : u^\ell \in \tilde{\mathcal{U}}\right\}$ to obtain a (transmission) subcode with probability of error $\varepsilon + \lambda_1$ and rate arbitrarily close to $\frac{1}{n} \log \frac{|\mathcal{U}|}{\overline{L}_{\tilde{\mathcal{U}}}}$, when ℓ is arbitrarily large (depending on m) and λ_1 and ε in (30) are arbitrarily small. Since m is fixed when $n = m\ell + r$, $r < m$, we asymptotically loose nothing, if we add r dummy letters. This completes our proof.

4 Proof of Theorem 70

(i) This is an exercise in [7, Problem 11(c), Page 226] and a very easy consequence of Theorem 67 as well.

The proofs of (ii) and (iii) are essentially the same and so we only prove (ii).

So we are given a system $(\mathcal{U}, Q)$ with $\mathcal{U} \subset \mathcal{X}^n$ and $Q : \mathcal{Y}^n \to 2^{\mathcal{U}}$ such that for all $u, u' \in \mathcal{U}$, $s^n \in \mathcal{S}^n$

$$\sum_{A: u \in A} \sum_{y^n \in \mathcal{Y}^n} Q(A|y^n) W^n(y^n|u, s^n) > 1 - \lambda_1, \tag{31}$$

$$\sum_{A: u' \in A} \sum_{y^n \in \mathcal{Y}^n} Q(A|y^n) W^n(y^n|u, s^n) < \lambda_2. \tag{32}$$

(Here we note that A's in (31) and (32) are "decoding sets" for u and u', respectively.)

We extract an NRI code $(u, \mathcal{D}_u)_{u \in \mathcal{U}}$ with error probability λ_1', λ_2' by letting $\mathcal{D}_u = \left\{y^n : \sum_{A: u \in A} Q(A|y^n) \geq \alpha\right\}$. Then by (31) for all s^n

$$\begin{aligned} 1 - \lambda_1 &< \sum_{y^n \in \mathcal{D}_u} \sum_{A: u \in A} Q(A|y^n) W^n(y^n|u, s^n) + \sum_{y^n \in \mathcal{D}_u^c} \sum_{A: u \in A} Q(A|y^n) W^n(y^n|u, s^n) \\ &< \sum_{y^n \in \mathcal{D}_u} 1 \cdot W^n(y^n|u, s^n) + \sum_{y^n \in \mathcal{D}_u^c} \alpha W^n(y^n|u, s^n) \\ &\leq W^n(\mathcal{D}_u|u, s^n) + \alpha \end{aligned} \tag{33}$$

written otherwise as

$$W^n(\mathcal{D}_u|u, s^n) > 1 - \lambda_1 - \alpha. \tag{34}$$

On the other hand (32) implies that for all s^n

$$\lambda_2 > \sum_{y^n \in \mathcal{D}_{u'}} \sum_{A: u' \in A} Q(A|y^n) W^n(y^n|u, s^n)$$

$$\geq \alpha \; W^n(\mathcal{D}_{u'}|u, s^n) \quad \text{or} \quad W^n(\mathcal{D}_{u'}|u, s^n) < \frac{\lambda_2}{\alpha}. \tag{35}$$

Finally, for instance with the choice $\alpha = \sqrt{\lambda_2}$ we conclude from (34) and (35) that we can achieve

$$\lambda_1' = \lambda_2 + \sqrt{\lambda_2} \text{ and } \lambda_2' = \sqrt{\lambda_2}.$$

Since for $(\lambda_1, \lambda_2) \to (0, 0)$ also $(\lambda_1', \lambda_2') \to (0, 0)$ we have established equality of the (weak) capacities.

5 Proof of Theorem 71

Fix $R < \hat{I}(P, \mathcal{W}) - \frac{\varepsilon}{2}$ and let

$$\mathcal{Q}_\delta(P, \mathcal{W}) \triangleq \left\{(X, X', Y) : P_{Y|X}, P_{Y|X'} \in \overline{\overline{\mathcal{W}}}, I(X \wedge Y|X') \leq \delta\right\}. \tag{36}$$

Then

$$\mathcal{Q}(P, \mathcal{W}) = \bigcap_{\delta > 0} \mathcal{Q}_\delta(P, \mathcal{W}), \tag{37}$$

and, by the continuity of the mutual information, there are $\alpha, \delta > 0$ such that for all $(X, X', Y) \in \mathcal{Q}_\delta(P, \mathcal{W})$

$$R < I(X' \wedge XY) - \alpha. \tag{38}$$

Next we apply the large deviation method in the standard way or directly use Lemma 3 in [9] to obtain a set of codewords, $\mathcal{U}' \subset \mathcal{X}^n$ such that $\frac{1}{n} \log |\mathcal{U}'| \sim R$ and for all $U \in \mathcal{U}'$, $P_{XX'} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{X})$ (where $\mathcal{P}_n(\mathcal{Z})$ is the set of n-types over $\mathcal{Z}$)

$$\frac{1}{n} \log |\left\{u' \in \mathcal{U}' : (u, u') \in T^n_{XX'}\right\}| < (R - I(X \wedge X'))^+ + \theta, \tag{39}$$

where θ is a positive number and can be chosen arbitrarily small and $a^+ = \max\{0, a\}$, if n is arbitrarily large. Thus by deleting the "bad codewords" from the

neighbourhoods of the codewords, we can obtain a subset $\mathcal{U} \subset \mathcal{U}'$ (for sufficiently large n) such that

$$\frac{1}{n} \log |\mathcal{U}| \geq \frac{1}{n} \log |\mathcal{U}'| - \frac{\varepsilon}{2} \tag{40}$$

and there is no pair (u, u') of codewords in $\mathcal{U}$ with $(u, u') \in \mathcal{T}^n_{XX'}$ for RV's X and X' with $R \leq I(X \wedge X')$. We choose $\mathcal{U}$ as our set of codewords and

$$\mathcal{D}_u = B(u) \setminus E(u) \tag{41}$$

as decoding set for $u \in \mathcal{U}$, where

$$B(u) = \bigcup_{\overline{\overline{W}} \in \overline{\overline{\mathcal{W}}}} \mathcal{T}^n_{\overline{\overline{W}}, \delta}(u) \text{ for } \delta \text{ in (38)}, \tag{42}$$

and $E(u)$ is the set of y^n's (in $\mathcal{Y}^n$) such that there exist a $u' \neq u$ and a triple $(X, X', Y) \in \mathcal{Q}_\delta$ with $(u, u', y^n) \in \mathcal{T}^n_{XX'Y}$.

Analysis

1. To show that for all $s^n \in \mathcal{S}^n$

$$W^n(\mathcal{D}_u | u', s^n) < \lambda_2 \qquad \text{if } u \neq u' \tag{43}$$

we partition $\mathcal{D}_u$ into polynomially many subsets according to the conditional ED's of y^n's, $P_{Y|XX'}(\cdot|u, u')$, for the u, u' in (43). By (41) $\mathcal{T}^n_{Y|XX'}(u, u') \cap \mathcal{D}_u \neq \emptyset$ implies $(X, X', Y) \notin \mathcal{Q}_\delta(P, \mathcal{W})$, or by (36) $I(X \wedge Y'|X') > \delta$, if $P_{Y|X}, P_{Y|X'} \in \overline{\overline{\mathcal{W}}}$. Thus, because the number of conditional ED's is a polynomial in n, (43) follows from the fact that for $(X, X', Y) \notin \mathcal{Q}_\delta(P)$, $P_{Y|X}, P_{Y|X'} \in \overline{\overline{\mathcal{W}}}$

$$\frac{1}{n} \log \overline{\overline{W}}^n \left(\mathcal{T}^n_{Y|XX'}(u, u') \lesssim H(Y|XX') - H(Y|X') = -I(X \wedge Y|X') < -\delta \right. \tag{44}$$

and $W^n\big(B(u')|u', s^n\big) > 1 - 2^{-n\eta}$ for all s^n and suitable $\eta > 0$.

2. We have to show that for all u and s^n

$$W^n(\mathcal{D}_u | u, s^n) > 1 - \lambda_1. \tag{45}$$

Since for all $s^n \in \mathcal{S}^n$, by (42) $W^n\big(B(u)|u, s^n\big) > 1 - 2^{-u\eta}$ for suitable $\eta > 0$, by (41) it is sufficient for (45) to show $W^n\big(E(u)|u, s^n\big)$ is exponentially small. Indeed,

by the definition of $E(u)$ and (39)

$$\begin{aligned}
&\frac{1}{n}\log W^n\left(E(u)|u,s^n\right)\\
&\lesssim \frac{1}{n}\max_{(X,X',Y)\in\mathcal{Q}_\delta(P,\mathcal{W})}\log\left|\left\{u':u'\in T^n_{X'|X}(u)\right\}\right||T^n_{Y|XX'}(u,u')|2^{-nH(Y|X)}\\
&\sim R-I(X\wedge X')+\theta+H(Y|XX')-H(Y|X)\\
&=R-I(X'\wedge XY)+\theta<-(\alpha-\theta)<0,
\end{aligned}\tag{46}$$

if we chose $\theta<\alpha$.

6 Proof of Lemma 73

Denote by $A_{n,\lambda_1,\lambda_2}(\mathcal{W})$ $\left(\mathcal{A}_{n,\lambda_1,\lambda_2}(\mathcal{W}^*)\right)$, the maximal M such that an $(n, M, \lambda_1, \lambda_2)$ NRI code for $\mathcal{W}$ (for $\mathcal{W}^*$) exists.

(i) $C_{\mathrm{NRI}}(\mathcal{W}^*)\le\max_p\left[h(p)+pC_{\mathrm{NRI}}(\mathcal{W})\right]$.

Let $\left\{(u,\mathcal{D}_u):u\in\mathcal{U}\right\}$ be an $(n, M, \lambda_1, \lambda_2)$ NRI code for $\mathcal{W}^*$. We partition $\mathcal{U}$ into subsets $\{\mathcal{U}_k\}_{k=0}^n$ according to the number of zeros in the codewords. Then there must be a k such that

$$|\mathcal{U}_k|\ge\frac{1}{n}\log|\mathcal{U}_k|.\tag{47}$$

Moreover, the relation $u\sim u'$ in $\mathcal{U}_k$ defined by the rule "$u\sim u'$ if $x_t=0$ exactly if $x'_t=0$ for $u=(x_1,\ldots,x_n)$, $u'=(x'_t,\ldots,x'_n)\in\mathcal{U}_k$" is an equivalent relation, which further partitions $\mathcal{U}_k$ into at most $\binom{n}{k}$ equivalence classes $\{\mathcal{V}_{k,j}\}_{j=1}^J$, $J\le\binom{n}{k}$.

All codewords in a fixed $\mathcal{V}_{k,j}$ have k zero-components at the same coordinates. By our assumption at all these coordinates the outputs are zeros whenever the inputs fall into $\mathcal{V}_{k,j}$. So we can obtain an $(n-k,|\mathcal{V}_{k,j}|,\lambda_1,\lambda_2)$ NRI code by deleting the k components from codewords in $\mathcal{V}_{k,j}$ (and corresponding components from decoding sets). Therefore

$$A_{n,\lambda_1,\lambda_2}(\mathcal{W}^*)\le n\binom{n}{n-k}A_{n-k,\lambda_1,\lambda_2}(\mathcal{W}).$$

(ii) $C_{\mathrm{NRI}}(\mathcal{W}^*) \geq \max_p \left[h(p) + pC_{\mathrm{NRI}}(\mathcal{W})\right]$

We have to find an $(n, M, \lambda_1, \lambda_2)$ code for $\mathcal{W}^*$ with

$$M \geq 2^{-n\eta}\binom{n}{n-k} A_{n-k,\lambda_1,\lambda_2}(\mathcal{W}), \tag{48}$$

for an arbitrarily small η.

We first find, by a greedy algorithm, a set B of binary sequences with Hamming weight $n-k$, pairwise Hamming distance not less than $2n\varepsilon$ and size

$$|B| \geq 2^{-n\eta}\binom{n}{k}, \tag{49}$$

where η is a positive constant depending on ε and $\eta \to 0$ as $\varepsilon \to 0$. Let $\left\{(u, \mathcal{D}_u) : u \in \mathcal{U}\right\}$ be an NRI code of length $n-k$ for $\mathcal{W}$ achieving $A_{n-k,\lambda_1,\lambda_2}(\mathcal{W})$. We define for $b^n \in B$ a subset $\mathcal{U}^*(b^n)$ in $\mathcal{X}^{*n}$,

$$\mathcal{U}^*(b^n) = \left\{x^n : x_t = 0 \text{ if } t \neq t_j, (x_{t_1}, \ldots, x_{t_{n-k}}) \in \mathcal{U}\right\} \tag{50}$$

if

$$b_{t_j} = 1 \text{ for } 1 \leq t_1 < t_2 < \cdots \leq t_{n-k} \leq n, \tag{51}$$

and let $\mathcal{U}^* = \bigcup_{b^n \in B} \mathcal{U}^*(b^n)$.

For $u^* = (x_1, \ldots, x_n) \in \mathcal{U}^*(b^n)$, the decoding set is defined by

$$\mathcal{D}^*_{u^*} = \left\{y^n : y_t = 0 \text{ if } t \neq t_j \text{ and } (y_{t_1}, \ldots, y_{t_{n-k}}) \in \mathcal{D}_u\right\}$$

for $t_1, \ldots, t_{n-k}$ in (51) and $u = (x_{t_1}, \ldots, x_{t_{n-1}})$.

Then for all s^n $W^{*n}(\mathcal{D}^*_{u^*}|u^*, s^n) > 1 - \lambda_1$, and for all $u^*, u^{*\prime} \in \mathcal{U}^*(b^n)$, $s^n \in \mathcal{S}^n$, $W^n(\mathcal{D}^*_{u^*}|u^{*\prime}, s^n) < \lambda_2$, since $\left\{(u, \mathcal{D}_u) : u \in \mathcal{U}\right\}$ has error probability (λ_1, λ_2). Moreover, for all $s^n \in \mathcal{S}^n$, $u^* \in \mathcal{U}^*(b^n)$, $u^{*\prime} \in \mathcal{U}^*(b'^n)$, $b^n, b'^n \in B$, and $b^n \neq b'^n$

$$W^n(\mathcal{D}^*_{u^*}|u^{*\prime}, s^n) \leq \overline{w}^{\frac{1}{2}d_H(b^n, b'^n)} \leq \overline{w}^{n\varepsilon} < \lambda_2$$

for $\overline{w} \triangleq \max_{s\in\mathcal{S}} \max_{x\in\mathcal{X}} W(0|x, s)$, if n is sufficiently large.

Thus $\left\{(u^*, \mathcal{D}^*_{u^*}) : u^* \in \mathcal{U}^*\right\}$ is a desired code.

7 Proof of Theorem 74

Without loss of generality assume that for $s \neq s'$, $V(\cdot|\cdot, s) \neq V(\cdot|\cdot, s')$.

It was shown in [3] that for any channel $\tilde{V} : \mathcal{X} \to \mathcal{Y}$ without two identical rows, any $u_1\ u_2, \varepsilon > 0$, sufficiently large n, and any $\mathcal{U} \subset \mathcal{X}^n$ such that for all $u, u' \in \mathcal{U}$, $d_H(u, u') \geq n\varepsilon$, there exists a family of subsets of $\mathcal{Y}^n$, say $\mathcal{D}_u$, $u \in \mathcal{U}$, such that $\tilde{V}^n(\mathcal{D}_u|u) > 1 - u_1$ and $\tilde{V}^n(\mathcal{D}_u|u') < u_2$ for all $u' \neq u$, where d_H. Let us fix a family $\big\{\{\mathcal{X}(1|s), \ldots, \mathcal{X}(j_s|s)\} : s \in \mathcal{S}\big\}$ of partitions in the last subsection of Sect. 1. For $x^n, x'^n \in \mathcal{X}^n$, $s \in \mathcal{S}$, we define

$$d_s(x^n, x'^n) = |\big\{t : x_t \in \mathcal{X}(j|s), x'_t \in \mathcal{X}(j'|s) \text{ with } j \neq j'\big\}|. \tag{52}$$

Thus, by the above auxiliary result, we have that for any $\lambda_1, \lambda_2, \varepsilon > 0$, sufficiently large n, and any $\mathcal{U} \subset \mathcal{X}^n$ such that for all $s \in \mathcal{S}$, $u, u' \in \mathcal{U}$

$$d_s(u, u') \geq n\varepsilon, \tag{53}$$

there is a family of subsets in $\mathcal{Y}^n$, say $\mathcal{D}_u(s)$, $u \in \mathcal{U}$, $s \in \mathcal{S}$, such that for all $u, u' \in \mathcal{U}$, $u \neq u'$, $s \in \mathcal{S}$,

$$V^n\big(\mathcal{D}_u(s)|u, s\big) > 1 - \frac{\lambda_1}{2} \text{ and } V^n\big(\mathcal{D}_u(s)|u', s\big) < \frac{\lambda_2}{2}. \tag{54}$$

To find a good NRI code for $\mathcal{V}$, we first find a $\mathcal{U}$ satisfying (53). Let X be the RV achieving the extremal value in the theorem. Then for any fixed $u \in \mathcal{T}_X^n$, if (53) is violated for s, u, and $u' \in \mathcal{T}_X^n$, then there exists a pair (X, X') such that with $u' \in \mathcal{T}_{X'|X}^n(u)$, $P_{X'} = P_X$, and $\mathbb{E}d_s(X, X') < \varepsilon$. For such (X, X'),

$$\frac{1}{n} \log |\mathcal{T}_{X'|X}^n(u)| = H(X'|X) + o(1), \tag{55}$$

and by the data processing inequality (Lemma 49) and by Fano's inequality (Lemma 48)

$$I(X \wedge X') \geq I\big(\hat{X}(s) \wedge \hat{X}'(s)\big) \geq H\big(\hat{X}(s)\big) - \alpha(\varepsilon) \tag{56}$$

as $E\ d_s(X, X') < \varepsilon$ implies that $\Pr\big(\hat{X}(s) \neq \hat{X}(s)\big) < \varepsilon$, where $\alpha(\varepsilon)$ is a constant depending on ε and $\alpha \to 0$ as $\varepsilon \to 0$.

Denote by $\mathcal{Q} = \big\{(X, X') : P_{X'} = P_X \text{ and } d_s(X, X') < \varepsilon \text{ for some } s \in \mathcal{S}\big\}$.

Then the total number of u's in $\mathcal{T}_X^n$, such that for an $s \in \mathcal{S}$ (53) does not hold, are not larger than $\exp\{n \Big[\max_{(X,X')\in\mathcal{Q}} H(X'|X) + o(1)\Big]\}$. Consequently, by the greedy

algorithm one can find a $\mathcal{U} \subset \mathcal{T}_X^n$ satisfying (53) such that

$$\begin{aligned}\frac{1}{n}\log|\mathcal{U}| &\geq H(X) - \max_{(X,X')\in\mathcal{Q}} H(X'|X) + o(1)\\ &= \min_{(X,X')\in\mathcal{Q}} I(X \wedge X') + o(1) \qquad (\text{since } H(X') = H(X))\\ &\geq \min_s H\big(\hat{X}(s)\big) - \alpha(s) + o(1) \qquad (\text{by (56)}). \end{aligned} \tag{57}$$

Then the following procedure works.

1. For all $a \in \mathcal{X}$ define $a^\ell = (a, \ldots, a)$. Choose a sufficiently small δ and a sufficiently large ℓ such that for all $a \in \mathcal{X}$, $V \in \mathcal{V}$, $V^\ell\big(\mathcal{T}^\ell_{V,\delta}(a^\ell)|a^b\big) > 1 - \frac{1}{2|\mathcal{X}|}\lambda$ and for all $V, V' \in \mathcal{V}$ there is an $a \in \mathcal{X}$ such that $\mathcal{T}^\ell_{V,\delta}(a^\ell) \cap \mathcal{T}^\ell_{V',\delta}(a^\ell) = \emptyset$, where $\lambda \triangleq \min(\lambda_1, \lambda_2)$. Then the encoder uses $|\mathcal{X}|$ blocks of length ℓ to send a^ℓ for all $a \in \mathcal{X}$. The decoder tries to find a $V \in \mathcal{V}$ (and the corresponding state $s \in \mathcal{S}$) such that for all $a \in \mathcal{X}$, the ath block output of length ℓ falls into $\mathcal{T}^\ell_{V,\delta}(a^\ell)$. If he can find it, it must be unique by our construction, otherwise the decoder just declares an error. When any $V \in \mathcal{V}$ governs the transmission, the decoder successfully estimates V with probability at least $1 - \frac{1}{2}\lambda$.
2. Knowing the state s governing the transmission, the decoder uses the decoding sets $\big\{\mathcal{D}_u(s) : s \in \mathcal{S}\big\}$ in (54) to identify the message for which the two kind of error probabilities are $\frac{\lambda_1}{2}$ and $\frac{\lambda_2}{2}$, respectively. Thus the two kind of error probabilities totally do not exceed λ_1 and λ_2, respectively.

   This and (57) complete the proof of the direct part (by choosing $\frac{\ell}{n}$ arbitrarily small).

   To prove the converse we partition the set $\mathcal{U}$ of codewords of a given NRI-code of length n according to the ED's. Then we can find a RV X and a $\mathcal{U}' \subset \mathcal{U}$ such that

$$\mathcal{U}' \subset \mathcal{T}_X^n \text{ and } |\mathcal{U}'| \geq (n+1)^{-|\mathcal{X}|}|\mathcal{U}|.$$

   Let φ_s be the mapping $\mathcal{X}^n \to \{1, 2, \ldots, j_s\}$ for a fixed $s \in \mathcal{S}$ such that $\varphi_s(x^n) = (i_1, \ldots, i_n)$, if $x_t \in \mathcal{X}(i_t|s)$. Then for all $s \in \mathcal{S}$ there are no $u, u' \in \mathcal{U}$ with $\varphi_s(u) = \varphi_s(u')$ and $u \neq u'$. Furthermore, the mapping φ_s sends $\mathcal{T}_X^n$ to $\mathcal{T}^n_{\hat{X}(s)}$. Consequently for all s,

$$\frac{1}{n}\log|\mathcal{U}'| \leq \frac{1}{n}\log|\mathcal{T}^n_{\hat{X}(s)}| = H\big(\hat{X}(s)\big) + o(1).$$

   Thus the converse holds.

References

1. R. Ahlswede, A note on the existence of the weak capacity for channels with arbitrarily varying channel probability functions and its relation to Shannon's zero error capacity. Ann. Math. Stat. **41**(3), 1027–1033 (1970)
2. R. Ahlswede, Elimination of correlation in random codes for arbitrarily varying channels. Z. Wahrscheinlichkeitstheorie u. verw. Gebiete **44**(2), 159–175 (1978)
3. R. Ahlswede, A method of coding and an application to arbitrarily varying channels. J. Comb. Inf. Syst. Sci. **5**(1), 10–35 (1980)
4. R. Ahlswede, A general theory of information transfer. IEEE Int. Symp. Inf. Theory 396–396 (1993). General theory of information transfer: updated, (Original version: General theory of information transfer, Preprint 97–118, SFB 343 "Diskrete Strukturen in der Mathematik", Universität Bielefeld) General Theory of Information Transfer and Combinatorics, Special Issue of Discrete Applied Mathematics **156**(9), 1348–1388 (2008). https://doi.org/10.1016/j.dam.2007.07.007 [pdf]
5. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inform. Theory **35**(1) (1989)
6. R. Ahlswede, B. Balkenhol, C. Kleinewächter, Identification for Sources, in *General Theory of Information Transfer and Combinatorics*, ed. by R. Ahlswede et al. Lecture Notes in Computer Science, vol. 4123 (2006)
7. I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Channels* (Academic Press, New York, 1981)
8. I. Csiszár, J. Körner, On the capacity of the arbitrarily varying channel for maximum probability of error. Z. Wahrscheinlichkeitstheorie u. verw. Gebiete **57**, 87–101 (1981)
9. I. Csiszár, P. Narayan, The capacity of arbitrarily varying channels revisited: positivity, constraints. IEEE Trans. Inform. Theory **34**(2), 181–193 (1988)
10. M.J. Salariseddigh, U. Pereg, H. Boche, C. Deppe, Deterministic identification over channels with power constraints. arXiv:2010.04239 [cs.IT] (2020)

Identification via Channels with Noisy Feedback

We consider here identification via channels with noisy feedback. Whereas for Shannon's transmission problem the capacity of a discrete memoryless channel does not change with feedback, we know from [2] and [3] (chapters "Identification via Channels" and "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas") that the identification capacity is affected by feedback. We study its dependence on the feedback channel. We prove both a direct and a converse coding theorem. Although a gap exists between the upper and lower bounds provided by these two theorems, the results of chapters "Identification via Channels" and "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas", namely the result for channels without feedback and the result for channels with complete feedback, are all special cases of these two new theorems, because in these cases the bounds coincide.

1 Introduction

In this lecture (see [4]) we introduce and study identification via channels with *noisy* feedback, which is a model that unifies also the case of channels without feedback and the case of channels with complete (or noiseless) feedback. The identification problems for these two cases were studied in the papers [2] and [3], which contain the most basic results in this area.

A communication channel with noisy feedback is denoted here by a quadruple

$$\{\mathcal{X}, W, \mathcal{Y}, \mathcal{Z}\} \tag{1}$$

where $\mathcal{X}$ is the input alphabet, $\mathcal{Y}$ is the output alphabet, $\mathcal{Z}$ is the output alphabet for the feedback and $W = \{W(y, z|x) : x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}\}$ is a stochastic matrix which gives the conditional probability of the output letters y and z when the input

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_5

letter is x. The transmission probability for n-sequences $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$, $y^n = (y_1, \ldots, y_n) \in \mathcal{Y}^n$, $z^n = (z_1, \ldots, z_n) \in \mathcal{Z}^n$ is given by

$$W^n(y^n, z^n|x^n) = \prod_{t=1}^{n} W(y_t, z_t|x_t) \tag{2}$$

for $n = 1, 2, 3, \ldots$. That is, the channel is assumed to be memoryless.

To define identification feedback codes (IDF) in the sense of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" for this channel we let $\mathcal{F}_n$ be the set of all possible vector valued functions

$$f = [f^1, \ldots, f^n] \tag{3}$$

where for $t \in \{2, \ldots, n\}$ f^t is defined on $\mathcal{Z}^{t-1}$ and takes values in $\mathcal{X}$. f^1 is an element of $\mathcal{X}$. It is understood that, when f is used for the transmission over the channel, after the feedback signals $z_1, z_2, \ldots, z_{t-1}$ have been made known to the sender by the feedback channel, the sender transmits $f^t(z_1, \ldots, z_{t-1})$. When $t = 1$, the sender transmits f^1. The joint distribution of the output RV's $Y_1, \ldots, Y_n$ and the feedback random variables $Z_1, \ldots, Z_n$ is determined by the function f used and by W as follows. For $y^n \in \mathcal{Y}^n$, $z^n \in \mathcal{Z}^n$

$$\Pr(Y^n = y^n, Z^n = z^n|f) = W^n(y^n, z^n|f) = \prod_{t=1}^{n} W\big(y_t, z_t|f^t(z_1, \ldots, z_{t-1})\big). \tag{4}$$

We set

$$W^n(y^n|f) = \sum_{z^n \in \mathcal{Z}^n} W^n(y^n, z^n|f) \tag{5}$$

and describe now the feedback codes with randomized encoding strategies, that is, elements of $\mathcal{P}(\mathcal{F}_n)$, the set of probability distributions on $\mathcal{F}_n$.

We remind the reader of Definition 50 Let $N_F(n, \lambda)$ be the maximum integer N for which a randomized (n, N, λ) IDF code for W exists.

We also use

$$\tilde{W}(y|x) = \sum_{z} W(y, z|x)$$

and call $\{\mathcal{X}, \tilde{W}, \mathcal{Y}\}$ the main channel. Our goal is to determine the double exponential growth of $N_F(n, \lambda)$. Insofar we have the following result.

Theorem 75 *If the transmission capacity C of the main channel $\tilde{W}$ is positive, then we have for all $\lambda \in \left(0, \frac{1}{2}\right)$:*

$$\liminf_{n\to\infty} \frac{1}{n} \log\log N_F(n, \lambda) \geq \max I(XU \wedge Y), \tag{6}$$

where the maximum is taken over all joint distributions P_{XYZU} with

$$P_{XYZU}(x, y, z, u) = p(x)W(y, z|x)q(u|x, z)$$

satisfying

$$I(U \wedge Z|XY) < I(X \wedge Y).$$

Furthermore

$$\limsup_{n\to\infty} \frac{1}{n} \log\log N_F(n, \lambda) \leq \max I(XZ \wedge Y), \tag{7}$$

where the maximum is taken over all joint distributions P_{XYZ} with

$$P_{XYZ}(x, y, z) = p(x)W(y, z|x).$$

Remarks

1. This theorem implies the results of chapters "Identification via Channels" and "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas". To see this, observe that in the case without feedback $Z = 0$ and both bounds equal Shannon's transmission capacity. This is the result of chapter "Identification via Channels". In the complete feedback case we have $Z = Y$ and both bounds equal the maximum entropy $H(Y)$. This is the result of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas". Therefore, Theorem 75 can be viewed as a unification of the results of chapters "Identification via Channels" and "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas".
2. A challenging task is to close the gap between the two bounds. We guess that the lower bound is tight, however, a converse proof technique more powerful than those of [2, 6] is needed!

2 Proof of Theorem 75

In this section we prove Theorem 75. The converse part is proved in the first subsection, the direct part is proved in the second subsection and deterministic IDF codes are briefly discussed in the last subsection.

Converse Part of Theorem 75

For identification via channels without feedback a so-called soft converse was proved in the original paper [2] (chapter "Identification via Channels"). The method was refined and strengthened in [6] to give a proof of the strong converse theorem. The reader should study again Sect. 3 before reading this section. The techniques used in [2] and [6] are needed in this section. Although, the proofs in [6] were already simplified compared with those in [2], it is still too long and too complicated to reproduce all the details here. Therefore, in this section, we briefly review some of the key steps and key definitions and present modifications necessary for our purposes. The details can be found in Sect. 3.

We start with a review of some definitions. A probability distribution Q on $\mathcal{A}$ is called an n-type if for any $a \in \mathcal{A}$ $Q(a) \in \left\{\frac{1}{n}, \ldots, \frac{i}{n}, \ldots, 1\right\}$. Let $\mathcal{P}_n$ be the set of all possible n-type's. Recall that for any $Q \in \mathcal{P}_n$

$$\mathcal{T}_Q^n = \left\{a^n : \forall a \in \mathcal{A}, \frac{|\{i : a_i = a\}|}{n} = Q(a)\right\}. \tag{8}$$

Let $\mathcal{B}$ be a subset of $\mathcal{A}^n$ and let $Q_{\mathcal{B}}$ denote the uniform distribution on $\mathcal{B}$. Finally, an ID code $\left\{(Q_i, \mathcal{D}_i) : 1 \le i \le M\right\}$ is called homogeneous (see also Definition 22) if for every $P \in \mathcal{P}_n$

$$Q_1(\mathcal{T}_P^n) = \cdots = Q_M(\mathcal{T}_P^n).$$

Furthermore, recall that for a distribution Q on $\mathcal{X}^n$ and every $P \in \mathcal{P}_n$,

$$Q^P(x^n) = \frac{Q(x^n)}{Q(\mathcal{T}_P^n)}.$$

An ID code is called M-regular if for every $P \in \mathcal{P}_n$ and all i, Q_i^P is of M-type (see Definition 24).

For the proof we need the following propositions which are simple conclusions of Sect. 3.

Proposition 76 *For every (n, N, λ) ID code, $\delta > 0$, $\lambda' > \lambda$ and all sufficiently large n, there exists a homogeneous (n, N', λ') ID code satisfying $N' > N \exp\left\{-\delta n(n+1)^{|\mathcal{X}|}\right\}$, where $\mathcal{X}$ is the input alphabet.*

(Conclusion of Lemma 27)

Proposition 77 *A homogeneous M-regular (n, N, λ) ID code with $\lambda < \frac{1}{2}$ satisfies*

$$\log N \leq n(n+1)M \log |\mathcal{X}|.$$

(Conclusion of Lemma 27)

The main result of Sect. 3, that is the strong converse for the channels without feedback, follows easily from the following result.

Proposition 78 *For every homogeneous (n, N, λ) ID code, $\lambda' > \lambda$, $\gamma > 0$, and for all sufficiently large n, there exists a homogeneous $\exp\{nC + n\gamma\}$-regular (n, N, λ') ID code, where C is the Shannon channel capacity.*

(Conclusion of Lemma 26)

In the proof of this proposition the following lemma is needed.

Lemma 79 *Let $P \in \mathcal{P}_n$ and let Q be a probability distribution on $\mathcal{T}_P^n$. For every $\varepsilon \in [0, \varepsilon_0]$, $\delta \in [0, \delta_0]$ and for all sufficiently large n, there exists an $\exp\{nC + n\gamma\}$-type distribution Q defined on $\mathcal{T}_P^n$, where C is the Shannon channel capacity, such that for every $\mathcal{D} \subset \mathcal{Y}^n$, where $\mathcal{Y}$ is the output alphabet,*

$$\tilde{Q}W^n(\mathcal{D}) \leq (1+\varepsilon)(1-e^{-n\delta})^{-1} QW^n(\mathcal{D}) + e^{-n\delta}, \tag{9}$$

$$\tilde{Q}W^n(\mathcal{D}) \geq (1-\varepsilon)(1-e^{-n\delta}) QW^n(\mathcal{D}) - e^{-n\delta}, \tag{10}$$

where W is the channel transition probability matrix and where $\gamma = \rho(\delta)$, and $\rho : [0, \delta_0] \to R^+$ is a continuous strictly increasing function such that $\rho(0) = 0$.

(Conclusion of Lemma 25)

By checking the proof of this lemma in Sect. 3 we can find that actually the following stronger result was proved:

Lemma 80 *Let $P \in \mathcal{P}_n$ and let Q be a probability distribution on $\mathcal{T}_P^n$. For every $\varepsilon \in [0, \varepsilon_0]$, $\delta \in [0, \delta_0]$, let $\overline{U} = \{U_1, \ldots, U_{M'}\}$ be a random code having independent codewords and with codeword distribution Q. For every $R > I(P, W)$, $M' = e^{nR+n\gamma}$, where γ is defined in Lemma 79, the probability of the event that the following conditions are satisfied approaches 1 as n goes to infinity: For all $\mathcal{D} \subset \mathcal{Y}^n$*

$$\tilde{Q}W^n(\mathcal{D}) \leq (1+\varepsilon)(1-2^{-n\delta})^{-1} QW^n(\mathcal{D}) + e^{-n\delta}, \tag{11}$$

$$\tilde{Q}W^n(\mathcal{D}) \geq (1-\varepsilon)(1-e^{-n\delta}) QW^n(\mathcal{D}) - e^{-n\delta}, \tag{12}$$

where $\tilde{Q}$ is the uniform distribution on $\overline{U}$.

Since the original proof is extremely complicated, instead of copying it step by step we just point out the modifications needed to reach the current conclusions. In this new version, the lemma is strengthened in two points:

1. C is replaced by any $R \geq I(P, W)$

2. the existence of such a code is replaced by the conclusion that the random code satisfies (11) and (12) with probability approaching 1.

The first conclusion can be justified by noticing that Eq. (40) of chapter "Identification via Channels" of the proof is really unnecessary for the lemma. We need only

$$\sup_{V\in\mathcal{P}_{n_\delta P}} I(P,V)+\delta \le I(P,W)+\rho(\delta) \le R+\rho(\delta).$$

This is so, because we are considering a fixed P anyway.

The second conclusion comes from the following refinement of Lemma 32.

Lemma 81 *Let $(\tilde{u}_1,\dots,\tilde{u}_M)$ be the realization of the i.i.d. RV's $(U_1,\dots,U_M)$ with common distribution Q. Let $\mathcal{E}$ be the event that the following conditions are satisfied:*

$$\frac{1}{M}\sum_{i=1}^{M} 1\{\tilde{u}_i \in H_V^P(y^n)\} \le (1+\varepsilon)Q\big(H_V^P(y^n)\big), \qquad \text{for all } y^n \in G_V^P,$$

$$\frac{1}{M}\sum_{i=1}^{M} 1\{\tilde{u}_i \in H_V^P(y^n)\} \ge (1-\varepsilon)Q\big(H_V^P(y^n)\big), \qquad \text{for all } y^n \in G_V^P,$$

and

$$\frac{1}{M}\sum_{i=1}^{M} W_V^P\big((G_V^P)^c|\tilde{u}_i\big) \le e^{-\frac{n\delta}{3}}.$$

Then we have

$$\Pr(\mathcal{E}) \le e^{-\frac{n\delta}{3}} + 2e^{-\frac{\delta^2}{3}e^{n\delta}}.$$

A careful check of the proof of Lemma 32 shows that the conclusion is what is actually proved, although the statement of the lemma is slightly weaker.

In our proofs we use these definitions and results.

The converse part can be proved by following the argument in Sect. 3 with certain modifications. We present in this subsection only the modifications without going into all the details. Since the proofs of the two converses are very similar, pointing out these modifications is good enough for the readers to complete the proof by going through the proof in Sect. 3.

Let f be a feedback coding function. With the function f a set of pairs $\mathcal{S}_f = \{(x^n,z^n) : f(z^n) = x^n\}$ is associated. The probability of a pair $(x^n,z^n) \in \mathcal{S}_f$ is $W^n(z^n|x^n) = \sum_{y^n} W^n(y^n,z^n|x^n)$. This gives a distribution on the set $\mathcal{X}^n \times \mathcal{Z}^n$.

We denote it by $P_f(x^n, z^n)$. Let $Q(\cdot|i) \in \mathcal{P}(\mathcal{F}_n)$ be the distribution of the user i. This induces a distribution on $\mathcal{X}^n \times \mathcal{Z}^n$ defined as follows:

$$P_i(x^n, z^n) = \sum_{f \in \mathcal{F}_n} Q(f|i) P_f(x^n, z^n).$$

An n-type P on $\mathcal{X} \times \mathcal{Z}$ is called ε-typical if for any $x \in \mathcal{X}$ and any $z \in \mathcal{Z}$

$$\left| \frac{P(x, z)}{\sum_{z'} P(x, z')} - W(z|x) \right| \leq \varepsilon.$$

Let $\mathcal{P}_n^{\varepsilon}$ be the set of all possible ε-typical n-types then we have from the weak law of large numbers

$$\lim_{n \to \infty} P_i \left(\bigcup_{P \in \mathcal{P}_n^{\varepsilon}} \mathcal{T}_P^n \right) = 1.$$

The idea of the proof is the following: a feedback code $\left\{(Q(\cdot|i), \mathcal{D}_i) : 1 \leq i \leq N\right\}$ induces an identification code without feedback for the channel from (x, z) to y with transition probability $\tilde{W}(y|x, z) = \frac{W(y,z|x)}{\sum_{y'} W(y',z|x)}$ of the form

$$\left\{(P_i, \mathcal{D}_i) : 1 \leq i \leq N\right\}.$$

Therefore, the proofs of Sect. 3 can be easily modified and applied to this induced code. The following proposition is modified from Lemma 23.

Proposition 82 *For every (n, N, λ)-feedback identification code and $\lambda' > \lambda$, $\gamma > 0$, $\delta > 0$ there exists a homogeneous $\exp\{nT + n\gamma\}$-regular (n, N', λ') ID code, where $N' > N \exp\left\{-\delta n(n+1)^{|\mathcal{X}||\mathcal{Z}|}\right\}$, and $T = \max_p I(XZ \wedge Y)$ where the joint distribution P_{XYZ} satisfies $P_{XYZ}(x, y, z) = p(x)W(y, z|x)$.*

This proposition is proved by an argument as that used in the proof of Lemma 23. The only difference is that Lemma 25 is now replaced by Lemma 80 of this lecture. Therefore C is replaced by

$$\max_{P \in \mathcal{P}_n^{\varepsilon}} I(P, \tilde{W}) = T + \nu,$$

where ν is a continuous function of ε satisfying $\nu(0) = 0$.

The converse is proved now by the same argument as in [6] with Lemma 23 of Sect. 3 replaced by this new proposition.

Proof of the Direct Part of Theorem 75

The proof of the direct part of Theorem 75 is based on two ideas.

The first one is the idea presented in Sect. 3 of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas", where the identification code is constructed by means of two fundamental codes. One code is of block length n and the other one is of block length m, which is much smaller than n. The task of the first code is to set up a common random experiment. The result of the experiment, which is known with high probability to both, the encoder and the decoder, serves as a "public key". According to this key a codeword of the second code is transmitted in the second step. Two different users use the same codeword for the same public key with very small probability. Therefore the goal of identification is achieved.

The second idea is the well-known superposition coding scheme introduced in [5]. In this coding scheme, there are K steps. In each step, a codeword is sent to transmit a new message as well as to resolve an uncertainty left over from the previous step.

For given $\delta > 0$ and $\varepsilon > 0$, let P_{XYZU} be a probability distribution $P_{XYZU}(x, y, z, u) = p(x)W(y, z|x)q(u|x, z)$ that achieves $\max I(U \wedge Z|X)$ under the constraint

$$I(U \wedge Z|XY) \leq I(X \wedge Y) - \delta.$$

We construct three codes of block length n using p and q of the form:

Code $\mathcal{C}_1$ Code $\mathcal{C}_1$ is an $(n, M_n, 2^{-n\alpha})$ channel code for the channel with $W(y|x) = \sum_z W(y, z|x)$. The codewords are assumed to be in $\mathcal{T}_P^n$, where we assume without loss of generality that P is an n-type. The cardinality of the code is $M_n = 2^{nI(X\wedge Y)-\varepsilon n}$, where ε is the given positive number which is assumed to be sufficiently small. Let $\{\mathcal{D}_i^{(n)} : 1 \leq i \leq M\}$ be the decoding regions of the codewords of $\mathcal{C}_1$ with maximum decoding error at most $2^{-\alpha n}$, where $\alpha > 0$ is a continuous function of ε satisfying $\alpha(0) = 0$.

Code Family $\mathcal{C}_2(\mathbf{c})$ Code family $\mathcal{C}_2(c) \subset \mathcal{U}^n$, where $\mathcal{U}$ is the alphabet of the RV U, is a family of source codes indexed by the codewords $c \in \mathcal{C}_1$. This family of codes are required to satisfy the following conditions:

1. Given $c \in \mathcal{C}_1$ the codewords in $\mathcal{C}_2(c)$ are jointly ε-typical with c with respect to the joint distribution P_{XU}, a marginal of P_{XYZU}.
2. The cardinalities of the codes are $N_n = 2^{nI(U\wedge Z|X)+\varepsilon n}$ for all $c \in \mathcal{C}_1$.
3. For each $c \in \mathcal{C}_1$, there exists a mapping

$$f_c : \mathcal{Z}^n \to \mathcal{C}_2(c)$$

satisfying

(i) If $f_c(z^n) \neq 0$, then $f_c(z^n)$, c and z^n are jointly ε-typical.
(ii) $\Pr\left(f_c(z^n) = 0|c\right) \leq 2^{-n\beta}$, where β is a continuous function of ε satisfying $\beta(0) = 0$.

Code Family $\mathcal{C}_3(\mathbf{c})$ The code family $\mathcal{C}_3(c)$ consists of an integer set $\{1, \ldots, L_n\}$, where $L_n = 2^{n(I(Z\wedge U|XY)+n\gamma)}$ and γ is a continuous function of ε satisfying $\gamma(0) = 0$, and two mappings defined as follows:

$$\Phi_c : \mathcal{C}_2(c) \to \{1, \ldots, L_n\} \text{ and } \Psi_c : \mathcal{Y}^n \times \{1, \ldots, L_n\} \to \mathcal{C}_2(c),$$

satisfying

$$\Pr\left(\Psi_c\left(Y^n, \Phi_c(f_c(Z^n))\right) \neq f_c(Z^n)|c\right) \leq 2^{-n\sigma},$$

where σ is a continuous function of ε with $\sigma(0) = 0$.

Lemma 83 *The codes $\mathcal{C}_1$, $\mathcal{C}_2(c)$ and $\mathcal{C}_3(c)$ exist.*

Proof The existence of the code $\mathcal{C}_1$ is based on the channel coding theorem with maximum error criterion [1].

The existence of the code family $\mathcal{C}_2(c)$ is proved by the random coding method. Since the method is classical, we give only a brief outline of the proof. The code is selected randomly according to the distribution $r^n(u^n|c) = \sum_{y^n,z^n} q^n(u^n|c, z^n)W^n(y^n, z^n|c)$. The N_n codewords are selected independently. The mapping f_c is defined by using joint ε-typicality as follows:

1. If there exists a unique codeword $c_2(c, i)$ in $\mathcal{C}_2(c)$ which is jointly ε-typical with z^n and c, then let $f_c(z^n) = i$,
2. otherwise, let $f_c(z^n) = 0$.

The properties of f_c are proved by using the properties of the joint ε-typical sequences. These proofs are standard and therefore omitted.

The existence of the code family $\mathcal{C}_3(c)$ is proved by using the source coding theorem with side information and by noticing the following fact. Since the joint distribution of c, Y^n, U^n, Z^n and $f_c(Z^n)$ are given by the joint distribution of X^n,Y^n, Z^n, U^n, the code $\mathcal{C}_2(c)$ and the mapping f_c, then

$$\begin{aligned} H\left(f_c(Z^n)|c, Y^n\right) &= I\left(Z^n \wedge f_c(Z^n)|c, Y^n\right) \\ &= H\left(Z^n|c, Y^n\right) - H\left(Z^n|c, Y^n, f_c(Z^n)\right) \\ &= H(Z^n|c, Y^n) - H\left(Z^n|c, Y^n, U^n, f_c(Z^n)\right) \end{aligned}$$

(where U^n is the codeword of $\mathcal{C}_2(c)$ whose index is the value of $f_c(Z^n)$)

$$\begin{aligned} &\geq H(Z^n|c, Y^n) - H(Z^n|c, Y^n, U^n) \\ &= \sum_i H(Z_i|x_i, Y_i) - H(Z_i|x_i, Y_i, U_i) \\ &= nI(Z \wedge U|X, Y) + \beta n, \end{aligned}$$

where β goes to zero as ε goes to zero. In the last step of the derivation we used the typicality of the codewords. Applying the source coding theorem with side information (if necessary, we may repeat the same code N times and use nN in place of n) to this case gives the existence of the code family $\mathcal{C}_3(c)$. □

Using these three codes (code families), the coding scheme can be described. It includes two steps. In the first step, the sender and the receiver set up with high probability a common random experiment. In the second step, based on the result of the common random experiment, the sender sends a codeword to the receiver.

We formulate the two steps as follows:

1. The coding is done in K blocks. Each block is of length n. The code $\mathcal{C}_1$ is partitioned into $L_n = 2^{nI(U\wedge Z|XY)+\gamma n}$ subcodes of equal size (roughly) $B_n = 2^{n(I(X\wedge Y)-I(U\wedge Z|X,Y)-\gamma-\varepsilon)}$, which are denoted by $\mathcal{C}_1^m$ for $m = 1, \ldots, L_n$. The codewords of the subcodes are indexed by the numbers in $\{1, \ldots, B_n\}$. Since $I(X \wedge Y) > I(U \wedge Z|X, Y) + \delta$, this is possible for ε small enough. We send a fixed codeword, say c_1 in $\mathcal{C}_1$, in the first block. In the second block, based on the feedback signal Z_1^n in the first block, we send a $c_2 \in \mathcal{C}_1^m$ where $m = \Phi_{c_1}\big(f_{c_1}(Z_1^n)\big)$ is determined by the channel and where c_2 is selected from the code $\mathcal{C}_1^m$ randomly with respect to the uniform distribution. In the following steps i for $i = 3$ to K, if the feedback signal in the previous step is Z_{i-1}^n, then the sender sends $c_i \in \mathcal{C}_1^{m_i}$ where $m_i = \Phi_{c_{i-1}}\big(f_{c_{i-1}}(Z_{i-1}^n)\big)$. c_i is selected randomly with respect to the uniform distribution in $\mathcal{C}_1^{m_i}$. The codewords $c_1, \ldots, c_K$ can be correctly decoded with probabilities at least $1 - K2^{-n\alpha}$. Then the codewords of the second code $\big\{f_{c_1}(Z_1^n), \ldots, f_{c_{K-1}}(Z_{K-1}^n)\big\}$ can be recovered with probability at least $1 - (K-1)2^{-n\sigma}$ under the condition that the codewords from the code $\mathcal{C}_1$ are correctly decoded. The overall misdecoding probability is at most

$$P_e = K(2^{-n\alpha} + 2^{-n\gamma}).$$

This means that with probability at least $1 - P_e$ the sender and the receiver have a common knowledge of $\big\{f_{c_1}(Z_1^n), \ldots, f_{c_{K-1}}(Z_{K-1}^n)\big\}$ and the indices b_i of the codewords c_i in their corresponding subcodes $\mathcal{C}_1^m$ of the code $\mathcal{C}_1$, which are numbers from the set $\{1, \ldots, B_n\}$, at the end of the first K blocks. They are viewed as the result of the common random experiment.

2. Let

$$\mathcal{F} = \left\{F \middle| F : \{1, \ldots, N_n\}^{K-1} \times \{1, \ldots, B_n\}^K \to \mathcal{C}_1\right\}. \tag{13}$$

Each user is assigned a mapping in $\mathcal{F}$. Let F_j be the mapping assigned to the user j. Once $\big\{f_{c_1}(Z_1^n), \ldots, f_{c_{K-1}}(Z_{K-1}^n)\big\}$ and $\{b_1, \ldots, b_K\}$ from the first K steps are available to the sender (and with probability at least $1 - P_e$ correctly to the receiver), the user j selects a codeword $F_j\big(f_{c_1}(Z_1^n), \ldots, f_{c_{K-1}}(Z_{K-1}^n), b_1, \ldots, b_K\big) \in \mathcal{C}_1$, and sends it through the channel. This codeword can be decoded correctly with probability at least $2^{-n\alpha}$.

We now prove that, if the user number satisfies a certain condition, then there exists mappings F_j for the users such that the two kinds of error probabilities of the code described above satisfy the requirement of the identification code.

Obviously, the misrejection probability is at most $1-(K+1)(2^{-n\alpha}+2^{-n\gamma})$, which goes to zero as n goes to infinity for a fixed K.

The misacceptance probability can be estimated as follows: we assume that the mapping is selected according to the uniform distribution on $\mathcal{F}$ and the selection for different users are independent. Let F_0 be the mapping assigned to the user to be identified, let F_i be the mapping assigned to the user i. For a particular $\overline{v} = \big(f_{c_1}(Z_1^n),\ldots,f_{c_{K-1}}(Z_{K-1}^n),b_1,\ldots,b_K\big)$,

$$F_0(\overline{v}) = F_i(\overline{v})$$

with probability $2^{-nI(X\wedge Y)+n\varepsilon} = M_n^{-1}$. The misacceptance probability, when the user is i, is greater than $\lambda + p_e$ (where p_e is the probability that the receiver and the sender can not reach a common result of the random experiment and which goes to zero as n goes to infinity) with probability at most

$$N_n^{K-1}B_n^K(1-2^{-nI(X\wedge Y)+n\varepsilon})^{N_n^{K-1}B_n^K-G}(2^{-nI(X\wedge Y)+n\varepsilon})^G\binom{N_n^{K-1}B_n^K}{G},$$

where

$$G = \max\big\{|\mathcal{V}| : \Pr(\overline{v}\in\mathcal{V}) \le \lambda\big\}.$$

Since any set $\mathcal{V}$ with cardinality at most

$$2^{-2Kn\varepsilon}N_n^{K-1}B_n^K = 2^{n(K-1)I(U\wedge Z|X)-n\varepsilon}B_n^K$$

has a vanishing probability as n goes to infinity, we have

$$G \ge 2^{-2Kn\varepsilon}N_n^{K-1}B_n^K.$$

Therefore, when $2K\varepsilon < I(X\wedge Y)-\varepsilon$ this probability has an upper bound

$$\begin{aligned}&\exp\{-\exp\{n(K-1)I(U\wedge Z|X)+nK(I(X\wedge Y)-I(U\wedge Z|X,Y))\\ &\qquad -n(K+1)\varepsilon - nK\gamma + o(nK)\}\}\\ &\le \exp\{-\exp\{n(K-1)I(XY\wedge Y)-n(K+1)\varepsilon-nK\gamma+o(nK)\}\}.\end{aligned}$$

When user i is to be identified, the probability that there exists a user $j\neq i$ having misacceptance probability at least $\lambda+p_e$ is at most

$$1-(1-\exp\{-\exp\{n(K-1)I(XU\wedge Y)-n(K+1)\varepsilon-nK\gamma+o(nK)\}\})^M.$$

This goes to zero when

$$M \leq \exp\{\exp\{n(K-1)I(XY \wedge Y) - n(K+1)\varepsilon - nK\gamma + o(nK)\} - \varepsilon n\}.$$

If we delete users for whom there exists at least one different user having misacceptance probability at least $\lambda + p_e$, then the number of users deleted is in average a vanishing portion of all M users. Therefore, there exists a set of mappings for M' users out of the M users, where

$$M' = \exp\{\exp\{n(K-1)I(XU \wedge Y) - n(K+1)\varepsilon - nK\gamma + o(nK)\} - 2\varepsilon n\},$$

such that the subcode of these users satisfies all requirements of the identification code. The rate of the code is at least $\frac{K-1}{K+1} I(XY \wedge Y) - 2\varepsilon - \gamma$. For large K, this is greater than $I(XU \wedge Y) - 3\varepsilon - \gamma$. Letting ε go to zero the direct part of Theorem 75 is proved.

Deterministic Identification Codes

Another result of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" is for deterministic feedback identification codes. Actually, the same concept can be defined for channels with noisy feedback. In this subsection, we present only the definitions and results for this concept without detailed proofs. These results are proved by the method used for the randomized identification code with some modifications.

Recall that a deterministic (n, N, λ) IDF code for W is a system

$$\left\{(f_i, \mathcal{D}_i) : i = 1, \ldots, N\right\} \text{ with } f_i \in \mathcal{F}_i, \mathcal{D}_i \subset \mathcal{Y}^n \text{ for } i \in \{1, \ldots, N\},$$

and

$$W^n(\mathcal{D}_i | f_i) \geq 1 - \lambda, \tag{14}$$

$$W^n(\mathcal{D}_j | f_i) \leq \lambda, \tag{15}$$

for all $i, j \in \{1, \ldots, N\}$ and $i \neq j$.

Let $N_f(n, \lambda)$ be the maximum integer N for which a deterministic (n, N, λ) IDF code exists. Here are our results for this quantity.

Theorem 84 *If the transmission capacity C of W is positive, then we have for all $\lambda \in \left(0, \frac{1}{2}\right)$:*

$$\liminf_{n\to\infty} \frac{1}{n} \log\log N_f(n, \lambda) \geq \max I(Z \wedge U|X), \tag{16}$$

where the maximum is over all joint distributions P_{XYZU} with

$$P_{XYZU}(x, y, z, u) = p(x)W(y, z|x)q(u|x, z)$$

satisfying

$$I(U \wedge Z|XY) < I(X \wedge Y).$$

Furthermore

$$\limsup_{n\to\infty} \frac{1}{n} \log\log N_f(n,\lambda) \le \min\left\{ \max I(XZ \wedge Y), \max H(Z|X)\right\}, \tag{17}$$

where the maximum is taken over all joint distributions P_{XYZ} with

$$P_{XYZ}(x,y,z) = p(x)W(y,z|x).$$

References

1. R. Ahlswede, A method of coding and an application to arbitrarily varying channels. J. Comb. Inf. Syst. Sci. **5**(1), 10–35 (1980)
2. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**(1), 15–29 (1989)
3. R. Ahlswede, G. Dueck, Identification in the presence of feedback—a discovery of new capacity formulas. IEEE Trans. Inf. Theory **35**(1), 30–36 (1989)
4. R. Ahlswede, Z. Zhang, New directions in the theory of identification via channels. IEEE Trans. Inf. Theory **41**(4), 1040–1050 (1995)
5. T.M. Cover, C.S.K. Leung, An achievable rate region for the multiple access channel with feedback. IEEE Trans. Inf. Theory **27**(3), 292–298 (1981)
6. T.S. Han, S. Verdú, New results in the theory of identification via channels. IEEE Trans. Inf. Theory **38**(1), 14–25 (1992)

Identification via Discrete Memoryless Wiretap Channels

We consider here identification via wiretap channels. A "Dichotomy Theorem" is proved which says here that the second order secrecy identification capacity is the same as Shannon's capacity for the main channel as long as the secrecy transmission capacity of the wiretap channel is not zero, and zero otherwise.

Equivalently we can say that the identification capacity is not lowered by the presence of a wiretapper as long as one bit can be transmitted (or identified) correctly with arbitrarily small error probability. This is in strong contrast to the case of transmission.

1 Introduction

We consider here identification via a wiretap channel. This channel was introduced by A.D. Wyner [2]. It can be viewed as a probabilistic model for cryptography.

The channel has two outputs. One is for the legitimate receiver and the other, which is a degraded version of the first output, is for the wiretapper. The goal of the communication is to send messages to the legitimate receiver while the wiretapper must be kept ignorant. A more general version of the wiretap channel was studied in [1], where the assumption that the output for the wiretapper is a degraded version of the output for the legitimate receiver is dropped. We address here right away this general model (Fig. 1).

Definition 85 As defined in [1], a wiretap channel is a quintuple

$$\{\mathcal{X}, W, V, \mathcal{Y}, \mathcal{Z}\}, \tag{1}$$

where $\mathcal{X}$ is the input alphabet, $\mathcal{Y}$ is the output alphabet for the legitimate receiver, $\mathcal{Z}$ is the output for the wiretapper, $W = \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ is the channel

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_6

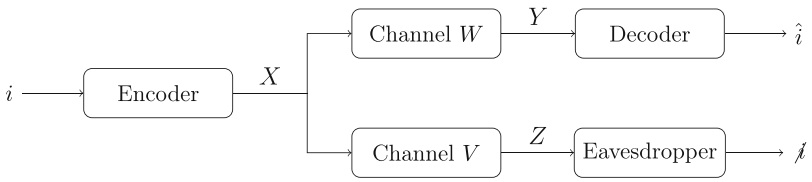


Fig. 1 The wiretap channel

transmission matrix, whose output is available to the legitimate receiver, and $V = \{V(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ is the channel transmission matrix, whose output is available to the wiretapper. The channel is assumed to be memoryless, that is, the conditional probabilities of the output word y^n and z^n given the input word x^n are $W^n(y^n|x^n) = \prod_{t=1}^n W(y_t|x_t)$ and $V^n(z^n|x^n) = \prod_{t=1}^n V(z_t|x_t)$.

In the classical transmission problem, an (n, M, ε)-code for the wiretap channel is defined as a system

$$\{(c_i, \mathcal{D}_i) : 1 \le i \le M\}, \tag{2}$$

where for all i, $c_i \in \mathcal{X}^n$ are the codewords and $\mathcal{D}_i \subset \mathcal{Y}^n$ are the disjoint decoding sets. It is required that for any i

$$\lambda_i \triangleq W^n(\mathcal{D}_i^c|c_i) \le \varepsilon, \tag{3}$$

and if X^n has uniform distribution over $\{c_i : 1 \le i \le M\}$, then

$$\frac{1}{n} I(X^n \wedge Z^n) \le \varepsilon. \tag{4}$$

The secrecy capacity of the wiretap channel is defined as the maximum rate of any code which satisfies these conditions. Formally, let

$$M(n, \varepsilon) = \max\{M : \exists \text{ an } (n, M, \varepsilon) \text{ code}\}, \tag{5}$$

then the secrecy capacity of the wiretap channel is defined as

$$C_s = \max\{R : \forall\, \varepsilon > 0\ \exists\, n(\varepsilon)\ \forall\, n \ge n(\varepsilon)\ \ M(n, \varepsilon) \ge 2^{nR}\}. \tag{6}$$

The secrecy capacity of the general wiretap channel was determined in [1]. It is

$$C_s = \max_{U \to X \to YZ} I(U \wedge Y) - I(U \wedge Z). \tag{7}$$

The problem of identification via this channel in the sense of chapter "Identification via Channels" can be formulated as follows: For any finite set A let $\mathcal{P}(A)$ stand for the set of all probability distributions on A.

Definition 86 A randomized (n, N, λ)-identification code for the wiretap channel is a system

$$\big\{(Q(\cdot|i), \mathcal{D}_i) : 1 \le i \le N\big\}, \tag{8}$$

where, for all i, $Q(\cdot|i) \in \mathcal{P}(\mathcal{X}^n)$ and $\mathcal{D}_i \subset \mathcal{Y}^n$, which satisfies the following three conditions:

(i) for all i

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|i) W^n(\mathcal{D}_i|x^n) \ge 1 - \lambda, \tag{9}$$

(ii) for all pairs (i, j) with $i \ne j$

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|j) W^n(\mathcal{D}_i|x^n) \le \lambda, \tag{10}$$

and

(iii) for any pair (i, j) with $i \ne j$ and any $\mathcal{V} \subset \mathcal{Z}^n$

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|j) V^n(\mathcal{V}|x^n) + \sum_{x^n \in \mathcal{X}^n} Q(x^n|i) V^n(\mathcal{V}^c|x^n) \ge 1 - \lambda. \tag{11}$$

In contrast to the transmission problem, the decoding sets for the identification problem are not necessarily disjoint.

Condition (iii) enforces that the wiretapper is kept with his error probability close to $\frac{1}{2}$. This is the highest possible value, because the wiretapper could just accept an i of his interest with probability $\frac{1}{2}$. Mathematically, condition (iii) means of course that the output distributions for the wiretap channel are almost the same for any two input distributions $Q(\cdot|i)$ and $Q(\cdot|j)$.

The maximum N for which a randomized (n, N, λ)-identification code exists is denoted by $N(n, \lambda)$. Define the secrecy identification capacity of the wiretap channel by letting

$$C_{\mathrm{SI}} = \max\big\{R : \forall\, \lambda > 0\ \exists\, n(\lambda)\ \forall\, n \ge n(\lambda)\ \ N(n, \lambda) \ge 2^{2^{nR}}\big\}.$$

The main result on this problem is the following:

Theorem 87 (Dichotomy Theorem) *Let C be the Shannon capacity of the channel W and let C_s be the secrecy transmission capacity of the wiretap channel, then*

$$C_{\mathrm{SI}} = C, \quad \textit{if } C_s > 0, \tag{12}$$

and

$$C_{\mathrm{SI}} = 0, \quad \textit{if } C_s = 0. \tag{13}$$

Remark Still it may be of interest to know whether the strong converse holds.

2 Proof of Theorem 87

In this section we prove Theorem 87. The direct part is proved in the first three subsections and the converse part is proved in the last subsection.

Preparations for the Proof of the Direct Part
In the proof of the direct part of Theorem 87, we use a coding technique introduced in Sect. 3 of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas", where the identification code is constructed by means of two fundamental codes. This coding technique has been already used for channels with noisy feedback. By Shannon's coding theorem, we know that for every $\varepsilon > 0$, $\varepsilon < C$, where C is the Shannon capacity of the main channel W, there is a $\delta = \delta(\varepsilon) > 0$ and an $n_0(\varepsilon)$ such that for $n > n_0(\varepsilon)$, there exists an $(n, M, 2^{-n\delta})$ code

$$\mathcal{C}_1 = \left\{(\tilde{c}_j, \tilde{\mathcal{C}}_j) : j = 1, \ldots, M\right\} \tag{14}$$

where $M = 2^{n(C-\varepsilon)}$. This code serves as the first fundamental code which will be used in the construction of the identification code.

For the wiretap channel, in place of the second fundamental code, we use a code system which consists of a code of length m and a collection of subcodes of this code. This code system should satisfy certain conditions described later. To construct this code system, we use a RV U jointly distributed with RV's X, Y and Z. The joint distribution of these RV's is of the form

$$P_{UXYZ}(u, x, y, z) = q(u)r(x|u)W(y, z|x), \tag{15}$$

which satisfies the condition

$$I(U \wedge Y) > I(U \wedge Z). \tag{16}$$

The following proposition gives the existence of a code system which will be used in the construction of the identification code. Let $\tilde{W}(y|u) = \sum_x r(x|u)W(y|x)$. $\tilde{W}$ is called the u, y-channel. Let $\tilde{V}(z|u) = \sum_x r(x|u)V(z|x)$. $\tilde{V}$ is called the u, z-channel.

Proposition 88 *For any $\varepsilon > 0$ there exists a $\delta(\varepsilon) > 0$ and an m_0 such that for any $m > m_0$ there exist an $(m, M', 2^{-m\delta})$ code*

$$\mathcal{C}_2 = \left\{(c_i', \mathcal{D}_i') : 1 \le i \le M'\right\} \tag{17}$$

for the u, y-channel $\tilde{W}$, where $M' = 2^{m(I(U\wedge Y)-\varepsilon)}$, and $L = 2^{m\varepsilon}$ subcodes

$$\mathcal{L} = \{\mathcal{C}_i^* : i = 1, \ldots, L\} \tag{18}$$

of the code $\mathcal{C}_2$ with a common cardinality $M^* = 2^{m(I(U\wedge Z)+\varepsilon)}$ having the following two properties:

(i) The number of common codewords of any two different subcodes is at most εM^*.

(ii) Let Q_i be the uniform distribution on $\mathcal{C}_i^*$. Then for every pair i and $j : i \ne j$,

$$D(Q_i \tilde{V}^m \| Q_j \tilde{V}^m) \le \varepsilon.$$

This proposition will be proved in the third subsection.

Proof of the Direct Part of Theorem 87

Using these two fundamental codes, we can construct the identification code as follows:

Let $\{1, \ldots, L\}$ be the index set of $\mathcal{L}$, the set of subcodes of the second fundamental code $\mathcal{C}_2$. We consider mappings of the form

$$\phi : \mathcal{C}_1 \to \{1, \ldots, L\}. \tag{19}$$

Let Φ be the set of all possible mappings ϕ. The Hamming distance of two mappings ϕ and ψ is defined as the number of codewords of $\mathcal{C}_1$ at which ϕ and ψ have different values. It can easily be seen that we can construct by the greedy algorithm a set of mappings of cardinality at least

$$N = \frac{L^M}{\sum_{k=\varepsilon M}^{M} \binom{M}{k}(L-1)^{M-k}} \ge \frac{2^{M D(\varepsilon\|L^{-1})}}{M} \ge 2^{2^{nC-2\varepsilon n}},$$

where

$$D(\varepsilon\|L^{-1}) = \varepsilon \log \frac{\varepsilon}{L^{-1}} + (1-\varepsilon)\log\frac{1-\varepsilon}{1-L^{-1}},$$

satisfying the property that the Hamming distance between any pair of different mappings in the set is at least $M - \varepsilon M$. Let this set of mappings be

$$\Phi^* = \{\phi_i : 1 \le i \le N\}, \tag{20}$$

which will be used in the construction of the identification codes. Let P be the uniform distribution over the code $\mathcal{C}_1$, q_i^* be the uniform distribution on the subcode $\mathcal{C}_i^*$, and let $Q_i^* = q_i^* r^m$, which is a distribution on the alphabet $\mathcal{X}^m$. Let mapping ϕ_i be assigned to user i, then the distribution $Q(\cdot|i)$ in the identification code is defined as follows: for $x^n \in \mathcal{X}^n$ and $x^m \in \mathcal{X}^m$

$$Q\big((x^n, x^m)|i\big) = P(x^n) Q^*_{\phi_i(x^n)}(x^m). \tag{21}$$

The decoding set $\mathcal{D}_i$ is defined as

$$\mathcal{D}_i = \bigcup_{t=1}^{M} \tilde{\mathcal{D}}_t \times \mathcal{D}^{\phi_i(\tilde{c}_t)},$$

where

$$\mathcal{D}^{(s)} = \bigcup_{c' \in \mathcal{C}_s^*} \mathcal{D}'_{c'},$$

and where $\tilde{\mathcal{D}}_t$ was defined in (14) and $\mathcal{D}'_i$ when $c' = c'_i$ (which was defined in (17)). We now estimate the first and the second kinds of error probabilities of this code.

We have for user i, the first kind of error probability is

$$\sum_{(x^n, x^m) \in \mathcal{X}^{n+m}} Q\big((x^n, x^m)|i\big) W^{n+m}(\mathcal{D}_i|x^n, x^m) \tag{22}$$

$$\geq 1 - 1 + \sum_{t=1}^{M} P(\tilde{c}_t) W^n(\tilde{\mathcal{D}}_t|\tilde{c}_t) - 1 \tag{23}$$

$$+ \sum_{t=1}^{M} P(\tilde{c}_t) \sum_{c' \in \mathcal{C}_{\Phi_i(\tilde{c}_t)}} q^*_{\phi_i(\tilde{c}_t)}(c') \tilde{W}^m\big(\mathcal{D}^{(\phi_i(\tilde{c}_t))}|c'\big) \tag{24}$$

$$\geq 1 - 2^{-n\delta} - 2^{-m\delta}. \tag{25}$$

If $i \neq j$, then the second kind error probability for the user j, when user i is to be identified, is

$$\sum_{(x^n, x^m) \in \mathcal{X}^{n+m}} Q\big((x^n, x^m)|j\big) W^n(\mathcal{D}_i|x^n, x^m)$$

$$\leq 1 - \sum_{t=1}^{M} P(\tilde{c}_t) W^n(\tilde{\mathcal{D}}_t|\tilde{c}_t) + \sum_{t=1}^{M} P(\tilde{c}_t) \sum_{c' \in \mathcal{C}_{\phi_j(\tilde{c}_t)}} q^*_{\phi_j(\tilde{c}_t)}(c') \tilde{W}^m(\mathcal{D}^{(\phi_i(\tilde{c}_t))}|c')$$

$$\le 2^{-n\delta} + \sum_{t:\phi_i(\tilde{c}_t)=\phi_j(\tilde{c}_t)} P(\tilde{c}_t) + \sum_{t:\phi_i(\tilde{c}_t)\neq\phi_j(\tilde{c}_t)} P(\tilde{c}_t) \sum_{c'\in\mathcal{C}_{\phi_j(\tilde{c}_t)}} q^*_{\phi_j(\tilde{c}_t)}(c')\tilde{W}^m(\mathcal{D}^{(\phi_i(\tilde{c}_t))}|c')$$

$$\le 2^{-n\delta} + \varepsilon + \sum_{t:\phi_i(\tilde{c}_t)\neq\phi_j(\tilde{c}_t)} P(\tilde{c}_t) \sum_{c'\in\mathcal{C}_{\phi_j(\tilde{c}_t)}} q^*_{\phi_j(\tilde{c}_t)}(c') \sum_{c''\in\mathcal{C}_{\phi_i(\tilde{c}_t)}} \tilde{W}^m(\mathcal{D}'_{c''}|c')$$

$$\le 2^{-n\delta} + \varepsilon + \sum_{t:\phi_i(\tilde{c}_t)\neq\phi_j(\tilde{c}_t)} P(\tilde{c}_t) \sum_{c'\in\mathcal{C}_{\phi_j(\tilde{c}_t)}\cap\mathcal{C}_{\phi_i(\tilde{c}_t)}} q^*_{\phi_j(\tilde{c}_t)}(c')\tilde{W}^m(\mathcal{D}'_{c'}|c')$$

$$+ \sum_{t:\phi_i(\tilde{c}_t)\neq\phi_j(\tilde{c}_t)} P(\tilde{c}_t) \sum_{c'\in\mathcal{C}_{\phi_j(\tilde{c}_t)}} q^*_{\phi_j(\tilde{c}_t)}(c')\tilde{W}^m\left((\mathcal{D}'_{c'})^c|c'\right)$$

$$\le 2^{-n\delta} + 2\varepsilon + 2^{-m\delta}. \tag{26}$$

For a fixed $\lambda < \frac{1}{2}$, let n be sufficiently large and then m be sufficiently large and ε sufficiently small, the requirement for these two error probabilities in the definition of the identification code can be satisfied.

The next problem is to prove that the wiretapper can not identify, that is, we need to prove (11).

We see that for any pair $i \neq j$,

$$\begin{aligned} &D\big(Q(\cdot|i)V^{n+m}\|Q(\cdot|j)V^{n+m}\big) \\ &= \sum_{\tilde{c}\in\mathcal{C}_1} P(\tilde{c})D(Q_{\phi_i(\tilde{c})}V^m\|Q_{\phi_j(\tilde{c})}V^m) \le \sum_{\tilde{c}\in\mathcal{C}_1} P(\tilde{c})\varepsilon = \varepsilon. \end{aligned} \tag{27}$$

Therefore, for any region $\mathcal{V} \subset \mathcal{Z}^{n+m}$, denoting by V_i the distribution $Q(\cdot|i)V^{n+m}$ and by V_j the distribution $Q(\cdot|j)V^{n+m}$, we obtain

$$V_i(\mathcal{V})\log\frac{V_i(\mathcal{V})}{V_j(\mathcal{V})} + V_i(\mathcal{V}^c)\log\frac{V_i(\mathcal{V}^c)}{V_j(\mathcal{V}^c)} \le D(V_i\|V_j) \le \varepsilon.$$

Since

$$\begin{aligned} D(\alpha\|\beta) &= \alpha\log\frac{\alpha}{\beta} + (1-\alpha)\log\frac{1-\alpha}{1-\beta} \\ &= -\alpha\log\left(1+\frac{\beta-\alpha}{\alpha}\right) - (1-\alpha)\log\left(1+\frac{\beta-\alpha}{1-\alpha}\right) \\ &\ge -\alpha\frac{\beta-\alpha}{\alpha} - (1-\alpha)\frac{\beta-\alpha}{1-\alpha} \\ &= 2(\beta-\alpha). \end{aligned}$$

Similarly

$$D(\beta\|\alpha) \geq 2(\alpha - \beta).$$

From

$$D(V_i\|V_j) \leq \varepsilon,$$

and

$$D(V_j\|V_i) \leq \varepsilon,$$

we obtain

$$V_i(\mathcal{V}) \log \frac{V_i(\mathcal{V})}{V_j(\mathcal{V})} + V_i(\mathcal{V}^c) \log \frac{V_i(\mathcal{V}^c)}{V_j(\mathcal{V}^c)} \leq \varepsilon,$$

and

$$V_j(\mathcal{V}) \log \frac{V_j(\mathcal{V})}{V_i(\mathcal{V})} + V_j(\mathcal{V}^c) \log \frac{V_j(\mathcal{V}^c)}{V_i(\mathcal{V}^c)} \leq \varepsilon.$$

This implies

$$|V_i(\mathcal{V}) - V_j(\mathcal{V})| \leq \frac{\varepsilon}{2}.$$

We can see that the last inequality implies the last requirement for the identification code. The direct part of Theorem 2 is proved.

Proof of Proposition 88
To construct the second fundamental code with the required structure and properties, we use the random coding method. It is well known that there exists an $(m, M', 2^{-m\delta})$ code $\mathcal{C}_2$. Without loss of generality, we can assume that the distribution q of U is an m-type and $\mathcal{C}_2 \subset \mathcal{T}_q^m$. To construct the random family of subcodes of this code $\{C_i^* : i = 1, \ldots, L\}$ satisfying the required properties, we proceed as follows: The size of the code is $M' = 2^{m(I(U\wedge Y)-\varepsilon)}$ and the common size of the subcodes is $M^* = 2^{m(I(U\wedge Z)+\gamma)}$, where the number γ is introduced in Lemma 79. This is possible because $I(U \wedge Y) > I(U \wedge Z)$. A subcode of this code can be selected by using a binary M'-sequence $s = (s_1, \ldots, s_{M'})$, where s_i is either 0 or 1. A codeword c_i^* is in the subcode $\mathcal{C}_s$ if and only if $s_i = 1$. After the code $\mathcal{C}_2$ is selected, we select $L = 2^{m\varepsilon}$ subcodes of $\mathcal{C}_2$. Let these subcodes be $\mathcal{C}_i^*$. These codes are chosen randomly by letting for any s of weight M^*

$$\Pr(\mathcal{C}_i^* = \mathcal{C}_s) = \frac{1}{\binom{M'}{M^*}},$$

and the selections for different i are done independently. We are going to prove that for the random code chosen as above with probability approaching 1 as m goes to infinity the code has the required properties.

From Lemma 80, the subcode selected satisfies the following condition with probability approaching 1: for every $\mathcal{D} \subset \mathcal{Z}^m$, let $\tilde{Q}$ be the uniform distribution on the subcode, then

$$\tilde{Q}\tilde{V}^m(\mathcal{D}) \leq (1+\varepsilon)(1-e^{-m\delta})^{-1} Q\tilde{V}^m(\mathcal{D}) + e^{-m\delta}, \tag{28}$$

$$\tilde{Q}\tilde{V}^m(\mathcal{D}) \geq (1-\varepsilon)(1-e^{-m\delta}) Q\tilde{V}^m(\mathcal{D}) - e^{-m\delta}. \tag{29}$$

For sufficiently large m, we may assume that this probability is at most ε. We prove that if two subcodes $\mathcal{C}_i$ and $\mathcal{C}_j$ both satisfy this condition, then

$$D(Q_i\tilde{V} \| Q_j\tilde{V}) \leq \beta(\varepsilon), \tag{30}$$

where $\beta(\varepsilon)$ goes to zero as ε does and where Q_i denotes the uniform distribution on the code $\mathcal{C}_i^*$.

This is proved as follows: Let

$$\mathcal{D}_t = \left\{ z^m : \sum_{u^m} Q_i \tilde{V}^m(z^m|u^m) > t \sum_{u^m} Q_j \tilde{V}^m(z^m|u^m) \right\},$$

then

$$(1+\varepsilon)(1-e^{-m\delta})^{-1} Q\tilde{V}^m(\mathcal{D}_t) + e^{-m\delta} > t(1-\varepsilon)(1-e^{-m\delta}) Q\tilde{V}^m(\mathcal{D}_t) - te^{-m\delta}.$$

This implies

$$(1+t)e^{-m\delta} > \left(t(1-\varepsilon)(1-e-m\delta) - (1+\varepsilon)(1-e^{-m\delta})^{-1}\right) Q\tilde{V}^m(\mathcal{D}_t),$$

that is

$$Q\tilde{V}^m(\mathcal{D}_t) < \frac{(1+t)e^{-m\delta}}{\left(t(1-\varepsilon)(1-e^{-m\delta}) - (1+\varepsilon)(1-e^{-m\delta})^{-1}\right)}.$$

We know that

$$\frac{\sum_{u^m} Q_i(u^m)\tilde{V}^m(z^m|u^m)}{\sum_{u^m} Q_j(u^m)\tilde{V}^m(z^m|u^m)}$$

is at most $e^{\alpha m}$, where $\alpha = \log \frac{\max \tilde{V}(z|u)}{\min \tilde{V}(z|u)}$. Letting $t = \frac{(1+\varepsilon)^2(1-e^{-m\delta})^{-1}}{(1-\varepsilon)(1-e^{-m\delta})}$, we obtain

$$D(Q_i \tilde{V} \| Q_j \tilde{V}) \leq \frac{(1+t)e^{-m\delta}(1-e^{-m\delta})}{\varepsilon(1+\varepsilon)} m\alpha + \log t. \tag{31}$$

We can easily see that the right hand side of (31) approaches zero as m goes to infinity and then ε goes to zero.

By randomly selecting L subcodes, then deleting those subcodes for which the conditions in the lemma are not satisfied, the number of remaining subcodes is in average at least $L(1-\varepsilon)$. This is enough for our purpose.

We now prove that the intersection of two subcodes has more than εM^* codewords with doubly exponentially small probability. This is done by the following calculation:

$$\Pr(|\mathcal{C}_i \cap \mathcal{C}_j| > \varepsilon M^*) \leq M^* \frac{\binom{M^*}{\varepsilon M^*}\binom{M'-M^*}{(1-\varepsilon)M^*}}{\binom{M'}{M^*}},$$

which is doubly exponentially small. This proves that with probability approaching 1, any pair of the subcodes satisfy the condition that their intersection has a size at most εM^*. The proposition is proved.

Proof of the Converse Part

We begin with the following lemma.

Lemma 89 *Let Q_1 and Q_2 be two distributions on $\mathcal{Z}^m$ and for any $\mathcal{V} \subset \mathcal{Z}^m$,*

$$Q_1(\mathcal{V}) + Q_2(\mathcal{V}^c) > 1 - \varepsilon, \tag{32}$$

and let U be a binary RV with uniform distribution and $V(z^m|U=i) = Q_i$ for $i = 1, 2$ then

$$I(U \wedge Z^m) \leq \inf_{x>0} \left\{ \frac{2}{x} + \log \frac{1}{1 - \frac{1}{2}x\varepsilon} \right\}. \tag{33}$$

Proof Let

$$\mathcal{V}_t = \left\{ z^m : Q_1(z^m) < t\, Q_2(z^m) \right\},$$

then

$$Q_1(\mathcal{V}_1) + Q_2(\mathcal{V}_1^c) \leq t\, Q_2(\mathcal{V}_t) + 1 - Q_2(\mathcal{V}_t).$$

Therefore,

$$t\,Q_2(\mathcal{V}_t) + 1 - Q_2(\mathcal{V}_t) > 1 - \varepsilon.$$

This implies

$$(1-t)\,Q_2(\mathcal{V}_t) < \varepsilon,$$

that is, for $0 \le t < 1$

$$Q_2(\mathcal{V}_t) < \frac{\varepsilon}{1-t}.$$

Similarly, for $t > 1$,

$$Q_1(\mathcal{V}_1^c) < \frac{t\varepsilon}{t-1}.$$

Therefore, for any $0 \le t < 1$,

$$\begin{aligned} I(U \wedge Z^m) = \sum_{z^m} \frac{1}{2} Q_1(z^m) \log \frac{\frac{1}{2} Q_1(z^m)}{\frac{1}{4}(Q_1(z^m) + Q_2(z^m))} \\ + \frac{1}{2} Q_2(z^m) \log \frac{\frac{1}{2} Q_2(z^m)}{\frac{1}{4}(Q_1(z^m) + Q_2(z^m))}. \end{aligned}$$

For any t, $0 \le t < 1$, we have for $z^m \notin \mathcal{V}_t \cup \mathcal{V}_{\frac{1}{t}}^c$

$$\left| \log \frac{\frac{1}{2} Q_1(z^m)}{\frac{1}{4}(Q_1(z^m) + q_2(z^m))} \right| \le 1 + \log \frac{1}{1+t},$$

and for any z^m

$$\begin{aligned} \frac{1}{2} Q_1(z^m) \log \frac{\frac{1}{2} Q_1(z^m)}{\frac{1}{4}(Q_1(z^m) + Q_2(z^m))} + \frac{1}{2} Q_2(z^m) \log \frac{\frac{1}{2} Q_2(z^m)}{\frac{1}{4}(Q_1(z^m) + Q_2(z^m))} \\ \le \frac{1}{2}\big(Q_1(z^m) + Q_2(z^m)\big). \end{aligned}$$

Therefore

$$
\begin{aligned}
I(U \wedge Z^m) &\leq \sum_{z^m \in \mathcal{V}_t \cup \mathcal{V}_{\frac{1}{t}}^c} Q_1(z^m) \log \frac{\frac{1}{2}Q_1(z^m)}{\frac{1}{4}(Q_1(z^m)+Q_2(z^m))} \\
&\quad + \frac{1}{2}Q_2(z^m) \log \frac{\frac{1}{2}Q_2(z^m)}{\frac{1}{4}(Q_1(z^m)+Q_2(z^m))} + 1 + \log \frac{1}{1+t} \\
&\leq \frac{1}{2}\big(Q_1(\mathcal{V}_t) + Q_2(\mathcal{V}_t) + Q_1(\mathcal{V}_{\frac{1}{t}}^c)\big) + 1 + \log \frac{1}{1+t} \\
&\leq (1+t)\frac{\varepsilon}{1-t} + 1 + \log \frac{1}{1+t}.
\end{aligned}
$$

Taking $t = 1 - x\varepsilon$, we obtain

$$
I(U \wedge Z^m) \leq \frac{2}{x} + \log \frac{1}{1 - \frac{1}{2}x\varepsilon}.
$$

This proves the lemma. □

Lemma 90 *Let Q_1 and Q_2 be two distributions on Z^m for which there exists a $\mathcal{V} \subset \mathcal{Z}^m$ such that*

$$
Q_1(\mathcal{V}) + Q_2(\mathcal{V}^c) < \varepsilon, \tag{34}
$$

and let U be a binary RV with uniform distribution and $V(z^m|U = i) = Q_i$ for $i = 1, 2$, then

$$
I(U \wedge Z^m) \geq h\left(\frac{1}{2}(1-\varepsilon)\right). \tag{35}
$$

Proof

$$
\begin{aligned}
&I(U \wedge Z^m) \qquad (36)\\
&\geq -\left[\frac{1}{2}Q_1(\mathcal{V}) \log \frac{Q_1(\mathcal{V})+Q_2(\mathcal{V})}{2Q_1(\mathcal{V})} + \frac{1}{2}Q_2(\mathcal{V}) \log \frac{Q_1(\mathcal{V})+Q_2(\mathcal{V})}{2Q_2(\mathcal{V})}\right. \\
&\quad \left. + \frac{1}{2}Q_1(\mathcal{V}^c) \log \frac{Q_1(\mathcal{V}^c)+Q_2(\mathcal{V}^c)}{2Q_1(\mathcal{V}^c)} + \frac{1}{2}Q_2(\mathcal{V}^c) \log \frac{Q_1(\mathcal{V}^c)+Q_2(\mathcal{V}^c)}{2Q_2(\mathcal{V}^c)}\right] \\
&\geq h\left(\frac{1}{2}(1-\varepsilon)\right). \qquad (37)
\end{aligned}
$$

□

The existence of the identification code at a positive rate implies the existence of distributions $Q(\cdot|i)$ and decoding regions $\mathcal{D}_i$ such that

1. for all i

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|i) W^n(\mathcal{D}_i|x^n) \geq 1 - \lambda, \tag{38}$$

2. for any pair $i \neq j$

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|j) W^n(\mathcal{D}_i|x^n) \leq \lambda, \tag{39}$$

and

3. for any pair $i \neq j$ and any $\mathcal{V} \subset \mathcal{Z}^n$

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|j) V^n(\mathcal{V}|x^n) + \sum_{x^n \in \mathcal{X}^n} Q(x^n|i) V^n(\mathcal{V}^c|x^n) \geq 1 - \lambda. \tag{40}$$

From the first two properties, we obtain

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|i) W^n(\mathcal{D}_i^c|x^n) + \sum_{x^n \in \mathcal{X}^n} Q(x^n|j) W^n(\mathcal{D}_i|x^n) \leq 2\lambda,$$

which implies by Lemma 89 that for the RV U defined there we have

$$I(U \wedge Y^n) \geq h\left(\frac{1}{2}(1 - 2\lambda)\right).$$

By the third property and Lemma 81, the same RV U satisfies

$$I(U \wedge Z^n) \leq \inf_{x>0} \left\{ \frac{2}{x} + \log \frac{1}{1 - \frac{1}{2}x\lambda} \right\}.$$

Then λ is small enough, we obtain the following conclusion: there exists a RV U satisfying

(i) $U \to X^m \to Y^m Z^m$
(ii) $I(U \wedge Y^m) > I(U \wedge Z^m)$.

The converse part is proved by noticing the following fact.

Proposition 91 *If there exists an m and a U satisfying the requirements (i) and (ii), then*

$$C_s > 0.$$

Proof

$$0 < I(U \wedge Y^m) - I(U \wedge Z^m)$$

$$= \sum_{t=1}^{m} I(U \wedge Y_t | Y_1, \ldots, Y_{t-1}, Z_{t+1}, \ldots, Z_m) - I(U \wedge Z_t | Y_1, \ldots, Y_{t-1}, Z_{t+1}, \ldots, Z_m).$$

Therefore, there exists a t such that

$$I(U \wedge Y_1 | Y_1, \ldots, Y_{t-1}, Z_{t+1}, \ldots, Z_m) - I(U \wedge Z_t | Y_1, \ldots, Y_{t-1}, Z_{t+1}, \ldots, Z_m) > 0,$$

which implies

$$C_2 > 0.$$

The converse is proved. □

Remark 92 Inspection of the proof shows that the possibility of "safe" identification for two options (or for one bit) implies "safe" identification at rate equal to Shannon's capacity.

References

1. I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Channels* (Academic Press, New York, 1981)
2. A.D. Wyner, The wire-tap channel. Bell Syst. Tech. J. **54**, 1355–1387 (1975)

Part II
A General Theory of Information Transfer

Introduction

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

The original paper [3] gave to and received from a ZIF-project[1] essential stimulations which resulted in contributions added as supplements "Search and channels with feedback" and "Noiseless coding for multiple purposes: a combinatorial model" to [4].

Other contributions—also to areas initiated—are published in the book [4].

We have included in the references several articles and books [6–9, 12], which deal with information not just in a more or less technical engineering sense. They are meant to enlarge our horizon, stimulate our awareness of what is unknown about "information", and to bring us into the spirit for new adventures. Some questions from [6] give indications of the kind of thoughts which took us into their chains.

In the Appendix of [6] one finds the following definition or explication of the concept "communication":

[1]Zentrum für interdisziplinäre Forschung (Center for Interdisciplinary Research), Bielefeld University.

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_7

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

Again in [6] on page 41 we read:

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

and on page 43:

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

This concept of information is closely related to the idea of selection, or discrimination and therefore sometimes called selective-information. It is also at the very basis of Shannon's celebrated statistical theory of communication [10].

This theory has by now been developed into a sophisticated mathematical discipline with many branches and facets. Sometimes more concrete engineering problems led to or gave the incentive to new directions of research and in other cases new discoveries were made by exploring inherent properties of the mathematical structures. Some of our views on the state of this theory, to which we also shall refer as the "Shannon Island", are expressed in [1].

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

"We live in a world vibrating with information" and in most cases we don't know how the information is processed or even what it is at the semantic and pragmatic levels. How does our brain deal with information? It is still worthwhile to read von Neumann's ideas about this [9].

Cherry writes on page of [6]:

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

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

There are various stimuli concerning the concepts of communication and information from the sciences, for instance from quantum theory in physics, the theory of learning in psychology [8], theories in linguistics [11], etc.

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

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

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

1. Mathematical Statistics
2. Communication Networks
3. Computer Storage and Distributive Computing
4. Memory Cells

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

A first attempt to explore connections between multi-user source coding and hypothesis testing was made in [5]. For interesting ideas about relations between multiple-access channels and communication networks see Gallager [7]. A multitude of challenges to information theory comes from computer science. A proper frame for storage in memory cells is our abstract coding theory [1]. Our work on identification has led us to reconsider the basic assumptions of Shannon's theory. It deals with "messages", which are elements of a *prescribed set of objects*, known to the communicators. The receiver wants to know the true message. This basic model occurring in all engineering work on communication channels and networks addresses a very special communication situation. More generally they are characterized by

(I) The questions of the receivers concerning the given "ensemble", to be answered by the sender(s)
(II) The prior knowledge of the receivers
(III) The senders prior knowledge.

Accordingly Part II consists of three lecture.

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

We continue in chapter "Identification and Transmission with Multi-way Channels" with (promised) capacity theorems for identification via multi-way channels. We also study identification in conjunction with transmission.

The proof of the "polynomial" weak converse is new even for the discrete memoryless channel (DMC).

In chapter "Data Compression" we discuss a new direction of research on sources, which goes back to a problem of [2]: noiseless coding for multiple purposes. It stimulated to go for a new concept: identification for sources.

Chapter "Perspectives" concludes with striking results on the relation of identification and common randomness and a general discussion.

References

1. R. Ahlswede, Coloring hypergraphs: a new approach to multi–user source coding, Part I. J. Comb. Inf. System Sci. **1**, 76–115 (1979). Part II, **5**(3), 220–268 (1980)
2. R. Ahlswede, in *Eight Roblems in Information Theory in Open Problems in Communication and Computation*, ed. by T.M. Cover, B. Gopinath (Springer, Berlin, 1987)
3. R. Ahlswede, General theory of information transfer, in *Preprint 97-118, SFB 343 Diskrete Strukturen in der Mathematik* (Universität Bielefeld, Bielefeld, 1997)
4. R. Ahlswede et al. (eds), *General Theory of Information Transfer and Combinatorics*. Lecture Notes in Computer Science, vol. 4123 (Springer, Berlin, 2006)
5. R. Ahlswede, I. Csiszár, Hypothesis testing under communication constraints. IEEE Trans. Inf. Theory **32**(4), 533–543 (1986)
6. C. Cherry, *On Human Communication, A Review, a Survey and a Criticism* (MIT Press, New York, 1978)
7. R.G. Gallager, A perspective on multi–access channels. IEEE Trans. Inf. Theory IT **31**(2), 124–142 (1985)
8. E. Mittenecker, E. Raab, Informationstheorie für Psychologen, Verlag für Psychologie, Dr. C. Hofgrefe, Göttingen (1973)
9. J. Von Neumann, *The Computer and the Brain* (Yale University, Yale, 1958)
10. C.E. Shannon, A mathematical theory of communication. Bell Syst. Techn. J. **27**, 339–425, 623–656 (1948)
11. J. Singh, *Great Ideas in Information Theory, Language and Cybernetics* (Dover Publication, Inc. New York, 1966)
12. P. Stucki, Advances in digital image processing, Theory, Application, Implentation, in *The IBM Research Symposium Series* (1979)

One Sender Answering Several Questions of Receivers

1 A General Communication Model for One Sender

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

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

$$\mathcal{D}_i \cap \mathcal{D}_{i'} = \emptyset \qquad \text{for all } i \ne i', \tag{1}$$

$$W^n(\mathcal{D}_i^c | u_i) \le \lambda \qquad \text{for all } i = 1, \ldots, M. \tag{2}$$

Given a set of messages $\mathcal{M} = \{1, \ldots, M\}$, by assigning i to codeword u_i we can transmit a message from $\mathcal{M}$ in blocklength n over the channel with a maximal error probability less than λ. Notice that the underlying assumption in this classical transmission problem is that both, sender and receiver, know that the message is from a specified set $\mathcal{M}$. They also know the code. The receiver's goal is to get to know the message sent. Having received an element in decoding set $\mathcal{D}_i$ he decides for codeword u_i and then for message i. By the assumptions his (maximal) error probability is bounded by λ.

An (n, M, λ) transmission code *with randomization* assigns to message i a probability distribution P_i on $\mathcal{X}^n$, for which

$$\sum_{x^n \in \mathcal{X}^n} W^n(\mathcal{D}_i^c | x^n) P_i(x^n) \le \lambda.$$

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_8

Observe that for some $v_i \in \mathcal{X}^n$

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

and that therefore the code $\{(P_i, \mathcal{D}_i) : 1 \leq i \leq M\}$ with randomization in the encoding can be replaced by the (deterministic) code $\{(v_i, \mathcal{D}_i) : 1 \leq i \leq M\}$ satisfying also the bound λ on the error probability. Obviously the same reduction holds for channels without time structure.

This implies that randomization is of no advantage for transmission over one-way channels like the DMC. However, *it has a dramatic effect on performance for identification*. To fix ideas, transmission concerns the question "How many messages can we transmit over a noisy channel?" One tries to give an answer to the question "What is the actual message from $\mathcal{M} = \{1, \ldots, M\}$?"

On the other hand in identification it is asked "How many possible messages can the receiver of a noisy channel identify?" One tries to give an answer to the question "Is the actual message i?" Here i can be any member of the set of possible messages $\mathcal{N} = \{1, 2, \ldots, N\}$.

Certain error probabilities are again permitted. From the theory of transmission one cannot derive answers for these questions in the theory of identification, which therefore goes beyond Shannon's theory.

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

$$\sum_{x^n} P_i(x^n) W^n(\mathcal{D}_i|x^n) > 1 - \lambda \qquad \text{for all } i$$

and

$$\sum_{x^n} P_i(x^n) W^n(\mathcal{D}_j|x^n) < \lambda \qquad \text{for all } i \neq j.$$

We know from [5] (see chapter "Identification via Channels", Part I) that any (second order) rate $R < C_p olSh = C$ is achievable for any $\lambda > 0$ and all large n, that is, there are (n, N, λ) codes with $R \leq \frac{1}{n} \log \log N$.

It is convenient to introduce the maximal code size

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

Already in [5] it was shown that for any exponentially small sequence of error probabilities $\lambda_n = e^{-\varepsilon n}$ $(\varepsilon > 0)$

$$\limsup_{n\to\infty} \frac{1}{n} \log\log N(n, \lambda_n) \le C.$$

This converse was named soft converse in [5]. We use here the more instructive name "exponential weak converse".

The (classical) weak converse states that

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

As a statement between these two we introduce now a *polynomial weak converse*: For some $\alpha > 0$

$$\limsup_{n\to\infty} \frac{1}{n} \log\log N\left(n, \frac{1}{n^\alpha}\right) \le C.$$

Such a statement was derived for $\alpha = 1$ (see Sect. 3).

Again already in [5] a version of the strong converse was conjectured:

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

In case of feedback this was proved in [6] and the conjecture of [5] was established by Han/Verdú [11] and with a simpler proof in [12].

Remark The capacity concept used in [5, 6] (chapters "Identification via Channels" and "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas", Part I) is often called *pessimistic capacity*, that is, the maximal rate achievable with arbitrary small *constant* error probability λ. Sometimes in the literature also the optimistic capacity $\bar{C}$ is used. Actually for many channels (like for instance non-stationary memoryless channels) other performance criteria like *capacity functions* say more about them. This is discussed in great detail in [3]. In this lecture we discuss only pessimistic capacities C, $C_{pol\,pol}$, and $C_{\exp}$ where the latter are defined as optimal rates achievable for all polynomial error probabilities $\lambda_n = n^{-\alpha}$, $\alpha > 0$, resp. exponential error probabilities $\lambda_n = 2^{-\epsilon n}$ with some small $\epsilon > 0$. It is important to notice that in order to establish a number as the (pessimistic) capacity neither strong nor weak converses are necessary. Furthermore, $C \ge C_{pol\,pol} \ge C_{\exp}$ and for instance for the DMC it is easy to prove that $C_{\exp} \ge C$ and these capacities are equal. The same holds for regions of the multiple access channel (MAC) and can also be shown for *regions for identification* following the direct proofs of chapters "Identification via Channels" and "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas", Part I, which are based on transmission codes and for maximal errors can be improved also by the

Ahlswede/Dueck local converse [4]. It is essential that one stays near to memoryless channels; in general the concepts go apart.

One can conceive of many situations in which the receiver has (or many receivers have) different goals. They lead to decoding rules with not necessarily disjoint decoding sets.

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

We describe now some seemingly natural families of partitions.

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

Model 2: $\Pi_I = \{\pi_m : m \in \mathcal{M}\}$ with $\pi_m = \big\{\{m\}, \mathcal{M} \setminus \{m\}\big\}$. Here decoder π_m wants to know whether m occurred or not. This is the identification problem introduced in chapter "Identification via Channels".

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

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

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

Model 5: $\Pi_B = \{\pi_{\mathcal{A}} : \mathcal{A} \subset \mathcal{M}\}$. A receiver associated with $\pi_{\mathcal{A}} = \{\mathcal{A}, \mathcal{M} \setminus \mathcal{A}\}$ wants to know the answer to the binary question "Is m in $\mathcal{A}$?" (Here, of course, $\pi_{\mathcal{A}}$ and $\pi_{\mathcal{M}\setminus\mathcal{A}}$ can be viewed as the same questions).

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

In all these models we can consider the first (or second) order capacities, defined analogously to those in models 1, 2, where they are known from chapters "Identification via Channels" and "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas". It is shown in Sect. 3 that for models 4 and 5 the capacities equal Shannon's transmission capacity.

The most challenging problem is the general K-identification problem of Model 3. For convenience, we define

$$\binom{\mathcal{M}}{K} \triangleq \{\mathcal{S} \subset \mathcal{M} : |\mathcal{S}| = K\}$$

as the set of subsets of size K. Here an (n, N, K, λ)-code is a family of pairs $\big\{(P(\cdot|i), \mathcal{D}_\pi) : 1 \le i \le N, \pi \in \Pi_K\big\}$, where the $P(\cdot|i)$'s are PD's on $\mathcal{X}^n$,

$\mathcal{D}_\pi \subset \mathcal{Y}^n$, and where for all $\pi = \{S, \mathcal{M} \setminus S\}$ with $S \in \binom{\mathcal{M}}{K}$

$$\sum_{x^n} P(x^n|i) W^n(\mathcal{D}_\pi^c|x^n) \le \lambda \qquad \text{for all } i \in S,$$

$$\sum_{x^n} P(x^n|i) W^n(\mathcal{D}_\pi|x^n) \le \lambda \qquad \text{for all } i \notin S. \tag{3}$$

We also write $\mathcal{D}_S$ instead of $\mathcal{D}_\pi$. A coding theorem is established in Sect. 2.

Remarks

1. K-identification applies whenever persons want to know whether a winner is among their favorite teams or lottery numbers or friends.
2. Most models fall into the following category of regular transfer models. By this we mean that the set of partitions Π of $\mathcal{M}$ is invariant under all permutations $\sigma : \mathcal{M} \to \mathcal{M}$:

$$\pi = (A_1, \ldots, A_r) \in \Pi \qquad \Longrightarrow \qquad \sigma\pi = \big(\sigma(A_1), \ldots, \sigma(A_r)\big) \in \Pi.$$

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

Examples

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

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

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

   How many bits have to be transmitted in a randomized encoding, so that every player knows with high probability, whether he won.
2. Lets view the elements of $\{1, \ldots, a\}^n$ as sequences of events. Historians (or observers of stockmarkets) have each their subsequence of events, say,

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

   The ℓ persons are to be informed with high probability correctly about the correct sequence of events. (Idea of binning, see [1, 2, 17]).
3. In some countries 40% of the healthy men of an age-class are drafted by random selection. Every candidate wants to know with high probability correctly whether he is among them. This falls under Model 6.

2 Analysis of a Specific Model: K-Identification

A Relation to Standard Identification

Recall the definition of an (n, N, K, λ) code given in Sect. 1. For reasons, which become apparent soon, we assume K to grow exponentially in the blocklength n, that is,

$$K = 2^{\kappa \cdot n}, \tag{4}$$

where κ is called a first order rate.

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

$$N = 2^{2^{Rn}}, R > 0, \tag{5}$$

where R is called a second order rate.

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

Proposition 93 *For every DMC the set $\mathcal{K}$ of all achievable rate pairs contains*

$$\left\{(R, \kappa) : 0 \leq R, \kappa; R + 2\kappa \leq C_p ol Sh\right\},$$

where $C_p ol Sh$ is Shannon's familiar capacity of the DMC.

Proof In chapter "Identification via Channels", Part I, the achievable triples (R, η_1, η_2) of second order rate R and error exponents η_1, η_2 have been investigated. Theorem 13 completely characterizes the set of achievable pairs (R, η_2) in the limit $\eta_1 \to 0$ as follows:

$$\lim_{\eta_1 \to 0} \left\{(R, \eta_2) : (R_1, \eta_1, \eta_2) \text{ is achievable}\right\} = \left\{(R, \eta_2) : R \leq C_p ol Sh - 2\eta_2\right\}. \tag{6}$$

Now, any identification code $\left\{(P_i, \mathcal{D}_i) : 1 \leq i \leq N\right\}$ with parameters (R, η_1, η_2) has an associated K-identification code $\left\{\mathcal{P}_i, \mathcal{D}_S : 1 \leq i \leq N, S \in \binom{\mathcal{N}}{K}\right\}$, where

$$\mathcal{D}_S = \bigcup_{i \in S} \mathcal{D}_i, \tag{7}$$

meeting the parameters $(R, \kappa, \eta_1, \eta_2 - \kappa)$.

This means that

$$\sum_{x^n} P_i(x^n) W^n(\mathcal{D}_S | x^n) \geq 1 - 2^{-n\eta_1} \qquad \text{for all } i \in S$$

and

$$\sum_{x^n} P_i(x^n) W^n(\mathcal{D}_S|x^n) \le K 2^{-n\eta_2} = 2^{-n(\eta_2-\kappa)} \qquad \text{for all } i \notin S.$$

These inequalities and (6) imply that for sufficiently small η_1 there exists for all pairs of rates (R, κ) with $R \le C_p olSh - 2\kappa - \delta$ an $\eta_2 > \kappa$ satisfying (6) such that for n large enough all error probabilities above fall below any $\lambda > 0$. □

Remark Especially, for $\kappa = 0$, Proposition 93 gives the standard coding theorem for identification.

There is a very important connection to r-cover-free families.

A family of sets $\mathcal{F}$ is called *r-cover-free* if $A_0 \not\subset A_1 \cup A_2 \cup \cdots \cup A_r$ holds for all distinct $A_0, A_1, \ldots, A_r \in \mathcal{F}$. Let $M(n, r)$ denote the maximum cardinality of such an $\mathcal{F}$ over an n-element underlying set. This notion was introduced in terms of superimposed codes in [14], where for suitable constants c_1, c_2 the inequalities

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

were proved. This result was rediscovered several times. In [9], with a rather complicated proof, the upper bound was improved to

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

After the purely combinatorial proof of [10] by a simpler argument (implicitly contained in [9]) the slightly weaker bound

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

was obtained in [16]. Let $a = |\mathcal{X}|$. With the replacements $r \to a^{\kappa n}$, $n \to a^n$ we obtain

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

and thus

$$R_n \triangleq \frac{\log\log M(a^n, a^{\kappa n})}{n} \le (1-2\kappa)\log a + o(1). \tag{8}$$

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

This raises the question of optimality of the bound in Proposition 1. For its answer one needs a suitable bound for r-cover-free uniform families $\mathcal{F}$ of subsets, each of cardinality ℓ exponential in n. However, the existing bounds are too rough!

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

Proposition 94 *In the noiseless case and for zero error probability of first kind the bound in Proposition 93 is tight.*

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

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

This terminology is consistent with the terminology for converses, which we introduced in Sect. 1. Further qualifications for several kinds of probabilities are given when needed. Actually for many coding problems several regions coincide. However, as long as we don't know this it is convenient to have this flexible language.

An Equivalence of Two Coding Problems

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

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

$$\sum_{x^n} W^n(E_s|x^n) P_s(x^n) \ge 1 - \lambda^* \quad \text{for all} \quad s \in S. \tag{9}$$

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

Theorem 95 (Equivalence Theorem 1) *For every DMC*

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

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

$$\lambda^*(n) \leq \lceil \kappa n \rceil \lambda(n). \tag{10}$$

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

$$\varphi : S \to \{0, 1\}^{\lceil \kappa n \rceil}. \tag{11}$$

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

$$\underline{S}_j = \big\{ s \in S : \varphi(s)_j = 1 \big\} \tag{12}$$

and

$$S_j = \underline{S}_j \cup \overline{S} \text{ for all } \overline{S} \subset \mathcal{N} \setminus S, |\overline{S}| = \frac{1}{2} K. \tag{13}$$

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

$$s \in S_j \text{ and } s' \notin S_j. \tag{14}$$

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

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

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

Then the sets

$$E_s \triangleq \bigcap_{j=1}^{\lceil \kappa n \rceil} (\mathcal{D}_{S_j})^{\varepsilon_j(s)}, \qquad s \in S, \tag{15}$$

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

Finally, we have by the properties of the original code

$$\sum_{x^n} W^n(E_s|x^n) P_s(x^n) \geq 1 - \lceil n\kappa \rceil \lambda(n), \qquad s \in S. \tag{16}$$

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

Furthermore, it becomes an exercise to show that the same argument also yields for a DMC a relation weaker than Proposition 93, namely

$$\mathcal{K} \supset \left\{(R, \kappa) : R + 2\kappa \leq C_p oler\right\},$$

where $C_p oler$ is the erasure capacity (c.f. [8]).

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

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

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

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

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

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

Remark

- $\lambda - K$-identification, λ^*-decodable codes give rise to associated identification codes with error probabilities smaller than $\lambda + \lambda^*$ by assigning to every $i \in \mathcal{N}$ a K-element subset S_i containing i and the decoding set $\mathcal{D}_i = E_i \in \mathcal{E}_{S_i}$. Therefore $R < C_p olSh$, and by Shannon's coding theorem also $\kappa \leq C_p olSh$.
- There is another instructive relation. Let us view $\binom{\mathcal{N}}{K}$ as set $\mathcal{M}$ of objects, one of which, say S, is given to the sender for encoding. The receiver wants to know whether it equals S' (any element of $\mathcal{M}$) or not. This is a standard identification problem with $|\mathcal{M}| = \binom{N}{K}$.

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

$$2^{2^{C_p olSh n}} \gtrsim \binom{N}{K} = \binom{2^{\varepsilon n}}{2^{\kappa n}} \geq 2^{(\varepsilon n - \kappa n) 2^{\kappa n}} \geq 2^{2^{\kappa n}} \text{ gives } \kappa \leq C_p olSh.$$

An Outer Bound on the Capacity Region $\mathcal{K}$

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

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

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

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

Therefore

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

and

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

We summarize this.

Lemma 96 *For every $\alpha \in (0,1)$ there is a family $\mathcal{S} = \{S_1, \ldots, S_{\tilde{N}}\} \subset \binom{\mathcal{N}}{K}$ with*

(i) $|S_i \triangle S_j| \geq (1-\alpha)2K$ *and* $|S_i \cap S_j| \leq \alpha K$.
(ii) $R + \kappa - \frac{1}{n}|\log \alpha| \leq \frac{1}{n} \log\log |\mathcal{S}| \leq \frac{1}{n} \log\log \binom{N}{K} \leq R + \kappa$.

We can therefore by (ii) upperbound $\binom{N}{K}$ by upperbounding $|\mathcal{S}|$. For this we relate $\mathcal{S}$ to a standard identification problem. For $S \in \mathcal{S}$ define $P_S \in \mathcal{P}(\mathcal{X}^n)$ by

$$P_S(x^n) = \frac{1}{K} \sum_{i \in S} P(x^n|i), \qquad x^n \in \mathcal{X}^n, \tag{17}$$

if $P(\cdot|i)$ is the randomized encoding for i. Now by Lemma 96(i) and the code definition in (1) and (2) we have for $S, S' \in \mathcal{S}$, $S \neq S'$,

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

and

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

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

$$\lambda' = \lambda + \alpha \geq \lambda.$$

By the weak converse in Sect. 2 of chapter "Identification and Transmission with Multi-way Channels" and Lemma 96 (ii) we get the desired bound for $\mathcal{K}$.

The same proof works for the K-separating codes of Sect. 2 of chapter "Models with Prior Knowledge of the Receiver", if we define $\mathcal{D}_E = \bigcup_{i \in E} \mathcal{D}_{E,i}$.

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

Proposition 97 $\mathcal{K} \subset \{(R, \kappa) : R + \kappa \leq C_{pol}Sh\}$.

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

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

Remark The two Propositions 93 and 94 imply for $\kappa = 0$ the standard identification capacity theorem.

Remark Using also Theorem 95 we see that for $R = 0$ we get the converse to Shannon's Coding Theorem and only the achievable rate $\frac{1}{2}C_{pol}Sh$!

On K-identification in Case of Noiseless Feedback

As in chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas", Part I, we assume the presence of a letter by letter noiseless feedback link. Again deterministic encoding functions for i are denoted by f_i^n and randomized encoding functions for i are denoted by F_i^n. The corresponding regions of achievable rate pairs are denoted by $\mathcal{K}_f$ and $\mathcal{K}_F$. Analogously, if all $S \in \binom{\mathcal{N}}{K}$ are λ-decodable we denote the regions by $\mathcal{K}_f^*$ and $\mathcal{K}_F^*$. We formulate now results, which are analog to those in the two previous subsections. Notice that the argument leading to (16) applies also in cases of deterministic and randomized feedback strategies. The results in [7] (see chapter "On Identification via Multi-Way Channels with Feedback: Mystery Numbers"), including constructive coding strategies, go considerably beyond [6] (see chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas") and also, if necessary, chapter "On Identification via Multi-Way Channels with Feedback: Mystery Numbers" can be consulted for detailed definitions of all concepts used in this section, when they are not immediately clear.

Theorem 98 (Equivalence Theorem 2) *For every DMC*

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

Proposition 99 *For every DMC W*

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

where $Q = PW$.

We use our entropy property for all discrete distributions.

Lemma 100 (Included in [7], chapter "On Identification via Multi-Way Channels with Feedback: Mystery Numbers") *For* $P = (P_1, P_2, \dots) \in \mathcal{P}(\mathbb{N})$ *define*

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

and set

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

Then

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

Proof (Proof of Proposition 99) In any (n, N, K, λ)-code with feedback

$$\left\{(F_i, \mathcal{D}_S) : 1 \leq i \leq N; S \in \binom{\mathcal{N}}{K}\right\}$$

let Y_i^n be the output process generated by F_i via the channel. Furthermore define the process Y_S^n by the distribution

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

By assumption

$$\Pr(Y_i^n \in \mathcal{D}_S) \geq 1 - \lambda, \quad \text{if } i \in S, \tag{18}$$

$$\Pr(Y_i^n \in \mathcal{D}_{S'}) \leq \lambda, \quad \text{if } i \notin S'. \tag{19}$$

By Lemma 100 there are sets $\mathcal{E}_S \subset \mathcal{Y}^n \left(S \in \binom{\mathcal{N}}{K}\right)$ with

$$\Pr(Y_S^n \in \mathcal{E}_S) \geq 1 - \frac{1}{d}, \tag{20}$$

$$|\mathcal{E}_S| \leq 2^{\lceil d\ H(Y_S^n)\rceil+1}. \tag{21}$$

We show later that the net $\mathcal{S} \subset \binom{\mathcal{N}}{K}$ with the properties (i), (ii) in Lemma 96 satisfies

$$\mathcal{D}_S \cap \mathcal{E}_S \neq \mathcal{D}_{S'} \cap \mathcal{E}_{S'} \text{ for all } S, S' \in \mathcal{S}; S \neq S', \tag{22}$$

provided that λ is sufficiently small.

We know from chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas", Part I, that

$$H(Y_S^n) \leq n \max_{P:Q=PW} H(Q) = \overline{H} \text{ (say)}. \tag{23}$$

Therefore by (21) and (22)

$$|\mathcal{S}| \leq 2^{2^{dn\overline{H}}} \tag{24}$$

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

$$|\mathcal{S}| \simeq N^K \lesssim 2^{2^{g(\lambda)n\overline{H}}} \tag{25}$$

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

Therefore

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

It remains to be seen that (22) holds.

Suppose that for $S, S' \in \mathcal{S}$, $\mathcal{E}_S \cap \mathcal{D}_S = \mathcal{E}_{S'} \cap \mathcal{D}_{S'}$. Then by (18) and (20)

$$\Pr(Y_{S'}^n \in \mathcal{E}_{S'} \cap \mathcal{D}_{S'}) = \Pr(Y_{S'}^n \in \mathcal{E}_S \cap \mathcal{D}_S) \geq 1 - \frac{1}{d} - \lambda. \tag{26}$$

On the other hand, by (19)

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

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

This contradicts (26), if

$$\lambda + \alpha < 1 - \frac{1}{d} - \lambda. \tag{27}$$

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

So in order to show that for any $\varepsilon > 0$, $\kappa + R \leq \overline{H} + \varepsilon$, choose first d so that $d > 1$ and $d\overline{H} \leq \overline{H} + \varepsilon$, then choose λ smaller than $\frac{1}{4}\left(1 - \frac{1}{d}\right)$, and finally choose α smaller than $\frac{1}{2}\left(1 - \frac{1}{d}\right)$. □

Remark

1. Notice that we have used that $\frac{1}{K}\sum_{i\in S} f_i$ defines a *randomized* feedback strategy F_S. So this approach does not work for the case of deterministic feedback strategies!
2. We have upperbounded $\binom{N}{K}$ via upperbounding $|\mathcal{S}|$, for which we used our old idea of "distinct carriers". Instead we could also follow the approach under the second subsection of Sect. 2, in which we relate the modified K-identification problem with a standard identification problem. In case of feedback we get the upper bound for randomized strategies by the strong converse of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas".
3. For small K, say for constant K while n grows, K-identification reduces of course to K identifications and thus to identification.

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

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

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

The converses in case of feedback show that

$$|\mathcal{E}| < 2^{2^{\overline{H}n}}. \tag{28}$$

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

By Sect. 3 decoding all subsets, gives optimal rate $C_{pol}Sh$. So the bound in (28) is not achievable.

Problem Does the Theorem 95 hold for general hypergraphs? ▲

A Combinatorial Consequence

It is remarkable that a result for K-identification (Proposition 93) has an important consequence for r-cover-free families in relation to packings. We use a result of Kuzjurin [15].

A family $\mathcal{A}$ of k-subsets of $[m] = \{1, 2, \ldots, m\}$ is called (m, k, ℓ)-*packing* iff each ℓ-subset of $[m]$ is contained in at most one member $A \in \mathcal{A}$. Therefore two members of $\mathcal{A}$ intersect in at most $\ell - 1$ elements. (In other words $\mathcal{A}$ can be viewed as a code with constant weight k, word length m and distance $d_H = 2(k - \ell) + 2$.)

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

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

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

Roughly speaking the result of [15] says that $k = \sqrt{m}$ is the threshold function for the existence of asymptotically good packings. Here is the precise result.

Theorem 101 *Let α be the minimum constant such that for every $\varepsilon > 0$ and sufficiently large n every interval $[n, n+n^{\alpha+\varepsilon}]$ contains a prime number. It is known that $\alpha \leq \frac{23}{43}$. The following bounds hold:*

(i) Let $c < 1$ and $k(m) < c\sqrt{m}$, where $\lim_{n\to\infty} k(m) = \infty$. Further, let for some $\varepsilon > o$ $\ell(m) = o\big(\sqrt{k(m)}\big)$ and $\ell(m) = o\left(\binom{m}{k(m)}^{1-\alpha-\varepsilon}\right)$.
Then asymptotically good (m, k, ℓ)-packings exist.

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

Corollary 102

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

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

We derive from the assumptions on μ, γ, β

$$\mu > 2\gamma, \qquad \gamma > 2\beta, \qquad \mu > \gamma + \frac{43}{20}\beta. \tag{29}$$

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

$$N^*(n, I, \kappa) \lesssim \frac{\binom{e^{In}}{e^{\beta n}}}{\binom{e^{\gamma n}}{e^{\beta n}}} = \frac{\binom{e^{In}}{e^{\beta n}}}{\binom{e^{(\beta+\kappa)n}}{e^{\beta n}}}, \tag{30}$$

$$\frac{1}{n}\log\log N^* \lesssim \beta, \tag{31}$$

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

However, our bound $R = I - 2\kappa = \frac{20}{43}I$ in Proposition 93 is much better!

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

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

We know now that the quantities can be very different!

3 Models with Capacity Equal to the Ordinary Capacity

Some of the cases considered here were first treated by Já Já [13] for non-randomized encoding on the BSC. If randomization is permitted, the analysis is somewhat more complicated. In this section we describe the various codes and capacities by words.

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

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

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

$$\left\{(P(\cdot|i), \mathcal{D}_i) : i = 1, 2, \ldots, N\right\}$$

of pairs with

$$P(\cdot|i) \in \mathcal{P}(\mathcal{X}^n), \mathcal{D}_i \subset \mathcal{Y}^n \text{ for all } i = 1, 2, \ldots, N \tag{32}$$

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

$$\sum_{x^n \in \mathcal{X}^n} P(x^n|i) W^n(\mathcal{D}_j|x^n) \geq 1 - \lambda_1 \text{ for all } i = 1, \ldots, j \tag{33}$$

and

$$\sum_{x^n \in \mathcal{X}^n} P(x^n|i) W^n(\mathcal{D}_j|x^n) \leq \lambda_2 \text{ for all } i > j. \tag{34}$$

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

Theorem 103 *Even for randomized encoding the polynomial ordering problem capacity does not exceed the transmission capacity. The same holds in case of noiseless feedback.*

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

The ordering problem code gives rise to a transmission code as follows:

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

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

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

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

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

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

All Binary Questions

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

Theorem 104 *Even for randomized encoding the binary questions capacity does not exceed the transmission capacity. The same holds in case of noiseless feedback.*

Identification of a Component

In model 6, the number of components is linear in the blocklength. For exponentially small error probability words can therefore be reproduced with small error probability. (For small, but constant error probabilities, rate-distortion theory is to be used).

Theorem 105 *Even for randomized encoding the component identification capacity does not exceed the transmission capacity. The same holds in case of feedback.*

References

1. R. Ahlswede, Channel capacities for list codes. J. Appl. Prob. **10**, 824–836 (1973)
2. R. Ahlswede, Coloring hypergraphs: a new approach to multi–user source coding, Part I. J. Comb. Inf. Syst. Sci. **1**, 76–115 (1979). Part II **5**(3), 220–268 (1980)
3. R. Ahlswede, On concepts of performance parameters for channels, in *General Theory of Information Transfer and Combinatorics*. Lecture Notes in Computer Science, vol. 4123 (Springer, Berlin, 2006), pp. 639-663
4. R. Ahlswede, G. Dueck, Every bad code has a good subcode: a local converse to the coding theorem. Z. Wahrsch. und verw. Geb. **34**, 179-182 (1976)
5. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**, 15–29 (1989)
6. R. Ahlswede, G. Dueck, Identification in the presence of feedback—a discovery of new capacity formulas. IEEE Trans. Inf. Theory **35**, 30–39 (1989)
7. R. Ahlswede, B. Verboven, On identification via multi–way channels with feedback. IEEE Trans. Inf. Theory **37**(5), 1519–1526 (1991)
8. R. Ahlswede, N. Cai, Z. Zhang, Erasure, list, and detection zero-error capacities for low noise and a relation to identification. IEEE Trans. Inf. Theory **42**(1), 55-62 (1996)
9. A.G. Dyachkov, V.V. Rykov, Bounds on the length of disjunctive codes. Problemy Peredachi Informatsii **18**(3), 7–13 (1982, in Russian)
10. P. Erdös, P. Frankl, Z. Füredi, Families of finite sets in which no set is covered by the union of r others. Isr. J. Math. **51**, 79–89 (1985)
11. T.S. Han, S. Verdú, New results in the theory and application of identification via channels. IEEE Trans. Inf. Theory **38**, 14–25 (1992)
12. T.S. Han, S. Verdú, Approximation theory of output statistics. IEEE Trans. Inf. Theory **39**(3), 752–772 (1993)
13. J. Ja'Ja', Identification is easier than decoding, in *Proceedings of the 26th Annual Symposium on Foundations of Computer Science (SFCS 1985)* (1985), pp. 43–50
14. W.H. Kautz, R.C. Singleton, Nonrandom binary superimposed codes. IEEE Trans. Inf. Theory **10**, 363–377 (1964)
15. N.N. Kuzjurin, On the difference between asymptotically good packings and coverings. Eur. J. Comb. **16**, 35–40 (1995)
16. M. Ruszinko, On the upper bound of the size of the r–cover–free families. J. Comb. Theory, Ser. A **66**, 302–310 (1994)
17. D. Slepian, J.K. Wolf, Noiseless coding of correlated information sources. IEEE Trans. Inf. Theory **19**, 471–480 (1973)

Models with Prior Knowledge of the Receiver

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

We consider first abstract hypergraphs.

1 Zero-error Decodable Hypergraphs

If the decoder wants to know $v \in E$, then any two vertices $x, y \in E$ must be separable for instance by different colors assigned to them.

Definition 106 The separability graph $\mathcal{G}(\mathcal{H}) = (\mathcal{V}, \mathcal{E}^*)$ is defined by

$$\{x, y\} \in \mathcal{E}^* \Leftrightarrow \exists\, F \in \mathcal{E} : \{x, y\} \subset F. \tag{1}$$

Let $\Psi(\mathcal{G})$ be the chromatic number of $\mathcal{G}$, then $\mathcal{H}$ is 0-error decodable iff $\Psi(\mathcal{G}) \leq 2^{C_0 n}$, where C_0 is the zero-error capacity of the channel W used for the transmission of this color. Now $\mathcal{H}$ is λ-identifiable iff $\Psi(\mathcal{G}) \lesssim 2^{2^{C(W)n}}$.

Remark Also if 2-separable only within edges by the results of [3, 4] the answer is the same.

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_9

2 K-Separating Codes

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

We call $\{(P_i, \mathcal{D}_{E,i}) : E \in \mathcal{E}, i \in E\}$ an (n, N, λ) code for $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and W, if $P_i \in \mathcal{P}(\mathcal{X}^n)$ for $i \in \mathcal{V} = \{1, 2, \ldots, N\}$, $\mathcal{D}_{E,i} \subset \mathcal{Y}^n$, and for all $E \in \mathcal{E}$

$$\mathcal{D}_{E,i} \cap \mathcal{D}_{E,i'} = \emptyset \text{ for all } i, i' \in E, i \neq i' \tag{2}$$

$$\sum_{x^n} P_i(x^n) W^n(\mathcal{D}_{E,i}|x^n) \geq 1 - \lambda \text{ for all } i \in E. \tag{3}$$

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

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

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

Example $\mathcal{E} = \binom{\mathcal{V}}{K}$, the family of all K-element subsets of $\mathcal{V}$, defines the complete K-uniform hypergraph. The codes defined above are denoted here by (n, N, K, λ) and called *K-separating codes*. ▲

Clearly, their capacity region $\mathcal{K}^{++}$ contains $\mathcal{K}^*$ and by Theorem 95 also $\mathcal{K}_{pol\,pol}$.

Moreover, the same proof as for Proposition 97 in 2 of chapter "One Sender Answering Several Questions of Receivers" works for K-separating codes, if we define $\mathcal{D}_E = \bigcup_{i \in E} \mathcal{D}_{E,i}$.

Corollary 107

(i) $\mathcal{K}^{++} \supset \mathcal{K}^* \supset \mathcal{K}_{pol\,pol}$.
(ii) $\mathcal{K}^{++} \subset \{(R, \kappa) : R + \kappa \leq C_{pol\,Sh}\}$.

Problem Determine $\mathcal{K}^{++}$! ▲

Second Order 2-Separation Capacity Without and with Feedback

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

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

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

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

Notice that an (n, N, λ) ID code $\{(P_i, \mathcal{D}_i) : 1 \leq i \leq N\}$ satisfies

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

$$\sum_{x^n} P_j(x^n) W^n(\mathcal{D}_i|x^n) \leq \lambda \qquad (i \neq j).$$

Therefore setting $\mathcal{D}_{ij} = \mathcal{D}_i \setminus \mathcal{D}_j$ and $\mathcal{D}_{ji} = \mathcal{D}_j \setminus \mathcal{D}_i$ we see that i and j are 2λ-separable. It immediately follows that the second order capacity for $K = 2$, say C_2, is not smaller than the ID-capacity $C_{pol}Sh$. Whereas in ID-codes the decoding sets carry one index, 2-separating codes carry two indices. The decoding sets for two messages are adapted for these two and no other message. Therefore *2-separation* is a weaker notion than identification (except, perhaps, for a small shift in error probability caused by the disjointness of the two decoding sets).

Theorem 108

(i) The 2-separation capacity of second order C_2 equals the second order identification capacity $C_{pol}Sh$.

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

Proof The issues are the converses.

(i) Here we can be brief, because inspection of the strong converse proof for identification of Han/Verdú [4] shows that it is actually designed for 2-separation. The key fact, called resolvability in [5], is this:

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

$$\|Q - Q^*\| \leq \varepsilon \text{ for all } n \geq n(\varepsilon). \tag{4}$$

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

In this way to every encoding distribution $P_i (1 \leq i \leq N)$ we can find a distribution P_i^* such that the corresponding output distribution is close to that of P_i. By the code properties the Q_i's and also the Q_i^*'s are distinct. Therefore the P_i^*'s must be distinct and their number in second order rate does not exceed $C_{pol}Sh$.

(ii) Let us consider the deterministic case. For the randomized case we just have to replace $\underline{H} = \max_x H(W(\cdot|x))$ by $\overline{H} = \max_P H(PW)$.

We know from Lemma 100 in Sect. 2 of chapter "One Sender Answering Several Questions of Receivers", that for encoding function f_i there exists an $\mathcal{E}_i \subset \mathcal{Y}^n$ such that for $Q_i = W^n(\cdot|f_i)$, $Q_i(\mathcal{E}_i) \geq 1 - \frac{1}{d}$, and $|\mathcal{E}_i| \leq 2^{\lceil d\underline{H}n\rceil+1}$. Omit from $\mathcal{E}_i$ the elements with smallest probability until we get a set $\mathcal{E}_i^* \subset \mathcal{E}_i$ with $Q_i(\mathcal{E}_i^*) \geq 1 - \frac{1}{d}$ and which is minimal with this property.

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

$$\left|\binom{\mathcal{Y}^n}{T}\right| \leq 2^{(n\log|\mathcal{Y}|)2^{\lceil d\underline{H}n\rceil+1}}. \tag{5}$$

This is the desired upper bound. However, not all $\mathcal{E}_i^*$'s are necessarily different. Therefore, we have to upperbound the multiplicity with which a set,

say $\mathcal{F}$, occurs among the $\mathcal{E}_i^*$'s. W.l.o.g. we label them $\mathcal{E}_1^*, \ldots, \mathcal{E}_M^*$. By our definitions

$$1 - \frac{1}{d} + \frac{Q_i(\mathcal{E}_i^*)}{|\mathcal{F}|} \geq Q_i(\mathcal{E}_i^*) \geq 1 - \frac{1}{d}. \tag{6}$$

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

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

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

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

If we now set $\mathcal{D}'_{\ell k} = \mathcal{F} \cap \mathcal{D}_{\ell k}$ and renormalize the measure Q_i on $\mathcal{F}$ from total measure $\sim 1 - \frac{1}{d}$ (see (6)) to 1, then we have a 2-separating code of size M with output space $\mathcal{F}$.

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

$$M \leq 2^{2^{d\underline{H}n}}. \tag{7}$$

Together with (5) we get

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

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

□

Strong Converses by the Method of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" for 2-Separation in Case of Feedback

We begin with Theorem 108 (ii). By Lemma 43 for any $\varepsilon \in (0, 1)$ we can find sets $\mathcal{E}_i^* (i = 1, \ldots, N)$ of minimal size with

$$1 \geq W^n(\mathcal{E}_i^* | f_i) \geq 1 - \varepsilon, \tag{8}$$

$$|\mathcal{E}_i^*| \leq 2^{\left(\underline{H} + \frac{c(\varepsilon)}{\sqrt{n}}\right)n}. \tag{9}$$

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

Now just repeat the previous proof in the previous subsection. Now (the sharper) (8) takes the role of (6). Instead of (7) we get now the stronger

$$M \leq 2^{2^{\left(\underline{H}+\frac{c(\varepsilon)}{\sqrt{n}}\right)n}} \tag{10}$$

and finally

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

and thus

$$\frac{1}{n}\log\log N \leq \underline{H} + \frac{c(\varepsilon)}{\sqrt{n}} \text{ (strong converse).} \tag{11}$$

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

Remark The results obviously generalize to any constant K.

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

3 Analysis of a Model with Specific Constraints: 2-Separation and Rényi's Entropy H_2

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

$$|A_i| = P_i\, N;\ i = 1, \ldots, \alpha. \tag{12}$$

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

Let now $\Pi_1, \Pi_2, \ldots, \Pi_m$ be a sequence of independent random partitions taking as values a partition of type P with equal probabilities. Equivalently we can say that a person $z \in \mathcal{N}$ belongs to the randomly chosen A_i with probability P_i independently of what happens to the other persons. (At discrete time points $1, 2, \ldots$ the partition of the persons in several rooms is reported.)

Imagine now that somebody, the interrogator, has difficulties to distinguish any two persons in his interest at the party, but is reported the sequence of partitions

described. So he knows at every time instance the set of persons in all rooms, but he cannot identify the persons in a set.

Let now $\lambda_{N,m}$ denote the probability that m such partitions separate any two persons in $\mathcal{N}$. Rényi [6] has shown that $m_2(N, \varepsilon)$, the smallest m with $\lambda_{N,m} \geq 1-\varepsilon$, satisfies

$$m_2(N, \varepsilon) \sim \frac{2\log_2 N + o(\varepsilon)}{H_2(P)}, \tag{13}$$

where H_2 is Rényi's entropy of order 2.

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

$$F_i(z) = j, \quad \text{if } z \in A_{ij}. \tag{14}$$

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

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

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

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

4 Binning via Channels

In Sect. 1 we considered vertex colorings with different colors in each edge. They have been called strict colorings in [1, 2]. Other colorings discussed there are

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

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

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

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

Indeed, let us look at K-separation. We know from Proposition 93 that we can choose N with second order rate R and K with rate κ, $R + 2\kappa \leq C_p olSh$, and achieve K-identification.

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

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

Furthermore, the theory in [1, 2] works, if the number of vertices, the number of edges, and the edge sizes are roughly of the same growth, *namely exponential in n*.

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

5 K-Identifiability, K-Separability and Related Notions

We discuss here connections between code concepts.

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

$$\mathcal{Q} = \{Q_i : i \in \mathcal{N}\}. \tag{15}$$

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

In case of identification the decoding structure is

$$\mathcal{D} = \{\mathcal{D}_i : i \in \mathcal{N}\}. \tag{16}$$

It is of *precision* λ, if

$$Q_i(\mathcal{D}_i) \geq 1 - \lambda \qquad (i \in \mathcal{N}) \tag{17}$$

$$Q_i(\mathcal{D}_j) \leq \lambda \qquad (i \neq j). \tag{18}$$

The precision relates to the whole encoding structure $\mathcal{Q}$, however, in a pairwise fashion (as specified in (18)).

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

$$\mathcal{D}_{Si} \cap \mathcal{D}_{Sj} = \emptyset, \qquad \text{and} \qquad Q_i(\mathcal{D}_{Si}), Q_j(\mathcal{D}_{Sj}) \geq 1 - \lambda. \tag{19}$$

These sets relate only to i and j.

Lemma 109 *1-identifiable with precision* λ *implies 2-separable with precision* 2λ.

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

There is also a general connection.

Lemma 110 *K-identifiable with precision $\lambda(n)$ implies K-separable with precision $\lambda'(n) = \lceil n\kappa \rceil \lambda(n)$, where*

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

Proof See proof of Theorem 95 in Sect. 2 of chapter "One Sender Answering Several Questions of Receivers". □

Problem For $L \geq K$, how does K-identifiability relate to L-separability? ▲

Finally we mention related concepts.

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

$$\mathcal{D}_S = \bigcup_{i \in S} \mathcal{D}_i, \qquad S \in \binom{\mathcal{N}}{K} \tag{20}$$

and

$$Q_i(\mathcal{D}_i) \geq 1 - \lambda \qquad \text{for all } \ i \in \mathcal{N}, \tag{21}$$

$$Q_i(\mathcal{D}_S) \leq \lambda \qquad \text{for all } \ i \notin S. \tag{22}$$

For the disjoint sets

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

we have

$$Q_i(\mathcal{D}_{Si}) \geq 1 - 2\lambda \ \text{ for all } \ i \in S, \tag{23}$$

a generalization of Lemma 109.

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

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

and

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

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

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

and

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

for $\varepsilon = 0, 1$.

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

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

References

1. R. Ahlswede, Channel capacities for list codes. J. Appl. Prob. **10**, 824–836 (1973)
2. R. Ahlswede, Coloring hypergraphs: a new approach to multi–user source coding. Part I J. Comb. Inf. Syst Sci **1**, 76–115 (1979). Part II **5**(3), 220–268 (1980)
3. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**, 15–29 (1989)
4. T.S. Han, S. Verdú, New results in the theory and application of identification via channels. IEEE Trans. Inf. Theory **38**, 14–25 (1992)
5. T.S. Han, S. Verdú, Approximation theory of output statistics. IEEE Trans. Inf. Theory **39**(3), 752–772 (1993)
6. A. Renýi, On the foundations of information theory. Rev. Inst. Internat. Stat. **33**, 1–14 (1965)

Models with Prior Knowledge at the Sender

1 Identification via Group Testing and a Stronger Form of the Rate-Distortion Theorem

Suppose that from the set $\mathcal{N} = \{1, 2, \ldots, N\}$ of persons any subset $\mathcal{S} \subset \mathcal{N}$ of persons may be the set of sick persons. Moreover it is known that with probability q a person is sick and that the RV S has the distribution

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

For each subset of the test subjects, $(B \subseteq \mathcal{N})$, the binary, error-free test, which determines whether at least one person in B is sick or not, is admissible. In the group testing model introduced in [5] the goal is to determine the expected number of tests $L(N, q)$ for an optimal sequential strategy to diagnose all sick persons (see also [3], pp. 112–117).

Theorem 111 (Ungar 1960, [5]) $Nh(q) \leq L(N, q) \leq N$.

In our model the decoder (person) s wants to know whether he is sick. Any other information is of much less relevance to him. In particular he does not care who the other sick persons are. In terms of partitions

$$\pi_s = \big\{\{\mathcal{S} \subset \mathcal{N} : s \in \mathcal{S}\}, \{\mathcal{S} \subset \mathcal{N} : s \notin \mathcal{S}\}\big\} \tag{2}$$

he wants to know which member of π_s occurred.

We can reformulate this problem by identifying $\mathcal{S} \subset \mathcal{N}$ with a word $x_{\mathcal{S}} = (x_1, \ldots, x_N) \in \{0,1\}^N$, $x_s = 1$ iff $s \in \mathcal{S}$. Thus the distribution defined in (1)

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_10

describes a discrete memoryless source (DMS) $\big(\{0,1\}^N, Q^N, X^N\big)$ with $Q^N(x^N) = \prod_{t=1}^N Q(x_t)$, where

$$Q(x_t) = \begin{cases} q & \text{for } x_t = 1 \\ 1-q & \text{for } x_t = 0, \end{cases} \tag{3}$$

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

$$\Pr(X^N = x^N) = Q^N(x^N). \tag{4}$$

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

$$\hat{X}_t = g_t\big(f_N(X^N)\big) \tag{5}$$

and consider the error probability

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

where d is the Hamming distance.

Now the rate-distortion theorem tells us how small a rate $R(q,\lambda)$ we can achieve with rate$(f_N) = \frac{1}{N} \log$ (Number of values of f_N) under the constraint

$$\sum_{t=1}^{N} \mathbb{E}\, d(X_t, \hat{X}_t) \le \lambda\, N. \tag{6}$$

However, we are interested in the stronger condition

$$\mathbb{E}\, d(X_t, \hat{X}_t) \le \lambda \quad \text{for all } 1 \le t \le N \tag{7}$$

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

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

and therefore as $\lambda \to 0$ by the source coding theorem also $\lim_{\lambda \to 0} R^*(q,\lambda) = h(q)$.

When λ is kept at a prescribed level we have the following result.

Theorem 112 *The identification after group testing in a group of N persons, everyone being independently sick with probability q, can be performed at error probability λ with $R(q,\lambda)N$ bits. Here $R(q,\lambda)$ is the rate-distortion function for the Bernoulli source with generic distribution $(q, 1-q)$ evaluated at distortion level λ.*

Remark Since space does not permit we leave the proof as an exercise using balanced hypergraph covering, which we started in [1]. The lemma in Section VI of [2] can be used for q-typical N sequences as vertex set $\mathcal{V}$ and p-typical N sequences as edge set $\mathcal{E}$ for covering or approximation. The exceptional set $\mathcal{V}_0$ in that lemma can be kept empty (see Lemma 9 of [4]). Now in addition to hypergraph $(\mathcal{V}, \mathcal{E})$ use also hypergraph $(\mathcal{V}_1, \mathcal{E})$, where $\mathcal{V}_1 = [N]$. There is a selection of edges $E_1, \ldots, E_L \in \mathcal{E}$ which simultaneously covers $\mathcal{V}$ and $\mathcal{V}_1$ in balanced ways. The second means (7), of course after polynomially many pairs (q', p') with q' close to q have been used.

Instead of two properties (sick and not sick) there can be any finite number of properties k defining k classes and every person wants to know its class. This leads to a *rate-distortion theorem for a DMS stronger than Shannon's.*

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

References

1. R. Ahlswede, Coloring hypergraphs: a new approach to multi–user source coding. Part I J. Comb. Inf. Syst. Sci. **1**, 76–115 (1979). Part II **5**(3), 220–268 (1980)
2. R. Ahlswede, Towards a general theory of information transfer, in *Shannon Lecture at ISIT in Seattle 13th July 2006*. IEEE Information Theory Society Newsletter (2007)
3. R. Ahlswede, I. Wegener, Search problems, in *Wiley-Interscience Series in Discrete Mathematics and Optimization* (1987)
4. R. Ahlswede, C. Mauduit, A. Sárközy, Large families of pseudorandom sequences of k symbols and their complexity, Part II, in *General Theory of Information Transfer and Combinatorics*. Lecture Notes in Computer Science, vol. 4123 (Springer, Berlin, 2006), pp. 308–325
5. P. Ungar, The cut-off point for group testing. Commun. Pure Appl. Math. **13**, 49–54 (1960)

Identification and Transmission with Multi-way Channels

1 Simultaneous Transfer: Transmission and Identification

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

A. Let us discuss now a specific example. Suppose that one DMC is used *simultaneously* for transmission and identification. Since both, the transmission capacity and the (second order) identification capacity, equal $C_{pol}Sh$, here is the best we can do: We use an (n, M) transmission code $\{(u_i, \mathcal{D}_i) : 1 \leq i \leq M\}$ with *average* error $\overline{\lambda} = \frac{1}{M} \sum_{i=1}^{M} W^n(\mathcal{D}_i^c | u_i)$. The randomness in the messages produces via this code a common random experiment for sender and receiver. Adding a few, say, $\sqrt{n}$ letters, we can get the desired identification code $(n + \sqrt{n}, N, \lambda)$ as in chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" (see [7] and also [5]) by the following approach.

From common randomness (also called shared randomness in physics) to identification: The $\sqrt{n}$-trick

Let $[M] = \{1, 2, \ldots, M\}$, $[M'] = \{1, 2, \ldots, M'\}$ and let $\mathcal{T} = \{T_i : i = 1, \ldots, N\}$ be a family of maps $T_i : [M] \to [M']$ and consider for $i = 1, 2, \ldots, N$ the sets

$$K_i = \{(m, T_i(m)) : m \in [M]\}$$

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_11

and on $[M] \times [M']$ the PD's

$$Q_i((m, m')) = \frac{1}{M} \text{ for all } (m, m') \in K_i.$$

Lemma 113 (Transformator Lemma) *Given* $M, M' = \exp\{\sqrt{\log M}\}$ *and* $\epsilon > 0$ *there exists a family* $\mathcal{T} = \mathcal{T}(\epsilon, M)$ *such that* $|\mathcal{T}| = N \geq \exp\{M - c(\epsilon)\sqrt{\log M}\}$, $Q_i(K_i) = 1$ *for* $i = 1, \ldots, N$, *and* $Q_i(K_j) \leq \epsilon \ \forall i \neq j$.

Hence, (C_polSh, C_polSh) is achievable.

Next suppose that there is a noiseless feedback channel and we use the same code as before. This generates an input process $X^n = (X_1, \ldots, X_n)$ and an output process $Y^n = (Y_1, \ldots, Y_n)$, which is known also to the sender by the feedback. So we get a common random experiment of rate $\frac{1}{n}H(Y^n)$. Again by the identification trick of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" now

$$R_{\text{transm.}} \sim \frac{1}{n} I(X^n \wedge Y^n)$$

$$R_{\text{ident.}} \sim \frac{1}{n} H(Y^n), \text{ second order.}$$

It is now easy to show the direct part in the following theorem.

Theorem 114 $\mathcal{R} = \text{conv}\left\{\left(I(X \wedge Y), H(Y)\right) : P_X \in \mathcal{P}(\mathcal{X})\right\}$ *is the set of achievable pairs of rates for the simultaneous transmission and identification over the DMC with noiseless feedback.*

Proof The direct part follows from the lemma and remarks above. It is thus missing only to show the converse part. We prove a weak converse.

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

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

and for identification with decoding sets $\mathcal{D}_i^*$

$$\frac{1}{M} \sum_{j=1}^{M} W^n\left(\mathcal{D}_i^* | F_i(u_j)\right) \geq 1 - \lambda \text{ for all } i = 1, \ldots, N \tag{2}$$

and

$$\frac{1}{M}\sum_{j=1}^{M} W^n\left(\mathcal{D}_k^*|F_i(u_j)\right) \le \lambda \quad \text{for all} \quad i \ne k. \tag{3}$$

For every $i = 1, 2, \ldots, N$ we get input variables $X_i^n = (X_{i1}, \ldots, X_{in})$ and output variables $Y_i^n = (Y_{i1}, \ldots, Y_{in})$.

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

$$\log M \le \frac{I(X_i^n, Y_i^n)}{1-\overline{\lambda}} \quad \text{for all} \quad i \tag{4}$$

and by the weak converse proof for identification on the DMC with feedback (chapter "On Identification via Multi-Way Channels with Feedback: Mystery Numbers")

$$\log\log N \le \max_i H(Y_i^n). \tag{5}$$

Therefore for some i

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

if we use the concavity of I and of H. This completes the weak converse proof. □

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

Finally we propose as the following problem.

Problem This proof assumes a deterministic transmission code. Can randomized transmission codes give better overall performance? ▲

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

Not just transmission and identification, but any collection of the models in Sect. 1 of chapter "One Sender Answering Several Questions of Receivers" can

occur in various combinations. For example consider a MAC with three senders. For a given sportsman sender 1 says from which country he comes, sender 2 informs about the age groups, and sender 3 is concerned about the fields of activities.

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

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

$$\max_{x^n} H\left(W^n(\cdot|x^n)\right) \leq \sum_{t=1}^{n} \max_{x_t} H\left(W(\cdot|x_t)\right)$$

(β) For a randomized strategy F

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

and

$$H(Y_n|Y_1, \ldots, Y_{n-1}) = \sum_{y^{n-1}} \Pr(Y^{n-1} = y^{n-1}) \cdot H(Y_n|y_1, \ldots, y_{n-1})$$

$$= H\left(\sum_x W_n(\cdot|x) \sum_{y^{n-1}} \Pr\left(F_n(y_1 \ldots y_{n-1}) = x\right)\right)$$

$$\leq \max_{P_X} H(P_X W_n).$$

2 A Proof of the Weak Converse to the Identification Coding Theorem for the DMC

We present here a new approach to polynomial converses for identification, which are explained in Sect. 1 of chapter "One Sender Answering Several Questions of Receivers". We consider the proof being simpler than its predecessors. (Except for those in case of feedback [7] (chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas"), [9] (chapter "On Identification via Multi-Way Channels with Feedback: Mystery Numbers").)

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

Furthermore, in contrast to the proofs in [11, 12] the approach works also for channels without a strong converse for transmission.

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

$$\sum_x P_i(x) W(\mathcal{D}_i|x) > 1 - \lambda \qquad \text{for all } i$$

$$\sum_x P_i(x) W(\mathcal{D}_j|x) < \lambda \qquad \text{for all } i \neq j.$$

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

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

$$\rho(G) = \min_{(x,y)\in G} \frac{P_{XY}(x, y)}{P_X(x) P_Y(y)} \tag{6}$$

and

$$\sigma(G) = \max_{(x,y)\in G} \frac{P_{XY}(x, y)}{P_X(x) P_Y(y)}. \tag{7}$$

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

We state now our key results.

Lemma 115 (Codes in Informationally Balanced Sets) *For any $G \subset \mathcal{X} \times \mathcal{Y}$, $P_{XY} = P_X W$, and any $\delta' < P_{XY}(G)$ there exists a transmission code $\{(u_i, \mathcal{E}_i) : 1 \le i \le M\}$ with*

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

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

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

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

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

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

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

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

$$W\left(G_x \setminus \bigcup_{i=1}^{M} \mathcal{E}_i \middle| x\right) \le \delta'. \tag{8}$$

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

$$W(y|x) \geq \rho(G) P_Y(y) \tag{9}$$

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

$$P_Y(\mathcal{E}_i) \leq \rho(G)^{-1}. \tag{10}$$

It follows from (8) that

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

and therefore also with (10)

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

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

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

This gives (iv):

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

Further, (8) leads to $W(G_x|x) - W\left(\bigcup_{i=1}^{M} \mathcal{E}_i \middle| x\right) < \delta'$, which implies

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

and hence (v).
Finally, by definition of ρ for $y \in \mathcal{E}_i \subset G_{u_i}$

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

and by (iv)

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

Therefore

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

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

The freedom in the choice of G or even several G's makes the power of this approach. We explain this in Sects. 3, 4, and 5.

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

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

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

Introduce

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

and for suitable $\theta > 0$ and positive integer L, to be specified below, the partition

$$G(I + \beta) = \bigcup_{\ell=0}^{L-1} G^{\ell}_{XY}(I + \beta),$$

where

$$G^{\ell}_{XY}(I + \beta) = G_{XY}(I + \beta - \ell\theta) - G_{XY}\big(I + \beta - (\ell + 1)\theta\big).$$

Its atoms are balanced, because

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

For the further analysis we need a simple fact about relative entropies.

Lemma 116 *For any PD's p, q on $\mathcal{Z}$ and any $\mathcal{Z}' \subset \mathcal{Z}$*

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

Proof

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

□

We apply this fact to the PD's P_{XY} and $P_X \times P_Y$ and $\mathcal{Z}' = G(I+\beta)$. Thus

$$\begin{aligned}
I &= \sum_{(x,y)\in\mathcal{Z}'} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} + \sum_{(x,y)\notin\mathcal{Z}'}\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \\
&\geq -c + \big(1 - P_{XY}(G(I+\beta))\big)(I+\beta)
\end{aligned}$$

or

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

We can choose ℓ such that

$$P_{XY}\big(G^{\ell}_{XY}(I+\beta)\big) \geq \frac{\beta - c}{(\beta+I)L}. \tag{11}$$

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

Lemma 117 *For any distribution P_{XY} and set $D \subset \mathcal{Y}$ with*

$$P_Y(D) = \sum_{x\in\mathcal{X}} P_X(x)W(D|x) \geq 1 - \lambda$$

consider for any $\beta > 0$ and positive integer L the representative $G^{\ell}_{XY}(I+\beta)$. Then we have for $G = G^{\ell}_{XY}(I+\beta) \cap \mathcal{X} \times D$

(i) $P_{XY}(G) \geq \frac{\beta - c}{(\beta + I)L} - \lambda = \delta$, *say.*
For any $\delta' < \delta$ there is a code

$$\{(u_i, \mathcal{E}_i) : 1 \leq j \leq M\} \text{ with } \mathcal{E}_j \subset G_{u_j} \subset D \text{ for all } j = 1, \ldots, M$$

and the properties

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

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

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

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

Proof (i) is a consequence of (11) and the assumption on D. Inequality (ii) follows from (iv) in Lemma 115 and inequality (iii) follows from (v) in Lemma 115 (and (i) above). Finally, this and (vi) in Lemma 115 give (iv) and (v). □

Theorem 118 *Let the discrete (not necessarily memoryless) channel $W^n : \mathcal{X}^n \to \mathcal{Y}^n$ have an (n, N, λ_n) identification code $\{(P_i, \mathcal{D}_i) : 1 \leq i \leq N\}$, then for pairs of RV's (X_i^n, Y_i^n) with distribution $P_i \times W^n$*

$$\log\log N \leq \max_i I(X_i^n \wedge Y_i^n) + o(n) \text{ if } \lambda_n \leq n^{-7}.$$

Proof Consider any pair $(P_i, \mathcal{D}_i)$ and apply Lemma 117 for $D = \mathcal{D}_i$, $P_X = P_i$. However, we write now P_{X^n} instead of P_X. Also, for $P_i \times W^n$ we write $P_{X^nY^n}$ (instead of P_{XY}) and thus we write the representation for $P_{X^nY^n}$ as $G = G^{\ell}_{X^nY^n}(I + \beta) \cap (\mathcal{X}^n \times D)$.

Our goal is to choose parameters so that M in (ii) of Lemma 117 becomes small and δ^* in (v) of Lemma 117 becomes large. The first property guarantees that $\binom{|\mathcal{X}^n|}{M}$ is so small that the number of representing encoding sets $\{u_j : 1 \leq j \leq M\}$ meets the desired double exponential bound.

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

Accordingly the proof goes in two steps.

Step 1: We choose for $\varepsilon > 0$ $\beta = \varepsilon n$ and for convenience we choose $\delta' = \delta/2$. Clearly, for n large by Lemma 117, (i) since c is constant

$$P_{X^nY^n}(G) \geq \frac{\beta}{(\beta + I)2L} - \lambda_n = \delta_n^*. \tag{12}$$

We choose $\theta = \frac{\beta+I}{2L}$.
Using (12) and Lemma 117 (i), (v) we get now

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

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

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

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

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

By our choices and Lemma 117(ii)–(v), $\delta \geq \lambda_n$ and

$$M \leq 2n^3\ e^{I(X_i^n \wedge Y_i^n)+\varepsilon n}. \tag{13}$$

This is the first desired property. The others are

$$P_Y\left(\bigcup_{i=1}^{M} \mathcal{E}_i\right) \geq \frac{\delta}{2} \geq \frac{1}{4}n^{-3/2}. \tag{14}$$

For $\mathcal{U} = \{u_1, \ldots, u_n\}$

$$Q_{\mathcal{U}}(y) = \frac{1}{M}\sum_{j=1}^{M} W^n(y|u_i) \geq \frac{1}{2}n^{-3}P_Y(y) \tag{15}$$

and so

$$Q_{\mathcal{U}}\left(\bigcup_{i=1}^{M} \mathcal{E}_i\right) \geq \frac{1}{8}n^{-9/2}, \tag{16}$$

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

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

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

$$\mathcal{E}'^j = \mathcal{E}^j - \bigcup_{j' \neq j} \mathcal{E}^{j'};\ j = 1, 2, \ldots, K. \tag{17}$$

Since $\mathcal{E}^j \subset \mathcal{D}_{Y^j}$ and the identification code has error probabilities less than λ_n, we get from (14)

$$P_{Y^j}(\mathcal{E}'^j) \geq \frac{1}{4} n^{-3/2} - K\lambda_n \tag{18}$$

and thus by (15)

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

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

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

So $\mathcal{U}$ serves at most $16\, n^{9/2}$ times as representative and the result follows with (13). □

Remark When determining pessimistic capacities or capacity regions the observations in the first remark in chapter "One Sender Answering Several Questions of Receivers" are relevant.

3 Two Promised Results: Characterisation of the Capacity Regions for the MAC and the BC for Identification

We know from [1, 2] that the transmission capacity region $\mathcal{R}$ of a (classical: memoryless, stationary) MAC $W : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ can be characterised as the convex hull of the set of pairs $(R_{\mathcal{X}}, R_{\mathcal{Y}})$ of non-negative numbers which satisfy for some input distribution $P_{XY} = P_X \times P_Y$

$$R_{\mathcal{X}} \leq I(X \wedge Z|Y)$$

$$R_{\mathcal{Y}} \leq I(Y \wedge Z|X)$$

$$R_{\mathcal{X}} + R_{\mathcal{Y}} \leq I(XY \wedge Z). \tag{19}$$

Also, in [1] there is a non-single letter characterisation.

$$\mathcal{R} = \left\{ \frac{1}{n} \left(I(X^n \wedge Z^n), I(Y^n \wedge Z^n) \right) : n \in \mathbb{N}, P_{X^n Y^n} = P_{X^n} \times P_{Y^n} \right\}. \tag{20}$$

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

Theorem 119 *The second order identification capacity region for the MAC* equals *the first order transmission capacity region* $\mathcal{R}$.

The broadcast channel is a stochastic map

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

with components $W_1^n : \mathcal{X} \to \mathcal{Z}$ and $W_2^n : \mathcal{X} \to \mathcal{Z}$ and set of messages or the object space is

$$\mathcal{N} = \mathcal{N}_{\mathcal{Y}} \times \mathcal{N}_{\mathcal{Z}}, \qquad |\mathcal{N}_{\mathcal{Y}}| = N_{\mathcal{Y}}, \qquad |\mathcal{N}_{\mathcal{Z}}| = N_{\mathcal{Z}}$$

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

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

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

$$\sum_{x^n} W_1^n(\mathcal{D}_i|x^n) P_{ij}(x^n) \ge 1 - \lambda \qquad \text{for all } i \text{ and } j \tag{21}$$

$$\sum_{x^n} W_1^n(\mathcal{D}_{i'}|x^n) P_{ij}(x^n) \le \lambda \qquad \text{for all } i \ne i' \text{ and all } j \tag{22}$$

$$\sum_{x^n} W_2^n(\mathcal{F}_j|x^n) P_{ij}(x^n) \ge 1 - \lambda \qquad \text{for all } j \text{ and } i \tag{23}$$

$$\sum_{x^n} W_2^n(\mathcal{F}_{j'}|x^n) P_{ij}(x^n) \le \lambda \qquad \text{for all } j \ne j' \text{ and all } i. \tag{24}$$

Let $\mathcal{B}$ be the set of all achievable pairs $(R_{\mathcal{Y}}, R_{\mathcal{Z}})$ of second order rates. For its analysis we need the cones

$$\mathbb{R}_{\mathcal{Y}}^{2+} = \{(R_1, R_2) \in \mathbb{R}^2 : R_1 \ge R_2 \ge 0\}$$

and

$$\mathbb{R}_{\mathcal{Z}}^{2+} = \{(R_1, R_2) \in \mathbb{R}^2 : R_2 \ge R_1 \ge 0\}.$$

We can write $\mathcal{B}$ as a union $\mathcal{B} = \mathcal{B}^+_{\mathcal{Y}} \cup \mathcal{B}^+_{\mathcal{Z}}$, where

$$\mathcal{B}^+_{\mathcal{Y}} = \mathcal{B} \cap \mathbb{R}^{2+}_{\mathcal{Y}} \text{ and } \mathcal{B}^+_{\mathcal{Z}} = \mathcal{B} \cap \mathbb{R}^{2+}_{\mathcal{Z}}.$$

Our key observation is that for identification we can relate the capacity regions for identification of independent messages to the capacity regions for identification for degraded message sets, $\mathcal{A}_{\mathcal{Y}}$ and $\mathcal{A}_{\mathcal{Z}}$, where $\mathcal{A}_{\mathcal{Y}}$ (resp. $\mathcal{A}_{\mathcal{Z}}$) concerns the pairs of the rates of separate messages for $\mathcal{Y}$ (resp. $\mathcal{Z}$) and of common messages for $\mathcal{Y}$ and $\mathcal{Z}$. Since common messages can be interpreted as separated messages obviously

$$\mathcal{A}_{\mathcal{Y}}, \mathcal{A}_{\mathcal{Z}} \subset \mathcal{B}.$$

We can also write

$$\mathcal{A}^+_{\mathcal{Y}} = \mathcal{A}_{\mathcal{Y}} \cap \mathbb{R}^{2+}_{\mathcal{Y}} \text{ and } \mathcal{A}^+_{\mathcal{Z}} = \mathcal{A}_{\mathcal{Z}} \cap \mathbb{R}^{2+}_{\mathcal{Z}}$$

and notice that

$$\mathcal{A}^+_{\mathcal{Y}} \subset \mathcal{B}^+_{\mathcal{Y}}, \qquad \mathcal{A}^+_{\mathcal{Z}} \subset \mathcal{B}^+_{\mathcal{Z}}.$$

We come now to a key tool.

Lemma 120 (Reduction)

(i) $\mathcal{B}^+_{\mathcal{Y}} \subset \mathcal{A}^+_{\mathcal{Y}}$ *and* $\mathcal{B}^+_{\mathcal{Z}} \subset \mathcal{A}^+_{\mathcal{Z}}$.
(ii) $\mathcal{B}^+_{\mathcal{Y}} = \mathcal{A}^+_{\mathcal{Y}}$ *and* $\mathcal{B}^+_{\mathcal{Z}} = \mathcal{A}^+_{\mathcal{Z}}$.
(iii) $\mathcal{B} = \mathcal{A}$.

Proof By previous observations it remains to show (i) and by symmetry only its first part.

Let $\left\{(P_{ij}, \mathcal{D}_i, \mathcal{E}_j) : 1 \le i \le N_{\mathcal{Y}}, 1 \le j \le N_{\mathcal{Z}}\right\}$ be an identification code for the BC with error probabilities $\le \lambda$. Since $R_{\mathcal{Z}} \le R_{\mathcal{Y}}$ we can define for

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

(where w.l.o.g. divisibility of $N_{\mathcal{Y}}$ by $N_{\mathcal{Z}}$ can be assumed)

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

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

If $\mathcal{R}_{\mathcal{Y}} > \mathcal{R}_{\mathcal{Z}}$, then with error probabilities $\le \lambda$

$$2^{2^{R_y n}} \cdot 2^{-2^{R_z n}} \sim 2^{2^{R_y n}}.$$

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

We need the direct part of the ABC (asymmetric broadcast channel) coding theorem for transmission [10, 13, 14]. Here, there are separate messages for decoder $\mathcal{Y}$ (resp. $\mathcal{Z}$) and common messages for both decoders.

Let us denote Markov chains with the symbol $\ominus$, i.e. $X \ominus Y \ominus Z$ for RV's X, Y and Z. The achievable rates are (with maximal errors)

$$\begin{aligned}\mathcal{T}_{\mathcal{Y}} = \big\{(R_{\mathcal{Y}}, R_0) : \; & R_0 \le I(U \wedge Z), \\ & R_0 + R_{\mathcal{Y}} \le \min\big[I(X \wedge Y), I(X \wedge Y|U) + I(U \wedge Z)\big], \\ & U \ominus X \ominus YZ, \;\; \|U\| \le |\mathcal{X}| + 2\big\}\end{aligned}$$

and

$$\begin{aligned}\mathcal{T}_{\mathcal{Z}} = \big\{(R_0, R_{\mathcal{Z}}) : \; & R_0 \le I(U \wedge Y), \\ & R_0 + R_{\mathcal{Z}} \le \min\big[I(X \wedge Z), I(X \wedge Z|U) + I(U \wedge Y)\big], \\ & U \ominus X \ominus YZ, \;\; \|U\| \le |\mathcal{X}| + 2\big\},\end{aligned}$$

respectively.

This is our surprising result.

Theorem 121 *For the (general) BC the set of achievable pairs of second order rates is given by*

$$\mathcal{B} = \mathcal{T}'_{\mathcal{Y}} \cup \mathcal{T}'_{\mathcal{Z}},$$

where

$$\mathcal{T}'_{\mathcal{Y}} = \{(R'_{\mathcal{Y}}, R'_{\mathcal{Z}}) : \exists (R_{\mathcal{Y}}, R_0) \in \mathcal{T}_{\mathcal{Y}} \text{ with } R'_{\mathcal{Y}} = R_{\mathcal{Y}} + R_0, R'_{\mathcal{Z}} = R_0\}$$

and

$$\mathcal{T}'_{\mathcal{Z}} = \{(R'_{\mathcal{Y}}, R'_{\mathcal{Z}}) : \exists (R_0, R_{\mathcal{Z}}) \in \mathcal{T}_{\mathcal{Z}} \text{ with } R'_{\mathcal{Y}} = R_0, R'_{\mathcal{Z}} = R_0 + R_{\mathcal{Z}}\}.$$

4 The Proof for the MAC

Direct Proof

The proof of achievability is straightforward by the second method of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas",

Part I, that is, the transformator lemma. Indeed, use an *average* error transmission code in blocklength n

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

with

$$\frac{1}{M_{\mathcal{X}}}\frac{1}{M_{\mathcal{Y}}}\sum_{i,j} W^n(\mathcal{D}_{ij}^c|u_i, v_j) \le \lambda. \tag{25}$$

Then of course also

$$\frac{1}{M_{\mathcal{X}}}\sum_i \left(\frac{1}{M_{\mathcal{Y}}}\sum_j W^n\left(\left(\bigcup_{j'}\mathcal{D}_{ij'}\right)^c|u_i, v_j\right)\right) \le \lambda \tag{26}$$

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

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

As in chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" by the transformator lemma with relatively few, say $\sqrt{n}$, letters (actually even $o(\log n)$) identification of *second* order rate $\sim \frac{1}{n}\log M_{\mathcal{X}}$ can be performed from $S_{\mathcal{X}}$ to the receiver. Finally, with other $\sqrt{n}$ letters the identification of *second* order rate $\sim \frac{1}{n}\log M_{\mathcal{Y}}$ can be done from $S_{\mathcal{Y}}$ to the receiver.

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

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

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

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

Remark The decomposition principle (see [4]) does not hold for identification on the MAC. If both encoders have independent messages, but can cooperate, then

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

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

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

Remark Steinberg [15] did not use the transformator lemma, but followed the first approach in [6] (see chapter "Identification via Channels", Part I), which is based on a transmission code with small maximal errors. With deterministic maximal error transmission code the (average error) capacity region of a MAC cannot be achieved. However, it can be achieved if stochastic encoders are used (as shown in [3]) and for those coding the approach of [6] again applies.

Problem Develop a theory for identification of correlated data (see "correlated codes" in [8]). ▲

Problem Develop approximation of output statistics for the MAC to obtain a strong converse. Use random coding instead of maximal coding with rates

$$I(X \wedge Z) \leq R_{\mathcal{X}} \leq I(X \wedge Z|Y)$$

$$I(Y \wedge Z) < R_{\mathcal{Y}} \leq I(Y \wedge Z|X)$$

$$I(XY \wedge Z) \leq R_{\mathcal{X}} + R_{\mathcal{Y}}$$

and code structure $\{u_1, \ldots, u_{M_{\mathcal{X}}}\}$ and $\{v_{i1}, \ldots, v_{iM_{\mathcal{Y}}}\}$ for $i = 1, \ldots, M_{\mathcal{X}}$. ▲

Converse Proof

We follow closely the proof for a one-way channel. *Here it is essential that our approach treats general channels with memory.* Secondly we use *the characterisation (20) of the rate-region $\mathcal{R}$ for the MAC.*

In addition we partition our encoding pairs $(P_i \times Q_j)_{\substack{i=1,\ldots,N_{\mathcal{X}} \\ j=1,\ldots,N_{\mathcal{Z}}}}$ according to the values of their corresponding pairs of mutual informations $\left(I(X_i^n \wedge Z_{ij}^n), I(Y_j^n \wedge Z_{ij}^n)\right)$ where $P_{X_i^n} = P_i$, $P_{Y_j^n} = Q_j$, $P_{Z_{ij}^n} = (P_i \times Q_j)W^n$, as follows.

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

$$S = \left\{(R_1, R_2) : 0 \leq R_1 \leq \log|\mathcal{X}|, 0 \leq R_2 \leq \log|\mathcal{Y}|\right\}$$

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

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

$$\left(\frac{1}{n} I(X_i^n \wedge Z_{ij}^n), \frac{1}{n} I(Y_j^n \wedge Z_{ij}^n)\right) \in S_{a(i,j),b(i,j)}. \tag{27}$$

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

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

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

$$\left|\left\{(i^*, j) : (i^*, j) \in \mathcal{N}(\eta)\right\}\right| \geq \frac{|\mathcal{N}(\eta)|}{N_{\mathcal{X}}} \geq \frac{N_{\mathcal{Y}}}{g(\eta)}, \tag{28}$$

$$\left|\left\{(i, j^*) : (i, j^*) \in \mathcal{N}(\eta)\right\}\right| \geq \frac{|\mathcal{N}(\eta)|}{N_{\mathcal{Y}}} \geq \frac{N_{\mathcal{X}}}{g(\eta)}. \tag{29}$$

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

$$M_{i\mathcal{X}}^j \leq \exp\left\{I(X_i^n \wedge Z_{ij}^n) + o(n)\right\}$$

$$M_{j\mathcal{Y}}^i \leq \exp\left\{I(Y_j^n \wedge Z_{ij}^n) + o(n)\right\}$$

and (27) holds.

Moreover, for all indices

$$\frac{1}{M_{i\mathcal{X}}^j} \sum_{u\in\mathcal{U}_i^j} \sum_{y^n} W^n(E_{iu}^j \cap \mathcal{D}_{ij}|u, y^n) Q_j(y^n) \geq n^{-4} \tag{30}$$

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

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

(1) $\frac{1}{n}\log M_{i\mathcal{X}}^j \leq R_{\mathcal{X}}^* + \eta$ and $\frac{1}{n}\log M_{j\mathcal{Y}}^i \leq R_{\mathcal{Y}}^* + \eta$.

(2) By (28), (29) there are at most $\binom{|\mathcal{X}|^n}{2^{(R_{\mathcal{X}}^*+\eta)n}}$ different codes $\mathcal{U}_{i^*}^j$ in row i^* and at most $\binom{|\mathcal{Y}|^n}{2^{(R_{\mathcal{Y}}^*+\eta)n}}$ codes $\mathcal{V}_{j^*}^i$ in column j^*.

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

Finally, therefore

$$\frac{1}{n}\log\log N_{\mathcal{X}} \leq R_{\mathcal{X}}^* + 2\eta,$$

$$\frac{1}{n}\log\log \mathcal{N}_{\mathcal{Y}} \leq R_{\mathcal{Y}}^* + 2\eta.$$

Problem In [15] Steinberg strengthens our polynomial converse to a weaker converse. The main difference of his proof is a sharpening of the bound in Theorem 118, which is based on a generalization of [12, Lemma 5]. We suggest as a further improvement to establish a strong converse by our hypergraph lemma, which is presented in [5, Section VI]. Otherwise in his proof the same ideas are used, namely facts (19) and (20) and a suitable subcode selection. The whole proof with all auxiliary results exceeds the present one in length roughly by a factor 3. ▲

5 The Proof for the BC

The Direct Part We use the reduction lemma and the ABC coding theorem mentioned in Sect. 3. Even though that theorem holds for maximal errors we use average errors so that the transmission codes establish two common random experiments of the sender with both receivers, resp., with rates in $\mathcal{T}'_{\mathcal{Y}} \cup \mathcal{T}'_{\mathcal{Z}}$.

The Converse Part Suppose w.l.o.g. that $R_{\mathcal{Z}} < R_{\mathcal{Y}} + \varepsilon$, ε arbitrarily small, and that the $\mathcal{Y}$-decoder has a separate part coded into row numbers and that the common part for both decoders is coded into column numbers with the encodings $(P_{uv})_{u=1,\ldots,N_{\mathcal{Y}} v=1,\ldots,N_{\mathcal{Z}}}$.

Note that we can start with a smaller common rate, so that $M_{\mathcal{Y}} \sim M_{\mathcal{Z}} \cdot M_{\mathcal{Y}}$ (If the common rate is bigger in the ABC model, we can convert this by the Reduction Lemma 120).

We associate RV's and information quantities as follows:

Let U, V be auxiliary RV's with $\Pr((U,V) = (u,v)) = \frac{1}{N_{\mathcal{Y}} N_{\mathcal{Z}}}$ for $u = 1,\ldots,N_{\mathcal{Y}}$ and $v = 1,\ldots,N_{\mathcal{Z}}$. Furthermore let X^n take values in $\mathcal{X}^n$ with conditional PD $P_{X^n|U=u,V=v}(x^n) = P_{uv}(x^n)$, let Y^n take values in $\mathcal{Y}^n$ with conditional PD $P_{Y^n|U=u,V=v}(y^n) = \sum_{x^n} P_{uv}(x^n) W_1^n(y^n|x^n)$, and let Z^n take values in $\mathcal{Z}^n$ with conditional PD $P_{Z^n|U=u,V=v}(z^n) = \sum_{x^n} P_{uv}(x^n) W_2^n(z^n|x^n)$.

Thus we get information quantities

$$I(U \wedge Z^n|V=v), \qquad I(X^n \wedge Y^n|U,V=v), \qquad I(X^n \wedge Y^n|V=v),$$

and the Markov condition $(U,V) \ominus X^n \ominus (Y^n, Z^n)$.

As in the proof of Theorem 119 we make η-approximations, first for all $\frac{1}{n} I(X^n \wedge Y^n|V=v)$ with biggest class of value I_{η_3}.

This gives as in the one-way channel coding theorem for identification

$$\frac{1}{n} \log\log N_{\mathcal{Y}} \le I_{\eta_3}. \tag{31}$$

In the remaining matrix keep I_{η_2} for $I(X^n \wedge Y^n|U,V=v)$ and then all $I(U \wedge Z^n|V=v)$ approximately I_{η_1}.

We upper bound the number of columns by upper bounding the number of codes (via Lemma 117) representing triples $(P_{U|V=v}, P_{Z^n|U,V=v}, \mathcal{D}_v)$. Thus for $\lambda_n = n^{-6}$ (as usual)

$$\frac{1}{n} \log \log N_{\mathcal{Z}} \leq I_{\eta_1} + 2\eta. \tag{32}$$

Within column v^* a significant number of terms has

$$\frac{1}{n} I(X^n \wedge Y^n | U = u, V = v^*) \leq I_{\eta_2} + \beta^*.$$

This gives the desired row number estimate

$$\begin{aligned}
&\frac{1}{n} \log \log N_{\mathcal{Y}} \\
&\leq \min(I_{\eta_1} + I_{\eta_2}, I_{\eta_3}) + 2\eta + \beta^* \\
&= \min\left\{I(U \wedge Z^n | V = v^*) + I(X^n \wedge Y^n | U, V = v^*), I(X^n \wedge Y^n | V = v^*)\right\} \\
&+ 2\eta + \beta^*
\end{aligned}$$

and thus $(R_{\mathcal{Y}}, R_{\mathcal{Z}}) \in \mathcal{T}'_{\mathcal{Y}}$ by the converse in the ABC coding theorem, which shows that the information quantities single-letterize.

Remark Theorem 121 has an important consequence. Whereas for one-way channels the common randomness capacity equals the transmission capacity and the transmission capacity region is still unknown for general broadcast channels *we know now its common randomness capacity region*, where common random experiments for $\mathcal{X}$-encoder and $\mathcal{Y}$-decoder and, simultaneously, for $\mathcal{X}$-encoder and $\mathcal{Z}$-decoder are generated. *Indeed it equals the second order identification capacity region!*

That the latter includes the former is clear from our proof of the direct part. The reverse implication follows indirectly by the same argument.

Interesting here is that the outer bound for the common randomness capacity region is proved via identification.

The situation changes, if constraints like independency or security are imposed on the two common random experiments.

A transmission code with rates $(R_{\mathcal{Y}}, R_{\mathcal{Z}})$ can be used for independent common random experiments and thus the transmission capacity region for the general broadcast channel is contained in the identification capacity region.

Finally we mention that the identification capacity region $T'_{\mathcal{Y}} \cup T'_{\mathcal{Z}}$ is convex, because it equals the common randomness capacity region for which time sharing applies and thus convexity is given.

References

1. R. Ahlswede, Multi–way communication channels, in *Proceedings of the 2nd International Symposium on Information Theory, Tsahkadsor Armenian SSR, 1971* (Akadémiai Kiadó, Budapest, 1973), pp. 23–52
2. R. Ahlswede, The capacity region of a channel with two senders and two receivers. Ann. Probab. **2**(5), 805–814 (1974)
3. R. Ahlswede, Elimination of correlation in random codes for arbitrarily varying channels. Z. Wahrsch. und verw. Geb. **44**, 159–175 (1978)
4. R. Ahlswede, Coloring hypergraphs: a new approach to multi–user source coding. Part I J. Comb. Inf. Syst Sci **1**, 76–115 (1979). Part II **5**(3), 220–268 (1980)
5. R. Ahlswede, Towards a general theory of information transfer, in *Shannon Lecture at ISIT in Seattle 13th July 2006*. IEEE Information Theory Society Newsletter (2007)
6. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**, 15–29 (1989)
7. R. Ahlswede, G. Dueck, Identification in the presence of feedback—a discovery of new capacity formulas. IEEE Trans. Inf. Theory **35**, 30–39 (1989)
8. R. Ahlswede, T.S. Han, On source coding with side information via a multiple–access channel and related problems. IEEE Trans. Inf. Theory **29**(3), 396–412 (1983)
9. R. Ahlswede, B. Verboven, On identification via multi–way channels with feedback. IEEE Trans. Inf. Theory **37**(5), 1519–1526 (1991)
10. T.M. Cover, An achievable rate region for the broadcast channel. IEEE Trans. Inf. Theory **21**, 399–401 (1975)
11. T.S. Han, S. Verdú, New results in the theory and application of identification via channels. IEEE Trans. Inf. Theory **38**, 14–25 (1992)
12. T.S. Han, S. Verdú, Approximation theory of output statistics, IEEE Trans. Inf. Theory **39**(3), 752–772 (1993)
13. J. Körner, K. Marton, General broadcast channels with degraded message sets. IEEE Trans. Inf. Theory IT **23**(1), 60–64 (1977)
14. E.C. van der Meulen, Random coding theorems for the general discrete memoryless broadcast channel. IEEE Trans. Inf. Theory IT **21**, 180–190 (1975)
15. Y. Steinberg, New converses in the theory of identification via channels. IEEE Trans. Inf. Theory **44**(3), 984–998 (1998)

Data Compression

1 Noiseless Coding for Identification

For this section we recall some basic definitions, which are introduced in [4].

Definition 122 A **code** (of variable length) is a function $c : \mathcal{X} \to \mathcal{Y}^*$,

$\mathcal{X} = \{1, \ldots, a\}$. So $\{c(1), c(2), \ldots, c(a)\}$ is the set of **codewords**, where for $x = 1, \ldots, a$ $c(x) = \big(c_1(x), c_2(x), \ldots, c_{L(x)}(x)\big)$ and $L(x)$ is denoted as the **length** of the codeword $c(x)$.

Definition 123 A code c is a **prefix code**, if for any two codewords $c(x)$ and $c(y)$, $x \neq y$, with $L(x) \leq L(y)$ holds $\big(c_1(x), c_2(x), \ldots, c_{L(x)}(x)\big) \neq \big(c_1(y), c_2(y), \ldots, c_{L(x)}(y)\big)$. So in at least one of the first $L(x)$ components $c(x)$ and $c(y)$ differ.

A *discrete source* is a pair $(\mathcal{U}, P)$, where the *output space* $\mathcal{U}$ is a finite set of cardinality N and P is a probability distribution on $\mathcal{U}$. We call the discrete source binary, if $N = 2$. It is called symmetric, if all probabilities are the same. Further, a *discrete memoryless source* is a pair $(\mathcal{U}^n, P^n)$, where $\mathcal{U}^n$ is the cartesian product of a finite set $\mathcal{U}$. P^n is a probability distribution on $\mathcal{U}^n$, where the probability of an element $u^n \in \mathcal{U}^n$ is product of the probabilities of its individual components.

We abbreviate a discrete memoryless binary symmetric source by BSS.

Let $(\mathcal{U}, P)$ be a discrete source, where $\mathcal{U} = \{1, 2, \ldots, N\}$, $P = (P_1, \ldots, P_N)$, and let $\mathcal{C} = \{c_1, \ldots, c_N\}$ be a binary prefix code (PC) for this source with $\|c_u\|$ as length of c_u.

Introduce the RV U with $\Pr(U = u) = p_u$ for $u = 1, 2, \ldots, N$ and the RV C with $C = c_u = (c_{u_1}, c_{u_2}, \ldots, c_{u\|c_u\|})$ if $U = u$.

We use the PC for noiseless identification, that is user u wants to know whether the source output equals u, that is, whether C equals c_u or not. He iteratively checks whether $C = (C_1, C_2, \ldots)$ coincides with c_u in the first, second, etc. letter and stops when the first different letter occurs or when $C = c_u$.

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_12

What is the expected number $L_{\mathcal{C}}(P,u)$ of checkings?

In order to calculate this quantity we introduce for the binary tree $T_{\mathcal{C}}$, whose leaves are the codewords $c_1,\ldots,c_N$, the sets of leaves $\mathcal{C}_{ik}(1 \le i \le N; 1 \le k)$, where $\mathcal{C}_{ik} = \{c \in \mathcal{C} : c$ coincides with c_i exactly until the k'th letter of $c_i\}$. If C takes a value in $\mathcal{C}_{uk}, 0 \le k \le \|c_u\| - 1$, the answers are k times "Yes" and 1 time "No". For $C = c_u$ the answers are $\|c_u\|$ times "Yes". Thus

$$L_{\mathcal{C}}(P,u) = \sum_{k=0}^{\|c_u\|-1} P(C \in \mathcal{C}_{uk})(k+1) + \|c_u\| P_u.$$

For code $\mathcal{C}$ $L_{\mathcal{C}}(P) = \max_{1\le u\le N} L_{\mathcal{C}}(P,u)$ is the expected number of checkings in the worst case and $L(P) = \min_{\mathcal{C}} L_{\mathcal{C}}(P)$ is this number for a best code.

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

$$L_{\tilde{\mathcal{C}}}(P,u),\ L_{\tilde{\mathcal{C}}}(P) \text{ and } \tilde{L}(P).$$

What are the properties of $L(P)$ and $\tilde{L}(P)$? We call for a kind of "identification entropies" serving as bounds like Boltzmann's entropy does in Shannon's source coding. Notice that every user comes with the same fixed code much faster to his goal to know "it's me—it's not me" than the one person in Shannon's model, who wants to use the outcome of the source always.

Moreover, as in [5] one can replace the lengths $||c_u||$ by $\varphi(||c_u||)$ where $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous and strictly monotone increasing.

Thus one gets functionals

$$L(P,\varphi) \text{ and } \tilde{L}(P,\varphi).$$

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

Let us start with $P_N = \left(\frac{1}{N},\ldots,\frac{1}{N}\right)$ and set $f(N) = L(P_N)$. Clearly

$$f(2^k) \le 1 + \frac{1}{2} f(2^{k-1}),\ f(2) = 1$$

and therefore

$$f(2^k) \le 2 - 2^{-(k-1)}. \tag{1}$$

On the other hand it can be verified that

$$f(9) = 1 + \frac{10}{9} > 2 \text{ and more generally, } f(2^k+1) > 2.$$

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

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

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

2 Noiseless Coding for Multiple Purposes

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

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

$$L(n) = \min_f \max_{1 \leq t \leq n} \mathbb{E}\ell(f,t). \tag{2}$$

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

Since $\frac{1}{n}\sum_{t=1}^{n} \ell(f,t) = \frac{n+1}{2}$, one should do better. In [1] we stated the problem to determine $L(n)$. Don Coppersmith [6] gave a rather precise bound.

Theorem 124 $\frac{n+1}{2} \leq L(n) \leq \frac{n+2}{2}$.

Proof The lower bound is obvious, because

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

and

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

For the upper bound set $f_1(X^n) = X_1$ and for $2 \leq i \leq n$ set

$$f_i(X^n) = \begin{cases} X_i & \text{if } X_1 = 0 \\ X_{n+2-i} & \text{if } X_1 = 1. \end{cases}$$

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

Remark A weaker upper bound, but more uniform distribution of the stopping times is obtained as follows: Let the first $\lceil \log_2 n \rceil$ bits be

$$\big(f_1(X^n), f_2(X^n), \ldots, f_{\lceil \log n \rceil}(X^n)\big) = (X_1, X_2, \ldots, X_{\lceil \log n \rceil})$$

and let these $\log n$ bits index a cyclic shift of the remaining $n - \log n$ bits so that the distribution of stopping times is approximately uniform between $\log n$ and n for $t > \lceil \log n \rceil$. This leads to the weaker upper bound

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

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

Noiseless Source Coding Problems of Infinite Order: Ordering and Identification

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

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

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

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

Lemma 125 $T_n = 1 + \frac{1}{2}T_{n-1} = \frac{2^n-1}{2^{n-1}}$, $n \geq 1$.

This is a simple exercise. Notice that

$$\lim_{n\to\infty} T_n = 2. \tag{3}$$

So the compression rate exceeds any finite order.

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

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

Problems

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

   An instructive source is given by the distribution which assigns probability $\frac{1}{n}$ to the sequences starting with k 1's and continuing with 0's only. For $u^n = (1, 1, \ldots, 1)$ the running time of the previous encoding function is always n. However, by choosing $f_1(X^n) = X_{\lceil \frac{n}{2} \rceil}$ etc. the worst case expected running time is still less than 2.
2. For any distribution P on $\{0, 1\}^n$, is the worst case expected running time less than 2? In case the answer is negative, determine the best constant (independent of n) upper bound! An obvious algorithm: number probabilities in decreasing order; $P_1 \geq P_2 \geq \cdots \geq P_N$ and divide as equally as possible $P_1 + P_2 + \cdots + P_{N_1}$, $P_{N_1+1} + \cdots + P_N$. $f_1(X^n)$ says whether $i \in \{1, \ldots, N_1\}$ or not, etc.

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

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

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

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

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

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

Let F choose with probabilities $\frac{1}{3}$ the partitions

$$\big\{\{1\},\{2,3\}\big\},\big\{\{1,3\},\{2\}\big\},\big\{\{3\},\{1,2\}\big\}$$

in the first step, the second step is canonical. Then

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

3. It is also reasonable to study alphabetical source codes for identification. For example for different intervals of a pipeline different repairman are responsible. They want to know whether a defect occurred in their interval or not.
4. Suppose that $N=2^k$ numbers are stored in 0–1 bits in a machine. Upon request a further bit is revealed by the machine. What is the average number of requests so that person i knows whether i occurred or not?
5. One can study multiple purpose coding problems with noise (see [3], which gives a common generalization of Shannon's noiseless coding theorem and coding theorems for noisy channels). What are the generalizations (there is one in [3]) of Kraft's inequality?
6. These source coding problems open a whole area of research. Are there coding problems of an order between first order (as in the component problem) and infinite order (as in the ordering problem)?
7. It is remarkable in this context also that the ordering problem *via channels* is not easier than transmission, if maximal errors are used. However, if for the second kind error probability the average is taken, then the ordering problem becomes of infinite order (similar as the identification problem does). Indeed just map the numbers $1,\dots,N$ onto codewords of a transmission code $\big\{(u_i,\mathcal{D}_i):1\le i\le\big\}$ as follows:

   For any $K<N$ write $j\in\{1,\dots,N\}$ as $j=rK+s$, $0\le s<K$, and map j on u_r. Now just let N go to infinity and choose $K=\lceil\frac{N}{M}\rceil$.
8. It is also interesting that for maximal second kind error probabilities the identification problem via channels has second order behaviour whereas—as mentioned before—the ordering problem has first order behaviour.

   We therefore ask the following question:

   Is there a reasonable coding problem with average error of second kind as performance criterion which is neither of first order nor of infinite order behaviour? In the positive case, what is the hierarchy of all orders?
9. If $\kappa<\frac{1}{2}C_{pol}Sh$, then first order capacity R_1 equals infinity. However, if $\kappa>\frac{1}{2}C_{pol}Sh$, is then $R_1>C_{pol}Sh$ possible?

References

1. R. Ahlswede, Eight problems in information theory, in *Open Problems in Communication and Computation* (eds.) by, T.M. Cover, B. Gopinath (Springer, New York, 1987)
2. R. Ahlswede, I. Csiszár, Common randomness in information theory and cryptography, Part I: Secret sharing. IEEE Trans. Inf. Theory **39**(4), 1121–1132 (1993). R. Ahlswede, I. Csiszár, Common randomness in information theory and cryptography, Part II: CR capacity. IEEE Trans. Inf. Theory **44**(1), 225–240 (1998)
3. R. Ahlswede, P. Gács, Two contributions to information theory, Colloquia Mathematica Societatis János Bolyai, 16, in *Topics in Information Theory* (eds.) by, I. Csiszár, P. Elias (Keszthely, Hungaria, 1975), pp. 17–40
4. A. Ahlswede, I. Althöfer, C. Deppe, U. Tamm (eds.) *Storing and Transmitting Data: Rudolf Ahlswede's Lectures on Information Theory 1*. Foundations in Signal Processing, Communications and Networking, vol. 10 (Springer, Berlin, 2014)
5. C.C. Campbell, Definition of entropy by means of a coding problem. Z. Wahrsch. und verw. Geb. **6**(2), 113–119 (1966)
6. D. Coppersmith, Private Communication (1987)

Perspectives

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

Right now the most visible influences are new approximation problems (like approximation of output statistics [14] or entropy approximations based on Schur-convexity [10] etc.), a new emphasis on random number generation [1] and, above all, an understanding of the concept of common randomness [9], in identification [10, 11, 13], cryptography [7], and classical transmission problems of arbitrarily varying channels [3, 5, 12], and the paper [6], with a novel capacity formula, which could not be derived before.

It is also fascinating to discover how transmission problems and identification problems in multi-user theory show often some kind of duality. Often identification problems are mathematically more complex and in other cases we encounter the opposite: there is a rather *complete* capacity theory for identification via multi-way channels in case of complete feedback [10, Lecture 3], whereas for transmission with feedback we don't even understand the multiple access channel.

We conclude with three more recently encountered directions of research.

1 Comparison of Identification Rate and Common Randomness Capacity: Identification Rate can Exceed Common Randomness Capacity and Vice Versa

One of the observations of [9] (chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas") was that random experiments, to whom the communicators have access, essentially influence the value of the identification

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_13

capacity $C_{pol}ID$. We introduce now *common randomness capacity*, which was called mystery number in [10] (chapter "On Identification via Multi-Way Channels with Feedback: Mystery Numbers"), and has subsequently been called by us in lectures and papers by its present name.

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

$$\Pr\{K \neq L\} < \epsilon \qquad \text{and} \qquad \frac{H(K)}{n} > \nu - \delta.$$

Actually, if sender and receiver have a common randomness capacity $C_{pol}CR$ then by the so called $\sqrt{n}$-trick of chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas", that is, the transformator lemma (discussed in [4]), always

$$C_{pol}ID \geq C_{pol}CR \text{ if } C_{pol}ID > 0. \tag{1}$$

For many channels (see [7, 9]), in particular for channels with feedback [9, 10], equality has been proved.

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

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

Example In [15] one can find also an example with $0 < C_{pol}ID < C_{pol}CR$)

Example We will now prove the existence of a sequence of channels (not a sequence of discrete memoryless channels) with $C_{pol}ID = 1, C_{pol}CR = 0$.

We use a Gilbert type construction of error correcting codes with constant weight words. This was done for certain parameters in [8] (see chapter "Identification via Channels", Part I). The same arguments give for parameters needed here the following auxiliary result.

Proposition 126 *Let $\mathcal{Z}$ be a finite set and let $\lambda \in (0, 1/2)$ be given. For $(2^{3/\lambda})^{-1} < \varepsilon < (2^{2/\lambda} + 1)^{-1}$ a family $A_1, \ldots, A_N$ of subsets of $\mathcal{Z}$ exists with the properties*

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

and

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

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

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

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

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

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

By Proposition 126 and $3/\lambda > \ell > 2/\lambda$

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

and

$$C_{pol}ID \ge \liminf_{n\to\infty} \frac{1}{n} \log\log N(n,\lambda) \ge 1.$$

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

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

Remark The case of bounded input alphabets remains to be analyzed. What are "natural" candidates for equality of $C_{pol}ID$ and $C_{pol}CR$?

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

2 Robustness, Common Randomness and Identification

It is understood now [6, 7] how the theory of AV-channels is *intimately* related to the concept of robust common randomness. A key tool is the balanced hypergraph coloring [2]. We sketch now another direction concerning robustness and identification.

For more robust channel models, for instance in jamming situations, where the jammer knows the word to be sent (c.f. AV-channels with maximal error criterion), the communicators are forced to use the maximal error concept. In case of identification this makes the randomization in the encoding (see [8, Lecture 1]) superfluous. Now, for a DMC W it was mentioned in chapter "Identification via Channels" that in the absence of randomization the identification capacity, say

$C_I^*(W)$, equals the logarithm of the number of different row-vectors in W. This is easy to show, however, a formidable problem arises if the DMC W is replaced by the AVC $\mathcal{W}$. In fact, for 0-1-matrices only in $\mathcal{W}$ we are—exactly as for transmission—led to the equivalent Shannon-zero-capacity problem. But for general $\mathcal{W}$ the identification problem is quite different from the transmission problem.

In so far there is a lower bound on $C_I^*(\mathcal{W})$, which implies for

$$\mathcal{W} = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ \delta & 1-\delta \end{pmatrix} \right\}, \qquad \delta \in (0,1)$$

that $C_I^*(\mathcal{W}) = 1$, which is obviously tight. It exceeds the known capacity for transmission. The capacity for

$$\mathcal{W} = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix} \right\}$$

is unknown.

3 Beyond Information Theory: Identification as a New Concept of Solution for Probabilistic Algorithms

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

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

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

References

1. R. Ahlswede, The capacity region of a channel with two senders and two receivers. Ann. Probab. **2**(5), 805–814 (1974)
2. R. Ahlswede, Coloring hypergraphs: a new approach to multi–user source coding. Part I, J. Comb. Inf. Syst. Sci. **1**, 76–115 (1979). Part II **5**(3), 220–268 (1980)
3. R. Ahlswede, General theory of information transfer, in *Preprint 97–118, SFB 343 Diskrete Strukturen in der Mathematik* (Universität Bielefeld, Bielefeld, 1997)
4. R. Ahlswede, Towards a general theory of information transfer, in *Shannon Lecture at ISIT in Seattle 13th July 2006*. IEEE Information Theory Society Newsletter (2007)
5. R. Ahlswede, N. Cai, Arbitrarily varying multiple–access channels, in *Part I: Ericson's Symmetrizability is Adequate, Gubner's Conjecture is True, Preprint 96–068, SFB Diskrete*

Strukturen in der Mathematik (Universität Bielefeld, Bielefeld). *Part II: Correlated Sender's Side Information, Correlated Messages and Ambiguous Transmission*. Preprint 97–006, SFB 343 Diskrete Strukturen in der Mathematik, Universität Bielefeld. IEEE Trans. Inf. Theory, vol. 45(2), 749–756 (1999)

6. R. Ahlswede, N. Cai, The AVC with noiseless feedback and maximal error probability: a capacity formula with a trichotomy. Preprint 96–064, SFB 343 Diskrete Strukturen in der Mathematik, Universität Bielefeld, Numbers, Information and Complexity, Special volume in honour of R. Ahlswede on occasion of his 60th birthday, editors I. Althöfer, N. Cai, G. Dueck, L.H. Khachatrian, M. Pinsker, A. Sárközy, I. Wegener, Z. Zhang (Kluwer Acad. Publication, Boston, Dordrecht, London), pp. 151–176 (1996)
7. R. Ahlswede, I. Csiszár, Common randomness in information theory and cryptography, Part I: Secret sharing. IEEE Trans. Inf. Theory **39**(4), 1121–1132 (1993). R. Ahlswede, I. Csiszár, Common randomness in information theory and cryptography, Part II: CR capacity. IEEE Trans. Inf. Theory **44**(1), 225–240 (1998)
8. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**, 15–29 (1989)
9. R. Ahlswede, G. Dueck, Identification in the presence of feedback—a discovery of new capacity formulas. IEEE Trans. Inf. Theory **35**, 30–39 (1989)
10. R. Ahlswede, B. Verboven, On identification via multi–way channels with feedback. IEEE Trans. Inf. Theory **37**(5), 1519–1526 (1991)
11. R. Ahlswede, Z. Zhang, New directions in the theory of identification via channels, in *SFB 343 Diskrete Strukturen in der Mathematik, Bielefeld, Preprint 94–010* (1994). IEEE Trans. Inf. Theory **41**(4), 1040–1050 (1995)
12. R. Ahlswede, B. Balkenhol, C. Kleinewächter, Identification for Sources, in *General Theory of Information Transfer and Combinatorics* (eds.) by R. Ahlswede, et al. Lecture Notes in Computer Science, vol. 4123 (2006)
13. T.S. Han, *Information-spectrum Methods in Information Theory*, vol. 50 (Springer, Berlin, 2013)
14. T.S. Han, S. Verdú, Approximation theory of output statistics. IEEE Trans. Inf. Theory **39**(3), 752–772 (1993)
15. C. Kleinewächter, On identification, in *General Theory of Information Transfer and Combinatorics*. Lecture Notes in Computer Science, vol. 4123 (Springer, New York, 2006), pp. 62–83

Part III
Identification, Mystery Numbers, or Common Randomness

The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints

1 Introduction

The first multi-user model of communication subject to secrecy constraints was Wyner's [13] "wiretap channel". This model involves two channels with common input, the output of the first channel being available to the "legitimate receiver" and the output of the second channel to the "wiretapper". The question is at which rate can messages be sent to the legitimate receiver while keeping the wiretapper ignorant. Wyner [13] considered and solved this problem for the case when the wiretapper's channel was a degraded version of the legitimate receiver's channel. Csiszár and Körner [5] gave a solution for the general case of any two (discrete memoryless) channels. They showed, in particular, that the "secrecy capacity" was always positive unless the wiretapper's channel was less noisy than the legitimate receiver's one (cf. also the book Csiszár and Körner [6], pp. 407–411).

Recently, Maurer [9] demonstrated that the availability of a public feedback channel could make secret transmission possible even in such cases when the secrecy capacity without feedback was zero. In fact, Maurer proposed a scheme that enabled the legitimate receiver to share a random key with the sender, using transmissions over the public feedback channel in such a way that no information about the key was given away to the wiretapper. In this scheme, both the legitimate receiver's and the wiretapper's channel were assumed to be binary symmetric, with independent but otherwise arbitrary noise. Since the key generated by the receiver and shared with the sender could be used to encrypt messages, secret transmission became possible even if the wiretapper's channel was the better one. Maurer also hinted at a source-type model. His presentation gave an important motivation for this work.

One goal of this chapter (see [2]) is to propose a systematic study of the related problems of secret sharing and secret transmission on the basis of an information theoretic model. By secret sharing we mean generating common randomness at two

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_14

(or more) terminals, without giving information about it to a third part. This may be realized by generating a random message at either terminal and transmitting it over a secure channel to the other one but also in more complex ways that may include communication over a public channel and using side information that may be available. Of course, once secret sharing has taken place, it can always be used to achieve secret transmission via encryption.

While this general problem area has been intensively researched, it has hardly been looked at from the information theoretic point of view. The popular computational complexity approach (Diffie and Hellman [8], Rivest, Shamir and Adleman [12]) certainly appears very fruitful. Still, we argue that a general information theoretic approach to this field is also needed. Even though it may not lead to the emergence of new cryptosystems, it is likely to lead to new insights, complementing the more practical complexity approach in much the same way as Shannon's theory, in general, complements communication theory and coding theory.

It is to be mentioned that the problem of generating common randomness is an important one even without the secrecy requirement. E.g., for arbitrarily varying channels, the possibility of reliable transmission often depends on whether sender and receiver have common randomness available (of arbitrarily small positive rate). Indeed, in the presence of such common randomness reliable transmission is possible at random coding capacity (Ahlswede [1]) that may be positive even if otherwise the capacity is zero (for more on the capacity of arbitrarily varying channels cf. Csiszár and Narayan [7]). Common randomness shared by sender and receiver plays a key role also in the theory of identification capacity as opposed to transmission capacity, developed by Ahlswede and Dueck [3, 4] (see chapters "Identification via Channels" and "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" in Part I). In this first part of the chapter, however, attention will be restricted to generating common randomness subject to a secrecy constraint, i.e., secret sharing.

Some problems that immediately present themselves will be fully or partially solved. The results demonstrate that secret sharing can be effectively dealt with by techniques originally developed for multiterminal communication problems without secrecy constraints, cf. Csiszár and Körner [6, Chapter 3]. The large variety of related further problems will be left for future research.

Two kinds of models will be considered, one having the flavor of source coding and the other of channel coding. These two models are closely related. To facilitate understanding, first the simplest versions of these models will be treated (Sect. 2). The main results of the chapter are stated in Sect. 3 and proved in Sect. 4.

Throughout this chapter, the terminology of the book Csiszár and Körner [6] will be used.

After the original paper [2] that this chapter is based on had been submitted, the authors learned of more recent results of Maurer on generating a shared key, that partially overlap with results in the paper. Maurer's results will be published in full in [11], and some of them appear already in [10]. In particular, Maurer [10] gave lower bounds on what we call key-capacity for the channel-type model with wiretapper, in the binary case. He also showed that the key rate he had obtained in

[7] was best possible for that model. This proof relied on a general upper bound stated but not proved in [10], which was the same as ours in Theorem 133. Maurer [11] addresses general source-type and channel-type models with wiretapper (in our terminology) and gives 'lower and upper bounds on key-capacity, including a proof of the upper bound stated in [10]. He also obtains the results of the corollaries of our Theorems 131 and 133. Maurer's results do not include a single-letter characterization of key-capacity with a one-way use of the public channel (our Theorem 131 and 133) and neither our Theorem 135. On the other hand, his papers [10, 11] contain some other results which we do not have in this chapter.

2 Generating a Shared Secret Key When the Third Party Has No Side Information

The main results of this chapter will be stated in Sect. 3. Here we introduce ampler versions of the problems treated there, in order to facilitate their understanding.

In both models below, we consider secret sharing between two terminals, to be called Terminal $\mathcal{X}$ and Terminal $\mathcal{Y}$. Both models involve an unspecified integer n (the block-length), and we will be interested in the case when n is large.

Model S: (source-type model) We are given a DMMS (discrete memoryless multiple source) with two component sources and generic variables (X, Y). Terminal $\mathcal{X}$ "can see" the source outputs $X^n = (X_1, \ldots, X_n)$ and Terminal $\mathcal{Y}$ "can see" the source outputs $Y^n = (Y_1, \ldots, Y_n)$. Further, a noiseless public channel of unlimited capacity is available for communication between the two terminals.

Model C: (Channel-type model) We are given a DMC (discrete memoryless channel) $\{W : \mathcal{X} \to \mathcal{Y}\}$. Terminal $\mathcal{X}$ can govern the input of this DMC while Terminal $\mathcal{Y}$ observes the output. In addition to transmissions of length n over this DMC, which is considered a secure channel, also a noiseless public channel of unlimited capacity may be used for communication between the two terminals.

Remark In Model C, we chose to denote the input and output alphabets by the same symbols as the corresponding terminals, believing that this will be intuitive rather then ambiguous. Similarly, in Model S, the alphabets of the two component sources will also be denoted by $\mathcal{X}$ and $\mathcal{Y}$.

Next we describe what we mean by permissible secret sharing strategies. Since we want to allow for all strategies that are abstractly conceivable (even very complex ones that may be quite impractical), this description is somewhat cumbersome. The main result of this section will be that recurse to those complex secret sharing strategies is not necessary because optimum secret sharing can be achieved in a very simple way.

Communication over the public channel will be visualized as an exchange of messages or codewords Φ_i, generated by terminal $\mathcal{X}$ and Ψ_i, generated by terminal

$\mathcal{Y}$, at consecutive instances $i = 1, \ldots, k$. Here Φ_i and Ψ_i may depend on all information available at the corresponding terminal at instant i. For convenience, these Φ_i and Ψ_i will be referred to as *forward transmissions* and *backward transmissions*, respectively. Of course, our model includes the possibility of one-way communication, because Φ_i or Ψ_i (or both) may be set equal to the empty word.

It will be convenient to assume that as the zero's step of any secret sharing strategy, the two terminals generate independent RV's $M_{\mathcal{X}}$ and $M_{\mathcal{Y}}$, respectively, and all further steps are deterministic. This does not restrict generality, because any randomized operations at either terminal (at any step) may be equivalently regarded as deterministic operations that depend also on an initially chosen random variable $M_{\mathcal{X}}$ or $M_{\mathcal{Y}}$, respectively.

Now the formal definition of a permissible secret sharing strategy for Model S is as follows.

Step 0 The terminals generate RV's $M_{\mathcal{X}}$ and $M_{\mathcal{Y}}$ such that $M_{\mathcal{X}}, M_{\mathcal{Y}}$ and (X^n, Y^n) are mutually independent.
Step 1 The two terminals exchange messages Φ_1, Ψ_1, over the public channel, where $\Phi_1 = \Phi_1(M_{\mathcal{X}}, X^n)$, $\Psi_1 = \Psi_1(M_{\mathcal{Y}}, Y^n)$.
Step i The two terminals exchange messages Φ_i, Ψ_i over the public channel, where $\Phi_i = \Phi_i(M_{\mathcal{X}}, X^n, \Psi^{i-1})$, $\Psi_i = \Psi_i(M_{\mathcal{Y}}, Y^n, \Phi^{i-1})$, (with the usual shorthand that upper index denotes a sequence up to that index).
Final Step (after k "exchange steps" have taken place): Both terminals compute what they deem to be the key established by the secret sharing process, as a function of the information available to them

$$K = K(M_{\mathcal{X}}, X^n, \Psi^k), \; L = L(M_{\mathcal{Y}}, Y^n, \Phi^k), \tag{1}$$

where K and L take values in the same finite set $\mathcal{K}$.

Of course, K and L must satisfy certain conditions in order that we can speak of a successful secret sharing. Before stating these (viz. equations (4) and (5) in Definition 127 below), first we define the permissible strategies for Model C. Here the situation is more complex because two channels are available for communication (the secure DMC, however, in one direction only) and these may be used in an interactive way.

In the following formal definition of a permissible secret sharing strategy for Model C we assume that the n symbols transmitted over the DMC are sent at instants $\ell, 2\ell, \ldots, n\ell$ (where ℓ is an unspecified integer), and at the other instants the public channel is used. This does not restrict generality, because any or all of the Φ_i and Ψ_i below may be set equal to the empty word.

Step 0 The terminals generate independent RV's $M_{\mathcal{X}}$ and $M_{\mathcal{Y}}$.
Step i, $\mathbf{0 < i < \ell}$ The two terminals exchange messages Φ_i, Ψ_i over the public channel, where $\Phi_i = \Phi_i(M_{\mathcal{X}}, \Psi^{i-1})$, $\Psi_i = \Psi_i(M_{\mathcal{Y}}, \Phi^{i-1})$.

Step $\boldsymbol{\ell}$ Terminal $\mathcal{X}$ determines the first input X_1 to the DMC, $X_1 = X_1(M_{\mathcal{X}}, \Psi^{\ell-1})$, and terminal $\mathcal{Y}$ observes the corresponding output Y_1. Φ_ℓ and Ψ_ℓ are set void.
Step $\mathbf{i}$, $(\mathbf{j}-\mathbf{1})\ell < \mathbf{i} < \mathbf{j}\ell, \mathbf{j} \leq \mathbf{n}$ The terminals exchange messages Φ_i, Ψ_i, where

$$\Phi_i = \Phi_i(M_{\mathcal{X}}, \Psi^{i-1}), \ \Psi_i = \Psi_i(M_{\mathcal{Y}}, Y^{j-1}, \Phi^{i-1}). \tag{2}$$

Step $\mathbf{i} = \mathbf{j}\ell, \mathbf{j} \leq \mathbf{n}$ Terminal $\mathcal{X}$ determines the jth input X_j to the DMC, $X_j = X_j(M_{\mathcal{X}}, \Psi^{j\ell-1})$ and terminal $\mathcal{Y}$ observes the corresponding output Y_j. $\Phi_{j\ell}$ and $\Psi_{j\ell}$ are set void.
Step $\mathbf{i}$, $\mathbf{n}\ell < \mathbf{i} \leq \mathbf{k}$ The terminals exchange messages Φ_i, Ψ_i as in (2), with Y^{j-1} replaced by Y^n in the definition of Ψ_i.
Final Step Same as in Model S (now in Eq. (1) actually $K = K(M_{\mathcal{X}}, \Psi^k)$, because X^n is uniquely determined by $M_{\mathcal{X}}$ and Ψ^k).

Notice that a strategy as above always determines X_j as a function of $M_{\mathcal{X}}, M_{\mathcal{Y}}$ and Y^{j-1}. The formal meaning of saying that Y_j is the DMC output corresponding to input X_j is

$$\Pr\left\{Y_j = y | M_{\mathcal{X}} = m, M_{\mathcal{Y}} = m', Y^{j-1} = y^{j-1}\right\} = W\left(y | X_j(m, m', y^{j-1})\right) \tag{3}$$

where $X_j(m, m', y^{j-1})$ denotes the input X_j determined by $M_{\mathcal{X}} = m$, $M_{\mathcal{Y}} = m'$, $Y^{j-1} = y^{j-1}$. It is easy to see that the functional relationships in the description of the strategy and Eq. (3) uniquely determine the joint distribution of all RV's involved (once the distributions of $M_{\mathcal{X}}$ and $M_{\mathcal{Y}}$ are specified) as it is necessary for mathematical consistency.

Definition 127 For Model S or C, a number H will be called an ε-achievable key rate if for every $\delta > 0$ and sufficiently large n there exists a strategy (permissible for the given model) such that K and L of (1) satisfy

$$\Pr\{K \neq L\} < \varepsilon \tag{4}$$

$$\frac{1}{n} I(\Phi^k, \Psi^k \wedge K) < \varepsilon \tag{5}$$

$$\frac{1}{n} H(K) > H - \delta \tag{6}$$

$$\frac{1}{n} \log |\mathcal{K}| < \frac{1}{n} H(K) + \varepsilon. \tag{7}$$

H is an *achievable key rate* if it is ε-achievable for every $\varepsilon > 0$, and the largest achievable key rate is the key-capacity. Further, we define a-weakly ε-achievable

key rates as above by replacing condition (7) by

$$\frac{1}{n}\log|\mathcal{K}| < a. \tag{8}$$

Then the *weakly achievable key rates* are those that, for some fixed a, are a-weakly ε-achievable for every $\varepsilon > 0$, and the largest weakly achievable key rate is the *weak key-capacity*.

Here condition (4) means that the two terminals have indeed generated a common key (with a small probability of error), and (5) means that this is indeed a secret key: the exchange over the public channel gave away effectively no information about it. Condition (7) means that the distribution of the key is "nearly uniform" in an entropy sense; this certainly appears desirable if the key is to be used for encryption, the most likely purpose of secret sharing. Below we will prove (Lemma 128) that condition (7) indeed ensures the suitability of the key for encryption. Nevertheless, it is a question of some mathematical interest whether the key rate $\frac{1}{n}H(K)$ could be increased by dropping the "uniformity condition" (7). Condition (8) has been imposed in order to exclude a trivial positive answer to that question. Indeed, if the "key spaces" $\mathcal{K}_n$ could grow faster than exponential with n, $\frac{1}{n}H(K)$ could be made arbitrarily large simply by uniformly distributing an arbitrarily small probability on a set of superexponential size.

To be rigorous, the obvious implication of Definition 127 that weak key-capacity is at least as large as key-capacity, requires a proof. Indeed, it is necessary to show that (7) (together with (4), (5) and the definition of permissible secret sharing strategies) implies that $\frac{1}{n}H(K)$ is bounded. A proof of this fact is contained in Proposition 129 below.

In the simple models treated in this section it will be easily shown that weak key capacity actually equals key capacity. We expect that the same holds also for the more complex models and all variants of the concept of key capacity treated in Sect. 3; indeed, this will be established in all cases when we can determine the key capacity. Still, no attempt will be made to prove a general theorem about this equality, because this technical problem does not appear to be of primary interest.

Now we show that if H is an achievable key rate in the sense of Definition 127 then, using the established key for encryption, secure transmission at rate H is possible over the public channel. For this, we set (without restricting generality) $\mathcal{K}_n = \{1, \ldots, N\}$ and consider the encryption of a random message $M \in \{1, \ldots, N\}$ (generated at terminal $\mathcal{X}$, say) simply as $M + K \pmod N$. If $M + K$ is sent over the public channel, terminal $\mathcal{Y}$ can decode M with small probability of error (by (4)), and the next lemma shows that a cryptonalist having access to the public transmissions only, gets effectively no information about M.

Lemma 128 *For a RV M with values in $\{1, \ldots, N\}$ independent of (Φ^k, Ψ^k, K), (5) and (7) imply that*

$$\frac{1}{n}I(\Phi^k, \Psi^k, M + K \wedge M) < 2\varepsilon.$$

Proof

$$I(\Phi^k, \Psi^k, M + K \wedge M) = I(M + K \wedge M | \Phi^k, \Psi^k) \tag{9}$$

$$= H(M + K | \Phi^k, \Psi^k) - H(M + K | M, \Phi^k, \Psi^k)$$

$$\leq \log N - H(K | M, \Phi^k, \Psi^k)$$

$$\leq H(K) + n\varepsilon - H(K | M, \Phi^k, \Psi^k) \tag{10}$$

$$= I(K \wedge M, \Phi^k, \Psi^k) + n\varepsilon$$

$$\leq 2n\varepsilon. \tag{11}$$

Here (9) holds because M is independent of (Φ^k, Ψ^k), (10) follows from (7) where now $\mathcal{K} = \{1, \ldots, N\}$, and (11) follows from (5) because M is independent of (Φ^k, Ψ^k, K). □

Remark The same simple proof shows that if $I(\Phi^k, \Psi^k \wedge K)$ were exactly 0 and the distribution of K exactly uniform then we would have $I(\Phi^k, \Psi^k, M + K \wedge M) = 0$, i.e., perfect secrecy.

In the simple models of this section this is indeed attainable, but in the more complex models of Sect. 3 one probably has to be satisfied with the almost complete secrecy of Lemma 128. The main result in this section is the following.

Proposition 129

(i) *For Model S, key capacity and weak key-capacity both equal the mutual information $I(X \wedge Y)$ and this is attainable by using a single forward (or backward) transmission only.*

(ii) *For Model C, key capacity and weak key-capacity both equal the ordinary capacity $C(W)$ of the DMC $\{W\}$, and this is attainable without using the public channel at all.*

Proof First we prove the direct assertion of (i), i.e. that $I(X \wedge Y)$ is an achievable key rate by using a single forward transmission. The idea is to transmit a code of X^n of rate $\approx H(X|Y)$ that, in the knowledge of Y^n, makes the reproduction of X^n possible (with small probability of error). A closer look at the proof of the Slepian-Wolf theorem ([6], pp. 238–239, Theorem 1.2) reveals that this can be done in such a way that the desired secret sharing results. □

For a formal proof, consider the DMC $\{W : \mathcal{X} \to \mathcal{Y}\}$ where $W = P_{Y|X}$, fix $\varepsilon > 0, \delta > 0, \eta > 0$, and pick consecutively disjoint codeword sets $\mathcal{C}_i$ of (n, ε) codes for this DMC, each consisting of codewords of the same type, and each of

size

$$M = \left\lceil \exp\{n\,(I(X \wedge Y) - \delta)\} \right\rceil. \tag{12}$$

If this process can not be continued after having picked $\mathcal{C}_N$, say, then necessarily

$$P_X^n\left(\bigcup_{i=1}^{N} \mathcal{C}_i\right) > 1 - \eta \tag{13}$$

(providing n is sufficiently large). Indeed, any subset A of $\mathcal{X}^n$ with $P_X^n(A) \geq \eta$ contains a codeword set C with the desired properties ([6], p. 107; there the constant type property was not required, but it can obviously be attained by looking at the largest subcode with codewords of constant type).

Now let terminal $\mathcal{X}$ transmit

$$\Phi(X^n) = \begin{cases} i & \text{if } X^n \in \mathcal{C}_i,\ 1 \leq i \leq N \\ 0 & \text{if } X^n \notin \bigcup_{i=1}^{N} \mathcal{C}_i\ . \end{cases}$$

Enumerate (in any way) the elements of each $\mathcal{C}_i$, and set $K = j$ if X^n equals the jth element of some $\mathcal{C}_i$. Terminal $\mathcal{Y}$, knowing y^n and $\Phi(X^n) = i$, can use the decoder of the channel code with codeword set $\mathcal{C}_i$; set $L = j$ if this decoding results in the jth element of $\mathcal{C}_i$. Then, since $W = P_{Y|X}$, and an (n, ε) code was used for the DMC $\{W\}$, we have

$$\Pr\left\{L \neq K | X^n \in \mathcal{C}_i\right\} < \varepsilon,\ i = 1, \ldots, N.$$

This and (13) imply that

$$\Pr\{K \neq L\} < \varepsilon + \eta, \tag{14}$$

no matter how K and L are defined when $X^n \notin \bigcup_{i=1}^{N} \mathcal{C}_i$. Further, since each set $\mathcal{C}_i$ consists of sequences of the same type, the conditional distribution of K on the condition $X^n \in \mathcal{C}_i$ is uniform on $\{1, \ldots, M\}$, for every $i = 1, \ldots, N$. For convenience, for $X^n \notin \bigcup_{i=1}^{N} \mathcal{C}_i$ we set K equal to a RV uniformly distributed on $\{1, \ldots, M\}$ and independent of X^n.

By this simple scheme, a key has been obtained, shared by both terminals (Eq. (14)), that is both uniformly distributed and independent in the exact sense of Φ transmitted over the public channel, and such that $\frac{1}{n}H(K)$ is arbitrarily close to $I(X \wedge Y)$ (Eq. (12)).

Having proved the direct assertion of (i) and the direct assertion of (ii) being obvious from the DMC coding theorem, it remains to prove the converses, i.e., that a (weakly) achievable key rate can not exceed $I(X \wedge Y)$ in Model S or $C(W)$ in Model C.

To this we send forward a simple lemma.

Lemma 130 *Let U and V be arbitrary RV's, and let $\Phi_1, \ldots, \Phi_k$, $\Psi_1, \ldots, \Psi_k$ be such that for every $i \leq k$, Φ_i is a function of U and Ψ^{i-1}, and Ψ_i is a function of V and Φ^{i-1}. Then*

$$I(U \wedge V|\Phi^k, \Psi^k) \leq I(U \wedge V).$$

Proof

$$\begin{aligned} I(U \wedge V|\Phi^k, \Psi^k) &= I(U \wedge V|\Phi^{k-1}, \Phi_k, \Psi^{k-1}, \Psi_k) \\ &\leq I(U, \Phi_k \wedge V|\Phi^{k-1}, \Psi^{k-1}, \Psi_k) \\ &\leq I(U, \Phi_k \wedge V, \Psi_k|\Phi^{k-1}, \Psi^{k-1}) \\ &= I(U \wedge V|\Phi^{k-1}, \Psi^{k-1}); \end{aligned}$$

here the last step follows from the assumption that Φ_k is a function of (U, Ψ^{k-1}) and Ψ_k is a function of (V, Φ^{k-1}). Repeating this argument k times, the Lemma follows. □

Returning to the proof of the converse assertions of (i) and (ii) of Proposition 129, consider any strategy (permissible for either Model S or Model C) with the properties (4), (5). Then

$$H(K) \leq I(K \wedge L) + \varepsilon \log |\mathcal{K}_n| + 1 \tag{15}$$

$$\leq I(K \wedge L, \Phi^k, \Psi^k) + \varepsilon \log |\mathcal{K}_n| + 1$$

$$\leq I(K \wedge L|\Phi^k, \Psi^k) + \varepsilon \log |\mathcal{K}_n| + 1 + \varepsilon n. \tag{16}$$

Here (15) follows from (4) by Fano's Lemma (Lemma 48) and (16) from (5).
Now, for Model S, we have

$$I(K \wedge L|\Phi^k, \Psi^k) \leq I(M_{\mathcal{X}}, X^n \wedge M_y, Y^n|\Phi^k, \Psi^k) \tag{17}$$

$$\leq I(M_{\mathcal{X}}, X^n \wedge M_y, Y^n) \tag{18}$$

$$= I(X^n \wedge Y^n) = nI(X \wedge Y). \tag{19}$$

Here (17) follows by (1), (18) from Lemma 130, and (19) from the independence of $M_{\mathcal{X}}, M_{\mathcal{Y}}, (X^n, Y^n)$.
Substituting (19) into (16), under condition (7) we obtain that

$$(1 - \varepsilon)H(K) < nI(X \wedge Y) + n\varepsilon^2 + n\varepsilon + 1$$

and under condition (8) that

$$H(K) < nI(X \wedge Y) + \varepsilon a + n\varepsilon + 1.$$

This completes the proof of the converse for Model S.
For Model C, we have

$$I(K \wedge L|\Phi^k, \Psi^k) \leq I(M_{\mathcal{X}} \wedge M_{\mathcal{Y}}, Y^n|\Phi^k, \Psi^k) \tag{20}$$

$$\leq I(M_{\mathcal{X}} \wedge M_{\mathcal{Y}}, Y^n) \tag{21}$$

$$\leq \sum_{j=1}^{n} I(M_{\mathcal{X}} \wedge Y_j|M_{\mathcal{Y}}, Y^{j-1}). \tag{22}$$

Here (20) follows because for Model C in Eq. (1) we have $K = K(M_{\mathcal{X}}, \Psi^k)$, (21) follows by Lemma 130, and (22) is the chain rule, taking into account that $I(M_{\mathcal{X}} \wedge M_{\mathcal{Y}}) = 0$.
But on account of (3) we have

$$I(M_{\mathcal{X}} \wedge Y_j|M_{\mathcal{Y}}, Y^{j-1}) = H(Y_j|M_{\mathcal{Y}}, Y^{j-1}) - H(Y_j|M_{\mathcal{X}}, M_{\mathcal{Y}}, Y^{j-1})$$

$$= H(Y_j|M_{\mathcal{Y}}, Y^{j-1}) - H(Y_j|X_j) \leq I(X_j \wedge Y_j).$$

Hence the right side of (22) is upper bounded by $nC(W)$. Returning to (16), the proof for Model C can be completed as that for Model S.

3 Secret Sharing When the Third Party Has Side Information

In this section, we consider generalizations of the simple models treated in Sect. 2 to the case when the third party should be kept ignorant of the result of secret sharing, to be called the wiretapper, has access to more information than what is transmitted over the public channel.

Model SW: (source-type model with wiretapper) We are given a DMMS with three component sources and generic variables (X, Y, Z). Terminal $\mathcal{X}$ "sees" the source outputs X^n, Terminal $\mathcal{Y}$ "sees" the source outputs Y^n, and the wiretapper "sees" the source outputs Z^n.

Model CW: (channel-type model with wiretapper) We are given a DMC $\{W : \mathcal{X} \to \mathcal{Y} \times \mathcal{Z}\}$. Terminal $\mathcal{X}$ governs the input, Terminal $\mathcal{Y}$ "sees" the Y-outputs whereas the wiretapper "sees" the Z-outputs.

In both cases also a noiseless public channel of unlimited capacity is available for communication between Terminal $\mathcal{X}$ and $\mathcal{Y}$; communication over this channel is completely known to the wiretapper.

The permissible strategies for Models SW and CW are the same as for Model S and C in Sect. 2, with two formal modifications: For Model SW in Step 0 we have to postulate that $M_{\mathcal{X}}, M_{\mathcal{Y}}, (X^n, Y^n, Z^n)$ are mutually independent, and in Model CW it has to be taken into account that every DMC input X_j generates a pair of outputs Y_j, Z_j; the formal way of doing this is to replace Eq. (3) by

$$\Pr\left\{Y_j = y, Z_j = z | M_{\mathcal{X}} = m, M_{\mathcal{Y}} = m', Y^{j-1} = y^{j-1}, Z^{j-1} = y^{j-1}, Z^{j-1} = z^{j-1}\right\}$$

$$= W\left(y, z | X_j(m, m', y^{j-1})\right). \tag{23}$$

Definition 127 applies also to Model SW and CW, with the single change that condition (5) has to be replaced by

$$\frac{1}{n} I(\Phi^k, \Psi^k, Z^n \wedge K) < \varepsilon. \tag{24}$$

In order to deal more systematically with the question whether simple strategies suffice also in Models SW and CW to achieve the key-capacity, we will consider some variants of the concept of key-capacity obtained by restricting the class of permissible secret sharing strategies.

One possible restriction would be that not more than k exchanges are permitted over the public channel; the analogue of key-capacity under this restriction might be called k-key-capacity. In this chapter only the case $k = 1$ will be considered, moreover, the restriction will be made that only a forward or only a backward transmission is permitted (formally, all Φ_i and Ψ_i in the description of a permissible strategy in Sect. 2 equal the empty word, except for one Φ_i or Ψ_i; recall that "forward" means the direction $\mathcal{X} \to \mathcal{Y}$ and "backward" the direction $\mathcal{Y} \to \mathcal{X}$). Thus, for both models, SW and CW, we define the *forward key-capacity* and *backward key-capacity*, as well as their weak versions, analogously to the general definition of key-capacity (weak key-capacity) but by permitting the use of the public channel for a single forward transmission or a single backward transmission only. For Model SW these two notions are completely symmetric but for Model CW they differ essentially.

By Proposition 129, for Models S and C both the forward and backward key-capacities equal the key-capacity. We will see that for Models SW and CW this is no longer true, in general. It remains, however, open whether for either model key-capacity can ever be larger than what by the previous paragraph would be termed the 1-key-capacity.

Before stating our main results, let us briefly review previous literature related to our subject. The model "wiretap channel" introduced by Wyner [13] and generalized by Csiszár and Körner [5] (cf. also [6], p. 407) can be described as follows:

Given $W : \mathcal{X} \to \mathcal{Y} \times \mathcal{Z}$, the sender is required to encode a RV M, uniformly distributed over a possibly large set, into a channel input X^n so that M be decodable (with small probability of error) from the received sequence Y^n whereas the other output sequence Z^n should give negligibly small information about M. The supremum of the rates $\frac{1}{n}H(M)$ subject to these conditions is called the *secrecy capacity*. Since this coding problem depends on W only through its marginal channels $W_1 : \mathcal{X} \to \mathcal{Y}$ and $W_2 : \mathcal{X} \to \mathcal{Z}$, it is often stated – as in [5] – in terms of these two channels. Wyner considered and solved the case where W_2 is a degraded version of W_1, that is $W_2(Z|X) = \sum_Y W_1(Y|X)V(Z|Y)$ and Csiszár and Körner gave a single-letter characterization of secrecy capacity in the general case. Clearly, any code in the definition of secrecy capacity represents a secret sharing strategy for our Model CW, that does not use the public channel at all, but has the properties required in Definition 127. Hence both the forward and backward key-capacity for Model CW must be at least as large as the wiretap secrecy capacity. Not unexpectedly, we will see that the forward key-capacity for Model CW is actually equal to the corresponding wiretap secrecy capacity.

The general problem of secret sharing does not seem to have been considered before in an information theoretic context, but an important step in this direction was made by Maurer [9]. He considered a wiretap channel whose marginal channels W_1 and W_2 are both binary symmetric, and have "independent noise" which, in our terminology, means

$$W(y, z|x) = W_1(y|x)W_2(z|x). \tag{25}$$

For this case, Maurer proposed a scheme in which the sender transmits a $\frac{1}{2} - \frac{1}{2}$ i.i.d. sequence and the *receiver* can send back information, using a public channel, in such a way that the original sender can decode this information but the wiretapper remains in complete ignorance. This makes a key exchange at a positive rate possible even in those cases when the wiretap secrecy capacity is equal to zero. Clearly, in our terminology, the key rate achieved in this way is a lower bound to backward key-capacity and hence also to key-capacity. Our results will imply that the key rate achieved by Maurer's scheme is actually equal to the key-capacity, and thus optimal for his model in our (more demanding) sense.

Now we state our main results. Though not stated explicitly, all assertions are true also for the weak version of the corresponding key-capacities. The theorems stated in this section will be proved in Sect. 4.

Theorem 131 *For Model SW, the forward key-capacity equals the maximum of*

$$I(T \wedge Y|U) - I(T \wedge Z|U), \tag{26}$$

for all pairs of RV's T, U (taking values in a sufficiently large finite set) that satisfy the Markov condition

$$U \ominus T \ominus X \ominus YZ. \tag{27}$$

Further, the key-capacity is upper bounded by $I(X \wedge Y|Z)$, and this bound is tight if X, Y, Z form a Markov chain in any order.

Corollary 132 *If in the Model SW $X \ominus Y \ominus Z$, then forward key-capacity and key-capacity both equal $I(X \wedge Y) - I(X \wedge Z)$.*

Theorem 133 *For Model CW, the forward key-capacity is equal to the secrecy capacity of the corresponding wiretap channel, namely to the maximum of the expression (26) for all quadruples of RV's T, U, Y, Z to which there exists an X satisfying the Markov condition (27) and $P_{YZ|X} = W$. Further, the key-capacity is upper bounded by $\max_{P_{YZ|X}=W} I(X \wedge Y|Z)$, and this bound is tight if W has the property that $P_{YZ|X} = W$ implies that X, Y, Z from a Markov chain in some order.*

Corollary 134

(i) *If W has the form*

$$W(y, z|x) = W_1(y|x)V(z|y) \tag{28}$$

then forward key-capacity and key-capacity both equal $\max_{P_{YZ|X}=W}[I(X \wedge Y) - I(X \wedge Z)]$, and can be attained without any use of the public channel.

(ii) *If W is of form (25) then backward key-capacity and key-capacity both equal $\max_{P_{YZ|X}=W}[I(X \wedge Y) - I(Y \wedge Z)]$, and this is in general larger than forward key-capacity.*

Remark The cardinalities of the ranges of the auxiliary random variables T and U could be bounded in a standard way, using the Support Lemma ([6], P. 310).

Remark The upper bounds in Theorems 131 and 133 on the key-capacity may sometimes be poor, e.g., they may be larger than the key-capacity in the absence of the wiretapper, determined in Sect. 2. For Model SW, the bound may be improved by the simple observation that for any RV V satisfying the Markov condition $XY \ominus Z \ominus V$, the key-capacity for the DMMS with generic variables (X, Y, V) is at least as large as for that with (X, Y, Z). Hence the result of Theorem 131 implies that the minimum of $I(X \wedge Y|V)$ subject to $XY \ominus Z \ominus V$ is also an upper bound to key-capacity in Model SW. The bound for Model CW given in Theorem 133 could be improved in a similar way. Still, we have no reason to believe that even these improved bounds are tight.

Although the upper bounds on key-capacity for Models SW and CW given in Theorems 131 and 133 are not tight, in general, there is a natural modification of these models for which the above bounds actually give the exact answer. This modification consists in the assumption that terminal $\mathcal{X}$ (or terminal $\mathcal{Y}$) has access

to the wiretapper's side information, i.e., the source resp. channel output sequence Z^n is available to terminal $\mathcal{X}$ (or terminal $\mathcal{Y}$).

In this modification of Model SW or CW, for which we prefer not to introduce a new notation, the permissible strategies will differ from those of Model SW or CW only in the obvious way: the operations at that terminal where Z^n is available, may depend also on Z^n (or, in the channel-type model, on that part Z^j of Z^n that is already available).

It appears safe to save space by omitting formal definitions for the next statement.

Theorem 135 *If Model SW or CW is modified by letting either terminal $\mathcal{X}$ of terminal $\mathcal{Y}$ know the Z-outputs, the key-capacity for the source-type model will always equal $I(X \wedge Y|Z)$ and for the channel-type model $\max_{P_{YZ|X=W}} I(X \wedge Y|Z)$. Further, this also equals the backward or forward key-capacity, respectively, according as terminal $\mathcal{X}$ or terminal $\mathcal{Y}$ is informed.*

At first sight, the result of Theorem 135 appears counter-intuitive, because it means that in some cases we can do better with a "known wiretapper" than if there were no wiretapper at all. The answer is, of course, that access to the wiretapper's information does contribute to generating common randomness (what secret sharing is all about) and this benefit can more than balance out the negative effect that the wiretapper must be kept ignorant of this common randomness.

It may be instructive to consider the following example.

Example Let the DMMS with generic variables (X, Y, Z) be as follows: let X and Y be independent $\frac{1}{2} - \frac{1}{2}$ binary RV's, and $Z = X + Y \mod 2$. Then $I(X \wedge Y) = 0$, $I(X \wedge Y|Z) = 1$. Clearly, if the terminals $\mathcal{X}$ and $\mathcal{Y}$ have access to X^n and Y^n only, no secret sharing between them is possible. However, if terminal $\mathcal{X}$, say, knows also Z^n then he can compute Y^n. Then with $K = Y^n$, secret sharing with key rate equal to 1 has taken place; the wiretapper remains completely ignorant because $I(Z^n \wedge Y^n) = 0$. ▲

4 Proofs

Proof of Theorem 131 The direct part of the first assertion, i.e., the achievability of key rate (26) with a single forward transmission, can be proved by essentially the same idea as for the direct part of Proposition 129, part (i). However, whereas there $\mathcal{X}^n$ was partitioned (up to a small probability set) into codeword sets of DMC codes, in the present more complex case we need a partitioning into codeword sets of wiretap channel codes. The proof relies upon a basic construction in multi-user information theory and its details are rather technical.

The proof of the converse is very similar to the wiretap channel converse proof of Csiszár and Körner [5]. Since it is instructive and quite simple, we give the proof here. The key is an algebraic identity for information measures, namely the following lemma. □

Lemma 136 *For arbitrary RV's U, V and sequences of RV's Y^n, Z^n, we have*

$$I(U \wedge Y^n|V) - I(U \wedge Z^n|V)$$

$$= \sum_{i=1}^{n} \Big[I(U \wedge Y_i|Y^{i-1}Z_{i+1}\dots Z_n V) - I(U \wedge Z_i|Y^{i-1}Z_{i+1}\dots Z_n V) \Big]$$

Proof The ith term of the sum equals

$$H(U|Y^{i-1}Z_{i+1}\dots Z_n V) - H(U|Y^i Z_{i+1}\dots Z_n V)$$
$$- H(U|Y^{i-1}Z_{i+1}\dots Z_n V) + H(U|Y^{i-1}Z_i Z_{i+1}\dots Z_n V)$$
$$= H(U|Y^{i-1}Z_i\dots Z_n V) - H(U|Y^i Z_{i+1} Z_n V).$$

Summing these, after cancellations the result is

$$H(U|Z^n V) - H(U|Y^n V).$$

On the other hand,

$$I(U \wedge Y^n|V) - I(U \wedge Z^n|V)$$
$$= H(U|V) - H(U|Y^n V) - H(U|V) + H(U|Z^n V)$$
$$= H(U|Z^n V) - H(U|Y^n V). \qquad \square$$

Continuing the proof of Theorem 131, consider any secret sharing strategy for Model SW that enters the definition of weak forward key-capacity, i.e., that involves a single forward transmission, say $\Phi = \Phi(M_{\mathcal{X}}, X^n)$, and satisfies conditions (4), (24), (8). Since now $(\Phi^k, \Psi^k) = \Phi$, (23) becomes

$$K = K(M_{\mathcal{X}}, X^n), \quad L = L(M_{\mathcal{Y}}, Y^n, \Phi) \tag{29}$$

and the conditions (4), (24), (8) are

$$\Pr\{K \neq L\} < \varepsilon \tag{30}$$

$$\frac{1}{n} I(\Phi, Z^n \wedge K) < \varepsilon \tag{31}$$

$$\frac{1}{n} \log |\mathcal{K}| < a \tag{32}$$

(30) and (32) give by Fano's inequality (Lemma 48) that

$$H(K) \leq I(K \wedge L) + na\varepsilon + 1 \tag{33}$$

Now

$$\begin{aligned} I(K \wedge L) &\leq I(K \wedge M_{\mathcal{Y}}, Y^n, \Phi) && (34)\\ &= (K \wedge Y^n, \Phi) && (35)\\ &\leq I(K \wedge Y^n, \Phi) - I(K \wedge Z^n, \Phi) + n\varepsilon && (36)\\ &= I(K \wedge Y^n|\Phi) - I(K \wedge Z^n|\Phi) + n\varepsilon \\ &= \sum_{i=1}^{n} \big[I(K \wedge Y_i|Y^{i-1} Z_{i+1} \dots Z_n \Phi) && (37)\\ &\qquad - I(K \wedge Z_i|Y^{i-1} Z_{i+1} \dots Z_n \Phi) \big] + n\varepsilon && (38) \end{aligned}$$

here (34) holds by (29), (35) because of the independence of $M_{\mathcal{Y}}$ from (K, Y^n, Φ) implied by the mutual independence of $M_{\mathcal{X}}, M_{\mathcal{Y}}, (X^n, Y^n)$, (36) is from (31) and (37) is by Lemma 136.

The last sum can be written, in the usual way, as

$$I(K \wedge Y_J|U) - I(K \wedge Z_J|U)$$

where J is a RV independent of all the previous ones and uniformly distributed on $\{1, \dots, n\}$, and $U = Y^{J-1} Z_{J+1} \dots Z_n \Phi J$, and the difference of the conditional information quantities is the same as

$$I(T \wedge Y_J|U) - I(T \wedge Z_J|U).$$

Thus (34) becomes

$$I(K \wedge L) \leq I(T \wedge Y_J|U) - I(T \wedge Z_J|U) + \varepsilon \tag{39}$$

It is clear from the definitions of J, U and T – using also that $\Phi = \Phi(M_{\mathcal{X}}, X^n)$ – that the Markov property $U \ominus T \ominus X_J \ominus X_J Z_J$ holds and that the joint distribution of $X_J Y_J Z_J$ is the same as that of XYZ. Hence, substitution of (39) into (33) proves that the weak forward key-capacity can not be larger than the supremum of all expressions of form (26), with the Markov condition (27).

It can be shown by a standard argument (omitted here) that the set of expressions of this form does not change if the ranges of T and U are restricted to sets of (sufficiently large) fixed cardinality. This implies that this set is closed, and the previous supremum is indeed a maximum.

The upper bound on (weak) key-capacity stated in Theorem 131 follows by the simple argument in the proof of Proposition 129. Namely, similarly as (16) followed from (4), (5), we obtain from (4) and (24) that

$$H(K) \leq I(K \wedge L|\Phi^k, \Psi^k, Z^n) + \varepsilon \log |\mathcal{K}_n| + 1. \tag{40}$$

Further, to the exact analogy of the derivation of (19) we obtain that

$$I(K \wedge L|\Phi^k, \Psi^k, Z^n) \leq I(X^n \wedge Y^n|Z^n) = I(X \wedge Y|Z) \tag{41}$$

(a minor difference is that Lemma 130 has to be used in a "conditional" version; but clearly that Lemma remains valid if a conditioning RV is added on both sides).

Combining (40) and (41) shows that $I(X \wedge Y|Z)$ is an upper bound to (weak) key-capacity.

$I(X \wedge Y|Z)$ is equal to 0 and hence gives a tight bound if $X \ominus Z \ominus Y$. Suppose next that $X \ominus Y \ominus Z$. Then the forward key-capacity can evaluated as follows. From (27) we now get $U \ominus T \ominus X \ominus Y \ominus Z$ and therefore

$$\begin{aligned} &I(T \wedge Y|U) - I(T \wedge Z|U) = I(T \wedge YZ|U) - I(T \wedge Z|U) \\ &= I(T \wedge Y|ZU) \leq I(TU \wedge Y|Z) \\ &= I(T \wedge Y|Z) \leq I(X \wedge Y|Z) = I(X \wedge Y) - I(X \wedge Z). \end{aligned}$$

Since $I(X \wedge Y) - I(X \wedge Z)$ is also an expression of form (26), this shows that in this case the forward key-capacity equals $I(X \wedge Y) - I(X \wedge Z) = I(X \wedge Y|Z)$. As key-capacity could only be larger than forward key-capacity, this shows that the upper bound $I(X \wedge Y|Z)$ is tight in this case. Finally, the third possible Markovity $X \ominus X \ominus Z$ is not a new case, by symmetry.

The Corollary has already been proved.

Proof of Theorem 133 It is obvious that the forward key-capacity is at least as large as the wiretap channel secrecy capacity. To show that it can not be larger, we use again the method of Csiszár and Körner [5], as in the proof of Theorem 131.

Consider any secret sharing strategy for Model CW that enters the definition of weak forward key-capacity. Then, since there are now backward transmissions, X^n has to be a function of $M_{\mathcal{X}}$ alone, and so have also Φ (the forward transmission over the public channel) and K:

$$X^n = X^n(M_{\mathcal{X}}),\ \Phi = \Phi(M_{\mathcal{X}}),\ K = K(M_{\mathcal{X}}). \tag{42}$$

Unlike in the proof of Theorem 131, Y^n and Z^n are now the channel outputs corresponding to input X^n, but (33) and (38) still hold as there (for step (ii) of the derivation of (38) we have to make sure that $M_{\mathcal{Y}}$ is independent of (K, Y^n, Φ), but this is intuitively obvious from (42), and formally follows from (23) where now

$X_j(m, m', y^{j-1}) = X_j(m)$ by (42)). Also the rewriting of (38) as (39) works as there, and even though now the joint distribution of the resulting RV's $X_J Y_J Z_J$, the Markov condition $U \ominus T \ominus X_J \ominus Y_J Z_J$ still holds, and it follows that the weak forward key-capacity can not be larger than the supremum of all expressions of form (26) subject to the conditions stated in Theorem 133.

Now we turn to the proof of the upper bound on (weak) key-capacity for Model CW. This is more difficult than the proof of the similar bound for Model SW in Theorem 131, or for Model C in Proposition 129.

Consider any permissible strategy, as described in Sect. 2 for Model C; in particular $K = K(M_{\mathcal{X}}, \Psi^k)$, $L = L(M_{\mathcal{Y}}, Y^n, \Phi^k)$. Since (40) holds also in the present case and by the last functional relationships

$$I(K \wedge L|\Phi^k, \Psi^k, Z^n) \le I(M_{\mathcal{X}} \wedge M_{\mathcal{Y}}, Y^n|\Phi^k, \Psi^k, Z^n) \tag{43}$$

we have to bound the right-hand side of (44). We proceed as follows:

$$I(M_{\mathcal{X}} \wedge M_{\mathcal{Y}}, Y^n|\Phi^k, \Psi^k, Z^n) = I(M_{\mathcal{X}} \wedge M_{\mathcal{Y}} Y^n Z^n, \Phi^k \Psi^k) - I(M_{\mathcal{X}} \wedge Z^n \Phi^k \Psi^k) \tag{44}$$

where by the chain rule

$$I(M_{\mathcal{X}} \wedge M_{\mathcal{Y}} Y^n Z^n \Phi^k \Psi^k) = I(M_{\mathcal{X}} \wedge M_{\mathcal{Y}} \Phi^{\ell-1} \Psi^{\ell-1}) + \sum_{j=1}^{n} (F_j + G_j), \tag{45}$$

$$F_j = I(M_{\mathcal{X}} \wedge Y_j Z_j | M_{\mathcal{Y}} Y^{j-1} Z^{j-1} \Phi^{j\ell-1} \Psi^{j\ell-1}) \tag{46}$$

$$G_j = I(M_{\mathcal{X}} \wedge \Phi_{j\ell+1} \dots \Phi_{(j+1)\ell-1} \Psi_{j\ell+1} \dots \Psi_{(j+1)\ell-1} | M_{\mathcal{Y}} Y^j Z^j \Phi^{j\ell-1} \Psi^{j\ell-1}), \tag{47}$$

with an obvious modification of the definition of G_j for $j = n$ (the last index is k rather than $(n+1)\ell - 1$).

Similarly,

$$I(M_{\mathcal{X}} \wedge Z^n \Phi^k \Psi^k) = I(M_{\mathcal{X}} \wedge \Phi^{\ell-1} \Psi^{\ell-1}) + \sum_{j=1}^{n} (F'_j + G'_j), \tag{48}$$

with

$$F'_j = I(M_{\mathcal{X}} \wedge Z_j | Z^{j-1} \Phi^{\ell j-1} \Psi^{\ell j-1}) \tag{49}$$

$$G'_j = I(M_{\mathcal{X}} \wedge \Phi_{j\ell+1} \dots \Phi_{(j+1)\ell-1} \Psi_{j\ell+1} \dots \Psi_{(j+1)\ell-1} | Z^j \Phi^{\ell j-1} \Psi^{\ell j-1}), \tag{50}$$

where, again, G'_n is defined with the obvious modification as above.

Now

$$F_j = H(Y_jZ_j|M_{\mathcal{Y}}Y^{j-1}Z^{j-1}\Phi^{\ell j-1}\Psi^{\ell j-1}) - H(Y_jZ_j|M_{\mathcal{X}}M_{\mathcal{Y}}Y^{j-1}Z^{j-1}\Phi^{\ell j-1}\Psi^{\ell j-1})$$

$$\leq H(Y_jZ_j|Z^{j-1}\Phi^{\ell j-1}\Psi^{\ell j-1}) - H(Y_jZ_j|M_{\mathcal{X}}Z^{j-1}\Phi^{\ell j-1}\Psi^{\ell j-1})$$

$$= I(M_{\mathcal{X}} \wedge Y_jZ_j|Z^{j-1}\Phi^{\ell j-1}\Psi^{\ell j-1}), \tag{51}$$

where, in the second step, we have used that

$$H(Y_jZ_j|M_{\mathcal{X}}M_{\mathcal{Y}}Y^{j-1}Z^{j-1}\Phi^{\ell j-1}\Psi^{\ell j-1}) = H(Y_jZ_j|M_{\mathcal{X}}Z^{j-1}\Phi^{\ell j-1}\Psi^{\ell j-1}).$$

This identity follows from the fact that on account of (23), the conditional distribution of Y_jZ_j on either condition is the same as on the condition $X_j = X_j(M_{\mathcal{X}}, \Psi^{\ell j-1})$ alone.

From (49) and (51)

$$F_j - F'_j \leq I\left(M_{\mathcal{X}} \wedge Y_j|Z_jZ^{j-1}\Phi^{\ell j-1}\Psi^{\ell j-1}\right)$$

$$= H\left(Y_j|Z_jZ^{j-1}\Phi^{\ell j-1}\Psi^{\ell j-1}\right) - H\left(Y_j|M_{\mathcal{X}}Z_jZ^{j-1}\Phi^{\ell j-1}\Psi^{\ell j-1}\right)$$

$$\leq H(Y_j|Z_j) - H(Y_j|X_jZ_j) = I(X_j \wedge Y_j|Z_j).$$

Here, in the second step, we have used (23) as above.

Next we compare the terms G_j and G'_j. Equation (47) can be equivalently written as

$$G_j = I\left(M_{\mathcal{X}} \wedge M_{\mathcal{Y}}Y^j\Phi_{j\ell+1}\dots\Phi_{(j+1)\ell-1}\Psi_{j\ell+1}\dots\Psi_{(j+1)\ell-1}|Z^j\Phi^{j\ell-1}\Psi^{j\ell-1}\right)$$

$$- I(M_{\mathcal{X}} \wedge M_{\mathcal{Y}}Y^j|Z^j\Phi^{(j+1)\ell-1}\Psi^{(j+1)\ell-1}),$$

and hence

$$G_j - G'_j = I\left(M_{\mathcal{X}} \wedge M_{\mathcal{Y}}Y^j|Z^j\Phi^{j\ell-1}\Psi^{j\ell-1}\right)$$

$$- I(M_{\mathcal{X}} \wedge M_{\mathcal{Y}}Y^j|Z^j\Phi^{(j+1)\ell-1}\Psi^{(j+1)\ell-1}).$$

On account of Lemma 130 (conditional version), this shows that $G_j - G'_j \leq 0$.

It follows again by Lemma 130 that

$$I\left(M_{\mathcal{X}} \wedge M_{\mathcal{Y}}\Phi^{\ell-1}\Psi^{\ell-1}\right) - I\left(M_{\mathcal{X}} \wedge \Phi^{\ell-1}\Psi^{\ell-1}\right)$$

$$= I\left(M_{\mathcal{X}} \wedge M_{\mathcal{Y}}|\Phi^{\ell-1}\Psi^{\ell-1}\right) \leq I(M_{\mathcal{X}} \wedge M_{\mathcal{Y}}) = 0$$

(thus this difference actually equals 0). Thus, finally, from (44), (45), (48), (51) we obtain that

$$I\left(M_{\mathcal{X}} \wedge M_{\mathcal{Y}} Y^n | \Phi^k \Psi^k Z^n\right) \leq \sum_{i=1}^{n} I(X_i \wedge Y_i | Z_i).$$

Returning to (43) and (40), this completes the proof of our upper bound on (weak) key-capacity for Model CW.

To show that our bound on the weak key-capacity is tight if X, Y, Z form a Markov chain in some order whenever $P_{YZ|X} = W$, notice first that the bound gives 0 and hence it is automatically tight, if this order is $X \ominus Z \ominus Y$ Next consider the Markov chain is $X \ominus Y \ominus Z$. Then $I(X \wedge Y|Z) = I(X \wedge Y) - I(X \wedge Z)$, and by the wiretap channel coding theorem this is an achievable key rate without any use of the public channel; in particular our bound is tight in this case, too. Finally, to show that $I(X \wedge Y|Z)$ is an achievable key rate, when $Y \ominus X \ominus Z$, consider any strategy where Terminal $\mathcal{X}$ sends an i.i.d. random sequence with distribution P_X. Then X^n, Y^n, Z^n represent the outcomes of a DMMS with component variables X, Y, Z and by Theorem 131 it follows that $I(X \wedge Y|Z)$ is an achievable key rate (even by using a single backward transmission only).

This completes the proof of Theorem 133, and in the course of the proof the Corollary was established, too. □

Proof of Theorem 135 The formal meaning of the assumption that the Z-outputs are known at Terminal $\mathcal{X}$ is that in the description of a permissible secret sharing strategy, the condition $\Phi_i = \Phi_i(M_{\mathcal{X}}, X^n, \Psi^{i-1})$ is replaced by $\Phi_i(M_{\mathcal{X}}, X^n, Z^n)$ in the source-type model or $\Phi_i = \Phi_i(M_{\mathcal{X}}, \Psi^{i-1})$ is replaced by $\Phi_i = \Phi_i(M_{\mathcal{X}}, \Psi^{i-1}, Z^{i-1})$ in the channel-type model (or a similar modification of Ψ_i, respectively, if the *Z-outputs are known at Terminal $\mathcal{Y}$*), and that also (1) is modified accordingly. The same proof by which we proved that $I(X \wedge Y|Z)$ resp. $\max_{P_{YZ|X=W}} I(X \wedge Y|Z)$ is an upper bound to the (weak) key-capacity for Models SW and CW, applies also in the present more general case, and shows that they are still upper bounds when the Z-outputs are known at Terminal $\mathcal{X}$ (or Terminal $\mathcal{Y}$).

If the Z-outputs are known at Terminal $\mathcal{Y}$, we can replace the Y-outputs Y_i by the pairs $Y_i Z_i$ without any loss of generality. But these "new Y-outputs" satisfy the Markov condition $X \ominus YZ \ominus Z$. Hence, by Theorems 131 and 133, in this case we obtain that the bound $I(X \wedge YZ|Z) = I(X \wedge Y|Z)$ (for the source-type model) resp. the corresponding maximum (for the channel-type model) is tight.

On the other hand, if the Z-outputs are known at Terminal $\mathcal{X}$, in the source type model we have the same situation as before. For the channel-type model, we can let Terminal $\mathcal{X}$ send an i.i.d. random sequence. Then we are in the situation of the source-type model, and it follows that $I(XZ \wedge Y|Z) = I(X \wedge Y|Z)$ is an achievable key-rate whenever $P_{XZ|X} = W$. □

5 Conclusions

We have considered various models of generating common randomness at two distant terminals $\mathcal{X}$ and $\mathcal{Y}$, with the additional requirement that a third party, the wiretapper $\mathcal{Z}$, be kept ignorant of the generated common randomness. Then the latter could be used as an encryption key to make communication between $\mathcal{X}$ and $\mathcal{Y}$ secure from $\mathcal{Z}$. For some models of generating this common randomness or key, we were able to determine the largest achievable key rate, called the key-capacity. For other models we gave bounds on the key-capacity.

The problems can be studied for all multi-way channels and multi-terminal sources. One can conceive even situations with several wiretappers.

The mathematical tools used in this chapter were those of multi-user information theory, in particular the single-letterization technique developed by Csiszár and Körner [9] for the wiretap channel and its generalization called broadcast channel with confidential messages. Still, it should be emphasized that there is a conceptual difference between the wiretap channel problem of transmitting messages from $\mathcal{X}$ to $\mathcal{Y}$ without giving information about them to $\mathcal{Z}$, and the problem of generating common randomness shared by $\mathcal{X}$ and $\mathcal{Y}$, secret from $\mathcal{Z}$. Notice that this common randomness need not be generated at $\mathcal{X}$ and communicated to $\mathcal{Y}$, it may as well be generated at $\mathcal{Y}$ and communicated to $\mathcal{X}$, or cooperatively generated by $\mathcal{X}$ and $\mathcal{Y}$.

Following the suggestion of a reviewer, we now summarize our main results and mention some of the open problems.

Our models were of two main types. In Model SW (source-type model with wiretapper), a discrete memoryless multiple source with generic variables X, Y, Z was given, and $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ "could see" the length $-n$ outputs X^n, Y^n, Z^n, respectively. In Model CW (channel-type model with wiretapper), a discrete memoryless channel with one input and two outputs was given, $\mathcal{X}$ governed the input, and the outputs were seen by $\mathcal{Y}$ and $\mathcal{Z}$, respectively. Both models involved the availability of a noiseless public channel of unlimited capacity for communication between $\mathcal{X}$ and $\mathcal{Y}$. As to the permitted use of the public channel, we focused mainly on the extreme cases:

(i) A single transmission from $\mathcal{X}$ to $\mathcal{Y}$ or from $\mathcal{Y}$ to $\mathcal{X}$; the corresponding key-capacities were called the forward key-capacity and the backward key-capacity, respectively.
(ii) As many exchanges between $\mathcal{X}$ and $\mathcal{Y}$ as desired; the term "key-capacity", without qualification, has been used to refer to this case of unlimited conversation.

For Model SW, we gave a single-letter characterization of forward key-capacity, by symmetry, this provided a characterization of backward key-capacity, too. The key-capacity with unlimited conversation could not be determined in general, but it was always upper bounded by $I(X \wedge Y|Z)$. If X, Y, Z formed a Markov chain in some order, that bound was tight, and key-capacity with unlimited conversation was equal to the forward or backward key-capacity. In general, two-way communication

over the public channel could increase the key-capacity above both forward and backward key-capacity, even if only one exchange of messages was permitted. In our example demonstrating this, the key-capacity for one exchange of messages was the same as for unlimited conversation. We do not expect this to be always so, but our results do not rule out that contingency.

For Model CW, it may be possible for $\mathcal{X}$ and $\mathcal{Y}$ to share common randomness secret from $\mathcal{Z}$ without using the public channel: this is the wiretap channel situation when our key-capacity reduces to the wiretap channel secrecy capacity. Using the public channel from $\mathcal{X}$ to $\mathcal{Y}$ does not help: we have shown that the forward key-capacity for Model CW equals the wiretap channel secrecy capacity, determined in [9]. A single-letter characterization of backward key-capacity, as well as of key-capacity with unlimited conversation, remains elusive for Model CW. Still, the maximum of $I(X \wedge Y|Z)$ – where (Y, Z) is the pair of outputs for input X – was shown to be an upper bound to key-capacity with unlimited conversation. This bound is tight in two important special cases, viz. for channels of form (28) or (25), and in those cases key-capacity with unlimited conversation equals the forward or backward key-capacity, respectively. In the cases we could determine key-capacity with unlimited conversation for Model CW, it could be achieved with $\mathcal{X}$ producing i.i.d. channel inputs. It remains open whether this is true in general.

Finally, if the information available to $\mathcal{Z}$ was made available to $\mathcal{X}$ and/or $\mathcal{Y}$, the key-capacity with unlimited conversation for this modified model (of either type) was shown to always equal the upper bound obtained before, and also to equal the forward or backward key-capacity for the modified model. This appears to be the first coding theorem that provides a direct operational characterization of conditional mutual information.

References

1. R. Ahlswede, Elimination of correlation in random codes for arbitrarily varying channels. Zeitschrift Wahrscheinlichkeitstheorie und verw. Geb. **33**, 159–175 (1978)
2. R. Ahlswede, I. Csiszár, Common randomness in information theory and cryptography, Part I: Secret sharing. IEEE Trans. Inf. Theory **39**, 1121–1132 (1993)
3. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**, 15–29 (1989)
4. R. Ahlswede, G. Dueck, Identification in the presence of feedback – a discovery of new capacity formulas. IEEE Trans. Inf. Theory **35**, 30–39 (1989)
5. I. Csiszár, J. Körner, Broadcast channels with confidential messages. IEEE Trans. Inf. Theory **24**, 339–348 (1978)
6. I. Csiszár, J. K"orner, *Information Theory: Coding Theorem for Discrete Memoryless Systems* (Academic, New York, 1982)
7. I. Csiszár, P. Narayan, The capacity of the arbitrarily varying channel revisited: positivity, constraints. IEEE Trans. Inf. Theory **34**, 181–193 (1988)
8. W. Diffie, M.E. Hellman, New directions in cryptography. IEEE Trans. Inf. Theory **IT–22**, 644–654 (1976)
9. U.M. Maurer, Provably–secure key distribution based on independent channels. Presented at the 1990 IEEE Workshop on Information Theory, Eindhoven, 10–15 June

10. U.M. Maurer, Perfect cryptographic security from partially independent channels, in *Proceedings of 23rd ACM Symposium on the Theory of Computing*, New Orleans (1991), pp. 561–572
11. U.M. Maurer, Secret key agreement by public discussion based on common information. IEEE Trans. Inf. Theory **39**(3), 733–742, (1993)
12. R.L. Rivest, A. Shamir, L. Adleman, A method for obtaining digital signatures and public–key cryptosystems. Commun. ACM **21**, 120–126 (1978)
13. A.D. Wyner, The wire–tap channel. Bell Syst. Techn. J. **54**, 1355–1387 (1975)

Common Randomness in Information Theory and Cryptography CR Capacity

The CR capacity of a two-terminal model is defined as the maximum rate of common randomness that the terminals can generate using resources specified by the given model. We determine CR capacity for several models, including those whose statistics depend on unknown parameters. The CR capacity is shown to be achievable robustly, by common randomness of nearly uniform distribution no matter what the unknown parameters are. Our CR capacity results are relevant for the problem of identification capacity, and also yield a new result on the regular (transmission) capacity of arbitrarily varying channels with feedback.

1 Introduction

Suppose two terminals, called Terminal $\mathcal{X}$ and Terminal $\mathcal{Y}$, have resources such as access to side information and communication links that allow them to observe and (perhaps cooperatively) generate certain RV's. The permissible rules for this are specified by the particular model, but it is always assumed that the terminals have unrestricted computational power, thus the RV's that can be generated and observed at a terminal at a given time include, as a minimum, all functions of the RV's previously observed there. Common randomness (CR) of $\mathcal{X}$ and $\mathcal{Y}$ means, intuitively, a RV generated by them and observable to both, perhaps with a small probability of error. A RV generated by a terminal is not necessarily observable there, e.g., when Terminal $\mathcal{X}$ inputs a RV X into a noisy channel to Terminal $\mathcal{Y}$, he thereby generates an output Y observable only at $\mathcal{Y}$. If Terminal $\mathcal{X}$ suitably encodes the RV X he wants to transmit, enabling $\mathcal{Y}$ to decode, then this X will represent CR. If noiseless feedback from $\mathcal{Y}$ to $\mathcal{X}$ is available then the output Y will always represent CR.

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_15

In chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints" [6] we were interested in CR under an additional secrecy constraint, with the motivation that the generated CR will be used as an encryption key. In this chapter [7] we do not require secrecy, and just study the maximum amount of CR afforded by a given model, the amount measured by entropy. The most convenient form of CR is uniform common randomness (UCR), i.e., CR represented by a uniformly (or nearly uniformly) distributed RV. For the type of models we will consider, the maximum attainable amount of CR and UCR will be asymptotically the same.

As a very simple example, suppose that there is a DMC from Terminal $\mathcal{X}$ to Terminal $\mathcal{Y}$, Terminal $\mathcal{X}$ can randomize (= can generate RV's with arbitrary distributions), and can input into the DMC any random sequence X^n he has generated (of given "large" length n). Terminal $\mathcal{Y}$ can observe the output Y^n but the terminals have no other resources. It is intuitively clear that in this case $\mathcal{X}$ has to chose X^n to be uniformly distributed on the $\approx \exp(nC)$ codewords of an optimum code; then $\mathcal{Y}$ can decode, and the achieved $\approx nC$ amount of CR is best possible. If noiseless feedback is available, it is better for $\mathcal{X}$ to send independent repetitions of a RV X that produces maximum output entropy $H(Y)$. As now $\mathcal{X}$ can observe Y^n, in this way CR of amount $nH(Y)$ results, clearly the largest possible. Notice that here, too, the optimum could (almost) be attained by a nearly uniform RV, obtained by applying a compression code to Y^n. As a combinatorial example, let $\mathcal{G}$ be a bipartite graph with vertex sets $\mathcal{X}$ and $\mathcal{Y}$ (we continue the practice of chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints" that the symbols of the terminals also denote sets assigned to them). An example of the bipartite graph $\mathcal{G}$ is depicted in Fig. 1. Nature selects an edge $(x, y) \in \mathcal{G}$ at random. For example, the dashed edge is denoted by (x_3, y_2). Terminal $\mathcal{X}$ observes x, Terminal $\mathcal{Y}$ observes y. The terminals can communicate over a noiseless channel, but at most b binary digits may be transmitted, in any number of rounds. No other resources are available (above the minimum described in the first paragraph), in particular, neither terminal can randomize. Then, clearly, $\log|\mathcal{G}|$ is an upper bound to CR, which can be attained iff the communication complexity $C_\infty(\mathcal{H}, P_\mathcal{V}, P_\mathcal{E})$ does not exceed b (with the notation of [5]); here $\mathcal{H}$ denotes the hypergraph with vertex set $\mathcal{V} = \mathcal{X}$ and edge set $\mathcal{E}$ consisting of the sets $\{x : (x, y) \in \mathcal{G}\}$, $y \in \mathcal{Y}$. It may be an interesting study in communication complexity to determine the maximum amount of CR when $b < C_\infty(\mathcal{H}, P_\mathcal{E}, P_\mathcal{V})$, i.e., the maximum entropy of a function on $\mathcal{G}$ that may be computed at both terminals with communication of at most b bits.

One obvious motivation of our interest in CR is that if the two terminals have access to the outcome of the same random experiment, this knowledge may be used to implement correlated random protocols, perhaps leading to much faster algorithms than deterministic ones or those using independent randomization only. In information theory, in particular for arbitrarily varying channels (AVC's), correlated random codes may much outperform deterministic (or randomized) codes; indeed, they may be necessary to attain positive capacity [12]. E.g. for the additive Gaussian AVC with power constraints, (average error) capacity for

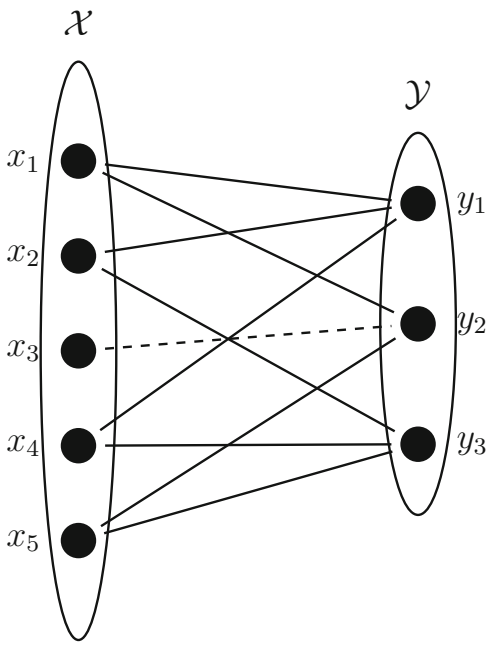


Fig. 1 Bipartite graph $\mathcal{G}$

deterministic codes equals random code capacity only if the sender's power exceeds the jammer's, otherwise the deterministic code capacity is 0 [16].

An even more striking application of CR appears in the theory of identification capacity [8] (see chapter "Identification via Channels" in Part I). It was shown in [9] (see chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas") that for any kind of channel, if sender and receiver can build up nR bits of UCR, this can be used to construct ID codes for $\approx 2^{2^{nR}}$ messages, with small probability of misidentification and misrejection, provided that the channel capacity is positive. The asymptotic optimality of this construction was also established in [9], for DMC's with no feedback and with complete feedback. Similar results for multi-user channels were obtained in [10] (see chapter "On Identification via Multi-Way Channels with Feedback: Mystery Numbers").

One feature of this chapter is that we also study "robust common randomness". This concept refers to models whose statistical properties are not completely specified but depend on certain parameters ("state") out of control of the two terminals and at least partially unknown to them. Then the distribution of the RV representing CR will depend on the actual state, and the minimum of its entropy (for all possible states) may be called the amount of robust CR. Most desirable is to have robust UCR, i.e., such RV representing CR whose distribution is nearly uniform, no matter what the actual state is. Again, for the type of models we will consider, the maximum attainable amount of robust CR and robust UCR will be asymptotically the same.

We will study robust UCR for AVC's and the results allow us to determine identification capacity for various AVC models. Quite remarkably, we also obtain a new result on regular (transmission) capacity, namely that the average error capacity of an AVC with complete feedback always equals the random code capacity of this AVC.

The problem of robust uniform randomness is of interest even if it is not required that distant terminals have access to it. Then the problem is that, if several probability distributions (PD's) are given on a set $\mathcal{V}$, how large can be the number of values of a function f on $\mathcal{V}$ whose distribution is nearly uniform under each of the given PD's. Here we state a simple combinatorial lemma, similar in spirit to the hypergraph coloring lemmas of [3]. It says that if the given PD's on $\mathcal{V}$ are uniform distributions on the edges $E \in \mathcal{E}$ of a hypergraph $(\mathcal{V}, \mathcal{E})$ then the maximum number of values of an f with the required property is not much smaller than the smallest edge size $|E|$. We believe that this lemma will help the reader to develop intuition, as it helped us to arrive at the results on robust VCR in Sect. 3.

Log's and exp's are to the base 2. Natural logarithms are denoted by ℓn.

Lemma 137 (Balanced coloring) *Let $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ be a hypergraph with $|\mathcal{E}| = N$ edges, each of size $|E| \geq d$. Then for any $0 < \varepsilon < \frac{1}{2}$ and $k < d\varepsilon^2/\ell n(2N)$ there exists an ε-balanced vertex coloring with k colors, i.e., a function $f : \mathcal{V} \to \{1, \ldots, k\}$, such that*

$$\left| \frac{|f^{-1}(i) \cap E|}{|E|} - \frac{1}{k} \right| < \frac{\varepsilon}{k} \text{ for all } 1 \leq i \leq k \text{ and } E \in \mathcal{E}. \tag{1}$$

Proof Let $\{Z(v), v \in \mathcal{V}\}$ be a family of i.i.d. RV's such that $Pr\{Z(v) = i\} = \frac{1}{k}$, $i = 1, \ldots, k$, and let $Z_i(v) = 1$ if $Z(v) = i$, and 0 otherwise. Then for the random coloring $f(v) = Z(v)$ we have $|f^{-1}(i) \cap E| = \sum_{v \in E} Z_i(v)$, and the standard large deviation bound for the binomial distribution gives, for every fixed $1 \leq i \leq k$ and $E \in \mathcal{E}$, that

$$\Pr\left\{ |f^{-1}(i) \cap E| < \frac{1-\varepsilon}{k}|E| \right\} \leq \exp\left\{ -|E| D\left(\frac{1-\varepsilon}{k} \| \frac{1}{k} \right) \right\}$$

$$\Pr\left\{ |f^{-1}(i) \cap E| > \frac{1+\varepsilon}{k}|E| \right\} \leq \exp\left\{ -|E| D\left(\frac{1+\varepsilon}{k} \| \frac{1}{k} \right) \right\},$$

where $D(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$.

Calculus shows that $D\left(\frac{1+\varepsilon}{k} \| \frac{1}{k} \right) - \frac{\varepsilon^2}{k\ell n 2}$ is a convex function of ε in the interval $-\frac{1}{2} \leq \varepsilon \leq \frac{1}{2}$, with minimum equal to 0 attained at $\varepsilon = 0$. It follows that the probability that (1) does not hold for the random coloring $f(v) = Z(v)$ is upper bounded by $N \cdot 2 \exp(-d\varepsilon^2/k\ell n 2)$. Under the hypothesis of Lemma 137, this bound is less than 1, and the assertion follows. □

2 Preliminaries

A key concept studied in this chapter for various models is what we call CR capacity. In this section, we first formally describe one model to be considered, and define achievable CR rates and CR capacity for that model. Then we indicate the changes needed for other models, including those where the underlying statistics depend on unknown parameters. (For all models considered, *alternative* definitions of capacities – or capacity functions – lead to the same values.) As one of our reasons for studying CR capacity is its relationship to ID capacity, at the end of this section we sketch how the latter concept can be defined for the type of models we are interested in, as a straightforward extension of the definition of ID capacity of a DMC without or with feedback, see chapters "One Sender Answering Several Questions of Receivers" and "Models with Prior Knowledge of the Receiver" in Part II. A general definition of transmission capacity is also included.

In Sect. 3 we will establish some general results, including the achievability of CR capacity with UCR, i.e., with nearly uniformly distributed RV's. For models where the statistics depend on unknown parameters, this UCR result holds in a robust sense. Then the result of Ahlswede and Dueck [9] referred to in the Introduction affords the conclusion that for the type of models considered in this chapter, CR capacity is always a lower bound to ID capacity, whenever the transmission capacity is positive.

Our results on the CR capacity of particular models will be stated and proved in Sects. 4 and 5.

As in chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints", we use the terminology of the book [14], and refer to it for notation not defined here.

One of the stimuli for this investigation came from [4], where first basic observations are made and first results are established for the binary symmetric case of the model we now describe.

2.1 *Model (i): Two-Source with One-Way Communication*

Given a discrete memoryless multiple source (DMMS) with two components, with alphabets $\mathcal{X}, \mathcal{Y}$ and generic variables X, Y, the n-length source outputs are observable at Terminals $\mathcal{X}$ and $\mathcal{Y}$, respectively. Moreover, $\mathcal{X}$ can send information to $\mathcal{Y}$ over a noiseless channel of capacity R, namely, he can noiselessly transmit any function $f(X^n)$ of X^n to $\mathcal{Y}$, subject to the rate constraint

$$\frac{1}{n} \log \| f \| \leq R. \tag{2}$$

Other resources are not available to the terminals. We will say that a pair of RV's (K, L) is permissible if K and L are functions of the data available at $\mathcal{X}$ resp. $\mathcal{Y}$,

i.e.,

$$K = K(X^n),\ L = L\big(Y^n, f(X^n)\big). \tag{3}$$

A permissible pair (K, L) represents ε-common randomness if

$$\Pr\{K \neq L\} < \varepsilon. \tag{4}$$

As K and L represent the same CR, intuition requires that the entropy rates $\frac{1}{n}H(K)$ and $\frac{1}{n}H(L)$ be arbitrary close if ε is small, independently of n. In order to ensure this, via Fano's inequality (Lemma 48), we impose the technical condition that K and L take values in the same set $\mathcal{K}$ whose cardinality satisfies

$$|\mathcal{K}| \leq \exp(cn) \tag{5}$$

for some c not depending on n.

For Model(i), we adopt the following definition that, with suitable interpretation, will be appropriate also for other models.

Definition 138 A number H is an achievable CR rate if for some constant c and every $\varepsilon > 0, \delta > 0$, for all sufficiently large n there exists a permissible pair of RV's (K, L) satisfying (4), (5), such that

$$\frac{1}{n}H(K) > H - \delta. \tag{6}$$

The largest achievable CR rate is the CR capacity.

Remark In chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints" we considered the related concept of key capacity, where also a secrecy requirement was imposed on the CR. There, nearly uniform distribution was also required in the sense that the entropy rate $\frac{1}{n}H(K)$ be close to $\frac{1}{n}\log|\mathcal{K}|$. As stated before, CR of nearly uniform distribution or UCR is desirable also in the present context. It turns out, however, that the CR capacity in the sense of Definition 138 can always be attained with nearly uniformly distributed RV's, even in the stronger sense of variation distance, i.e., with K satisfying

$$\sum_{k\in\mathcal{K}} \left| \Pr\{K = k\} - \frac{1}{|\mathcal{K}|} \right| < \varepsilon.$$

Actually, it will be seen in Theorem 143 that a still stronger kind of near uniformity can be attained, with the variation distance above going to 0 exponentially as $n \to \infty$. Of course, then also $H(K)$ will be exponentially close to $\log|\mathcal{K}|$.

For orientation notice that the CR capacity for Model (i) never exceeds $H(X)$. If $H(X|Y) < R$ then an f satisfying (2) can be chosen to let $\mathcal{Y}$ recover X^n from

$f(X^n)$ and Y^n with small probability of error (Slepian and Wolf, [20]). Thus in this case the CR capacity equals $H(X)$.

The question of how large CR rate can be attained in the extreme case $R = 0$ of Model (i), when no communication is permitted between $\mathcal{X}$ and $\mathcal{Y}$, was asked by the second author in 1970. It was answered by Gács and Körner [17] who showed that it was equal to the largest entropy of a common function of X and Y, hence always 0 if X and Y had indecomposable joint distribution. This paper was one of the starting points of multiuser information theory, at least for the Hungarian research group. It will turn out that the (by now) standard "multi-user" techniques permit to determine the CR capacity both for Model (i) and its extensions considered in Sect. 4.

In the model described above, randomization was not permitted. As in chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints", we will always regard randomization (at either terminal) as generating a RV at the very start, and let further actions depend on this RV, but already in a deterministic way. Thus, Model (i) with randomization at $\mathcal{X}$ means that a RV $M = M_{\mathcal{X}}$ (of arbitrary distribution, but independent of X^n, Y^n) may be generated at $\mathcal{X}$; then the information sent to $\mathcal{Y}$ may be $f(X^n, M)$ (still subject to (2)), and Definition 138 applies with the understanding of permissible pairs as $K = K(X^n, M)$, $L = L\big(Y^n, f(X^n, M)\big)$. Randomization at $\mathcal{Y}$ might also be permitted, then $\mathcal{Y}$ could generate a RV $M_{\mathcal{Y}}$ (independent of $X^n, Y^n, M_{\mathcal{X}}$), and let L be a function of $M_{\mathcal{Y}}$, too. Notice that whereas randomization at $\mathcal{X}$ may increase the CR capacity of Model (i), randomization at $\mathcal{Y}$ cannot.

A variant of Model (i) is when the given channel from $\mathcal{X}$ to $\mathcal{Y}$ is not noiseless but a DMC, say with the same word-length n as the observed source output. The input is selected by Terminal $\mathcal{X}$ as a function of X^n (or of X^n and M if randomization is permitted) and Terminal $\mathcal{Y}$ observes the output, say Z^n. Then the change required in the definition of permissible pairs (K, L) is that now $L = L(Y^n, Z^n)$.

A somewhat different model is

2.2 Model (ii): DMC with Active Feedback

Given a DMC $\{W : \mathcal{X} \to \mathcal{Y}\}$, Terminal $\mathcal{X}$ selects the inputs, Terminal $\mathcal{Y}$ observes the outputs, and $\mathcal{Y}$ can send back information to $\mathcal{X}$ over a noiseless channel of capacity R. We assume that $\mathcal{X}$ is permitted to randomize but $\mathcal{Y}$ is not. Formally, the terminals' permissible actions are as follows. Initially, $\mathcal{X}$ generates a randomization RV $M_{\mathcal{X}} = M$, then he inputs $X_1 = f_1(M)$ into the DMC. The output Y_1 is observed by $\mathcal{Y}$ who then noiselessly sends $\mathcal{X}$ a message $g_1(Y_1)$. Then $\mathcal{X}$ sends $X_2 = f_2\big(M, g_1(Y_1)\big)$ over the DMC, $\mathcal{Y}$ observes the output Y_2 and sends back $g_2(Y_1, Y_2)$. Next $\mathcal{X}$ sends $X_3 = f_3\big(M, g_1(Y_1), g_2(Y_1, Y_2)\big)$ and $\mathcal{Y}$ sends back $g_3(Y_1, Y_2, Y_3)$, etc. through n rounds.

The individual feedback messages may be arbitrary, but $g = (g_1, \ldots, g_n)$ is supposed to satisfy the global rate constraint

$$\frac{1}{n} \log \|g\| \leq R. \tag{7}$$

E.g., $g_1, \ldots, g_n$ may be binary words of variable length with prefix property, then (7) will mean that their total length is $\leq nR$.

In this model, the permissible pairs (K, L) (which will represent ε-common randomness if they satisfy (4)) are of the form

$$K = K(M, g_1, \ldots, g_n), \ L = L(Y^n). \tag{8}$$

With this understanding, Definition 138 of the achievable CR rates and CR capacity applies to the present model. Notice that some of the messages g_i may be empty, indeed it is permissible that Terminal $\mathcal{Y}$ sends only one message to $\mathcal{X}$ after having received the whole Y^n (of course, then the input X^n must be a function of M alone). We will show that the CR capacity for Model (ii) is always attainable that way.

In another version of Model (ii) also Terminal $\mathcal{Y}$ is permitted to randomize, which formally means that he, too, generates a randomization RV $M_{\mathcal{Y}}$ at the start (independent of $M_{\mathcal{X}}$), and then $g_1, \ldots, g_n$ as well as L may depend also on $M_{\mathcal{Y}}$. Still another version would be when neither terminal is allowed to randomize, but that will not be considered here.

Just as Model (i) could be modified replacing the noiseless channel from $\mathcal{Y}$ to $\mathcal{X}$ by a DMC, the same is possible also for Model (ii). Actually, several such variants of Model (ii) could be considered, one of them is when the i'th input of the backward channel is a function $g_i(Y_1, \ldots, Y_i)$ of the first i outputs of the forward channel, and Terminal $\mathcal{X}$ observes the corresponding output Z_i before selecting the input X_{i+1} to the forward channel. Then the permissible pairs (K, L) are defined by letting $K = K(M, Z^n)$, while $L = L(Y^n)$ as before.

Remark Our terminology "active feedback" refers to the freedom of Terminal $\mathcal{Y}$ to select the inputs of the backward channel. It differs from the terminology of the book [14] where "active feedback" means that $\mathcal{Y}$ is allowed to randomize. By "passive feedback" we mean that the inputs of the backward channel are equal to the outputs of the forward channel. In particular, noiseless passive feedback (also called complete feedback) means that the outputs of the DMC $\{W : \mathcal{X} \to \mathcal{Y}\}$ are observable not only to Terminal $\mathcal{Y}$ but also to Terminal $\mathcal{X}$. The variant of Model (ii) with complete feedback has been hinted at in the Introduction as a simple example for which the problem of CR-capacity is trivial. The variant with noisy passive feedback deserves interest, but will not be considered in this chapter.

The paper [21] came to our attention. It gives a detailed treatment of that version of our Model ii, where the backward channel is a DMC, including the case when both terminals have available a limited amount of private randomness.

A more complex version of the two-source model is the next one.

2.3 Model (iii): Two-Source with Two-Way Noiseless Communication

Given a DMMS as in Model (i), suppose that after Terminal $\mathcal{Y}$ received the message sent by $\mathcal{X}$ over a noiseless channel of capacity R_1, he in turn can send $\mathcal{X}$ a message over a noiseless channel of capacity R_2. This can be any function g of Y^n and the received $f(X^n)$ (or $f(X^n, M_{\mathcal{X}})$), subject to the rate constraint

$$\frac{1}{n}\log \|g\| \leq R_2. \tag{9}$$

If $\mathcal{Y}$ is permitted to randomize, g may also depend on $\mathcal{Y}$'s randomization RV $M_{\mathcal{Y}}$, chosen at the start, independently of $(M_{\mathcal{X}}, X^n, Y^n)$.

Now (K, L) is a permissible pair of RV's if $K = K(X^n, g)$ or $K = K(X^n, M_{\mathcal{X}}, g)$ and $L = L(Y^n, f)$ or $L = L(Y^n, M_{\mathcal{Y}}, f)$, according as randomization is permitted or not. With this understanding of the permissible pairs, Definition 138 applies as before.

It is obvious how to extend the model to permit several rounds of communication between $\mathcal{X}$ and $\mathcal{Y}$, each transmission subject to a rate constraint. Alternatively, the transmissions may not be constrained individually only their total rate is. The CR capacity can always be defined as in Definition 138, letting the permissible (K, L) pairs be functions of the data that become available at the corresponding terminals after having executed a protocol allowable by the particular model.

2.4 Models with Robust CR

The simplest model of this kind is that when both terminals can observe the output of an arbitrary varying source (AVS) but have no other resources whatsoever. An AVS with alphabet $\mathcal{X}$ and state set $\mathcal{S}$ (both finite) is determined by a family $\{P(\cdot|s), s \in \mathcal{S}\}$ of PD's on $\mathcal{X}$. The distribution of the n-length source output X^n depends on the state sequence $\mathbf{s} \in \mathcal{S}^n$, and equals

$$P(\cdot|\mathbf{s}) = P(\cdot|s_1) \times \cdots \times P(\cdot|s_n) \text{ if } \mathbf{s} = (s_1, \ldots, s_n). \tag{10}$$

In this model, any function $K = K(X^n)$ represents CR, thus the largest CR, for any fixed block-length n, is represented by $K = X^n$. In the definition of achievable CR rates, the condition (6) is now required to hold independently of the underlying statistics, i.e., for all $\mathbf{s} \in \mathcal{S}^n$. Thus the CR capacity for this model equals $H_{\min} = \min_{s\in\mathcal{S}} H\big(P(\cdot|s)\big)$. It is non-trivial but will be shown in Theorem 141 that this CR capacity can be attained with robust UCR, i.e., that $K = K(X^n)$ satisfying (7) for

all possible $\mathbf{s} \in \mathcal{S}^n$ can be given, such that

$$\frac{1}{n} \log |\mathcal{K}| > H_{\min} - \delta.$$

We will consider various AVC models in this chapter. An AVC with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$ and state set $\mathcal{S}$, each finite, is determined by a family $\mathcal{W} = \{W(\cdot|\cdot, s), s \in \mathcal{S}\}$ of channels $W(\cdot|\cdot, s) : \mathcal{X} \to \mathcal{Y}$. Terminal $\mathcal{X}$ selects the inputs, Terminal $\mathcal{Y}$ observes the outputs, and the state sequence $\mathbf{s} \in \mathcal{S}^n$ governing the n-length transmission may be arbitrary. Several different models are possible according to the availability of information to Terminal $\mathcal{X}$ about the states and the previous outputs when selecting the input X_i, and whether or not randomization is allowed.

We now formally describe two models, both with randomization permitted at $\mathcal{X}$, thus Terminal $\mathcal{X}$ first generates a randomization RV $M_{\mathcal{X}} = M$. In the "no feedback" model, Terminal $\mathcal{X}$ selects the input sequence X^n as a function of M. In the "complete feedback" model, the inputs $X_1, \ldots, X_n$ are selected successively as $X_i = f_i(M, Y_1, \ldots, Y_{i-1})$, where $Y_1, \ldots, Y_{i-1}$ are the previous outputs ("seen" by Terminal $\mathcal{X}$ through a noiseless feedback channel from $\mathcal{Y}$ to $\mathcal{X}$). In both models, the joint distribution of M and the output sequence Y^n, when the state sequence is $\mathbf{s} = (s_1, \ldots, s_n)$, is given by

$$\Pr\{M = m, Y^n = \mathbf{y}\} = \Pr\{M = m\} \prod_{i=1}^{n} W(y_i|x_i, s_i). \tag{11}$$

Here x_i denotes the i'th input symbol when $M = m$ (in the no feedback model) or when $M = m, Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}$ (in the complete feedback model). For both models, the CR capacity is defined as in Definition 138, requiring (4) and (6) to hold robustly, i.e., for every $\mathbf{s} \in \mathcal{S}^n$. The permissible pairs K, L are of form $K = K(M)$, $L = L(Y^n)$ in the no feedback case, and formally, K should be replaced by $K = K(M, Y^n)$ in the complete feedback case; for the latter model, however, $K = L = L(Y^n)$ may be taken, without restricting generality.

Both the "no feedback" and "complete feedback" AVC models can be modified by letting Terminal $\mathcal{X}$ know the state sequence $\mathbf{s}$. Then the inputs $X_1, \ldots, X_n$ and the RV K may depend also on $\mathbf{s}$. Also the AVC analogue of Model (ii), i.e., AVC with active feedback could be considered.

We will not attempt to give a general formal definition of the class of models we are interested in, but all our models involve the block-length n of observable source RV's or permissible channel transmissions (or both). For such models, Definition 138 always makes sense if we specify, for every n, the class of permissible pairs of RV's that may be generated by the terminals as functions of the data available to them. We now sketch how ID codes and ID capacity can be defined for arbitrary models of this kind, as a straightforward extension of the corresponding definitions for channels without or with feedback, see chapters "Identification via

Channels" and "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas".

Suppose one of N contingencies $k \in \{1, \ldots, N\}$ takes place, Terminal $\mathcal{X}$ knows this k, and the goal is to let Terminal $\mathcal{Y}$ reliably decide, for any $1 \leq j \leq N$ he may choose, whether or not $k = j$. To this end, the terminals perform a protocol permissible by the given model, for some block-length n, with the understanding that the actions of Terminal $\mathcal{X}$, but not those of Terminal $\mathcal{Y}$, may explicitly depend on k. E.g., for Model (ii), the functions $f_i\big(M, g_1(Y_1), \ldots, g_{i-1}(Y_1, \ldots, Y_{i-1})\big)$ specifying the channel inputs X_i, will depend on k, whereas for the feedback messages $g_i(Y_1, \ldots, Y_i)$ no such dependence is allowed, except for implicit dependence through the Y_i's. Let U denote all information available at Terminal $\mathcal{Y}$ after having performed the protocol, e.g., for Model (ii), $U = Y^n$. Then, if $\mathcal{Y}$ wants to decide whether or not $k = j$, he decides "yes" if $U \in D_j$ and "no" if $U \notin D_j$, where D_j, $1 \leq j \leq N$ are certain subsets of $\mathcal{U}$, the range of U.

Definition 139 A protocol as above together with a family $\{D_j, 1 \leq j \leq N\}$ of subsets of $\mathcal{U}$ is called an (N, n, ε) ID code for the given model if for each distinct k, j in $\{1, \ldots, N\}$

$$P_j(D_j^c) \leq \varepsilon; \quad P_k(D_j) \leq \varepsilon. \tag{12}$$

Here P_k denotes the distribution of U when contingency k has taken place. The ID capacity of the given model is the supremum of the numbers R such that for every $\varepsilon > 0$ and sufficiently large n there exists an (N, n, ε) ID code with $N \geq \exp\exp(nR)$.

For models whose statistics depend on unknown parameters ("state"), Definition 139 applies with the obvious modification. Namely, as then the distributions P_k also depend on the state, we require (12) to hold robustly, i.e., for all possible states. In particular, for an AVC without feedback (with $\mathcal{X}$ permitted to randomize) an (N, n, ε) ID code is defined by a family $\{Q_j, \quad 1 \leq j \leq N\}$ of PD's on $\mathcal{X}^n$, Q_j representing the distribution of the input sequence when contingency j takes place, together with a family $\{D_j, \quad 1 \leq j \leq N\}$ of subsets of $\mathcal{Y}^n$, such that for each distinct k, j in $\{1, \cdots, N\}$ and all $\mathbf{s} \in \mathcal{S}^n$

$$\sum_{\mathbf{x} \in \mathcal{X}^n} Q_j(x) W^n(D_j^c | \mathbf{x}, \mathbf{s}) \leq \varepsilon, \qquad \sum_{\mathbf{x} \in \mathcal{X}^n} Q_k(\mathbf{x}) W^n(D_j | \mathbf{x}, \mathbf{s}) \leq \varepsilon. \tag{13}$$

It is important to emphasize that the sets D_j in Definition 139 need not be disjoint. If they were, Terminal $\mathcal{Y}$ could infer k (as that j for which $U \in D_j$) with probability of error less than ε, thus the ID code would become a transmission code. Whereas for ID codes N can grow doubly exponentially with the block-length n, for transmission codes only exponential growth is possible.

As a straightforward generalization of the concept of channel capacity, we can define the transmission capacity of a general model as the supremum of numbers R such that for every $\varepsilon > 0$ and sufficiently large n there exists an (N, n, ε)

transmission code. Notice that for transmission codes, i.e., when the sets D_j, $1 \leq j \leq N$, are disjoint, it suffices to impose the first inequality in (12). More exactly, the transmission capacity defined in this way is that for the "maximum error" criterion, whereas transmission capacity for the "average error" criterion is obtained if the transmission codes are required to satisfy only

$$\frac{1}{N}\sum_{j=1}^{N} P_j(D_j^c) \leq \varepsilon, \tag{14}$$

a weaker condition than (12). Just as for standard channel capacity, these two concepts of transmission capacity coincide for models with uniquely determined statistics, but transmission capacity for average error can be larger than that for maximum error when the statistics depend on unknown states.

Remark For models with randomization allowed at Terminal $\mathcal{X}$, transmission capacity (for average error) is always a lower bound to CR capacity. Indeed, a trivial way of generating CR is that Terminal $\mathcal{X}$ generates a RV uniformly distributed on $\{1, \ldots, N\}$ and then transmits it to Terminal $\mathcal{Y}$ with probability of error less than ε. From the point of view of CR capacity, the interesting models are those for which this trivial scheme is not optimal.

3 Some General Results

Lemma 140 *Let $\mathcal{P}$ be any family of N PD's $P = \{p(v), v \in V\}$ on a finite set V, let $0 < \varepsilon \leq \frac{1}{9}$ and let $d > 0$ be such that for every $P \in \mathcal{P}$ the set*

$$E(P, d) = \left\{v : p(v) \leq \frac{1}{d}\right\} \tag{15}$$

has P-probability

$$P\big(E(P, d)\big) \geq 1 - \varepsilon. \tag{16}$$

Then for $k \leq \frac{\varepsilon^2}{3\log(2N)}d$, there exists $f : V \to \{1, \ldots, k\}$ such that for every $1 \leq i \leq k$ and $P \in \mathcal{P}$ the conditional P-probability of $f(v) = i$ on the condition $v \in E(P, d)$ differs from $\frac{1}{k}$ by less than $\frac{\varepsilon}{k}$, i.e.,

$$\left|\frac{P\big(f^{-1}(i) \cap E(P, d)\big)}{P\big(E(P, d)\big)} - \frac{1}{k}\right| < \frac{\varepsilon}{k}, \quad 1 \leq i \leq k, P \in \mathcal{P}. \tag{17}$$

In particular, the variation distance of the distribution of f from the uniform distribution on $\{1, \ldots, k\}$ is less than 3ε, i.e.,

$$\sum_{i=1}^{k} \left| P\big(f^{-1}(i)\big) - \frac{1}{k} \right| < 3\varepsilon, \tag{18}$$

for each of the PD's $P \in \mathcal{P}$.

Proof The proof is similar to that of Lemma 137, but requires a little more calculation.

Choosing f at random as in the proof of Lemma 137, with $Z_i(v)$ as there, we have

$$P\big(f^{-1}(i) \cap E(P,d)\big) = \sum_{v \in E(P,d)} p(v) Z_i(v). \tag{19}$$

Chernoff bounding gives that for any $A \subset \mathcal{V}$

$$\begin{aligned}
&\Pr\left\{ \sum_{v \in A} p(v) Z_i(v) > \frac{1+\varepsilon}{k} P(A) \right\} \\
&= \Pr\left\{ \exp\left[\beta \sum_{v \in A} p(v) Z_i(v) \right] > \exp\left(\beta \frac{1+\varepsilon}{k} P(A) \right) \right\} \\
&\le E\left(\exp\left[\beta \sum_{v \in A} p(v) Z_i(v) \right] \right) \exp\left(-\beta \frac{1+\varepsilon}{k} P(A) \right) \\
&= \exp\left(-\beta \frac{1+\varepsilon}{k} P(A) \right) \prod_{v \in A} \left[1 + \frac{1}{k} \big(\exp(\beta p(v)) - 1 \big) \right]
\end{aligned} \tag{20}$$

where $\beta > 0$ is arbitrary, and similarly

$$\begin{aligned}
&\Pr\left\{ \sum_{v \in A} p(v) Z_i(v) < \frac{1-\varepsilon}{k} P(A) \right\} \\
&\le \exp\left(\beta \frac{1-\varepsilon}{k} P(A) \right) \prod_{v \in A} \left[1 + \frac{1}{k} \big(\exp(-\beta p(v)) - 1 \big) \right].
\end{aligned} \tag{21}$$

Apply (20) to $A = E(P,d)$ with $\beta = \varepsilon d$. Then for $v \in A = E(P,d)$ we have $\beta p(v) \le \varepsilon$, by (15), and therefore

$$\begin{aligned}\exp\big(\beta p(v)\big) - 1 &= \sum_{j=1}^{\infty} \frac{\big(\beta p(v)\ell n2\big)^j}{j!} \\ &< \beta p(v)\left[1 + \frac{1}{2}\sum_{j=1}^{\infty}(\varepsilon \ell n2)^j\right]\ell n2 \\ &= \beta p(v)(1+\varepsilon^*)\ell n2,\end{aligned}$$

where $\varepsilon^* = \frac{\varepsilon \ell n2}{2(1-\varepsilon \ell n2)}$.

Using the inequality $1 + t\ell n2 \le \exp(t)$, it follows that the last product in (20) is upper bounded by

$$\exp\left[\sum_{v\in E(P,d)} \frac{1}{k}\beta p(v)(1+\varepsilon^*)\right] = \exp\left[\frac{\beta}{k}(1+\varepsilon^*)P\big(E(P,d)\big)\right].$$

Thus (20) gives, using the assumption (16) and recalling that $\beta = \varepsilon d$,

$$\begin{aligned}&\Pr\left\{\sum_{v\in E(P,d)} p(v)Z_i(v) > \frac{1+\varepsilon}{k}P\big(E(P,d)\big)\right\} \\ &< \exp\left[-\frac{\beta}{k}(\varepsilon-\varepsilon^*)P\big(E(P,d)\big)\right] < \exp\left(-\frac{\varepsilon d(\varepsilon-\varepsilon^*)(1-\varepsilon)}{k}\right) \\ &< \exp\left(-\frac{\varepsilon^2}{3k}d\right); \qquad (22)\end{aligned}$$

here, in the last step, we used that $(\varepsilon - \varepsilon^*)(1-\varepsilon) = \varepsilon\left(1 - \frac{\ell n2}{2(1-\ell\varepsilon n2)}\right)(1-\varepsilon) > \frac{\varepsilon}{3}$, if $\varepsilon < 3 - 2\log e$, and that condition does hold by the assumption $\varepsilon \le \frac{1}{9}$. It follows from (21) in a similar but even simpler way (as $\exp\big(-\beta p(v) - 1\big)$) can be bounded by $\beta p(v)\left(-1 + \frac{1}{2}\varepsilon\ell n2\right)\ell n2$ that the left hand side of (21) is also bounded by $\exp\left(-\frac{\varepsilon^2}{3k}d\right)$.

Recalling (19), we have thereby shown that the probability that (17) does not hold for a randomly chosen f is $< 2N\exp\left(-\frac{\varepsilon^2}{3k}d\right)$. Hence this probability is less than 1 if $k \le \frac{\varepsilon^2}{3\log(2N)}d$. This completes the proof of Lemma 140, because (18) is an immediate consequence of (17). □

Consider now the problem of robust uniform randomness obtainable by encoding the n-length output X^n of an AVS, where the distribution of X^n depending on the state sequence $\mathbf{s} \in \mathcal{S}^n$ is given by (10). We are interested in mappings $f : \mathcal{X}^n \to \mathcal{M}$ of possibly large rate $\frac{1}{n}\log|\mathcal{M}|$ for which $f(X^n)$ represents robust ε-uniform randomness, i.e., the variation distance of the distribution of $f(X^n)$ from the uniform distribution on $\mathcal{M}$ is less than ε, no matter what is the state sequence $\mathbf{s} \in \mathcal{S}^n$.

Theorem 141 *Let us be given an AVS by a set of PD's* $\{P(\cdot|s), s \in \mathcal{S}\}$ *on* $\mathcal{X}$*, such that* $H_{\min} = \min_{s\in\mathcal{S}} H\big(P(\cdot|s)\big) > 0$*. Then for every* $0 < \varepsilon < \frac{1}{3}$ *and every* n *there exists a mapping* $f : \mathcal{X}^n \to \mathcal{M}$ *of rate*

$$\frac{1}{n}\log|\mathcal{M}| > H_{\min} - \delta(\varepsilon, n) \tag{23}$$

such that $f(X^n)$ *represents robust* ε*-uniform randomness, where*

$$\delta(\varepsilon, n) = \sqrt{\frac{2\ell n 3/\varepsilon}{n}}\log|\mathcal{X}| + \frac{2\log 1/\varepsilon}{n} + \frac{\log\log(2|\mathcal{S}|)}{n} + 0\left(\frac{\log n}{n}\right) \tag{24}$$

if $|\mathcal{X}| \geq 3$*, and* $|\mathcal{X}|$ *should be replaced by* 3 *if* $|\mathcal{X}| = 2$*; the* $0\left(\frac{\log n}{n}\right)$ *term in* (24) *does not depend on* ε *and the AVS, not even on* $\mathcal{X}$ *and* $\mathcal{S}$*.*

Remark One feature of Theorem 141 that will be used in Theorem 143 below is that it brings out explicitly the dependence of $\delta(\varepsilon, n)$ on $\mathcal{X}$ and $\mathcal{S}$. For a fixed AVS, Theorem 141 shows that robust ε-uniform randomness for (arbitrarily small but) constant ε can be attained by mappings of rate approaching $H_{\min}$ with speed $0(n^{-\frac{1}{2}})$, and the rate will approach $H_{\min}$ even if $\varepsilon = \varepsilon_n \to 0$, providing it goes to 0 slower than exponentially. Moreover, robust ε-uniform randomness with rate $\frac{1}{n}\log|\mathcal{M}| > H_{\min} - \delta$ with an arbitrarily small but constant $\delta > 0$ is attainable even with ε going to 0 exponentially.

Proof of Theorem 141 Apply Lemma 140 to the family of PD's $P(\cdot|\mathbf{s})$, $\mathbf{s} \in \mathcal{S}^n$, on $V = \mathcal{X}^n$, with ε replaced by $\varepsilon/3$ (in order to get ε-uniform rather than 3ε-uniform randomness, cf. (18)). Then $N = |\mathcal{S}|^n$, and we will choose the number d in (15) as

$$d = \exp\big[n(H_{\min} - \xi)\big], \tag{25}$$

with $\xi > 0$ such that (16) (with ε replaced by $\varepsilon/3$) is fulfilled for each $P = P(\cdot|\mathbf{s})$. As shown in the Appendix,

$$\xi = \sqrt{\frac{2\ell n 3/\varepsilon}{n}}\log|\mathcal{X}| \tag{26}$$

is an adequate choice, with the understanding (as also in the rest of the proof) that $|\mathcal{X}|$ should be replaced by 3 if $|\mathcal{X}| = 2$. Then Lemma 140 gives that for

$$|\mathcal{M}| \leq \frac{(\varepsilon/3)^2}{3\log(2|\mathcal{S}|^n)} \exp\left[n\left(H_{\min} - \sqrt{\frac{2\ell n 3/\varepsilon}{n}}\log|\mathcal{X}|\right)\right] \tag{27}$$

there exists $f : \mathcal{X}^n \to \mathcal{M}$ such that $f(X^n)$ represents ε-uniform randomness, for each $\mathbf{s} \in \mathcal{S}^n$. Comparison of (23) and (27) shows that both can be satisfied with $\delta(\varepsilon, n)$ as in (23).

We have to show that the P-probability of the complementary set of (15) with $P = P(\cdot|\mathbf{s})$ defined by (10) is $\leq \varepsilon$ if d is as in (25), with ξ given by (26). This probability can be written as

$$\Pr\big\{P(X^n|\mathbf{s}) > \exp\big[-n(H_{\min} - \xi)\big]\big\} \tag{28}$$

where Pr denotes probability under $P(\cdot|\mathbf{s})$. Now, for every $t > 0$

$$\begin{aligned}
&\Pr\big\{P(X^n|\mathbf{s}) > \exp\big[-n(H_{\min} - \xi)\big]\big\} \\
&= \Pr\big\{P^t(X^n|\mathbf{s}) > \exp\big[-nt(H_{\min} - \xi)\big]\big\} \\
&< \exp\big[nt(H_{\min} - \xi)\big]E\big(P^t(X^n|\mathbf{s})\big) \\
&= \exp\big[nt(H_{\min} - \xi)\big]\prod_{i=1}^{n}\sum_{x\in\mathcal{X}} P^{1+t}(x|s_i),
\end{aligned} \tag{29}$$

where E denotes expectation under $P(\cdot|\mathbf{s})$. To bound the last product in (29), notice that for any PD $P = \{p(x)\}$ on $\mathcal{X}$

$$\begin{aligned}
\sum_{x\in\mathcal{X}} p^{1+t}(x) &= \sum_{x\in\mathcal{X}} p(x)\sum_{j=0}^{\infty}\frac{\big(t\ell n\ p(x)\big)^j}{j!} \\
&\leq 1 + t\sum_{x\in\mathcal{X}} p(x)\ell n\ p(x) + \frac{t^2}{2}\sum_{x\in\mathcal{X}} p(x)\big[\ell n\ p(x)\big]^2 \\
&= 1 - tH(P)\ell n\ 2 + \frac{t^2}{2}\sum_{x\in\mathcal{X}} p(x)\big[\ell n\ p(x)\big]^2.
\end{aligned} \tag{30}$$

Calculus shows that the last sum in (30) is maximum when P is the uniform distribution on $\mathcal{X}$, providing $|\mathcal{X}| \geq 3$. Hence

$$\sum_{x\in\mathcal{X}} p^{1+t}(x) \leq 1 - tH(P)\ell n\, 2 + \frac{t^2}{2}(\ell n|\mathcal{X}|)^2 \leq \exp\Big[-tH(P) + \frac{t^2}{2}(\log|\mathcal{X}|)^2\ell n\, 2\Big], \tag{31}$$

with the understanding (as also in the rest of the proof) that $|\mathcal{X}|$ should be replaced by 3 if $|\mathcal{X}| = 2$.

As $H\big(P(\cdot|s_i)\big) \geq H_{\min}$ by definition, (29) and (31) give that the probability (28) is upper bounded by $\exp\big[-nt\xi + n\frac{t^2}{2}(\log|\mathcal{X}|)^2\ell n\, 2\big]$, for each $t > 0$. Setting $t = \xi/(\log|\mathcal{X}|)^2\ell n\, 2$, we get

$$\Pr\big\{P(X^n|\mathbf{s}) > \exp\big[-n(H_{\min} - \xi)\big]\big\} < \exp\left[-n\frac{\xi^2}{2(\log|\mathcal{X}|)^2\ell n\, 2}\right]. \tag{32}$$

For ξ given by (26), the right hand side of (32) is equal to $\frac{\varepsilon}{3}$, establishing our claim. □

Having available Theorem 141, we now prove that for the type of models treated in this chapter, CR capacity can be attained with uniform CR. Although we did not give a formal definition of this class of models, we recall from Sect. 2 that all our models involve the specification of permissible pairs of RV's (K, L), for each block-length n. The following definition postulates a property common to all models we are interested in.

Definition 142 A model permits independent concatenations if for any pairs of RV's (K_1', L_1') and (K_2', L_2') permissible for block-lengths n_1 and n_2, there exists a pair (K, L) permissible for block-length $n_1 + n_2$ such that $K = (K_1, K_2)$, $L = (L_1, L_2)$, where (K_1, L_1) and (K_2, L_2) are independent and have the same distribution as (K_1', L_1') and (K_2', L_2'). When the underlying statistics are not uniquely determined but depend on some parameters ("state"), the last condition means that under any permissible statistics for block-length $n_1 + n_2$, (K_1, L_1) and (K_2, L_2) are independent, with distributions equal to those of (K_1', L_1') and (K_2', L_2') under one of permissible statistics for block-length n_1 resp. n_2.

For models with statistics depending on "states", let $\mathcal{S}(n)$ denote the set of possible states for block-length n. We will assume that this set does not grow faster than doubly exponentially, more exactly, that $\frac{1}{n}\log\log|\mathcal{S}(n)|$ is bounded by a constant. This holds for all models we are aware of, e.g. for the standard AVS and AVC models $|\mathcal{S}(n)| = |\mathcal{S}|^n$ grows only exponentially. Even for the variant of the AVC where the state sequence $\mathbf{s}$ may depend on the input sequence $\mathbf{x}$, in which case $\mathcal{S}(n)$ is the set of all mappings of $\mathcal{X}^n$ into $\mathcal{S}^n$, the growth rate of $|\mathcal{S}(n)|$ is "only" doubly exponential.

Theorem 143 *Let us be given a model permitting independent concatenations. If the statistics are not uniquely determined, we assume that* $\frac{1}{n}\log\log|\mathcal{S}(n)|$ *is bounded. Then for any fixed* $\varepsilon > 0$*, every* H *less than CR capacity, and sufficiently large* n*, there exists a permissible pair of RV's* (K, L)*, both distributed on a set* $\mathcal{M}$ *satisfying* $\frac{1}{n}\log|\mathcal{M}| \geq H$*, such that*

$$\Pr\{K \neq L\} < \varepsilon, \quad \sum_{k\in\mathcal{M}} \left|\Pr\{K = k\} - \frac{1}{|\mathcal{M}|}\right| < \varepsilon, \tag{33}$$

for every possible choice of the underlying statistics.

Remark It will be clear from the proof that the near uniformity of K can be attained also in a stronger sense, namely in the second inequality in (33), instead of a fixed $\varepsilon > 0$ one could take a sequence ε_n going to 0 exponentially as $n \to \infty$ (with a sufficiently small exponent). A similar improvement of the first inequality in (33) is possible providing in the definition of CR capacity, the fixed $\varepsilon > 0$ in (4) can be replaced by ε_n going to 0 exponentially; this holds for all the models treated in this chapter.

Proof of Theorem 143 As H is less than CR capacity, there exists $H' > H$ which is still an achievable CR rate. Applying Definition 138 to H' in the role of H, with $\delta' = \frac{H'-H}{2}$, and $\varepsilon' > 0$ specified later, it follows that for sufficiently large m there exists a pair (K', L') permissible for block-length m such that their common range $\mathcal{K}$ satisfies

$$|\mathcal{K}| \leq \exp(cm), \tag{34}$$

and

$$\Pr\{K' \neq L'\} < \varepsilon', \tag{35}$$

$$\frac{1}{m}H(K') > H' - \delta' = H + \delta', \tag{36}$$

for every choice of the underlying statistics. Clearly, the case of uniquely determined statistics $\big(|\mathcal{S}(m)| = 1\big)$ need not be considered separately.

As the model permits independent concatenations, for every r there exists a pair (K^r, L^r) permissible for block-length $n = rm$, with $K^r = K_1 \dots K_r$, $L^r = L_1 \dots L_r$, such that for every possible statistics for block-length n the pairs (K_i, L_i), $i = 1, \dots, r$ are independent, with distributions equal to that of (K', L') for some possible statistics for block-length m (possibly different for each i). In particular, K^r may be regarded as the r-length output of an AVS with alphabet $\mathcal{K}$ and state set $\mathcal{S}(m)$. For this AVS, $H_{\min} > m(H + \delta')$ by (36). Thus by Theorem 141

there exists a mapping $f : \mathcal{K}^r \to \mathcal{M}$ with

$$\frac{1}{r} \log |\mathcal{M}| > m(H + \delta') - \delta(\varepsilon, r) \tag{37}$$

such that the distribution of $f(K^r)$ is robustly ε-close to the uniform distribution on $\mathcal{M}$, where

$$\delta(\varepsilon, r) = \sqrt{\frac{2\ell n 1/\varepsilon}{r}} \log |\mathcal{K}| + \frac{2 \log 1/\varepsilon}{r} + \frac{\log \log |2\mathcal{S}(m)|}{r} + 0\left(\frac{\log r}{r}\right). \tag{38}$$

Using (34) and the assumption on the growth rate of $\mathcal{S}(m)$, it follows from (37) and (38) that $\frac{1}{rm} \log |\mathcal{M}| > H$ if r is sufficiently large, depending on ε but not on m.

With such an r we set $K = f(K^r)$, $L = f(L^r)$ for block-length $n = rm$. Then K and L are distributed on $\mathcal{M}$ satisfying $\frac{1}{n} \log |\mathcal{M}| > H$, and the second inequality in (33) holds for every possible choice of the underlying statistics. Finally, the first inequality in (33) follows from (35), if we choose $\varepsilon' = \varepsilon/r$. This completes the proof, because it clearly suffices to restrict attention to block-lengths n which are multiples of a constant r. □

Theorem 144 *For all models as in Theorem 143, the CR capacity is a lower bound to ID capacity, provided the transmission capacity (for the maximum error criterion) is positive.*

Proof Immediate from Theorem 143 and the result from chapter Identification in the Presence of Feedback: A Discovery of New Capacity Formulas". □

4 Common Randomness in Models (i), (ii), and (iii)

Theorem 145 *For Model (i) described in Sect. 2, the CR capacity equals*

$$C_1(R) = \max_U \left[I(U \wedge X) | I(U \wedge X) - I(U \wedge Y) \leq R \right] \tag{39}$$

if no randomization is permitted, and

$$\tilde{C}_1(R) = \begin{cases} C_1(R) & \text{if } R \leq H(X|Y) \\ R + I(X \wedge Y) & \text{if } R \geq H(X|Y) \end{cases} \tag{40}$$

if Terminal $\mathcal{X}$ is allowed to randomize. Here the maximum is for all RV's U that satisfy the Markov condition $U \ominus X \ominus Y$, and the range constraint $|\mathcal{U}| \leq |\mathcal{X}|$, and R is the capacity of the noiseless channel in the model. Moreover, the CR capacity of the variant of Model (i) where the noiseless channel is replaced by a DMC, is still given by (39) resp. (40), with R replaced by the capacity of that DMC.

Remark If $\mathcal{X}$ is permitted to randomize, a trivial way to create CR is that $\mathcal{X}$ generates nR random bits and transmits them to $\mathcal{Y}$, disregarding the DMMS. Theorem 145 shows that this is suboptimal, and for $R \geq H(X|Y)$ the CR capacity exceeds R (attained by the trivial scheme) by exactly $I(X \wedge Y)$. This means that although mutual information does not represent a "common information" (as shown in [17]), it does represent a kind of hidden common randomness that can be recovered if sufficient transmission capacity is available. It is interesting to compare this interpretation of mutual information with that obtained in chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints" in a context involving secrecy.

Proof of Theorem 145 A short proof is available using standard results of multi-user information theory, cf. the proof of Theorem 146 below. Here we prefer an independent proof, which later will be extended to the case of two-way communication.

We state, also for later reference, an identity also used in chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints" (Lemma 136; cf. also [14], p. 409): For arbitrary RV's S, T and sequences of RV's X^n, Y^n

$$I(S \wedge X^n|T) - I(S \wedge Y^n|T)$$

$$= \sum_{i=1}^{n} \big[I(S \wedge X_i|X_1 \dots X_{i-1}Y_{i+1} \dots Y_n T) - I(S \wedge Y_i|X_1 \dots X_{i-1}Y_{i+1} \dots Y_n T) \big]$$

$$= n\big[I(S \wedge X_J|V) - I(S \wedge Y_J|V) \big] \tag{41}$$

where J is a RV independent of all the previous ones, uniformly distributed on $\{1, \dots, n\}$, and

$$V = X_1 \dots X_{J-1}Y_{J+1} \dots Y_n T J. \tag{42}$$

1. *Converse part.* Consider first the "no randomization" case. Suppose (K, L) satisfy (2), (3), (4), and (5). Write

$$H(K|Y^n) = I\big(K \wedge f(X^n)|Y^n\big) + H\big(K|Y^n, f(X^n)\big). \tag{43}$$

Here the first term is $\leq nR$, by (2), and the second term is $\leq H(K|L) \leq \varepsilon cn + 1$ by (3) and Fano's inequality (Lemma 48), using (4), (5). Thus we have

$$H(K) - I(K \wedge Y^n) = H(K|Y^n) \leq nR + \varepsilon cn + 1. \tag{44}$$

Apply (41) to the present X^n, Y^n with $S = K$ and T trivial/absent. Then V in (42) is independent of (X_J, Y_J), hence the last line in (41) can also be written as

$$I(U \wedge X_J) - I(U \wedge Y_J), \quad \text{with } U = KV.$$

Thus

$$H(K) - I(K \wedge Y^n) = I(K \wedge X^n) - I(K \wedge Y^n) = n\big[I(U \wedge X_J) - I(U \wedge Y_J)\big] \tag{45}$$

where $U = KX_1 \dots X_{J-1} Y_{J+1} \dots Y_n$ satisfies the Markov condition $U \ominus X_J \ominus Y_J$. Notice also that

$$H(K) = I(K \wedge X^n) = \sum_{i=1}^{n} I(K \wedge X_i | X_1 \dots X_{i-1})$$

$$= nI(K \wedge X_J | X_1, \dots, X_{J-1}) \le nI(U \wedge X_J). \tag{46}$$

As X_J, Y_J may be identified with the generic variables X, Y of our DMMS, (44), (45) and (46) show that $\frac{1}{n} H(K)$ is upper bounded by the maximum of $I(U \wedge X)$ subject to $I(U \wedge X) - I(U \wedge Y) \le R + \varepsilon c + \frac{1}{n}$, for RV's U satisfying the Markov condition $U \ominus X \ominus Y$. It is routine to show that there exists U attaining the maximum that satisfies the range constraint $|\mathcal{U}| \le |\mathcal{X}|$ (direct application of the Support Lemma of [14], p. 310 gives only $|\mathcal{U}| \le |\mathcal{X}| + 1$, but for a U yielding an extremal value, this bound can be improved by 1, cf. [19]).

This completes the proof for the "no randomization" case. Notice that in (39) necessarily $I(U \wedge Y) \le I(X \wedge Y)$ hence

$$C_1(R) \le R + I(X \wedge Y), \quad \text{equality holds if } R = H(X|Y). \tag{47}$$

When $\mathcal{X}$ may randomize, we will conveniently regard his randomization RV $M_{\mathcal{X}}$ as an i.i.d. sequence M^n (of course, independent of X^n, Y^n). This reduces the present case to the previous one, replacing X by XM, where M is independent of X, Y. Thus we need to maximize $I(U \wedge XM)$ subject to

$$I(U \wedge XM) - I(U \wedge Y) \le R, \ U \ominus XM \ominus Y. \tag{48}$$

It follows similarly to (47) that (48) implies $I(U \wedge XM) \le R + I(X \wedge Y)$, thus for the case $R \ge H(X|Y)$ we are done. For $R < H(X|Y)$, notice that since $I(U \wedge XM) = I(U \wedge X) + I(U \wedge M|X)$, and the Markov condition in (48) implies $U \ominus X \ominus Y$, it follows from (47) that

$$I(U \wedge XM) \le C_1\big(R - I(U \wedge M|X)\big) + I(U \wedge M|X). \tag{49}$$

It is easy to check that the function defined by (39) is concave, hence, by (47), its slope is ≥ 1 if $R \leq H(X|Y)$. Thus the right hand side of (49) is $\leq C_1(R)$. This completes the proof for the randomized case.

Finally, if the channel from $\mathcal{X}$ to $\mathcal{Y}$ is not noiseless but a DMC, the only modification needed in the above proof is to replace $f(X^n)$ in Eq. (43) by the output of that DMC. Denote the input of this DMC by T^n and the output by Z^n. Whether or not Terminal $\mathcal{X}$ randomizes, the Markov condition $Y^n \ominus X^n K \ominus T^n \ominus Z^n$ must hold. Thus the first term in (43) with $f(X^n)$ replaced by Z^n can be bounded as

$$I(K \wedge Z^n|Y^n) \leq I(X^n K \wedge Z^n|Y^n) \leq I(T^n \wedge Z^n|Y^n) \leq nC,$$

establishing our claim.

2. *Direct part.* It suffices to consider the case $R \leq H(X|Y)$, with no randomization. By continuity, it suffices to show that $C_1(R')$ is an achievable CR rate for every $R' < R$. We are going to show this by exhibiting for arbitrary U satisfying

$$U \ominus X \ominus Y, \quad I(U \wedge X) - I(U \wedge Y) < R \tag{50}$$

and for any $\varepsilon > 0$, $\delta > 0$ and sufficiently large n, a permissible pair K, L as defined by (2), (3), such that K, L satisfy (4), (5), and (6) with $H = I(U \wedge X)$.

Assuming without any loss of generality that the distribution of U is a possible ED for block-length n, select at random $\exp\{n(I(U \wedge X) + \delta)\}$ sequences $\mathbf{u} \in \mathcal{U}^n$ of ED P_U, denoted as $\mathbf{u}_{ij}$, $1 \leq i \leq N_1$, $1 \leq j \leq N_2$, with

$$N_1 = \exp\{n(I(U \wedge X) - I(U \wedge Y) + 3\delta)\}, N_2 = \exp\{n(I(U \wedge Y) - 2\delta)\}. \tag{51}$$

Then for every X-typical $\mathbf{x} \in \mathcal{X}^n$ the probability that neither $\mathbf{u}_{ij}$ is jointly UX-typical with $\mathbf{x}$ is doubly exponentially small. Hence with probability close to 1, every typical $\mathbf{x}$ is jointly typical with some $\mathbf{u}_{ij}$.

Let $K(\mathbf{x})$ be equal to an $\mathbf{u}_{ij}$ jointly typical with $\mathbf{x}$ (either one if there are several), and let $f(\mathbf{x}) = i$ if $K(\mathbf{x}) = \mathbf{u}_{ij}$; both functions are set constant when $\mathbf{x}$ is not typical. Further let $L\big(\mathbf{y}, f(\mathbf{x})\big) = \mathbf{u}_{ij}$ if $f(\mathbf{x}) = i$ and $\mathbf{u}_{ij}, \mathbf{y}$ are jointly UY-typical. If there is no such $\mathbf{u}_{ij}$ or there are several, L is set equal to a constant. Then, by (50), (51), the rate constraint (2) on f is satisfied if δ is sufficiently small, and $K = K(X^n)$, $L = L\big(Y^n, f(X^n)\big)$ obviously satisfy (5). Also (6) is satisfied since

$$\begin{aligned} \Pr\{K = \mathbf{u}_{ij}\} &\leq P_X^n\big(\{\mathbf{x} : (\mathbf{u}_{ij}, \mathbf{x}) \text{ jointly typical}\}\big) \\ &= \exp\big(-nI(U \wedge X) + o(n)\big) \end{aligned} \tag{52}$$

implies that $H(K) \geq nI(U \wedge X) + o(n)$.

It remains to check (4), i.e., $\Pr\{K \neq L\} \leq \varepsilon$. Notice that for any jointly UX-typical pair $(\mathbf{u}, \mathbf{x})$, the set of $\mathbf{y}$'s jointly typical with $(\mathbf{u}, \mathbf{x})$ has conditional probability arbitrarily close to 1 on the condition $U^n = \mathbf{u}$, $X^n = \mathbf{x}$, and hence by Markovity, also on the condition $X^n = \mathbf{x}$. It follows that the set A of those pairs $(\mathbf{x}, \mathbf{y})$ for which

$\big(K(\mathbf{x}), \mathbf{x}, \mathbf{y}\big)$ are jointly UXY-typical has P_{XY}^n arbitrarily close to 1. Let us denote by B the set of those pairs $(\mathbf{x}, \mathbf{y}) \in A$ for which in addition to $\mathbf{u}_{ij} = K(\mathbf{x})$, some other $\mathbf{u}_{ij}$ (with the same first index i) is also jointly typical with $\mathbf{y}$. To complete the proof, it suffices to show that $P_{XY}^n(B)$ with be arbitrarily small, with large probability with respect to the random choice of $\{\mathbf{u}_{ij}\}$.

Now, for fixed $(\mathbf{x}, \mathbf{y})$, the probability that B determined by the random $\{\mathbf{u}_{ij}\}$ contains $(\mathbf{x}, \mathbf{y})$, is upper bounded by

$$\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{\substack{ell=1 \\ \ell \neq j}}^{N_2} \Pr\big\{(\mathbf{u}_{ij}, \mathbf{x}) \text{ jointly typical}, (\mathbf{u}_{ij}, \mathbf{y}) \text{ jointly typical}\big\}$$

$$= N_1 N_2^2 \exp\big[-nI(U \wedge X) + o(n)\big] \exp\big[-nI(U \wedge Y) + o(n)\big]$$

$$= \exp\big[-n\delta + o(n)\big]; \tag{53}$$

here we used (51) and that the $\mathbf{u}_{ij}$ are independent, chosen with uniform distribution from the sequences of ED P_U. Hence the expectation of $P_{XY}^n(A)$, as a RV depending on $\{\mathbf{u}_{ij}\}$, is also upper bounded by $\exp\big[-n\delta + o(n)\big]$. This completes the proof. □

Consider now the following generalization of Model (i) to generating CR at $r+1$ (rather than 2) terminals. Given a DMMS with $r+1$ components, with generic variables $X, Y_1, \ldots, Y_r$, Terminal $\mathcal{X}$ can observe X^n and send messages $f_i(X^n)$ to Terminals $\mathcal{Y}_i$, subject to rate constraints

$$\frac{1}{n} \log \|f_i\| \leq R_i, \ i = 1, \ldots, r. \tag{54}$$

Terminal $\mathcal{Y}_i$ can observe Y_i^n, and the message $f_i(X^n)$ sent him by Terminal $\mathcal{X}$. Achievable CR rates and CR capacity are defined by the natural extension of Definition 2.1, namely the role of permissible pairs (K, L) is now played by permissible $(r+1)$-tuples $(K, L_1, \ldots, L_r)$ defined in analogy to (3), and the role of condition (4) is played by r similar conditions $\Pr\{K \neq L_i\} < \varepsilon, i = 1, \ldots, r$.

Theorem 146 *For the above model, with no randomization permitted, the CR capacity equals the maximum of $I(U \wedge X)$ subject to the constraints*

$$I(U \wedge X) - I(U \wedge Y_i) \leq R_i, \ i = 1, \ldots, k, \ U \ominus X \ominus Y_1, \ldots, Y_k, \tag{55}$$

where U may be supposed to satisfy the range constraint $|\mathcal{U}| \leq |\mathcal{X}| + r - 1$. If Terminal $\mathcal{X}$ is permitted to randomize, the CR capacity is still the same if $R_i < H(X|Y_i)$ for some i, and it equals $\min_{1 \leq i \leq k}\big[R_i + I(X \wedge Y_i)\big]$ if $R_i \geq H(X|Y_i)$, $i = 1, \ldots, k$

Proof If H is an achievable CR rate and $\delta > 0$, take (for large n) $K = K(X^n)$ that can be ε-reproduced at each terminal by $L_i = L_i\big(Y_i^n, f_i(X^n)\big)$, $i = 1, \ldots, r$, and

such that

$$\left|\frac{1}{n}H(K) - H\right| < \delta. \tag{56}$$

Although the definition of achievable CR rates postulates $\frac{1}{n}H(K) > H - \delta$ only, it clearly does not restrict generality to require $\frac{1}{n}H(K) < H + \delta$, as well. In order that K could be ε-reproduced at $\mathcal{Y}_i$ it is necessary that

$$\frac{1}{n}H(K|Y_i^n) < R_i + \delta \tag{57}$$

(formally, $H(K|Y_i^n)$ can be written as a sum of two terms as in (43), and bounded as there).

On the other hand, if to a number H for all $\delta > 0$ sufficiently large n there exists a function $K = K(X^n)$ that satisfies (56), (57), then H is an achievable entropy rate. Indeed, from Y_i^{nr} and a suitable code $f_i(K^r)$ of rate $\frac{1}{nr}\log\|f_i\| \leq R_i + \delta$ of the r-fold repetition of K, Terminal $\mathcal{Y}_i$ can reproduce K^r with arbitrarily small probability of error, by Slepian-Wolf . Although the permissible rate is only R_i, this can be remedied by taking block-length $N = n'r$ with n' slightly larger than n, satisfying $n'R_i \geq n(R_i + \delta)$, $i = 1, \ldots, r$, and disregarding the last $N - nr$ source outputs. Thus for block-length N, the terminals can produce ε-common randomness of rate $\frac{1}{N}rH(K) = \frac{1}{n'}H(K)$, arbitrarily close to H.

Thus we have obtained a "product space characterization" of achievable CR rates, namely that H is achievable iff for every $\delta > 0$ and sufficiently large n there exists a function $K = K(X^n)$ satisfying (56), (57). This can be easily single-letterized, using results available in the literature. To this, notice that in the above product space characterization, (56), (57) may be replaced by

$$\left|\frac{1}{n}H(X^n|K) - \big(H(X) - H\big)\right| < \delta,\ \frac{1}{n}H(Y_i^n|K) \leq R_i - H + H(Y_i) + 2\delta. \tag{58}$$

Now, by [14], p. 352, an $(r+1)$-tuple $\tilde{R}_0, \tilde{R}_1, \ldots, \tilde{R}_r$ has the property that for every $\delta > 0$ and sufficiently large n there exists a function $f(X^n)$ satisfying

$$\left|\frac{1}{n}H\big(x^n|f(X^n)\big) - \tilde{R}_0\right| < \delta,\ \ \frac{1}{n}H\big(Y_i^n|f(X^n)\big) \leq \tilde{R}_i + \delta$$

iff there exists a RV U with $U \ominus X \ominus Y_1, \ldots, Y_r$ such that

$$H(X|U) = \tilde{R}_0,\ \ H(Y_i|U) \leq \tilde{R}_i.$$

Substituting here $\tilde{R}_0 = H(X) - H$, $\tilde{R}_i = R_i - H + H(Y_i)$, we get

$$I(U \wedge X) = H,\ I(U \wedge X) - I(U \wedge Y_i) \leq R_i,$$

and this completes the proof for the no randomization case (up to the routine range constraint).

If randomization is permitted, we replace X by XM and proceed as in the proof of Theorem 145. □

Theorem 147 *For Model (ii) described in Sect. 2, the CR capacity equals*

$$C_2(R) = \max_X \big[I(X \wedge Y) + \min\big(R, H(Y|X)\big)\big] \tag{59}$$

if randomization at $\mathcal{Y}$ is not permitted, and

$$\tilde{C}_2(R) = C(W) + R \tag{60}$$

if Terminal $\mathcal{Y}$ is allowed to randomize. In (59), the maximum is taken for random inputs X to the DMC $\{W\}$ given in the model, Y denoting the corresponding output. R is the capacity of the backward noiseless channel in the model, and $C(W)$ is the capacity of $\{W\}$. Moreover, the CR capacity of the variant of Model (ii) where the backward channel is replaced by a DMC, is still given by (59) resp. (60), with R replaced by the capacity of that DMC.

Remark Comparing (59) and (60) shows that if R is not larger than $H(Y|X)$ for a capacity achieving X then $\tilde{C}_2(R) = C_2(R)$. Thus, similarly to Model (i), randomization helps only when R is "large". For many DMC's, the maximum of $H(Y)$ is attained for a capacity achieving X. In those cases $C_2(R) = H(Y)$ for all R "small" in the above sense.

Remark One possible strategy of Terminal $\mathcal{X}$ in Model (ii) is to use an i.i.d. sequence X^n as channel input, which leads to the situation of Model (i), with the roles of $\mathcal{X}$ and $\mathcal{Y}$ reversed. Comparing Theorems 145 and 147 shows that this reduction to Model (i) suffices to achieve CR capacity for Model (ii) when the max in (59) is attained for some X with $H(Y|X) \leq R$, but not otherwise. (Namely, in that case, iq.'s (59) and (39), the lattes with X and Y reversed, give $C_2(R) = C_1(R) = H(Y)$.)

Proof

1. *Direct part.* If Terminal $\mathcal{Y}$ can randomize (recall that Terminal $\mathcal{X}$ always can in this model), a CR rate as in (60) can be attained in a trivial way: For large block-length n, Terminal $\mathcal{X}$ generates and transmits to $\mathcal{Y}$ a RV uniformly distributed on a set of size $\exp\big[(C(W) - \delta)\big]$; $\mathcal{Y}$ can decode it with small probability of error. Terminal $\mathcal{Y}$, in turn, generates a RV uniformly distributed on a set of size $\exp(nR)$, and transmits it to $\mathcal{X}$.

   If $\mathcal{Y}$ can not randomize, a CR rate as in (59) can be attained as follows. For large n, take (X, Y) almost attaining the maximum in (59) such that P_X is a possible ED for block-length n. Terminal $\mathcal{X}$ generates a RV M uniformly distributed on a set of size $\exp\big[n\big(I(X \wedge Y) - \delta\big)\big]$ and transmits it to $\mathcal{Y}$ using a code of fixed composition P_X. $\mathcal{Y}$ sends nothing back until he has received all

n outputs. Then $\mathcal{Y}$ decodes M. Terminal $\mathcal{Y}$ can decode with small probability of error, and he gets access to additional randomness from the channel output. Namely, he can numerate the words in each of his decoding sets from 1 to $\exp\big[nH(Y|X) + o(n)\big]$, then the RV Z equal to the number assigned to the observed output sequence will be almost independent of M and have entropy $\big[nH(Y|X) + o(n)\big]$. If $R > H(Y|X)$, this Z can be transmitted back to $\mathcal{X}$, and if $R < H(Y|X)$ then a suitable function of Z of entropy $nR + o(n)$ can be transmitted back.

2. *Converse part.* Let (K, L) be a permissible pair for block-length n, thus $K = K(M, g_1, \ldots, g_n)$, $L = L(Y^n)$, with $g = (g_1, \ldots, g_n)$ satisfying (7). Supposing that (K, L) satisfies the condition in Definition 138 we decompose $H(L|M)$ in analogy to (43), replacing $f(X^n)$ there by $g = (g_1, \ldots, g_n)$. Bounding as there we obtain

$$H(L|M) = I(L \wedge g|M) + H(L|M, g) \leq nR + \varepsilon nc + 1. \tag{61}$$

Further, (8) and the memoryless character of the DMC $\{W\}$ imply

$$\begin{aligned} I(L \wedge M) \leq I(M \wedge Y^n) &= \sum_{i=1}^{n} I(M \wedge Y_i|Y^{i-1}) \\ &\leq \sum_{i=1}^{n} I(X_i \wedge Y_i|Y^{i-1}) \leq \sum_{i=1}^{n} I(X_i \wedge Y_i) \leq nI(X_J \wedge Y_J) \end{aligned} \tag{62}$$

where J is an auxiliary RV uniformly distributed on $\{1, \ldots, n\}$, independent of (M, Y^n). On the other hand,

$$H(L) \leq H(Y^n) \leq \sum_{i=1}^{n} H(Y_i) \leq H(Y_J).$$

Combining this with (61)–(62) we get that

$$\begin{aligned} \frac{1}{n}H(L) &\leq \min\Big[I(X_J \wedge Y_J) + R + \varepsilon c + \frac{1}{n}, H(Y_J)\Big] \\ &\leq I(X_J \wedge Y_J) + \min\big[R, H(Y_J|X_J)\big] + \varepsilon c + \frac{1}{n}. \end{aligned}$$

As Y_J is the channel output for input X_J, this completes the converse proof also for the no randomization case. When the backward channel is not noiseless but a DMC, then denoting its input and output by T^n and Z^n, the only difference will be that g in (61) has to be replaced by Z^n.

Then the first term will be bounded as

$$I(L \wedge Z^n|M) \le I(T^n \wedge Z^n|M) \le nC,$$

where C is the capacity of the backward channel.

□

Theorem 148 *For Model (iii) described in Sect. 2, the CR capacity without randomization is equal to*

$$C_3(R_1, R_2) = \max_{U,V} \big[I(U \wedge X) + I(V \wedge Y|U) \mid I(U \wedge X) - I(U \wedge Y) \le R_1,$$
$$I(V \wedge Y|U) - I(V \wedge X|U) \le R_2 \big] \tag{63}$$

where the maximization is for RV's U and V satisfying the Markov conditions

$$U \ominus X \ominus Y, \qquad X \ominus YU \ominus V. \tag{64}$$

Moreover, the range sizes of U and V can be bounded by $|\mathcal{X}| + 2$ and $|\mathcal{Y}|$, resp.

Remark It is reassuring to check that (63) reduces to the expected simple results when either $R_1 \ge H(X|Y)$ or $R_2 \ge H(Y|X)$. In the first case $U = X$ is a permissible choice, then the Markov condition for V becomes void, and it follows that

$$C_3(R_1, R_2) = H(X) + \min\big(R_2, H(Y|X)\big) \text{ if } R_1 \ge H(X|Y). \tag{65}$$

In the second case $V = Y$ is a permissible choice, which leads to

$$I(U \wedge X) + I(V \wedge Y|U) = I(U \wedge X) + H(Y|U) = I(U \wedge X) - I(U \wedge Y) + H(Y).$$

Hence

$$C_3(R_1, R_2) = H(Y) + \min\big(R_1, H(X|Y)\big) \text{ if } R_2 \ge H(Y|X). \tag{66}$$

Proof of Theorem 148

1. *Converse part.* Let (K, L) be a permissible pair for Model (iii) without randomization, i.e., $K = K(X^n, g)$, $L = L(Y^n, f)$, $f = f(X^n)$, $g = g(Y^n, f)$, where f and g satisfy the rate constraints (2), (9). Suppose that (K, L) satisfy the conditions (4), (5) of Definition 138.

   Our key tool is the identity (41), which will be applied twice. First we get

$$nR_1 \ge H(f) = I(f \wedge X^n) \ge I(f \wedge X^n) - I(f \wedge Y^n) = n\big(I(U \wedge X) - I(U \wedge Y)\big), \tag{67}$$

with

$$X = X_J, Y = Y_J,\ U = fX_1 \dots X_{J-1} Y_{J+1} \dots Y_n J \tag{68}$$

(where we proceed as in the derivation of Eq. (45), the role of K there now played by f).

Notice now that just as $H(K|Y^n)$ was bounded in Eq. (43), we have the bound

$$H(L|X^n) \leq nR_2 + \varepsilon cn + 1. \tag{69}$$

Applying the identity (41) again, we get

$$\begin{aligned} -H(L|X^n) &= -H(L|X^n, f) \\ &= I(L \wedge X^n|f) - H(L|f) \\ &= I(L \wedge X^n|f) - I(L \wedge Y^n|f) \\ &= n\big(I(L \wedge X|U) - I(L \wedge Y|U)\big) \end{aligned} \tag{70}$$

where X, Y, U are (luckily) the same as in (68).

By (69) (with sufficiently small ε) and (70) we have for any fixed $\delta > 0$

$$R_2 \geq I(L \wedge Y|U) - I(L \wedge X|U) - \delta. \tag{71}$$

Finally we can write

$$\begin{aligned} I(L \wedge X^n) &= \sum_{i=1}^{n} I(L \wedge X_i|X_1 \dots X_{i-1}) \\ &= \sum_{i=1}^{n} I(LX_1 \dots X_{i-1} \wedge X_i) \leq nI(LU \wedge X). \end{aligned} \tag{72}$$

Combining (70) and (72) gives

$$\begin{aligned} H(L) &= I(L \wedge X^n) + H(L|X^n) \\ &\leq n\big[I(LU \wedge X) + I(L \wedge Y|U) - I(L \wedge X|U)\big] \\ &= n\big[I(U \wedge X) + I(L \wedge Y|U)\big]. \end{aligned} \tag{73}$$

Replacing L with V, we have thus proved that achievable CR-rates are bounded above by an expression as in (63) (the Markov conditions (64) are easily

verified), perhaps with R_2 replaced by $R_2 + \delta$; the latter is inconsequential, by continuity.

2. *Direct part.* As in the proof of Theorem 145, it suffices to prove that $I(X \wedge U) + I(V \wedge Y|U)$ is an achievable CR rate whenever U and V satisfy (in addition to (64)) the inequalities in (63) with strict inequality. The form of Eq. (63) suggests that in the first round CR of rate $I(U \wedge X)$ ought to be generated, and in the second round additional CR of rate $I(V \wedge Y|U)$.

   We use the same construction as in the proof of Theorem 145. First we generate $\{\mathbf{u}_{ij}, 1 \le i \le N_1, 1 \le j \le N_2\}$ and associate with them functions $K_1(\mathbf{x})$ and $f(\mathbf{x})$ and $L_1(\mathbf{y}, i)$ as there (we write K_1 and L_1 rather than K and L, for now these functions will represent only the first part of the CR).

   Then, by the proof of Theorem 145, for every pair $(\mathbf{x}, \mathbf{y})$ not in the exceptional set $A^c \cup B$ of arbitrarily small probability P_{XY}^n, $K_1(\mathbf{x}) = L_1\big(\mathbf{y}, f(\mathbf{x})\big)$.

Next, to each $\mathbf{u}_{ij}$ as above, we generate at random $\exp\big[n\big(I(V \wedge Y|U) + \delta\big)\big]$ sequences $\mathbf{v} \in V^n$ of joint ED with $\mathbf{u}_{ij}$ equal to P_{UV}, denoted as $\mathbf{v}_{k\ell}^{(ij)}$, $1 \le k \le M_1$, $1 \le \ell \le M_2$, where

$$M_1 = \exp\big[n\big(I(V \wedge Y|U) - I(V \wedge X|U) + 3\delta\big)\big], \; M_2 = \exp\big[n\big(I(V \wedge X|U) - 2\delta\big)\big]. \tag{74}$$

Then for every $\mathbf{y} \in \mathcal{Y}^n$ jointly UY-typical with $\mathbf{u}_{ij}$, the probability that neither $\mathbf{v}_{k\ell}^{(ij)}$ is jointly UYV-typical with $(\mathbf{u}_{ij}, \mathbf{y})$ is doubly exponentially small. Hence with probability close to 1, to every jointly typical pair $(\mathbf{u}_{ij}, \mathbf{y})$ there is a $\mathbf{v}_{k\ell}^{(ij)}$ such that $(\mathbf{u}_{ij}, \mathbf{y}, \mathbf{v}_{k\ell}^{ij})$ is jointly typical; we denote by $L(\mathbf{y}, \mathbf{u}_{ij})$ such a $\mathbf{v}_{k\ell}^{(ij)}$ (either one if there are several). Then for each $\mathbf{y}$ and $1 \le i \le N_1$ we take for $\mathbf{u}_{ij} = L_1(\mathbf{y}, i)$ the unique $\mathbf{u}_{ij}$ with the given first index i which is jointly typical with $\mathbf{y}$, or a constant if no or several such $\mathbf{u}_{ij}$ exist, and define $L_2(\mathbf{y}, i)$ as the $\mathbf{v}_{k\ell}^{(ij)}$ selected to this $\mathbf{u}_{ij}$ and $\mathbf{y}$:

$$L_2(\mathbf{y}, i) = L\big(\mathbf{y}, L_1(\mathbf{y}, i)\big) = \mathbf{v}_{k\ell}^{(ij)}. \tag{75}$$

Moreover, we define $g(\mathbf{y}, i)$ to equal the first index k of $\mathbf{v}_{k\ell}^{(ij)}$ in (75). Finally, for $\mathbf{x} \in \mathcal{X}^n$ and $1 \le k \le M_1$ we define $K_2(\mathbf{x}, k)$ as the unique $\mathbf{v}_{k\ell}^{(ij)}$ jointly typical with $(\mathbf{u}_{ij}, \mathbf{x})$, where $\mathbf{u}_{ij} = K_1(\mathbf{x})$, or set $K_2(\mathcal{X}, k) = \text{const}$ if no or several such $\mathbf{v}$ exist.

Then, by (74), g satisfies the rate constraint (9) if δ is sufficiently small. It is also clear that

$$K = \big(K_1(X^n), K_2\big(X^n, g(Y^n, f(X^n))\big)\big)$$
$$L = \big(L_1\big(Y^n, f(X^n)\big), L_2\big(Y^n, f(X^n)\big)\big)$$

represent a permissible pair for Model (iii), satisfying (5), and one shows as in the proof of Theorem 145 that

$$\frac{1}{n}H(L) \geq I(U \wedge X) + I(V \wedge Y|U) - \delta.$$

It remains only to show that the condition (4), i.e., $\Pr\{K = L\} > 1 - \varepsilon$ is also satisfied, at least with large probability with respect to the random selections. $\Pr\{K_1 = L_1\} > 1 - \varepsilon$ has already been demonstrated in the proof of Theorem 141. The remaining part $\Pr\{K_2 = L_2\} > 1 - \varepsilon$ can be proved similarly, though with a little more work.

Here we show that to any RV's U, V satisfying the Markov conditions (64) there exist $\tilde{U}, \tilde{V}$ satisfying the same conditions, with range sizes $|\tilde{\mathcal{U}}| \leq |\mathcal{X}|+2, |\tilde{\mathcal{V}}| \leq |\mathcal{Y}|$ such that

$$I(\tilde{U} \wedge X) - I(\tilde{U} \wedge Y) = I(U \wedge X) - I(U \wedge Y) \tag{76}$$

$$I(\tilde{V} \wedge Y|\tilde{U}) - I(\tilde{V} \wedge X|\tilde{U}) \leq I(V \wedge Y|U) - I(V \wedge X|U) \tag{77}$$

$$I(\tilde{U} \wedge X) + I(\tilde{V} \wedge Y|\tilde{U}) \geq I(U \wedge X) + I(V \wedge Y|U). \tag{78}$$

(i) Given U, V satisfying (64), introduce an equivalence relation on $\mathcal{U}$ by letting $u_1 \sim u_2$ iff $P_{X|U=u_1} = P_{X|U=u_2}$. Our first claim is that U, V can be replaced by U', V' without changing the relevant mutual information, such that no distinct elements of the range of U' are equivalent in the above sense.

Let $f(u)$ denote the equivalence class of u. Then clearly

$$P_{XY|U=u} = P_{XY|f(U)=f(u)} \tag{79}$$

hence

$$I(U \wedge X) = I\big(f(U) \wedge X\big),\ I(U \wedge Y) = I\big(f(U) \wedge Y\big).$$

This, in turn, implies that

$$\begin{aligned} I(V \wedge X|U) &= I(UV \wedge X) - I(U \wedge X) \\ &= I\big(UVf(U) \wedge X\big) - I\big(f(U) \wedge X\big) \\ &= I\big(UV \wedge X|f(U)\big), \end{aligned}$$

and similarly

$$I(V \wedge Y|U) = I\big(UV \wedge Y|f(U)\big).$$

Thus $U' = f(U)$, $V' = UV$ satisfy our claim, since $U' \ominus X \ominus Y$ is obvious from $U \ominus X \ominus Y$ and (79), and $X \ominus YU' \ominus V'$ follows as

$$P_{X|Y=y,f(U)=f(u),U=u,V=v} = P_{X|Y=y,U=u,V=v}$$
$$= P_{X|Y=y,U=u} = P_{X|Y=y,f(U)=f(u)},$$

where the second equality holds by $X \ominus YU \ominus V$ and the third by (79).

(ii) By (i), it suffices to consider U, V (satisfying (64)) such that the PD's $P_u = P_{X|U=u}, u \in \mathcal{U}$ are all distinct. This will enable us to use the Support Lemma (see [19] or [14], p. 310) to reduce the range size of U. To this end, define the stochastic matrix valued function $F(P)$ for $P \in \{P_u, u \in \mathcal{U}\}$ by letting $\Pr\{V = v|Y = y, U = u\}$ be the (y, v) entropy of $F(P_u)$. Then extend $F(P)$ continuously but otherwise arbitrarily to the set $\mathcal{P}(\mathcal{X})$ of all PD's on $\mathcal{X}$. Now apply the Support Lemma to the following continuous functions on $\mathcal{P}(\mathcal{X})$:

$$f_1(P) = H(X) - H(Y) - H(P) + H(PW), \quad \text{where } W = P_{Y|X}$$

$$f_2(P) = H(X) - H(P) + I\big(PW, F(P)\big)$$

$$f_3(P) = I\big(PW, F(P)\big) - I\big(P, WF(P)\big)$$

$$f_i(P) = P(x_{j-3}),\ 4 \le j \le |\mathcal{X}| + 2, \quad \text{where } \mathcal{X} = \{x_1, \ldots, x_{|\mathcal{X}|}\}.$$

It follows that there exist PD's $P_i \in \mathcal{P}(\mathcal{X})$, $i = 1, \ldots, |\mathcal{X}| + 2$, and a PD $\{\alpha_1, \ldots, \alpha_{|\mathcal{X}|+2}\}$ on $\{1, \ldots, |\mathcal{X}| + 2\}$ such that

$$\sum_{u \in \mathcal{U}} \Pr\{U = u\} f_j(P_u) = \sum_{i=1}^{|\mathcal{X}|+2} a_i f_j(P_i),\ j = 1, \ldots, |\mathcal{X}| + 2. \tag{80}$$

The last $|\mathcal{X}| - 1$ identities in (80) mean that a RV $\tilde{U}$ with range $\tilde{\mathcal{U}} = \{1, \ldots, |\mathcal{X}| + 2\}$ and distribution $\{\alpha_1, \ldots, \alpha_{|\mathcal{X}|+2}\}$ exists such that $P_{X|\tilde{U}=i} = P_i, i = 1, \ldots, |\mathcal{X}|+2$. Letting this $\tilde{U}$ satisfy $\tilde{U} \ominus X \ominus Y$, the first identity in (80) gives (76). Further letting $\tilde{V}$ be such that $X \ominus Y\tilde{U} \ominus \tilde{V}$, $P_{\tilde{V}|Y,\tilde{U}=i} = F(P_i)$, the second and third identities in (80) mean that (77) and (78) hold with equality.

(iii) Finally, it remains to show that $\tilde{V}$ in (ii) can be replaced by some $\tilde{V}'$ with range size $\le |\mathcal{Y}|$ and $X \ominus Y\tilde{U} \ominus \tilde{V}'$ such that $\tilde{V}'$ still satisfies (77), (78). Now, by the range constraint result of Theorem 145, applied to RV's with joint distribution $P_{YX|\tilde{U}=u}$ in the role of X, Y, for each fixed $\tilde{u} \in \tilde{\mathcal{U}}$ there exists $V_{\tilde{u}}$ distributed on a set of size $\le |\mathcal{Y}|$ and conditionally independent of X on the conditions

$Y = y, \tilde{U} = \tilde{u}$, for every $y \in \mathcal{Y}$, such that

$$I(V_{\tilde{u}} \wedge Y|\tilde{U} = \tilde{u}) \geq I(\tilde{V} \wedge Y|\tilde{U} = \tilde{u})$$

$$I(V_{\tilde{u}} \wedge Y|\tilde{U} = \tilde{u}) - I(V_{\tilde{u}} \wedge X|\tilde{U} = \tilde{u}) \leq I(\tilde{V} \wedge Y|\tilde{U} = \tilde{u}) - I(\tilde{V} \wedge X|\tilde{U} = \tilde{u}).$$

But then we can define a RV $\tilde{V}'$ with $X \ominus YU \ominus \tilde{V}'$ such that $P_{\tilde{V}'|Y=y,\tilde{U}=\tilde{u}} = P_{V_{\tilde{u}}|Y=y,\tilde{U}=\tilde{u}}$ for every $y \in \mathcal{Y}$, $\tilde{u} \in \tilde{\mathcal{U}}$. This $\tilde{V}'$ of range size $\leq |\mathcal{Y}|$ will satisfy the last inequalities for every $\tilde{u} \in \tilde{\mathcal{U}}$, and hence also (77), (78), as required. □

5 Common Randomness, Identification, and Transmission for Arbitrarily Varying Channels

Recall the definition of an AVC in Sect. 2 by a class $\mathcal{W} = \{W(\cdot|\cdot, s), s \in \mathcal{S}\}$ of channels $W(\cdot|\cdot, s)\colon \mathcal{X} \to \mathcal{Y}$. There also CR capacities, ID capacities, and transmission capacities have been defined for various models involving an AVC. We present now our results.

5.1 Model (A): AVC Without Feedback and Any Other Side Information

First we recall some well-known results for transmission capacities, cf. [14].

A random code $(\mathcal{C}_1, \cdots, \mathcal{C}_M, Q)$ is defined by deterministic codes $\mathcal{C}_1, \cdots, \mathcal{C}_M$ of the same block-length n and a PD Q on $\{1, \cdots, M\}$, with the understanding that $\mathcal{C}_i$ will be used with probability $Q(i)$. The error criterion is that the maximum or the average (for k) of $\sum_{i=1}^{M} Q(i)e_k(i, \mathbf{s})$ be small for every $\mathbf{s} \in \mathcal{S}^n$, where $e_k(i, \mathbf{s})$ denotes the probability of not decoding correctly the message k when the code $\mathcal{C}_i$ is used and the state sequence is $\mathbf{s}$. Both criteria lead to the same random code capacity C_{CR}. Notice that random codes can be used for transmission only if sender and receiver have access to CR, the outcome of a random experiment with distribution Q.

It was shown in [12] that

$$C_{\mathrm{CR}} = \max_{P} \min_{W \in \mathcal{W}} I(P, W) = \min_{W \in \mathcal{W}} C(W). \tag{81}$$

Here $I(P, W)$ denotes the mutual information of input and output RV's with joint distribution $P(x)W(y|x)$, $C(W) = \max_P I(P, W)$ is the Shannon capacity of the channel W, and $\mathcal{W}$ denotes the convex hull of $\mathcal{W}$.

By an elimination technique – based on an idea called now "derandomization" in computer science – it was shown in [1] that C_{CR} can be attained by random codes $(\mathcal{C}_1, \cdots, \mathcal{C}_M, Q)$ with M not larger than the square of the block-length n and with uniform Q. As a consequence, the capacity for deterministic codes and the average probability of error criterion, denoted by $\overline{C}$, satisfies

$$\overline{C} = C_{\mathrm{CR}} \qquad \text{if} \qquad \overline{C} > 0. \tag{82}$$

Random codes should be distinguished from codes with randomized encoding, which do not need CR, the decoding being deterministic. It was also shown in [1] that with randomized encoding, both the maximum and average error criteria lead to the same capacity, and

$$\text{capacity under randomized encoding} = \overline{C}. \tag{83}$$

We note for later reference that (82) and (83) remain valid also for AVC's with noiseless feedback, if $\overline{C}$ is replaced by $\overline{C}_f$, the average error capacity for deterministic codes with feedback.

A necessary and sufficient condition for $\overline{C} > 0$, given in [1], is that for some n there exist PD's Q_1, Q_2 on $\mathcal{X}^n$ and disjoint subsets D_1, D_2 of $\mathcal{Y}^n$ such that

$$\min_{\mathbf{s}\in\mathcal{S}^n} \sum_{\mathbf{x}\in\mathcal{X}^n} Q_i(\mathbf{x}) W^n(D_i|\mathbf{x},\mathbf{s}) > \frac{1}{2}, \quad i = 1, 2. \tag{84}$$

A single-letter necessary and sufficient condition for $\overline{C} > 0$ was given in [15]: $\overline{C} > 0$ iff $\mathcal{W}$ is not symmetrizable, where symmetrizability of $\mathcal{W}$ means the existence of a channel $U : \mathcal{X} \to \mathcal{S}$ such that

$$\sum_{s\in\mathcal{S}} U(s|x') W(y|x,s) = \sum_{s\in\mathcal{S}} U(s|x) W(y|x',s) \tag{85}$$

for every x, x' in $\mathcal{X}$ and y in $\mathcal{Y}$.

With these results and Theorems 143 and 144 the following theorem is readily obtained.

Theorem 149 *For an AVC without feedback, both ID capacity and CR capacity with sender permitted to randomize are equal to average error transmission capacity for deterministic codes:*

$$C_{\mathrm{ID}} = C_{\mathrm{CR}} = \overline{C}. \tag{86}$$

Their common value equals C_{CR} given by (81) if $\mathcal{W}$ is not symmetrizable, and 0 otherwise.

Proof

(i) $C_{\mathrm{CR}} = \overline{C}$: the non-trivial part $C_{\mathrm{CR}} \leq \overline{C}$ follows from Theorem 143 and (83). Indeed, a permissible pair (K, L) that satisfies (33) for every choice of $\mathbf{s} \in \mathcal{S}^n$ gives rise to a code with randomized encoder of rate $\frac{1}{n} \log |\mathcal{M}|$ and average probability of error $< 2\varepsilon$.

(ii) $C_{\mathrm{ID}} \geq C_{\mathrm{CR}} = \overline{C}$: In the non-trivial case $\overline{C} > 0$, this is a consequence of Theorem 144 and of the fact that $\overline{C}$ equals the maximum error capacity for randomized encoding.

(iii) $C_{\mathrm{ID}} \leq \overline{C}$: Notice that $C_{\mathrm{ID}} > 0$ implies $\overline{C} > 0$, because Q_1 and Q_2 as in (13) with $D_1' = D_1 \setminus D_2$, $D_2' = D_2 \setminus D_1$ satisfy (84) if $\varepsilon < 1/4$ in (13). Thus, on account of (82), it suffices to show that $C_{\mathrm{ID}} \leq C_{\mathrm{CR}}$. It follows from (13) that an (N, n, ε) ID code for the AVC is, for each $W \in \mathcal{W}$, also an (N, n, ε) code for the DMC $\{W\}$. Since the ID capacity of a DMC equals its transmission capacity, this and (81) imply the claimed inequality. □

5.2 Model (B): AVC with Noiseless (Passive) Feedback

Let C_{CRF} and C_{CRf} denote the CR capacity and C_{IDF} and C_{IDf} the identification capacity of the AVC with noiseless (passive) feedback, according as Terminal $\mathcal{X}$ is permitted to randomize or not. As now $\mathcal{X}$ knows everything that $\mathcal{Y}$ does, C_{CRF} equals the limit as $n \to \infty$ of the maximum, for all protocols as described in the passage containing Eq. (11), of

$$\frac{1}{n} \min_{\mathbf{s} \in \mathcal{S}^n} H(Y^n). \tag{87}$$

C_{CRf} is obtained similarly, with the maximum taken for the deterministic protocols (formally, with $M = \mathrm{const}$ in (11)).

Theorem 150 *For an AVC with noiseless feedback,*

$$C_{\mathrm{CRF}} = \max_P \min_{W \in \mathcal{W}} H(PW), \tag{88}$$

$$C_{\mathrm{CRf}} = \max_P \min_{W \in \mathcal{W}} H(W|P) \qquad \text{if } C_{\mathrm{CRf}} > 0, \tag{89}$$

$$C_{\mathrm{CRf}} > 0 \qquad \text{iff } \exists\, x \in \mathcal{X} : \min_{W \in \mathcal{W}} H(W(\cdot|x)) > 0. \tag{90}$$

Here $H(PW)$ and $H(W|P)$ denote the entropy $H(Y)$ and conditional entropy $H(Y|X)$ for RV's X, Y with joint distribution $P(x)W(y|x)$.

Remark These single-letter characterizations have been obtained independently also by Ning Cai (1995, Personal communication).

Proof of Theorem 150

(i) For a protocol that disregards the feedback information and selects i.i.d. inputs $X_1, \cdots, X_n$ with distribution P, the quantity (87) becomes $\min_{W \in \mathcal{W}} H(PW)$. This proves that the right hand side of (88) is an achievable CR rate. For the converse, we prove by induction that for any given protocol,

$$\min_{\mathbf{s} \in \mathcal{S}^k} H(Y^k) \le k \max_P \min_{W \in \mathcal{W}} H(PW), \tag{91}$$

for $k = 1, \cdots, n$. Indeed, (91) clearly holds for $k = 1$. Now, if $\min_{\mathbf{s} \in \mathcal{S}^k} H(Y^k)$ is attained for $\tilde{\mathbf{s}} = \tilde{s}_1 \cdots \tilde{s}_k$, let $\widetilde{P}$ denote the distribution of X_{k+1} when (the given protocol is used and) $s_i = \tilde{s}_i$, $i = 1, \cdots, k$. Then

$$\begin{aligned} \min_{\mathbf{s} \in \mathcal{S}^{k+1}} H(Y^{k+1}) &\le \min_{\mathbf{s} \in \mathcal{S}^{k+1}} (H(Y^k) + H(Y_{k+1})) \\ &\le \min_{\mathbf{s} \in \mathcal{S}^k} H(Y^k) + \min_{W \in \mathcal{W}} H(\widetilde{P}W). \end{aligned} \tag{92}$$

Hence (91) holds for $(k+1)$ if it does for k.

(ii) For a deterministic protocol, when X_i is a function of Y^{i-1}, we have

$$H(Y^n) = \sum_{i=1}^n H(Y_i|Y^{i-1}) = \sum_{i=1}^n H(Y_i|Y^{i-1}X_i) = \sum_{i=1}^n H(Y_i|X_i). \tag{93}$$

Using (93), an induction as above shows that the right hand side of (89) is an upper bound to (87) for any deterministic protocol.

Now, let P^* be the PD achieving the maximum in (89). Supposing $C_{\mathrm{CRf}} > 0$, it follows from Theorem 143 that to any $\varepsilon > 0$ there exists $k = k(\varepsilon)$, a protocol of block-length k, and a mapping f of $\mathcal{Y}^k$ into $\mathcal{X}$, such that the distribution of $f(Y^k)$ differs by less than ε from P^*, in variation distance, no matter what is the state sequence $\mathbf{s} \in \mathcal{S}^k$. We extend this protocol to block-length n, by letting $X_i = f(Y^k)$ for $i = k+1, \cdots, n$. Then, by (93), the limit of (87) as $n \to \infty$ will be arbitrarily close to the right hand side of (89), if $\varepsilon > 0$ is sufficiently small.

(iii) Obviously, the condition in (90) is sufficient for $C_{\mathrm{CRf}} > 0$. To prove its necessity, suppose indirectly that to each $x \in \mathcal{X}$ there is an $s = s(x)$ such that $W(\cdot|x, s)$ is the point mass at some $y = y(x)$. Given any deterministic protocol, consider $\mathbf{x} \in \mathcal{X}^n$, $\mathbf{s} \in \mathcal{S}^n$, $\mathbf{y} \in \mathcal{Y}^n$ defined recursively such that $s_i = s(x_i)$, $y = y(x_i)$, and x_{i+1} is the input symbol that the given protocol specifies when the past output sequence is $y_1 \cdots y_i$. For this particular state sequence $\mathbf{s}$, the given protocol leads to a unique output sequence $\mathbf{y}$, proving that quantity (87) is equal to 0 for every deterministic protocol, hence $C_{\mathrm{CRf}} = 0$. □

Our result on the CR capacity leads to a noticeable conclusion for the classical transmission problem.

Theorem 151 *The average error capacity* $\overline{C}_f$ *of an AVC with noiseless feedback, for deterministic coding, is always equal to* C_{CR} *given by* (81)*. Further,*

$$C_{\mathrm{IDF}} = C_{\mathrm{CRF}},\ C_{\mathrm{IDf}} = C_{\mathrm{CRf}} \qquad \text{if } C_{\mathrm{CR}} > 0 \tag{94}$$

$$C_{\mathrm{IDF}} = C_{\mathrm{IDf}} = 0 \qquad \text{if } C_{\mathrm{CR}} = 0. \tag{95}$$

Proof

(i) The random code $(\mathcal{C}_1, \cdots, \mathcal{C}_M, Q)$ in the paragraph containing Eq. (82) can be used for transmission if $\mathcal{X}$ and $\mathcal{Y}$ have access to $2 \log n$ bits of robust UCR, i.e., to RV's K, L satisfying (33) for every $\mathbf{s} \in \mathcal{S}^n$ with $|\mathcal{M}| = n^2$. Since $C_{\mathrm{CR}} > 0$ implies $C_{\mathrm{CRF}} > 0$, cf. (81), (88), such UCR may be generated using a protocol of block-length $n' = c \log n$, by Theorem 143. This proves that C_{CR} is an achievable transmission rate, at least if randomization is permitted (randomization may be needed in the CR-generating protocol of negligible block-length $n' = c \log n$, whose outcome will identify the $\mathcal{C}_i$ actually used). The proof is completed by reference to the feedback versions of (82) and (86).

(ii) If $\overline{C}_f = C_{\mathrm{CR}} > 0$, the inequalities $C_{\mathrm{IDF}} \geq C_{\mathrm{CRF}}$, $C_{\mathrm{IDf}} \geq C_{\mathrm{CRf}}$ are proved analogously to the proof of Theorem 149, part (ii). The reversed inequalities follow by the method of chapter [9] (2), where the ID capacity of a DMC with feedback has been determined. If $C_{\mathrm{CR}} = 0$ then $C(W) = 0$ for some $W \in \mathcal{W}$. Then the feedback ID capacity of the DMC $\{W\}$ is 0 by [9], and (95) follows. □

5.3 Model (C): Strongly Arbitrarily Varying Channel (SAVC)

It is assumed here that the jammer can make his choice of $\mathbf{s} \in \mathcal{S}^n$ depend on the sent $\mathbf{x} \in \mathcal{X}^n$. Formally, the parameter determining the statistics is now an arbitrary mapping from $\mathcal{X}^n$ to $\mathcal{S}^n$.

Since the number of such mappings is doubly exponential in n, the hypothesis of Theorem 143 is still satisfied. The criterion (12) for an (N, n, ε) ID code becomes

$$\sum_{\mathbf{x} \in \mathcal{X}^n} Q_j(\mathbf{x}) \max_{\mathbf{s}} W^n(D_j^c | \mathbf{x}, \mathbf{s}) \leq \varepsilon, \quad \sum_{\mathbf{x} \in \mathcal{X}^n} Q_k(\mathbf{x}) \max_{\mathbf{s}} W^n(D_j | \mathbf{x}, \mathbf{s}) \leq \varepsilon. \tag{96}$$

The first inequalities here (with disjoint sets D_j) represent the maximum probability of error criterion for transmission codes with randomized encoding.

Any (N, n, ε) transmission code with randomized encoding gives rise to a deterministic (N, n, ε) code, with codewords $\mathbf{x}_j = \arg\min_{\mathbf{x}} (\max_{\mathbf{s}} W^n(D_j^c | \mathbf{x}, \mathbf{s}))$. Hence the maximum error capacity of a SAVC for deterministic and randomized

encoding is the same. It is also (well known and) easy to see that this capacity coincides with the average error capacity for deterministic codes, and it equals the maximum error capacity for deterministic codes of the AVC defined by the same $\mathcal{W}$. We shall denote this capacity by $\overline{\overline{C}}$. As shown in [18], $\overline{\overline{C}} > 0$ iff there exists x and x' in $\mathcal{X}$ with $\mathcal{T}(x) \cap \mathcal{T}(x') = \emptyset$ where $\mathcal{T}(x)$ denotes the convex hull of the set of PD's $W(\cdot|x, s)$, $s \in \mathcal{S}$.

The row-convex hull $\overline{\mathcal{W}}$ of $\mathcal{W}$ is the set of all channels $W\colon \mathcal{X} \to \mathcal{Y}$ such that $W(\cdot|x) \in \mathcal{T}(x)$, $x \in \mathcal{X}$. Write

$$D = \min_{W \in \overline{\mathcal{W}}} C(W). \tag{97}$$

Theorem 152 *For a SAVC, the CR capacity $C_{\mathrm{CR}}^{\mathrm{s}}$ and ID capacity $C_{\mathrm{ID}}^{\mathrm{s}}$ (with $\mathcal{X}$ permitted to randomize) satisfy*

$$\overline{\overline{C}} \le C_{\mathrm{CR}}^{\mathrm{s}} \le C_{\mathrm{ID}}^{\mathrm{s}} \le D \tag{98}$$

$$C_{\mathrm{ID}}^{\mathrm{s}} > 0 \qquad \text{iff } \overline{\overline{C}} > 0. \tag{99}$$

Remark Under not too restrictive hypotheses, $\overline{\overline{C}} = D$, cf. [2] for $\mathcal{W}$ satisfying $\mathcal{T}(x) \cap \mathcal{T}(x') = \emptyset$ whenever $x \neq x'$, and [13] under a weaker hypothesis; there are, however, examples of $0 < \overline{\overline{C}} < D$. For $\mathcal{W}$ with $\overline{\overline{C}} = D$, Theorem 152 gives a conclusive result, but we do not know whether $C_{\mathrm{ID}}^{\mathrm{s}} = C_{\mathrm{CR}}^{\mathrm{s}}$ and/or $C_{\mathrm{CR}}^{\mathrm{s}} = \overline{\overline{C}}$ hold for every SAVC. $C_{\mathrm{CR}}^{\mathrm{s}}$ always equals the average error capacity for randomized encoding, but it appears unknown whether the latter can ever be larger than $\overline{\overline{C}}$.

Proof of Theorem 152 The first inequality of (98) is obvious, and if $\overline{\overline{C}} > 0$, the second inequality follows from Theorem 144. It remains to prove that $C_{\mathrm{ID}}^{\mathrm{s}} \le D$ and that $\overline{\overline{C}} = 0$ implies $C_{\mathrm{CR}}^{\mathrm{s}} = C_{\mathrm{ID}}^{\mathrm{s}} = 0$.

Consider an auxiliary model where at each instant i the state s_i may depend on x_i but not on the other x_j's. Formally, this is an AVC model, with state set $\mathcal{S}^*$ consisting of all mappings $s^*\colon \mathcal{X} \to \mathcal{S}$, defined by the set of channels

$$\mathcal{W}^* = \{W^*(\cdot|\cdot, s^*), s^* \in \mathcal{S}^*\}, \quad W^*(\cdot|x, s^*) = W(\cdot|x, s^*(x)). \tag{100}$$

Clearly, the CR and ID capacities of this AVC are upper bounds to $C_{\mathrm{CR}}^{\mathrm{s}}$ and $C_{\mathrm{ID}}^{\mathrm{s}}$. Thus, on account of Theorem 149, it suffices to show that (i) the random code capacity of the AVC defined by (100) equals D and (ii) $\mathcal{W}^*$ is symmetrizable if $\overline{\overline{C}} = 0$.

(i) is obvious from (81) and (97) since $\overline{\mathcal{W}^*} = \overline{\mathcal{W}}$.

To prove (ii), use either the Strong Separation Lemma of [1] or, alternatively, recall that $\overline{\overline{C}} = 0$ iff $\mathcal{T}(x) \cap \mathcal{T}(x')$ is never empty, i.e., for suitable PD's $U(\cdot|x, x')$

on $\mathcal{S}$,

$$\sum_{s\in\mathcal{S}} U(s|x,x')W(y|x,s) = \sum_{s\in\mathcal{S}} U(s|x',x)W(y|x',s) \tag{101}$$

for every x, x' and y. (101) means that $\mathcal{W}^*$ satisfies (85), with $U^*\colon \mathcal{X} \to \mathcal{S}^*$ defined by

$$U^*(s^*|x) = \prod_{\tilde{x}\in\mathcal{X}} U(s^*(\tilde{x})|\tilde{x},x). \tag{102}$$

□

Remark Work relevant for problems concerning feedback with *noise* can be found in [11] (see chapter "Identification via Channels with Noisy Feedback", Part I).

References

1. R. Ahlswede, Elimination of correlation in random codes for arbitrarily varying channels. Zeitschrift Wahrscheinlichkeitstheorie und verw. Geb. **33**, 159–175 (1978)
2. R. Ahlswede, A method of coding and an application to arbitrarily varying channels. J. Combin. Inf. Syst. Sci. **5**, 1035 (1980)
3. R. Ahlswede, Coloring hypergraphs: a new approach to multiuser source coding. J. Combin. Inf. Syst. Sci. **1**, 76–115 (1979) and **2**, 220–268 (1980)
4. R. Ahlswede, V.B. Balakirsky, Identification under random processes. Preprint 95–098, SFB 343, "Diskrete Strukturen in der Mathematik", Universität Bielefeld. Problemy peredachii informatsii (special issue devoted to M.S. Pinsker), **32**(1), 144–160 (1996)
5. R. Ahlswede, N. Cai, Z. Zhang, On interactive communication. Preprint 93–066, SFB 343 "Diskrete Strukturen in der Mathematik", Universität Bielefeld. IEEE Trans. Inf. Theory **43**(1), 22–37 (1997)
6. R. Ahlswede, I. Csiszár, Common randomness in information theory and cryptography, Part I: Secret sharing. IEEE Trans. Inf. Theory **39**, 1121–1132 (1993)
7. R. Ahlswede, I. Csiszár, Common randomness in information theory and cryptography, Part II: CR capacity. IEEE Trans. Inf. Theory **44**(1), 225–240 (1998)
8. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**, 15–29 (1989)
9. R. Ahlswede, G. Dueck, Identification in the presence of feedback – a discovery of new capacity formulas. IEEE Trans. Inf. Theory **35**, 30–39 (1989)
10. R. Ahlswede, B. Verboven, On identification via multi-way channels with feedback. IEEE Trans. Inf. Theory **37**, 1519–1526 (1991)
11. 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. Inf. Theory **41**(4), 1040–1050 (1995)
12. D. Blackwell, L. Breiman, A.J. Thomasian, The capacities of certain channel classes under random coding. Ann. Math. Stat. **31**, 558–567 (1960)
13. I. Csiszár, J. Körner, On the capacity of the arbitrarily varying channels for maximum probability of error. Z. Wahrscheinlichkeitsth. Verw. Gebiete **57**, 87–101 (1981)
14. I. Csiszár, J. Körner, *Information Theory: Coding Theorem for Discrete Memoryless Systems* (Academic, New York, 1982)
15. I. Csiszár, P. Narayan, The capacity of the arbitrarily varying channel revisited: positivity, constraints. IEEE Trans. Inf. Theory **34**, 181–193 (1988)

16. I. Csiszár, P. Narayan, Capacity of the Gaussian arbitrarily varying channel. IEEE Trans. Inf. Theory **37**, 18–26 (1991)
17. P. Gács, J. Körner, Common information is far less than mutual information. Probl. Control Inf. Theory **21**, 149–162 (1973)
18. J. Kiefer, J. Wolfowitz, Channels with arbitrarily varying channel probability functions. Inf. Control **5**, 44–54 (1962)
19. M. Salehi, Cardinality bounds on auxiliary variables in multiple–user theory via the method of Ahlswede and Körner, Stanford Technical Report (1978)
20. D. Slepian, J.K. Wolf, Noiseless coding of correlated information sources. IEEE Trans. Inf. Theory **19**, 471–480 (1973)
21. S. Venkatesan, V. Anantharam, The common randomness capacity of independent discrete memoryless channels, Memorandum No. UCB/ERL M95/85 (1995)

Watermarking Identification Codes with Related Topics on Common Randomness

Watermarking identification codes were introduced by Steinberg and Merhav. In their model they assumed that

1. the attacker uses a single channel to attack the watermark and both, the information hider and the decoder, know the attack channel;
2. the decoder either completely knows the covertext or knows nothing about it.

Then instead of the first assumption they suggested to study more robust models and instead of the second assumption they suggested to consider the case where the information hider is allowed to send a secret key to the decoder according to the covertext.

In response to the first suggestion in this lecture (see [6]) we assume that the attacker chooses an unknown (for both information hider and decoder) channel from a set of channels or a compound channel, to attack the watermark. In response to the second suggestion we present two models. In the first model according to the output sequence of covertext the information hider generates side information componentwise as the secret key. In the second model the only constraint to the key space is an upper bound for its rate.

We present lower bounds for the identification capacities in the above models, which include the Steinberg and Merhav results on lower bounds. To obtain our lower bounds we introduce the corresponding models of common randomness. For the models with a single channel, we obtain the capacities of common randomness. For the models with a compound channel, we have lower and upper bounds and the differences of lower and upper bounds are due to the exchange and different orders of the max-min operations.

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_16

1 Introduction

Watermarking technique is a way to embed secret information into a given message, say image, that cannot be removed nor deciphered without access to a secret key. It can be used to protect copy right. Watermarking is now a major activity in audio, image, and video processing and standardization efforts for JPEG-2000, MPEG-4 and Digital Video Disks are underway.

One way to analyze watermarking problems is to regard them as communication systems e.g., [14, 17, 22–25, 27]. In these systems the messages, which are called covertext, are generated by an information source. An information hider, whom we often call encoder because of his role in the system, has full access to the information source of covertexts and the set of secret messages. These secret messages are independent of the covertext, they are uniformly generated from the set, and will be called watermark. The role of the information hider, or encoder, is to embed the watermark in the covertext. When the embedding changes the covertext, it disturbs the message. To guarantee the quality of the watermarked message, we certainly would like not too much distortion. That is, for a given distortion measure, the distortion between the original covertext and the watermarked message in average may not exceed a given constant. An attacker wants to remove the watermark from the watermarked message without distorting the message too much i.e., the distortion between the covertext and the message corrupted by the attacker is not too large with respect to a certain distortion measure. Finally a decoder tries to recover the watermark from the corrupted message correctly with high probability. As the attacker is allowed to use a random strategy, we assume that the attacker uses a noisy channel to attack the watermark. Depending on the models the attacker may choose various channels and the encoder and decoder share different resources (e.g., secret key, side information, etc.).

Among huge contributions on watermarking we here briefly review two of them. In [23] Moulin and O'sullivan obtained the capacity for the watermarking codes under the assumptions that the covertexts are generated from a memoryless source, the distortions are sum-type and the attack channels are compound channels whose states are known to the decoder but unknown to the encoder. The strategies of encoder-decoder and attacker are discussed as well.

Identification codes for noisy channels were introduced by Ahlswede and Dueck for the situation in which the receiver needs to identify whether the coming message equals a specified one. If not, then they don't care what it is [9]. It turned out that this weaker requirement dramatically increased the sizes of messages sets which could be handled: double exponential grown in the block lengths of codes. Identification is much faster than transmission!

Steinberg and Merhav notice that in most cases people check watermarks in order to identify them (e.g. copyright) rather than recognize them and so they introduced identification codes to watermarking models [27]. In their models the attack channels are single memoryless channels. That means the attacker's random strategy is known by information hider (encoder) and decoder. They notice that the

assumption is not robust and so suggested to study more robust models. As to the resources shared by encoders and decoders they consider two cases, the decoder either completely knows the covertext or he knows nothing about it. (In all cases the attacker must not know the covertext because otherwise there would be no safe watermarking.)

By considering common randomness between encoder and decoder, they obtained lower bounds to the capacities of watermarking identification in both cases and the upper bounds easily followed from a theorem in [26]. The lower and upper bounds are tight in the former case but not in the latter case. As Steinberg and Merhav only studied two extremal cases, they suggested to consider the more general case, that the decoder may obtain partial information, about the covertext, say key, from the encoder via a secure noiseless channel. The exponent of error probability was discussed as well.

In the present lecture we deal with these two problems. But before turning to our result, we draw readers' attention to common randomness, which – as noticed in [10] (chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas", Part I) – plays a central role in identification problems. It does so also in [27] and here Ahlswede and Dueck discovered in [10] that common randomness shared by encoder and decoder can be used to construct identification codes and therefore the rate of common randomness (in the sense of first order of logarithm) is not larger than the rate of identification codes (in the sense of the second order of logarithm). In general the capacities of common randomness shared by the encoder and the decoder may be smaller than the capacities of identification. Examples for discrete channels and Gaussian channels were presented in [12] and [15] respectively. Notice that the sizes of the input alphabets of the former channel is growing super exponentially as the length of codes and the sizes of the input alphabets of the latter is infinity. In fact it is seen from [26] that for any channel, whose input alphabet is exponentially increasing in the case that strong converse holds, the rates of common randomness and identification codes are the same.

The topic of common randomness has been become more and more popular e.g., [4, 7, 8, 19, 21, 28, 29], etc. Common randomness may be applied to cryptography, (e.g., [7, 16, 19, 21]), identification (e.g., [3, 8–10, 13, 16]), and arbitrarily varying channels (e.g., [1, 2, 5, 8]). For the first two applications the rates are important and the distributions of common randomnesses are required nearly uniformly. For cryptography certain secure conditions additionally needed. For the last application one has to face in the difficulty made by the jammer and find a smart way to generate the common randomness.

Now let us return to the two suggestions by Steinberg and Merhav. For the first suggestion we assume in our models, attackers are allowed to choose a channel arbitrarily from a set of memoryless channels to attack watermarks and neither encoders nor decoders know the attack channels. This is known as compound channel in Information Theory.

The assumption makes our models slightly more robust than that in [23] since in [23] the decoders are supposed to know the attack channels.

For the second suggestion we set up two models. In out first model we assume the encoder generates a RV at time t according to component at time t of the output sequence of covertext source and certain probability and sends it to decoder via a secure channel. In this case the "key" actually is a side information of covertext shared by encoder and decoder. We obtain the first and the second models in [27] if we choose the side information equal to covertext almost surely and independent of covertext respectively. So our first model contains both models in [27]. In our second model the encoder is allowed to generate a key according to the covertext (but independently on watermark) in arbitrary way and sends the key to decoder through a secure channel with rate R_{K}. Obviously in our second model the key can be generated in a more general way than in our first model. For all combinations of above assumptions, we obtain lower bounds to the identification capacities, which contains both lower bounds in [27] as special cases.

To obtain our lower bounds to identification capacities, for each combination, we introduce a corresponding model of common randomness and obtain lower and upper bound to its capacity. For the single channel, the two bounds agree. For compound channel, the gap between two bounds is up to the order of max-min. In addition, we show a lower bound to common randomness in [27] in fact is tight, which supports a conjecture in [27].

We must point out that our assumption of compound attack channels is still far from the most robust and practical assumption although according to our knowledge, it is most robust and practical existing assumption in this area. Actually the attacker has much more choices.

- He does not necessarily use a memoryless channel and stead he can chooses a channel with finite memory.
- The attacker may change the states time by time i.e., he may use an arbitrarily varying channel.
- The attacker knows output of the channel; even at time t, he know the output at time $t' > t$, since all outputs in fact are chosen by himself/herself. So the attacker may use this information to choose attack channel. This clearly makes the attack much more efficient.

So there is still a long way for us to achieve the most practical results and it provides a wide space for future research.

The rest part of the lecture is organized as follows. In the next section we present the notation used in the lecture. Our models and results are stated in Sects. 3 and 4 respectively. The direct parts of coding theorems of common randomness are proven in Sect. 5 and their converse parts are proven in Sect. 6. In Sect. 7 we briefly review the observation in [10] (chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas") on the relation of identification and common randomness and therefore the lower bounds to the identification capacities from capacities of common randomness. Finally the converse theorem for a model in [27] is proven in Sect. 8.

2 The Notation

Our notation in this lecture is fairly standard. log and ln stand for the logarithms with bases 2 and e respectively and a^z is often written as $exp_a[z]$. The RV's will be denoted by capital letters L, U, V, X, Y, Z etc. and their domains are often denoted by the correspondent script letters $\mathcal{L}, \mathcal{U}, \mathcal{V}, \mathcal{X}, \mathcal{Y}, \mathcal{Z}$ etc. But in some special cases it may be exceptional. When we denote a set by a script letter (for example, $\mathcal{X}$), its element is often denoted by the corresponding lower letter (for example x). $\mathcal{X}^n$ is the nth Cartesian power of the set $\mathcal{X}$ and $x^n = (x_1, x_2, \ldots x_n)$ is the sequence of length n. $\Pr\{\mathcal{E}\}$ is the probability of that the event $\mathcal{E}$ occurs and $\mathbb{E}[\cdot]$ is the operator of expectation. P_X, P_{XY}, $P_{Z|X}$ etc. will stand for the distribution of RV X, the joint distribution of the RV's (X, Y), the conditional distribution of RV Z under the condition that X is given respectively. When we write a probability distribution as P^n, we mean that it is a product distribution of P and similarly a discrete memoryless channel of length n with stochastic matrix W is written as W^n.

Throughout this lecture $\mathcal{T}^n_U$, $\mathcal{T}^n_{UV}$, $\mathcal{T}^n_{U|VL}(v^n l^n)$ etc. will denote the sets of typical, joint typical, and conditional typical sequences and the corresponding sets of δ- typical, joint typical, and conditional typical sequences are written as $\mathcal{T}^n_U(\delta)$, $\mathcal{T}^n_{UV}(\delta)$, $\mathcal{T}^n_{U|VL}(v^n l^n, \delta)$ etc. We always understand these sets are not empty when we use the notation. When we introduce a set of typical sequences (for example, say $\mathcal{T}^n_Z$), it is understood that the correspondent RV(s) (i.e., Z in the example) with the (joint) empirical distribution (ED) as distribution (P_Z) is introduced at the same time. For a subset $\mathcal{A}$ of sequences of length n we write $\mathcal{A}_U = \mathcal{A} \cap \mathcal{T}^n_U$ and analogously $\mathcal{A}_{UV}$, $\mathcal{A}_{U|VL}(v^n l^n)$, $\mathcal{A}_U(\delta)$, $\mathcal{A}_{UV}(\delta)$, $\mathcal{A}_{U|VL}(v^n l^n, \delta)$ etc. $|\mathcal{T}^n_U|$ and the common values of $|\mathcal{T}^n_{U|L}(l^n)|$, $l^n \in \mathcal{T}^n_L$ some times are written as t_U, $t_{U|L}$ etc. respectively (the length n of the sequences are understood by the context). Analogously $t_U(\delta)$, $t_{Y|X}(\delta)$ etc., also are used.

3 The Models

3.1 Watermarking Identification Codes

In this subsection, we state our models for the simpler case that the attacker chooses a single channel to attack the watermark and both the encoder (information hider) and the decoder know the attack channel. In the next subsection, we introduce the corresponding models of common randomness. In the last subsection of the section, we assume the attack chooses a channel unknown by both encoder and decoder from a set of channels and replace the single channel by a compound channel.

Let $\mathcal{V}$ be a finite set, and V be a RV taking values in $\mathcal{V}$. Then the covertext is assumed to be generated by an memoryless information source $\{V^n\}_{n=1}^{\infty}$ with generic V. The watermark is uniformly chosen from a finite set $\{1, 2, \ldots, M\}$ independently on the covertext. The encoder is fully accessed to the covertext and

source of watermark and encodes the outputs of covertext v^n and of watermark m jointly to a sequence $x^n\big(=x^n(v^n,m)\big)$ with the same length of sequence of covertext. The attacker uses a single discrete memoryless channel W to attack the watermarked sequence x^n i.e., to change x^n to y^n with probability $W^n(y^n|x^n)=\prod_{t=1}^{n}W(y_t|x_t)$. Usually for practical reason people assume that v^n, x^n, and y^n are chosen from the same finite alphabet, but for convenience of notation we assume they are from finite alphabets $\mathcal{V}$, $\mathcal{X}$, and $\mathcal{Y}$ respectively. The encoding mapping in general disturbs the covertext. To measure the distortion, we introduce a sum type distortion measure, watermarking distortion measure (WD-measure) ρ, such that for all $v^n=(v_1,\ldots,v_n)\in\mathcal{V}^n$, $x^n=(x_1,\ldots,x_n)\in\mathcal{X}^n$,

$$\rho(v^n,x^n)=\sum_{t=1}^{n}\rho(v_t,x_t), \tag{1}$$

where for all $v\in\mathcal{V}$, $x\in\mathcal{X}$ $0\le\rho(v,x)\le\Delta$, for a positive constant Δ.

By definition, there should be certain distortion constraint to the output of attack channel. But now we are given a memoryless attack channel and we may omit the constraint simply by assuming that the attack channel satisfies the constraint automatically. This clearly does not loss generality. Next we have to set up the key-resources shared by encoder and decoder, according to which we distinguish our watermarking identification codes into watermarking identification codes with side information (WIDSI codes) and watermarking identification codes with secure key (WIDK codes) as follows.

Watermarking identification codes with side information (WIDSI codes): In the first case, we assume that the encoder can generate "a component of a key", $L_t=l_t$ at the time t according to the current output of covertext $V_t=v_t$ and a given conditional distribution $P_{L|V}(\cdot|v)$. That is, the sender generates a sequence $L^n=(L_1,L_2,\ldots,L_n)=l^n=(l_1,l_2,\ldots,l_n)$ with probability $P^n_{L|V}(l^n|v^n)$ if the source outputs a sequence v^n of covertext and then sends it to the decoder. The latter try to recover the watermark from the invalidated message by the attacker with the help of the side information $L^n=l^n$. In this case the key-resource is actually governed by the conditional distribution $P_{L|V}$ or equivalently the joint probability distribution P_{VL}. So it can be understood as a pure side information at both sides of encoder and decoder instead of a " secure key". That is, if $\{V^n\}_{n=1}^{\infty}$ is a memoryless covertext with generic V, and $\{L^n\}_{n=1}^{\infty}$ is a side information observed by both encoder and decoder, then $\{(V^n,L^n)\}$ is a correlated memoryless source with generic (V,L). Thus the decoder can learn some thing about the covertext from the side information whereas the attacker knows nothing about it. A WIDSI code becomes a "watermarking identification code with side information at transmitter and receiver" in [27] when V and L have the same alphabet and equal to each other almost surely and it becomes a "watermarking identification code with side information at the transmitter only" in [27] if V and L are independent.So the two codes defined in [27] are really the extreme cases of WIDCI codes.

Watermarking identification codes with secure key (WIDK codes): In this case we assume the encoder may generate a key $K_n = K_n(v^n)$ according to the whole output sequence $V^n = v^n$ of the random covertext V^n in an arbitrary way and send it to the decoder through a secure (noiseless) channel so that the attacker has absolutely no knowledge about the covertext (except its distribution) nor the key. Since for given output v^n of the covertext the encoder may generate the K_n randomly, a WIDSI code is a special WIDK code. We shall see that in general the latter is more powerful. Notice that a deterministic key function of output of covertext is a special random key. Finally of course the size of the key must be constraint. We require it exponentially increasing with the length of the code and its rate upper bounded by the key rate R_{K}. When the key rate is larger than the covertext entropy $H(V)$ the encoder certainly may inform the receiver about the output of covertext. However "the rest part" of the key may serve as a common randomness between the communicators which increases the identification capacity (see chapters "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" and "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints", and [27]).

Definition 153 An $(n, R, \lambda_1, \lambda_2, D_1)$ WIDSI code is a system $\{Q_m, \mathcal{D}_m(l^n) : l^n \in \mathcal{L}^n, m \in \mathcal{M}\}$ for $\mathcal{M} = \{1, 2, \ldots, M\}$ satisfying the following conditions.

- $Q_m, m = 1, 2, \ldots, M$ are stochastic matrices $Q_m : \mathcal{V}^n \times \mathcal{L}^n \longrightarrow \mathcal{X}^n$ such that for $m = 1, 2, \ldots, M$,

$$\sum_{v^n \in \mathcal{V}^n, l^n \in \mathcal{L}^n} P_{VL}^n(v^n, l^n) \sum_{x^n \in \mathcal{X}^n} Q_m(x^n|v^n, l^n)\rho(v^n, x^n) \le D_1, \tag{2}$$

  where P_{VL} is the joint distribution of the generic (V, L).
- For all $l^n \in \mathcal{L}^n, m \in \mathcal{M}, \mathcal{D}_m(l^n) \subset \mathcal{Y}^n$ and for all $m \in \mathcal{M}$,

$$\sum_{v^n \in \mathcal{V}^n, l^n \in \mathcal{L}^n} P_{VL}^n(v^n, l^n) \sum_{x^n \in \mathcal{X}^n} Q_m(x^n|v^n, l^n) W^n(\mathcal{D}_m(l^n)|x^n) > 1 - \lambda_1, \tag{3}$$

  and for all $m, m' \in \mathcal{M}, m \ne m'$,

$$\sum_{v^n \in \mathcal{V}^n, l^n \in \mathcal{L}^n} P_{VL}^n(v^n, l^n) \sum_{x^n \in \mathcal{X}^n} Q_m(x^n|v^n, l^n) W^n(\mathcal{D}_{m'}(l^n)|x^n) < \lambda_2. \tag{4}$$

  λ_1 and λ_2 is called the errors of the first and the second kinds of the code
- The rate of the code is

$$R = \log\log M. \tag{5}$$

Next we define WIDK code.

Definition 154 Let $\{V^n\}_{n=1}^{\infty}$ be a memoryless covertext with generic V and alphabet $\mathcal{V}$, the attack channel W be memoryless, and WD-measure ρ be as (1). Then an $(n, R, R_{\mathrm{K}}, \lambda_1, \lambda_2, D_1)$ WIDK code is a system $\{Q_m^*, \mathcal{D}_m^*(k_n), W_{K_n} : m \in \mathcal{M}, k_n \in \mathcal{K}_n\}$ for $\mathcal{M} = \{1, 2, \ldots, M\}$ satisfying the following conditions.

– $\mathcal{K}_n$ is a finite set, which will be called the key book, with

$$\frac{1}{n} \log |\mathcal{K}_n| \leq R_{\mathrm{K}}. \tag{6}$$

R_{K} will be called key rate.

– W_{K_n} is a stochastic matrix, $W_{K_n} : \mathcal{V}^n \longrightarrow \mathcal{K}_n$. The output RV will be denoted by K_n when the random covertext V^n is input to the channel W_{K_n} i.e., the pair of RV's (V^n, K_n) have joint distribution $P_{V^n K^n}(v^n, k^n) = P_V^n(v^n) W_{K_n}(k_n|v^n)$, $v^n \in \mathcal{V}$, $k_n \in \mathcal{K}_n$. In particular K_n may be a deterministic function of output of covertext and in this case we write $K(\cdot)$ as a function defined on $\mathcal{V}^n$. Note that the choice of K_n does NOT depend on the message $m \in \mathcal{M}$ since the key should independent of the protected message.

– $Q_m^*, m = 1, 2, \ldots, M$ are stochastic matrices from $\mathcal{V}^n \times \mathcal{K}_n$ to $\mathcal{X}^n$, (the alphabet of the input of the attack channel), such that

$$\sum_{v^n \in \mathcal{V}^n} P_V^n(v^n) \sum_{k_n \in \mathcal{K}_n} W_{K_n}(k_n|v^n) \sum_{x^n \in \mathcal{X}^n} Q_m^*(x^n|v^n, k_n) \rho(v^n, x^n) \leq D_1. \tag{7}$$

– For all $k_n \in \mathcal{K}_n, m \in \mathcal{M}$, $\mathcal{D}_m(k_n) \subset \mathcal{Y}^n$ and for all $m \in \mathcal{M}$, the error of first kind satisfies

$$\sum_{v^n \in \mathcal{V}^n} P_V^n(v^n) \sum_{k_n \in \mathcal{K}_n} W_{K_n}(k_n|v^n) \sum_{x^n \in \mathcal{X}^n} Q_m^*(x^n|v^n, k_n) W^n(\mathcal{D}_m(k_n)|x^n) > 1 - \lambda_1 \tag{8}$$

and for all $m, m' \in \mathcal{M}$ $m \neq m'$,

$$\sum_{v^n \in \mathcal{V}^n} P_V^n(v^n) \sum_{k_n \in \mathcal{K}_n} W_{K_n}(k_n|v^n) \sum_{x^n \in \mathcal{X}^n} Q_m^*(x^n|v^n, k_n) W^n(\mathcal{D}_{m'}(k_n)|x^n) < \lambda_2. \tag{9}$$

– Finally the rate of the code is defined in (5).

The capacities of the codes of the two types are defined in the standard way and denoted by $C_{\mathrm{WIDSI}}((V, L), W, D_1)$ and $C_{\mathrm{WIDK}}(V, W, R_{\mathrm{K}}, D_1)$ respectively, where (V, L) and V are the generic of memoryless correlated source and source respectively, W is an attack memoryless channel, R_{K} is the key rate, and D_1 is the distortion criterion.

3.2 The Common Randomness

We speak of the common randomness between two (or among more than two) persons who share certain common resources, which may be correlated sources and/or (noisy or noiseless) channels.The common randomness capacity is defined as the maximum number of random bits per channel use that the two persons can generate. According to the resources, different models are established.

To build watermarking identification codes we need the following two kinds of common randomness. In the following two models of common randomness, the correlated source $\{(V^n, L^n)\}_{n=1}^{\infty}$ corresponds to the source of covertext and side information and the memoryless channel W corresponds to the attack channel in the models of watermarking identification. The K_n in the Model II corresponds to the key in the model of WIDK.

Model I: Two-source with a constraint noisy channel

Let $\{(V^n, L^n)\}_{n=1}^{\infty}$ be a correlated memoryless source with two components, alphabets $\mathcal{V}$ and $\mathcal{L}$, and generic (V, L). Assume that there are two persons, say sender (or encoder) and receiver (or decoder). The sender may observe the whole output of the source (V^n, L^n) whereas only the output of the component L^n is observable for the receiver. To establish common randomness the sender may send message through memoryless channels W with input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ under certain constraint condition (specified below). The receiver is not allowed to send any message to the sender. The sender first chooses a channel code with set of codewords $\mathcal{U} \subset \mathcal{X}^n$ with the same length n as output sequence of the source and generates a RV M, his/her "private randomness" taking values uniformly in a finite set $\mathcal{M}$ (which is exponentially increasing as the length n of the source sequences increases) and independent of the output of the source (V^n, L^n). Assume a (sum type) distortion measure ρ in (1) and a criterion of distortion D_1 are given. According to the output of the source $(V^n, L^n) = (v^n, l^n)$ and the output of his/her private randomness $M = m$, the sender chooses a codeword $x_m(v^n, l^n) \in \mathcal{U}(\subset \mathcal{X}^n)$ such that the average of the distortion between the codeword and the component $V^n = v^n$ of the correlated source may not exceed D_1. Namely,

$$\frac{1}{n} \sum_{m \in \mathcal{M}} P_M(m) \sum_{v^n \in \mathcal{V}^n} \sum_{l^n \in \mathcal{L}^n} P_{VL}(v^n, l^n) \rho(x_m(v^n, l^n), v^n) \le D_1. \tag{10}$$

If $x_m(v^n, l^n)$ is the channel input, he receiver receives an output sequence $y^n \in \mathcal{Y}^n$ with probability $W^n(y^n | x_m(v^n, l^n))$. We also allow to choose $x_m(v, l^n)$ as a random input sequence instead of deterministic one (it is more convenient in the proof). Finally for a finite set $\mathcal{A}$, which typically increases exponentially when the length n of the source increases, i.e., for a constant κ

$$\frac{1}{n} \log |\mathcal{A}| \le \kappa, \tag{11}$$

the sender creates a RV F with range $\mathcal{A}$, according to the outputs of (V^n, L^n) and M, through a function

$$F : \mathcal{V}^n \times \mathcal{L}^n \times \mathcal{M} \longrightarrow \mathcal{A} \tag{12}$$

and the receiver creates a RV G according to the output of the channel W^n and the output of the component L^n of the source, through a function

$$G : \mathcal{L}^n \times \mathcal{Y}^n \longrightarrow \mathcal{A}. \tag{13}$$

After the terminology in chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints" we called the pair of random variables (F, G) generated in the above way permissible and say that a permissible pair (F, G) represents λ-common randomness if

$$\Pr\{F \neq G\} < \lambda. \tag{14}$$

Typically λ should be an arbitrarily small but positive real number when length n of source sequences is arbitrarily large. It is not hard to see that under the conditions (11) and (14) by Fano inequality (Lemma 48), the entropy rates $\frac{1}{n}H(F)$ and $\frac{1}{n}H(G)$ are arbitrarily close if λ in (14) is arbitrarily small. This was observed in [8](chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints"). Thus we can choose any one from the pair of entropy rates, say $\frac{1}{n}H(F)$ as the rate of common randomness.

A pair of real numbers (r, D_1) is called achievable for common randomness if for arbitrary positive real numbers ϵ, λ, μ and sufficiently large n (depending on ϵ, λ and μ) there exists a λ-common randomness satisfying (10)–(14), such that

$$\frac{1}{n}H(F) > r - \epsilon \tag{15}$$

and

$$\sum_{a \in \mathcal{A}} \mid \Pr\{F = a\} - \frac{1}{|\mathcal{A}|} \mid < \mu. \tag{16}$$

The last condition says that the common randomness is required to be nearly uniform and we call it nearly uniform condition. We set it for reducing the errors of second kind of identification codes. The set of achievable pairs is called common randomness capacity region. For fixed D_1 the common randomness capacity (CR-capacity) is $C_{\mathrm{CRI}}((V, L), W, D_1) = \max\{r : (r, D_1) \text{ is achievable}\}$.

Notice that there is no limit to the amount of sender's private randomness in the present model and the next model, Model II. However, because of the limit of the capacity of the channel, the " extra" private randomness is useless.

We here remark that this model is different from the model (i) in chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints" in three points.First, the channel in the model (i) of chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints" is noiseless with rate $\leq R$, whereas the current model is in general noisy. More importantly, because of the distortion requirement, the source not only plays a role of " side information" but also a role of "constrainer". That is, to fight for reducing the distortion, the sender has to choose codewords properly. This makes the transformation more difficult. To see that, let us consider an extremal case where the component L^n of the source is a constant. In this case, the source makes no difference at all in the model (i) of chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints" and therefore the common randomness capacity is trivially equal to the capacity of the channel. But in the case of the present model, the source makes difference i.e., because of it the sender may not choose the codewords freely and therefore the common randomness is reduced. The evaluation of the CR-capacity region for this model is also absolutely non-trivial. Finally, in this model the sender and receiver observe the output $(V^n, L^n) = (v^n, l^n)$ and $L^n = l^n$ respectively. The common randomness before the transmission is equal to $H(L^n) = I(V^n, L^n; L^n)$; the mutual information between the two observations. Therefore, it is not surprising that our characterization in Theorem 156 is quite different from that in Theorem 145 of chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints" and can not simply be obtained by substituting the rate of the noiseless channel by the capacity of the noisy channel.

Model II: Two-source with a constraint noisy channel and a noiseless channel

It is clear that our goal to study the common randomness of the model I is for the construction of WIDSI codes. Next, to study WIDK codes, we introduced the Model II of common randomness. Actually our model is a little more general than that we really need. That is, we add "the side information". But for this we need to do almost no more work. Thus, to define the Model II we only add a noiseless channel between the sender and receiver based on th Model I.

Namely we assume that the correlated source $\{(V^n, L^n)\}_{n=1}^{\infty}$, the noisy channel W, the distortion constraint (10), and the sender's private randomness M are still available. Additionally the sender may send a message k_n from a set of message $\mathcal{K}_n$ with rate $\frac{1}{n}\log|\mathcal{K}| \leq R_{\mathrm{K}}$ to the receiver via noiseless channel. Again R_{K} is called key rate. Of course k_n is necessarily to be a function of the outputs of the source and sender's private randomness i.e., $k_n = k_n(v^n, m)$ for $v^n \in \mathcal{V}^n$, $m \in \mathcal{M}$. More generally, the sender may use random strategies i.e., treats k_n as output of a channel W_K with input (v^n, m). To define the common randomness for this model, we change (13) to

$$G : \mathcal{K}_n \times \mathcal{L}^n \times \mathcal{Y}^n \longrightarrow \mathcal{A}. \tag{17}$$

and keep the conditions (10), (11), (12), (14), (15), and (16) unchanged. Now, accordingly, the definition of function G has been changed.

Analogously, one can define CR-capacity $C_{\mathrm{CRII}}((V, L), W, R_{\mathrm{K}}, D_1)$ for memoryless correlated source with generic (V, L), memoryless channel W, key rate R_{K} and the distortion criterion D_1 of this model.

3.3 The Models for Compound Channels

In this subsection, we assume that the attacker employs a (stationary) memoryless channel from a family of channels satisfying attack distortion criterion to attack the watermark. Neither the sender nor the receiver knows which channel the attacker uses. These channels are known as compound channels in Information Theory. This assumption is slightly more robust and practical than that in [23] where the decoder has to know the attack channel in order to decode. In fact, to the best of our knowledge, it is a most robust assumption in this direction.

Definition 155 A compound channel is just a family of memoryless channels $\mathcal{W} = \{W(\cdot|\cdot, s) : s \in \mathcal{S}\}$ with common input and output alphabet $\mathcal{X}$ and $\mathcal{Y}$ respectively. $\mathcal{S}$ is an index set which is called state set and its members are called states. An output sequence $y^n \in \mathcal{Y}^n$ is output with the probability

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

when the channel is governed by the state s and $x^n \in \mathcal{X}^n$ is input.

Underlie assumption for the attacker to use a compound channel to attack a watermarking transmission or identification code is that the attacker knows the input distribution P_n generated by the code. He then may employ such a compound channel that for all $s \in \mathcal{S}$

$$\frac{1}{n} \sum_{x^n \in \mathcal{X}^n} P_n(x^n) \sum_{y^n \in \mathcal{Y}^n} W^n(y^n|x^n, s)\rho'(x^n, y^n) \le D_2,$$

where ρ' is a sum type distortion measure, attack distortion measure (AD-measure), may or may not be identify to WD-measure ρ and D_2 is the attack distortion criterion. In particular when the codewords are generated by an i.i.d. input distributions so that the input distribution generated by the code is an i.i.d. distribution

$$P^n(x^n) = \prod_{i=1}^{n} P(x_t)$$

a compound channel such that for all $s \in \mathcal{S}$

$$\sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}^n} W(y|x, s)\rho'(x, y) \leq D_2$$

may be used. We always assume that all compound channels under the consideration satisfy the condition of distortion and do not worry about it at all.

To adjust the models in the last two subsections to the compound channels the following modifications are necessary.

For *WIDSI code* for compound channels: replace (3) and (4) by for all $l^n \in \mathcal{L}^n, m \in \mathcal{M}, \mathcal{D}_m(l^n) \subset \mathcal{Y}^n$ such that for all $m \in \mathcal{M}$, and $s \in \mathcal{S}$,

$$\sum_{v^n \in \mathcal{V}^n, l^n \in \mathcal{L}^n} P_{VL}^n(v^n, l^n) \sum_{x^n \in \mathcal{X}^n} Q_m(x^n|v^n, l^n) W^n(\mathcal{D}_m(l^n)|x^n, s) > 1-\lambda_1, \tag{18}$$

and for all $m, m' \in \mathcal{M}\ m \neq m'$, and $s \in \mathcal{S}$

$$\sum_{v^n \in \mathcal{V}^n, \ell^n \in \mathcal{L}^n} P_{VL}^n(v^n, \ell^n) \sum_{x^n \in \mathcal{X}^n} Q_m(x^n|v^n, \ell^n) W^n(\mathcal{D}_{m'}(l^n)|x^n, s) < \lambda_2 \tag{19}$$

respectively.

For *WIDK* for compound channels: replace (8) and (9) by for all $k_n \in \mathcal{L}_n, m \in \mathcal{M}, \mathcal{D}_m(k_n) \subset \mathcal{Y}^n$ such that for all $m \in \mathcal{M}$, and $s \in \mathcal{S}$,

$$\sum_{v^n \in \mathcal{V}^n} P_V^n(v^n) \sum_{k_n \in \mathcal{K}_n} W_{K_n}(k_n|v^n) \sum_{x^n \in \mathcal{X}^n} Q_m^*(x^n|v^n, k_n) W^n(\mathcal{D}_m(k_n)|x^n, s) > 1 - \lambda_1, \tag{20}$$

and for all $m, m' \in \mathcal{M}\ m \neq m'$, and $s \in \mathcal{S}$,

$$\sum_{v^n \in \mathcal{V}^n} P_V^n(v^n) \sum_{k_n \in \mathcal{K}_n} W_{K_n}(k_n|v^n) \sum_{x^n \in \mathcal{X}^n} Q_m^*(x^n|v^n, k^n) W^n(\mathcal{D}_{m'}(k_n)|x^n, s) < \lambda_2. \tag{21}$$

Here the fact that Q_m, Q_m^*, $\mathcal{D}_m(l^n)$ and $\mathcal{D}_m(k_n)$ are independent of the states governing the channels reflects the requirement that neither encoder nor decoder knows the states and that (18)–(21) hold for all $s \in \mathcal{S}$ is because the worst case to the encoder and decoder is considered.

For the *common randomness in the models I and II:* for compound channels, replace (14) by whenever any state s governs the channel,

$$\Pr\{F \neq G|s\} < \lambda. \tag{22}$$

Again the functions F, G, codewords are independent of the states because the states are unknown for both encoder and the decoder.

For compound channel $\mathcal{W}$ the corresponding capacities of watermarking identification codes are denoted $C_{\mathrm{WIDSI}}((V, L), \mathcal{W}, D_1)$ and $C_{\mathrm{WIDK}}(V, \mathcal{W}, R_{\mathrm{K}}, D_1)$, and that of common randomness codes are denoted $C_{\mathrm{CRI}}((V, L), \mathcal{W}, D_1)$ and $C_{\mathrm{CRII}}((V, L)\mathcal{W}, R_{\mathrm{K}}, D_1)$.

4 The Results

4.1 The Results on Common Randomness

For a given correlated memoryless source $\{(V^n, L^n)\}_{n=1}^{\infty}$, whose generic has joint distribution P_{VL}, a memoryless channel W and distortion criterion D_1, let $\mathcal{Q}((V, L), W, D_1)$ be the set of RV (V, L, U, X, Y) with domain $\mathcal{V} \times \mathcal{L} \times \mathcal{U} \times \mathcal{X} \times \mathcal{Y}$ and the following properties, where $\mathcal{U}$ is a finite set with cardinality $|\mathcal{U}| \leq |\mathcal{V}||\mathcal{L}||\mathcal{X}|$ and $\mathcal{X}$ and $\mathcal{Y}$ are input and output alphabets of the channel W respectively.

For all $v \in \mathcal{V},\ l \in \mathcal{L},\ u \in \mathcal{U},\ x \in \mathcal{X}$, and $y \in \mathcal{Y}$

$$\begin{aligned}\Pr\{(V, L, U, X, Y) &= (v, l, u, x, y)\} \\ &= P_{VLUXY}(v, l, u, x, y) \\ &= P_{VL}(v, l) P_{UX|VL}(u, x|v, l) W(y|x). \end{aligned} \tag{23}$$

For the given distortion measure ρ

$$\mathbb{E}\rho(V, X) \leq D_1. \tag{24}$$

$$I(U; V, L) \leq I(U; L, Y). \tag{25}$$

Then we have the coding theorem of common randomness in the model I for single channel W.

Theorem 156

$$C_{\mathrm{CRI}}((V, L), W, D_1) = \max_{(V,L,U,X,Y) \in \mathcal{Q}((V,L),W,D_1)} [I(U; L, Y) + H(L|U)]. \tag{26}$$

For a given correlated source with generic (V, L) a channel W and positive real numbers R_{K} and D_1, we denote by $\mathcal{Q}^*((V, L), W, R_{\mathrm{K}}, D_1)$ the set of RV's

(V, L, U, X, Y) with domain as above and such that (23), (24) and

$$I(U; V, L) \leq I(U; L, Y) + R_{\mathrm{K}} \tag{27}$$

hold. Then

Theorem 157

$$C_{\mathrm{CRII}}((V, L), W, R_{\mathrm{K}}, D_1) = \max_{(V,L,U,X,Y) \in \mathcal{Q}^*((V,L),W,R_{\mathrm{K}},D_1)} [I(U; L, Y) + H(L|U)] + R_{\mathrm{K}}. \tag{28}$$

To state the coding theorem for compound channels we need new notation. For RV's (V, L, U, X) with alphabet $\mathcal{V} \times \mathcal{L} \times \mathcal{U} \times \mathcal{X}$ as above and the channel with input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ respectively, denote by $Y(W)$ the RV such that the joint distribution $P_{LVUXY(W)} = P_{LVUX}W$ (consequently, $LVU \leftrightarrow X \leftrightarrow Y$ form a Markov chain). For a compound channel $\mathcal{W}$ with set of states $\mathcal{S}$ and a state $s \in \mathcal{S}$ we also write $Y(W(\cdot|\cdot, s)) = Y(s)$. With the notation we write

$$I(U; L, Y(\mathcal{W})) = \inf_{s \in \mathcal{S}} I(U; L, Y(s))$$

and

$$I(U; Y(\mathcal{W})|L) = \inf_{s \in \mathcal{S}} I(U; Y(s)|L).$$

Sometimes just for the convenience, we also write $Y(s)$ as $\tilde{Y}(s)$ when we substitute P_{LVUX} by $P_{\tilde{L}\tilde{V}\tilde{U}\tilde{X}}$ and similarly $\tilde{Y}(\mathcal{W})$. Then

$$I(U; L, Y(\mathcal{W})) = I(U; L) + I(U; Y(\mathcal{W})|L). \tag{29}$$

Now for a compound channel we define $\mathcal{Q}_1((V, L), \mathcal{W}, D_1)$ as the set of RV's (V, L, U, X) such that its marginal distribution for the first two components is equal to the distribution P_{VL} and (24) and

$$I(U; V, L) \leq I(U; L, Y(\mathcal{W})) \tag{30}$$

hold. Analogously to set $\mathcal{Q}^*((V, L), W, R_{\mathrm{K}}, D_1)$, we define $\mathcal{Q}_1^*((V, L), \mathcal{W}, R_{\mathrm{K}}, D_1)$ the set of RV's (V, L, U, X) such that its marginal distribution for the first two

components is equal to the distribution P_{VL} and (24) and

$$I(U;V,L) \le I(U;L,Y(\mathcal{W})) + R_{\mathrm{K}}. \tag{31}$$

hold. Then

Theorem 158

$$C_{\mathrm{CRI}}((V,L),\mathcal{W},D_1) \ge \sup_{\substack{(V,L,U,X)\in \\ \mathcal{Q}_1((V,L),\mathcal{W},D_1)}} [I(U;L,Y(\mathcal{W})) + H(L|U)] \tag{32}$$

$$C_{\mathrm{CRI}}((V,L),\mathcal{W},D_1) \le \inf_{W\in\mathcal{W}} \max_{\substack{(V,L,U,X,Y)\in \\ \mathcal{Q}((V,L),W,D_1)}} [I(U;L,Y) + H(L|U)]. \tag{33}$$

Theorem 159

$$C_{\mathrm{CRII}}((V,L),\mathcal{W},R_{\mathrm{K}},D_1) \ge \sup_{\substack{(V,L,U,X)\in \\ \mathcal{Q}_1^*((V,L),\mathcal{W},R_{\mathrm{K}},D_1)}} [I(U;L,Y(\mathcal{W})) + H(L|U)] + R_{\mathrm{K}} \tag{34}$$

$$C_{\mathrm{CRII}}((V,L),\mathcal{W},R_{\mathrm{K}},D_1) \le \inf_{W\in\mathcal{W}} \max_{\substack{(V,L,U,X,Y)\in \\ \mathcal{Q}^*((V,L),W,R_{\mathrm{K}},D_1)}} [I(U;L,Y) + H(L|U)] + R_{\mathrm{K}}. \tag{35}$$

Notice the gaps of lower and upper bounds in both Theorems 158 and 159 are due to the orders of inf-sup.

4.2 The Results on Watermarking Identification Codes

We shall use the same notation as in the above part. Moreover for above sets $\mathcal{V}$, $\mathcal{X}$ and $\mathcal{Y}$ and a finite set $\mathcal{U}$ with cardinality bounded by $|\mathcal{V}||\mathcal{X}|$, a memoryless source with generic V, a memoryless channel W, and compound channel $\mathcal{W}$, we define the following sets. Let $\mathcal{Q}^{**}(V,W,R_{\mathrm{K}},D_1)$ be the set of random variables (V,U,X,Y) with domain $\mathcal{V}\times\mathcal{U}\times\mathcal{X}\times\mathcal{Y}$ such that for all $v\in\mathcal{V}$, $u\in\mathcal{U}$, $x\in\mathcal{X}$, and $y\in\mathcal{Y}$

$$P_{VUXY}(v,u,x,y) = P_V(v)P_{UX|V}(u,x|v)W(y|x), \tag{36}$$

$$I(U;V) \le I(U;Y) + R_{\mathrm{K}}, \tag{37}$$

and (24) hold. Let $\mathcal{Q}_1^{**}(V, \mathcal{W}, R_K, D_1)$ be set of RV's (V, U, X) with domain $\mathcal{V} \times \mathcal{U} \times \mathcal{X}$ such that for all $v \in \mathcal{V}$, $u \in \mathcal{U}$ and $x \in \mathcal{X}$,

$$P_{VUX}(v, u, x) = P_V(v) P_{UX|V}(u.x|v), \tag{38}$$

$$I(U; V) \leq I(U; Y(\mathcal{W})) + R_K, \tag{39}$$

and (24) hold, where $I(U; Y(\mathcal{W})) = \inf_{W \in \mathcal{W}} I(U; Y(W))$. In particular, when the second component L^n of the correlated source $\{(V^n, L^n)\}_{n=1}^{\infty}$ is a constant, $\mathcal{Q}^*((V, L), W, R_K, D_1)$ and $\mathcal{Q}_1^*((V, L), \mathcal{W}, R_K, D_1)$ become $\mathcal{Q}^{**}(V, W, R_K, D_1)$ and $\mathcal{Q}_1^{**}(V, \mathcal{W}, R_K, D_1)$ respectively.

Theorem 160

$$C_{\text{WIDSI}}((V, L), W, D_1) \geq \max_{(V,L,U,X,Y) \in \mathcal{Q}((V,L),W,D_1)} [I(U; L, Y) + H(L|U)]. \tag{40}$$

Theorem 161

$$C_{\text{WIDK}}(V, W, R_K, D_1) \geq \max_{(V,U,X,Y) \in \mathcal{Q}^{**}(V,W,R_K,D_1)} I(U; Y) + R_K. \tag{41}$$

Theorem 162

$$C_{\text{WIDSI}}((V, L), \mathcal{W}, D_1) \geq \sup_{(V,L,U,X) \in \mathcal{Q}_1((V,L),\mathcal{W},D_1)} [I(U; L, Y(\mathcal{W})) + H(L|U)]. \tag{42}$$

Theorem 163

$$C_{\text{WIDK}}(V, W, R_K) \geq \sup_{(V,U,X) \in \mathcal{Q}_1^{**}(V,\mathcal{W},R_K,D_1)} I(U; Y(\mathcal{W})) + R_K. \tag{43}$$

Note that in Theorems 161 and 163 one may add side information L^n, the second component of the correlated source and then one can obtain the corresponding lower bound almost without changing the proof.

4.3 A Result on Watermarking Transmission Code with a Common Experiment Introduced by Steinberg-Merhav

To construct watermarking identification code Y. Steinberg and Merhav in [27] introduced a code, which they call watermarking transmission code with common experiment, distortion measure ρ, and covertext P_V. They obtained there an inner bound to the its capacity region, which is sufficient for achieving their goal. We shall

show that their bound is tight and therefore actually the capacity region. Their definition and result on it and our proof will be presented it the last section.

5 The Direct Theorems for Common Randomness

In this section we prove the direct parts of Theorems 156–159. Since a DMC can be regarded as a special compound channel with a single member (i.e., $|\mathcal{S}| = 1$), we only have to show the direct parts of Theorems 158 and 159. To this end we need the following three lemmas for n-ED $P_{\tilde{V}\tilde{L}\tilde{U}}$ over the product set $\mathcal{V} \times \mathcal{L} \times \mathcal{U}$ of finite sets $\mathcal{V}$, $\mathcal{L}$ and $\mathcal{U}$.

Lemma 164 *[Uniformly covering] For $\ell^n \in \mathcal{T}^n_{\tilde{L}}$, let $U_i(\ell^n) i = 1, 2, \ldots, \lfloor 2^{n\alpha} \rfloor$ be a sequence of independent RV's with uniform distribution over $\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)$ and for any $v^n \in \mathcal{T}^n_{\tilde{V}|\tilde{L}}(\ell^n)$ let $\hat{\mathcal{U}}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)$ be the random set $\{U_i(\ell^n) : i = 1, 2, \ldots, \lfloor 2^{n\alpha} \rfloor\} \cap \mathcal{T}^n_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)$. Then for all $\varepsilon \in (0, 1]$*

$$\Pr\left\{\left|\left|\hat{\mathcal{U}}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)\right| - \lfloor 2^{n\alpha} \rfloor \frac{|\mathcal{T}^n_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)|}{|\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)|}\right| \geq \lfloor 2^{n\alpha} \rfloor \frac{|\mathcal{T}^n_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)|}{|\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)|}\varepsilon\right\} < 4 \cdot 2^{-\frac{\varepsilon^2}{4}2^{n\eta}} \tag{44}$$

for sufficiently large n if

$$\lfloor 2^{n\alpha} \rfloor > 2^{n\eta} \frac{|\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)|}{|\mathcal{T}^n_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)|}$$

Proof Let

$$Z_i(v^n, \ell^n) = \begin{cases} 1 & \text{if } U_i(\ell^n) \in \mathcal{T}^n_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n), \\ 0 & \text{else,} \end{cases} \tag{45}$$

and $q = \frac{|\mathcal{T}^n_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)|}{|\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)|}$. Then

$$|\hat{\mathcal{U}}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)| = \sum_{i=1}^{\lfloor 2^{n\alpha} \rfloor} Z_i(v^n\ell^n)$$

and for $i = 1, 2, \ldots, \lfloor 2^{n\alpha} \rfloor$

$$\Pr\{Z_i(v^n \ell^n) = z\} = \begin{cases} q & \text{if } z = 1 \\ 1 - q & \text{if } z = 0 \end{cases} \tag{46}$$

by the definitions of $U_i(\ell^n)$ and $Z_i(v^n, \ell^n)$.
Then by Chernoff's bound, we have that

$$\begin{aligned}
&\Pr\left\{ \sum_{i=1}^{\lfloor 2^{n\alpha} \rfloor} Z_i(v^n \ell^n) \geq \lfloor 2^{n\alpha} \rfloor q(1+\varepsilon) \right\} \\
&\leq e^{-\frac{\varepsilon}{2}\lfloor 2^{n\alpha} \rfloor q(1+\varepsilon)} \, E \, e^{\frac{\varepsilon}{2} \sum_{i=1}^{\lfloor 2^{n\alpha} \rfloor} Z_i(v^n, \ell^n)} \\
&= e^{-\frac{\varepsilon}{2}\lfloor 2^{n\alpha} \rfloor q(1+\varepsilon)} \prod_{i=1}^{\lfloor 2^{n\alpha} \rfloor} E \, e^{\frac{\varepsilon}{2} Z_i(v^n, \ell^n)} \\
&= e^{-\frac{\varepsilon}{2}\lfloor 2^{n\alpha} \rfloor q(1+\varepsilon)} [1 + (e^{\frac{\varepsilon}{2}} - 1)q]^{\lfloor 2^{n\alpha} \rfloor} \\
&\leq e^{-\frac{\varepsilon}{2}\lfloor 2^{n\alpha} \rfloor q(1+\varepsilon)} \left[1 + \left(\frac{\varepsilon}{2} + \left(\frac{\varepsilon}{2} \right)^2 \right) q \right]^{\lfloor 2^{n\alpha} \rfloor} \\
&\leq \exp_e \left\{ -\frac{\varepsilon}{2} \lfloor 2^{n\alpha} \rfloor q(1+\varepsilon) + \frac{\varepsilon}{2} \lfloor 2^{n\alpha} \rfloor q \left(1 + \frac{\varepsilon}{2} \right) \right\} \\
&= e^{-\frac{\varepsilon^2}{4}\lfloor 2^{n\alpha} \rfloor q} < 2e^{-\frac{\varepsilon^2}{4} 2^{n\eta}}
\end{aligned} \tag{47}$$

if $\lfloor 2^{n\alpha} \rfloor > 2^{n\eta} q^{-1}$.

Here the first inequality follows from Chernoff's bound; the second equality holds by (46); the second inequality holds because $e^{\frac{\varepsilon}{2}} < 1 + \frac{\varepsilon}{2} + \left(\frac{\varepsilon}{2}\right)^2$ by the assumption that $\varepsilon < 1$, $e^{\frac{\varepsilon}{2}} < e^{\frac{1}{2}} < 2$; and the third inequality follows from the well known inequality $1 + x < e^x$. Similarly one can obtain

$$\Pr\left\{ \sum_{i=1}^{\lfloor 2^{n\alpha} \rfloor} Z_i(v^n \ell^n) \leq \lfloor 2^{n\alpha} \rfloor q(1-\varepsilon) \right\} < 2e^{-\frac{\varepsilon^2}{4} 2^{n\eta}} \tag{48}$$

if $\lfloor 2^{n\alpha} \rfloor > 2^{n\eta} q^{-1}$.

Finally we obtain the lemma by combining (47) and (48). □

Lemma 165 (Packing) *Let $P_{\tilde{L}\tilde{U}}$ be an n-ED, let $U_i(\ell^n)$, $i = 1, 2, \ldots, \lfloor 2^{n\alpha} \rfloor$ be a sequence of independent RV's uniformly distributed on $\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)$ for an $\ell^n \in \mathcal{T}^n_{\tilde{L}}$, and let $\mathcal{Y}$ be a finite set. Then for all n-ED's $P_{\tilde{L}\tilde{U}\tilde{Y}}$ and $P_{\tilde{L}\tilde{U}\overline{Y}}$ with common marginal distributions $P_{\tilde{L}\tilde{U}}$ and $P_{\overline{Y}} = P_{\tilde{Y}}$, all i, $\gamma > 0$ and sufficiently large n,*

$$\Pr\left\{\frac{1}{\lfloor 2^{n\alpha}\rfloor}\sum_{i=1}^{\lfloor 2^{n\alpha}\rfloor}\left|\mathcal{T}^n_{\tilde{Y}|\tilde{L}\tilde{U}}\left(\ell^n U_i(\ell^n)\right)\cap\left[\bigcup_{j\neq i}\mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}\left(\ell^n U_j(\ell^n)\right)\right]\right| \geq t_{\tilde{Y}|\tilde{L}\tilde{U}}2^{-\frac{n}{2}\gamma}\right\} < 2^{-\frac{n}{2}\gamma} \tag{49}$$

if $\lfloor 2^{n\alpha}\rfloor \leq \frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{U}|\tilde{L}\overline{Y}}}2^{-n\gamma}$.

Here $t_{\tilde{Y}|\tilde{L}\tilde{U}}$, $t_{\tilde{U}|\tilde{L}}$, and $t_{\tilde{U}|\tilde{L}\overline{Y}}$ are the common values of $|\mathcal{T}^n_{\tilde{Y}|\tilde{L}\tilde{U}}(\ell^n u^n)|$ for $(\ell^n, u^n) \in \mathcal{T}^n_{\tilde{L}\tilde{U}}$, $|\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)|$ for $\ell^n \in \mathcal{T}^n_{\tilde{L}}$, and $|\mathcal{T}^n_{\tilde{U}|\tilde{L}\overline{Y}}(\ell^n y^n)|$ for $(\ell^n, y^n) \in \mathcal{T}^n_{\tilde{L}\overline{Y}}$, respectively.

Proof For $i = 1, 2, \ldots, \lfloor 2^{n\alpha}\rfloor$, $y^n \in \mathcal{T}^n_{\overline{Y}} = \mathcal{T}^n_{\tilde{Y}}$, let

$$\hat{Z}_i(y^n) = \begin{cases} 1 & \text{if } y^n \in \bigcup_{j\neq i}\mathcal{T}_{\overline{Y}|\tilde{L}\tilde{U}}\left(\ell^n U_j(\ell^n)\right) \\ 0 & \text{else} \end{cases} \tag{50}$$

and for all $u^n \in \mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)$

$$S_i(u^n) = \left|\mathcal{T}^n_{\tilde{Y}|\tilde{L}\tilde{U}}(\ell^n u^n)\cap\left[\bigcup_{j\neq i}\mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}\left(\ell^n U_j(\ell^n)\right)\right]\right|. \tag{51}$$

Then

$$S_i(u^n) = \sum_{y^n\in\mathcal{T}_{\tilde{Y}|\tilde{L}\tilde{U}}(\ell^n u^n)}\hat{Z}_j(y^n) \tag{52}$$

and

$$E\ \hat{Z}_i(y^n) = \Pr\left\{y^n \in \bigcup_{j\neq i}\mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}\left(\ell^n U_j(\ell^n)\right)\right\} \tag{53}$$

$$\leq \sum_{j\neq i}\Pr\left\{y^n \in \mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}\left(\ell^n U_j(\ell^n)\right)\right\} \tag{54}$$

$$= \sum_{j \neq i} \Pr\left\{U_j(\ell^n) \in \mathcal{T}^n_{\tilde{U}|\tilde{L}\overline{Y}}(\ell^n y^n)\right\} \tag{55}$$

$$= (2^{\lfloor n\alpha \rfloor} - 1)\frac{t_{\tilde{U}|\tilde{L}\overline{Y}}}{t_{\tilde{U}|\tilde{L}}} < 2^{-n\gamma} \tag{56}$$

if $\lfloor 2^{n\alpha} \rfloor \leq \frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{U}|\tilde{L}\overline{Y}}} 2^{-n\gamma}$.

Hence by (52) and (56) we have that $E\ S_i(u^n) \leq t_{\tilde{Y}|\tilde{L}\tilde{U}} 2^{-n\gamma}$ and i.e.,

$$E\big[S_i(U_i(\ell^n))|U_i(\ell^n)\big] < t_{\tilde{Y}|\tilde{L}\tilde{U}} 2^{-n\gamma} \text{ (a.s.)},$$

so

$$E\ S_i\big(U_i(\ell^n)\big) = E\big\{E\big[S_i(U_i(\ell^n))|U_i(\ell^n)\big]\big\} < t_{\tilde{Y}|\tilde{L}\tilde{U}} 2^{-n\gamma}. \tag{57}$$

Thus by Markov's inequality we have that

$$\Pr\left\{\frac{1}{\lfloor 2^{n\alpha} \rfloor}\sum_{i=1}^{\lfloor 2^{n\alpha} \rfloor} S_i\big(U_i(\ell^n)\big) \geq t_{\tilde{Y}|\tilde{L}\tilde{U}} 2^{-\frac{n}{2}\gamma}\right\} < 2^{-\frac{n}{2}\gamma},$$

i.e., (49). □

Lemma 166 (Multi-packing) *Under the conditions of the previous lemma, let $U_{i,k}(\ell^n)$, $i = 1, 2, \ldots, \lfloor 2^{n\beta_1} \rfloor$, $k = 1, 2, \ldots, \lfloor 2^{n\beta_2} \rfloor$, be a sequence of independent RV's uniformly distributed on $\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)$ for a given $\ell^n \in \mathcal{T}^n_{\tilde{L}}$. Then for all n-ED's $P_{\tilde{L}\tilde{U}\tilde{Y}}$ and $P_{\tilde{L}\tilde{U}\overline{Y}}$ in the previous lemma*

$$\Pr\left\{\frac{1}{\lfloor 2^{n\beta_2} \rfloor}\sum_{k=1}^{\lfloor 2^{n\beta_2} \rfloor}\frac{1}{\lfloor 2^{n\beta_1} \rfloor}\sum_{i=1}^{\lfloor 2^{n\beta_1} \rfloor}\left|\mathcal{T}^n_{\tilde{Y}|\tilde{L}\tilde{U}}\big(\ell^n U_{i,k}(\ell^n)\big) \cap \left[\bigcup_{j\neq i}\mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}\big(\ell^n U_{j,k}(\ell^n)\big)\right]\right|\right.$$
$$\left. \geq t_{\tilde{Y}|\tilde{L}\tilde{U}} 2^{-n\eta}\right\} < 2^{-\frac{n}{2}\gamma} \tag{58}$$

if $\lfloor 2^{n\alpha} \rfloor \leq \frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{U}|\tilde{L}\overline{Y}}} 2^{-n\gamma}$.

Proof For $u^n \in \mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)$, let

$$S_{i,k}(u^n) = \left|\mathcal{T}^n_{\tilde{Y}|\tilde{L}\tilde{U}}(\ell^n u^n) \cap \left[\bigcup_{j\neq i}\mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}\big(\ell^n U_{j,k}(\ell^n)\big)\right]\right|.$$

Then we have shown in the proof to the previous lemma (c.f. (57))

$$E\ S_{i,k}\big(U_i(\ell^n)\big) < t_{\tilde{Y}|\tilde{L}\tilde{U}} 2^{-n\gamma}.$$

Thus (58) follows from Markov's inequality. □

Now let us turn to the direct part of Theorem 158.

Lemma 167 (The direct part of Theorem 158) *For a compound channel $\mathcal{W}$,*

$$C_{\mathrm{CRI}}\big((V,L),\mathcal{W},D_1\big) \geq \sup_{(V,L,U,X)\in Q_1\left((V,L),\mathcal{W},D_1\right)} \big[I(U;L,Y(\mathcal{W})\big) + H(L|U)\big]. \tag{59}$$

Proof We have to show for a given correlated memoryless source with generic (V,L), a compound channel $\mathcal{W}$, $(V,L,U,X) \in Q_1\big((V,L),\mathcal{W},D_1\big)$ and sufficiently large n, the existence of the functions, F, G and $x_m(v^n,\ell^n)$ satisfying (10)–(13), (22), (15) and (16) with the rate arbitrarily close to $I\big(U;L,Y(\mathcal{W})\big) + H(L|U)$. Obviously the set of achievable rates of the common randomness is bounded and closed (i.e., compact). So without loss of generality, by uniform continuity of information quantities, we can assume that $E\rho(V,X) < D_1$, and $I(U;V,L) < I\big(U;L,Y(\mathcal{W})\big)$. Because $I(U;V,L) = I(U;L) + I(U;V|L)$ and $I\big(U;L,Y(\mathcal{W})\big) = I(U;L) + I\big(U;Y(\mathcal{W})|L\big)$, there exists a sufficiently small but positive constant ξ, such that

$$I\big(U;Y(\mathcal{W})|L\big) - I(U;V|L) > \xi. \tag{60}$$

Without loss of generality, we also assume P_U is an n-ED to simplify the notation. Then for arbitrary $\varepsilon_1 > 0$, by uniform continuity of information quantities, we can find $\delta_1, \delta_2 > 0$ with the following properties.

(i) For all $\ell^n \in \mathcal{T}_L^n(\delta_1)$ with ED $P_{\ell^n} = P_{\tilde{L}}$, there exists a $\delta' > 0$, such that $(v^n,\ell^n) \in \mathcal{T}_{VL}^n(\delta_2')$ yields that $\mathcal{T}_{\tilde{V}|\tilde{L}}^n(\ell^n) \subset \mathcal{T}_{\hat{V}|\tilde{L}}^n(\ell^n,\delta_2)$, where $P_{\tilde{V}\tilde{L}}$ is the joint ED of (v^n,ℓ^n) and $P_{\hat{V}\tilde{L}} = P_{\tilde{L}}P_{V|L}$.

We call a pair (v^n,ℓ^n) of sequences with $\ell^n \in \mathcal{T}_L^n(\delta_1)$, $(v^n,\ell^n) \in \mathcal{T}_{VL}^n(\delta_2)$, (δ_1,δ_2)-typical and denote the set of (δ_1,δ_2)-typical sequences by $\mathcal{T}^n(\delta_1,\delta_2)$.

Then we may require $\delta_2 \to 0$ as $\delta_1 \to 0$. Moreover (e.g., see [30]), there exist positive $\zeta_1 = \zeta_1(\delta_1)$, $\zeta_2 = \zeta_2(\delta_1,\delta_2)$, and $\zeta = \zeta(\delta_1,\delta_2)$ such that

$$P_L^n\big(\mathcal{T}_L^n(\delta_1)\big) > 1 - 2^{-n\zeta_1} \tag{61}$$

$$P_{V|L}^n\big\{v^n : (v^n,\ell^n) \in \mathcal{T}^n(\delta_1,\delta_2)|\ell^n\big) > 1 - 2^{-n\zeta_2} \tag{62}$$

for all $\ell^n \in \mathcal{T}_L^n(\delta_1)$ and

$$P_{VL}^n\big(\mathcal{T}^n(\delta_1,\delta_2)\big) > 1 - 2^{n\zeta}. \tag{63}$$

(ii) For all $\ell^n \in \mathcal{T}_L^n(\delta_1)$ with ED $P_{\ell^n} = P_{\tilde{L}}$ (say), one can find a joint ED of sequences in $L^n \times \mathcal{U}^n$, say $P_{\tilde{L}\tilde{U}}$, with marginal distributions $P_{\tilde{L}}$ and P_U, sufficiently close to P_{LU}, (which will be specified below). We say that $P_{\tilde{L}\tilde{U}}$ is generated by the ED $P_{\tilde{L}}$ of ℓ^n.

(iii) For all $(v^n,\ell^n) \in \mathcal{T}^n(\delta_1,\delta_2)$ with joint ED $P_{v^n\ell^n} = P_{\tilde{V}\tilde{L}}$ (say), one can find a joint ED $P_{\tilde{V}\tilde{L}\tilde{U}}$ of sequences in $\mathcal{V}^n \times \mathcal{L}^n \times \mathcal{U}^n$ with marginal distributions $P_{\tilde{V}\tilde{L}}$ and $P_{\tilde{L}\tilde{U}}$ and sufficiently close to P_{VLU} (which will be specified below), where $P_{\tilde{L}\tilde{U}}$ is the ED generated by $P_{\tilde{L}}$. We say $P_{\tilde{V}\tilde{L}\tilde{U}}$ is generated by the joint ED $P_{\tilde{V}\tilde{L}}$ of (v^n,ℓ^n).

(iv) For all (δ_1,δ_2)-typical sequences (v^n,ℓ^n) with joint ED $P_{\tilde{V}\tilde{L}}$ (say) and the joint ED $P_{\tilde{V}\tilde{L}\tilde{U}}$ generated by $P_{\tilde{V}\tilde{L}}$, we let $(\tilde{V},\tilde{L},\tilde{U},\tilde{X})$ be RV's with joint distribution $P_{\tilde{V}\tilde{L}\tilde{U}\tilde{X}}$ such that for all $v \in \mathcal{V}$, $\ell \in \mathcal{L}$, $u \in \mathcal{U}$ and $x \in \mathcal{X}$

$$P_{\tilde{V}\tilde{L}\tilde{U}\tilde{X}}(v,\ell,u,x) = P_{\tilde{V}\tilde{L}\tilde{U}}(v,\ell,u)P_{X|VLU}(x|v,\ell,u), \tag{64}$$

and let $\big(\tilde{V},\tilde{L},\tilde{U},\tilde{X},\tilde{Y}(W)\big)$ be RV's with joint distribution $P_{\tilde{V}\tilde{L}\tilde{U}\tilde{X}\tilde{Y}(\mathcal{W})}$ such that for all $v \in \mathcal{V}$, $\ell \in \mathcal{L}$, $u \in \mathcal{U}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$

$$P_{\tilde{V}\tilde{L}\tilde{U}\tilde{X}\tilde{Y}(W)}(v,\ell,u,x,y) = P_{\tilde{V}\tilde{L}\tilde{U}\tilde{X}}(v,\ell,u,x)W(x|y), \tag{65}$$

for any $W \in \mathcal{W}$ and $P_{\tilde{V}\tilde{L}\tilde{U}\tilde{X}}$ in (64). Then the following inequalities hold

$$E\rho(\tilde{V},\tilde{X}) < D_1, \tag{66}$$

$$|H(\tilde{L}) - H(L)| < \varepsilon_1, \tag{67}$$

$$|I(\tilde{U};\tilde{V}|\tilde{L}) - I(U;V|L)| < \varepsilon_1, \tag{68}$$

and

$$|I(\tilde{U};\tilde{Y}(\mathcal{W})|\tilde{L}) - I(U;Y(\mathcal{W})|L)| < \varepsilon_1, \tag{69}$$

where $I(\tilde{U};\tilde{Y}(\mathcal{W})|\tilde{L}) = \inf_{W\in\mathcal{W}} I(\tilde{U};\tilde{Y}(W)|\tilde{L})$.

For arbitrarily small fixed ε_2 with $0 < \varepsilon_2 < \frac{1}{2}\xi$, for ξ in (60), we choose ε_1 (and consequently, δ_1,δ_2) so small that $\varepsilon_1 < \frac{1}{2}\varepsilon_2$ and an α such that

$$I(U;Y(\mathcal{W})|L) - \frac{\xi}{2} < \alpha < I(U;Y(\mathcal{W})|L) - \varepsilon_2 \tag{70}$$

and $M = 2^{n\alpha}$ (say) is an integer. Notice that by (70) we may choose α arbitrarily close to $I(U; Y(\mathcal{W})|L) - \varepsilon_2$ and therefore arbitrarily close to $I(U; Y(\mathcal{W})|L)$ by choosing ε_2 arbitrarily small. Then by (60), (68) and (70) we have that

$$\alpha > I(U; V|L) + \frac{\xi}{2} > I(\tilde{U}; \tilde{V}|\tilde{L}) + \frac{\xi}{2} - \varepsilon_1 > I(\tilde{U}; \tilde{V}|\tilde{L}) + \frac{\xi}{4}, \tag{71}$$

where the last inequality holds by our choice $\varepsilon_1 < \frac{1}{2}\varepsilon_2 < \frac{1}{4}\xi$, and by (69) and (70) we have

$$\alpha < I(\tilde{U}; \tilde{Y}(\mathcal{W})|\tilde{L}) + \varepsilon_1 - \varepsilon_2 < I(\tilde{U}; \tilde{Y}(\mathcal{W})|\tilde{L}) - \frac{\varepsilon_2}{2}. \tag{72}$$

Denote by $t_{\tilde{U}|\tilde{L}}$ and $t_{\tilde{U}|\tilde{V}\tilde{L}}$ the common values of $|\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)|$, $\ell^n \in \mathcal{T}^n_{\tilde{L}}$ and $|\mathcal{T}^n_{\tilde{U}|\tilde{V}\tilde{L}}(v^n, \ell^n)|$, $(v^n, \ell^n) \in \mathcal{T}^n_{\tilde{V}\tilde{L}}$, respectively. Then it is well known that $\frac{1}{n}\log\frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{U}|\tilde{V}\tilde{L}}}$ arbitrarily close to $I(\tilde{U}; \tilde{V}|\tilde{L})$.

This means under our assumption that $\frac{1}{2}\varepsilon_2 < \frac{1}{4}\xi$, (71) implies that for all ED's $P_{\tilde{V}\tilde{L}\tilde{U}}$ generated by the joint ED's $P_{\tilde{V}\tilde{L}}$ of (δ_1, δ_2)-typical sequences

$$2^{\frac{n}{3}\varepsilon_2}\frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{U}|\tilde{V}\tilde{L}}} < 2^{n\alpha} = M. \tag{73}$$

Next we let $\mathcal{Q}_{\mathcal{W}}(\ell^n u^n, \tau)$ be the set of conditional ED $P_{\overline{Y}|\tilde{L}\tilde{U}}$, for a pair (ℓ^n, u^n) of sequences such that there exists a $W \in \mathcal{W}$ with $\mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}(\ell^n u^n) \subset \mathcal{T}^n_{\tilde{Y}(W)|\tilde{L}\tilde{U}}(\ell^n u^n, \tau)$, where $P_{\tilde{L}\tilde{U}}$ is the ED of (ℓ^n, u^n) and $P_{\tilde{L}\tilde{U}\tilde{Y}(\mathcal{W})}$ is the marginal distribution of the distribution in (65). Then

$$\bigcup_{P_{\overline{Y}|\tilde{L}\tilde{U}} \in \mathcal{Q}_{\mathcal{W}}(\ell^n u^n, \tau)} \mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}(\ell^n u^n, \tau) = \bigcup_{W \in \mathcal{W}} \mathcal{T}^n_{\tilde{Y}(\mathcal{W})|\tilde{L}\tilde{U}}(\ell^n u^n, \tau), \tag{74}$$

and

$$|\mathcal{Q}_{\mathcal{W}}(\ell^n u^n, \tau)| < (n+1)^{|\mathcal{L}||\mathcal{U}||\mathcal{Y}|}. \tag{75}$$

Again for the common values $t_{\tilde{U}|\tilde{L}}$ of $|\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)|$, $\ell^n \in \mathcal{T}^n_{\tilde{L}}$, $t_{\tilde{U}|\tilde{L}\overline{Y}}$ of $|\mathcal{T}^n_{\tilde{U}|\tilde{L}\overline{Y}}(\ell^n y^n)|$, $(\ell^n, y^n) \in \mathcal{T}^n_{\tilde{L}\overline{Y}}$, $\lim_{n\to\infty}\frac{1}{n}\log\frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{U}|\tilde{L}\overline{Y}}} = I(\tilde{U}; \overline{Y}|\tilde{L})$.

Thus, (72) yields that for all $P_{\tilde{V}\tilde{L}\tilde{U}}$ generated by the joint ED of (δ_1, δ_2)-typical sequences, $(\ell^n, u^n) \in \mathcal{T}^n_{\tilde{L}\tilde{U}}$, and $P_{\overline{Y}|\tilde{L}\tilde{U}} \in \mathcal{Q}_{\mathcal{W}}(\ell^n u^n, \tau)$,

$$M = 2^{n\alpha} < 2^{-\frac{n}{4}\varepsilon_2}\frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{U}|\tilde{L}\overline{Y}}}, \tag{76}$$

if we choose τ so small (depending on ε_2) that for all $P_{\overline{Y}|\tilde{L}\tilde{U}} \in \mathcal{Q}_{\mathcal{W}}(\ell^n u^n, \tau)$

$$I(\tilde{U}; \overline{Y}|\tilde{L}) > I(\tilde{U}; Y(\mathcal{W})|\tilde{L}) - \frac{1}{8}\varepsilon_2$$

(recalling that by its definition $I(\tilde{U}; \tilde{Y}(\mathcal{W})|\tilde{L}) = \inf_{W \in \mathcal{W}} I(\tilde{U}; \tilde{Y}(W)|\tilde{L})$).

Now we are ready to present our coding scheme at rate α, which may arbitrarily be close to $I(U; Y(\mathcal{W})|L)$.

Coding scheme.

1. Choosing Codebooks:

   For all $\ell^n \in \mathcal{T}_L^n(\delta_1)$ with ED $P_{\tilde{L}}$, $P_{\tilde{L}\tilde{U}}$ generated by $P_{\tilde{L}}$ (cf. condition (ii) above), we apply Lemma 164 with $\eta = \frac{\varepsilon_2}{3}$ and Lemma 165 with $\gamma = \frac{\varepsilon_2}{4}$ to random choice. Then since the numbers of sequences v^n, ℓ^n and the number of n-joint ED's are increasing exponentially and polynomially respectively, for all $\ell^n \in \mathcal{T}_L^n(\delta_1)$ with ED $P_{\tilde{L}\tilde{U}}$ generated by $P_{\tilde{L}}$, by (73), (76) we can find a subset $\mathcal{U}(\ell^n) \subset \mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)$ with the following property if n is sufficiently large.

   If $(v^n, \ell^n) \in \mathcal{T}^n(\delta_1, \delta_2)$ and has joint ED $P_{\tilde{V}\tilde{L}}$ and $P_{\tilde{V}\tilde{L}\tilde{U}}$ is generated by $P_{\tilde{V}\tilde{L}}$ (cf. condition (iii) above), then

$$\left| |\mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n \ell^n)| - M \frac{t_{\tilde{U}|\tilde{V}\tilde{L}}}{t_{\tilde{U}|\tilde{L}}} \right| < M \frac{t_{\tilde{U}|\tilde{V}\tilde{L}}}{t_{\tilde{U}|\tilde{L}}} \varepsilon \tag{77}$$

   for any $\varepsilon > 0$ (with $\varepsilon \to 0$ as $n \to \infty$), where

$$\mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n \ell^n) \triangleq \mathcal{U}(\ell^n) \cap \mathcal{T}^n_{\tilde{U}|\tilde{V}\tilde{L}}(v^n \ell^n). \tag{78}$$

   For any $P_{\tilde{V}\tilde{L}\tilde{U}}$ generated by a joint ED of (δ_1, δ_2)-typical sequence, (v^n, ℓ^n), and joint ED $P_{\tilde{L}\tilde{U}\tilde{Y}}$ with marginal distribution $P_{\tilde{L}\tilde{U}}$ and any $P_{\overline{Y}|\tilde{L}\tilde{U}} \in \mathcal{Q}_{\mathcal{W}}\big(\ell^n u^n_{m'}(\ell^n), \tau\big)$ (notice that $\mathcal{Q}_{\mathcal{W}}(\ell^n u^n, \tau)$ depends on $(\ell^n u^n)$ only through their joint ED $P_{\ell^n u^n}$!)

$$M^{-1} \sum_{m=1}^{M} \left| \mathcal{T}^n_{\tilde{Y}|\tilde{L}\tilde{U}}\big(\ell^n \tilde{u}^n_m(\ell^n)\big) \cap \left[\bigcup_{m' \neq m} \mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}\big(\ell^n \tilde{u}^n_{m'}(\ell^n)\big) \right] \right| < 2^{-\frac{n}{8}\varepsilon_2} t_{\tilde{Y}|\tilde{L}\tilde{U}} \tag{79}$$

if we label the members of $\mathcal{U}(\ell^n)$ as $\tilde{u}_1^n(\ell^n), \tilde{u}_2^n(\ell^n), \ldots, \tilde{u}_M^n(\ell^n)$. Consequently by (75) and the fact that (ℓ^n, u^n), (ℓ'^n, u'^n) have the same ED $\mathcal{Q}_{\mathcal{W}}(\ell^n u^n) = \mathcal{Q}_{\mathcal{W}}(\ell'^n u'^n)$,

$$M^{-1}\sum_{m=1}^{M}\left|\mathcal{T}^n_{\tilde{Y}|\tilde{L}\tilde{U}}\left(\ell^n, \tilde{u}_m^n(\ell^n)\right) \cap \left[\bigcup_{m'\neq m}\bigcup_{\substack{P_{\overline{Y}|\tilde{L}\tilde{U}}\in \\ \mathcal{Q}_{\mathcal{W}}(\ell^n u^n_{m'}(\ell^n))}} \mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}\left(\ell^n u^n_{m'}(\ell^n)\right)\right]\right| < 2^{-\frac{n}{9}\varepsilon_2} t_{\tilde{Y}|\tilde{L}\tilde{U}}. \tag{80}$$

We call the subset $\mathcal{U}(\ell^n)$ the codebook for ℓ^n and its members $\tilde{u}_m^n(\ell^n)$ $m = 1, 2, \ldots, M$ codewords.

2. Choosing Input Sequence to Send through the Channel:

The sender chooses an input sequence $x^n \in \mathcal{X}^n$ according to the output (v^n, ℓ^n) of the correlated source observed by him and his private randomness as follows.

— In the case that outcome of the source is a (δ_1, δ_2)-typical sequence (v^n, ℓ^n) with joint ED $P_{\tilde{V}\tilde{L}}$, the sender chooses a codeword in $\mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n, \ell^n)$ in (78) randomly uniformly (by using his private randomness), say

$$\tilde{u}_m(\ell^n) \in \mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n, \ell^n) \subset \mathcal{U}(\ell^n). \tag{81}$$

Then the sender chooses an input sequence $x^n \in \mathcal{X}^n$ with probability

$$P_{X|VLU}\left(x^n|v^n, \ell^n, \tilde{u}_m^n(\ell^n)\right) \tag{82}$$

by using the chosen $\tilde{u}_m^n(\ell^n)$ and his private randomness and sends it through the channel.

— In the other case i.e., a non-(δ_1, δ_2)-typical sequence is output, the sender chooses an arbitrarily fixed sequence, say x_e^n, and sends it through the channel.

— The codewords randomly chosen here and the random input of the channel generated here will be denoted by U'^n and X'^n in the part of analysis below.

3. Choosing the Common domain $\mathcal{A}$ of Functions F and G:

Let

$$J = \lfloor 2^{n(H(L)-2\varepsilon_1)} \rfloor \tag{83}$$

and let e be an abstract symbol (which stands for that "an error occurs"). Then we define

$$\mathcal{A} = \left\{\{1, 2, \ldots, M\} \times \{1, 2, \ldots, J\}\right\} \cup \{e\}. \tag{84}$$

4. Defining the Functions F and G:
To define functions F and G, we first partition each $\mathcal{T}^n_{\tilde{L}} \subset \mathcal{T}^n_L(\delta_1)$ into J subsets with nearly equal size i.e., each subset has cardinality $\left\lfloor \frac{|\mathcal{T}^n_{\tilde{L}}|}{J} \right\rfloor$ or $\left\lceil \frac{|\mathcal{T}^n_{\tilde{L}}|}{J} \right\rceil$. Then we take the union of the jth subsets in the partitions over all $\mathcal{T}^n_{\tilde{L}} \subset \mathcal{T}^n_L(\delta_1)$ and obtain a subset $\mathcal{L}_j$ of $\mathcal{T}^n_L(\delta_1)$. That is for $j = 1, 2, \ldots, J$

$$|\mathcal{L}_j \cap \mathcal{T}^n_{\tilde{L}}| = \left\lfloor \frac{|\mathcal{T}^n_{\tilde{L}}|}{J} \right\rfloor \text{ or } \left\lceil \frac{|\mathcal{T}^n_{\tilde{L}}|}{J} \right\rceil. \tag{85}$$

4(i) Defining Function F:
The sender observes the output of the source and decides on the value of function F.

— In the case that the source outputs a (δ_1, δ_2)-typical sequence (v^n, ℓ^n), F takes value (m, j) if $\ell^n \in \mathcal{L}_j$, according to sender's private randomness $\tilde{u}_m(\ell^n)$ in (81) is chosen in the step 2 of the coding scheme.
— In the other case $F = e$.

4(ii) Defining Function G:
The receiver observes the output ℓ^n of the component L^n (side information) of the correlated source and output of the channel y^n to decide on the value of function G. We use the abbreviation $\mathcal{Y}_m(\ell^n) = \bigcup_{P_{\overline{Y}|\tilde{L}\tilde{U}} \in \mathcal{Q}_{\mathcal{W}}(\ell^n \tilde{u}^n_m(\ell^n), \tau)} \mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}\big(\ell^n \tilde{u}^n_m(\ell_n), \tau\big)$.

— In the case that $\ell^n \in \mathcal{T}^n_L(\delta_1)$ and that there exists an $m \in \{1, 2, \ldots, M\}$ such that $y^n \in \mathcal{Y}_m(\ell^n) \setminus \left\{ \bigcup_{m' \neq m} \mathcal{Y}_m(\ell^n) \right\}$ G takes value (m, j) if $\ell^n \in \mathcal{L}_j$. Notice that this m must be unique if it exists.
— In the other case $G = e$.

Analysis.

1. Distortion Criterion:
First we recall our assumption that the watermarking distortion measure ρ is bounded i.e.

$$0 \leq \rho \leq \Delta. \tag{86}$$

Then by (63)

$$\frac{1}{n} \Pr\big((V^n, L^n) \notin \mathcal{T}^n_{VL}(\delta_2)\big) E\big[\rho(V'^n, X'^n) | (V^n, L^n \notin \mathcal{T}^n_{VL}(\delta_2)\big] < 2^{-n\xi} \Delta. \tag{87}$$

On the other hand, under the condition that $(V^n, L^n) \in \mathcal{T}^n_{\tilde{V}\tilde{L}} \subset \mathcal{T}^n_{VL}(\delta_2)$, by definition $(V^n, L^n, U'^n) \in \mathcal{T}^n_{\tilde{V}\tilde{L}\tilde{U}}$ with probability one for the joint ED $P_{\tilde{V}\tilde{L}\tilde{U}}$ generated by $P_{\tilde{V}\tilde{L}}$.

So, by (65), (66) and the definition of (U'^n, X'^n) we have that

$$\begin{aligned}
&\frac{1}{n}E\big[\rho(V'^n, X'^n)|(V^n, L^n) \in \mathcal{T}^n_{\tilde{L}\tilde{V}}\big] \\
&= \sum_{(v,\ell,u)\in\mathcal{V}\times\mathcal{L}\times\mathcal{U}} P_{\tilde{V}\tilde{L}\tilde{U}}(v, \ell, u) \sum_x P_{X|VLU}(x|v, \ell, u)\rho(v, x) \\
&= E\rho(\tilde{V}, \tilde{X}) < D_1. \qquad (88)
\end{aligned}$$

Thus it follows from (87) and (88) that

$$\begin{aligned}
&\frac{1}{n}E\rho(V^n, X'^n) \\
&= \Pr\big((V^n, L^n) \notin \mathcal{T}^n_{VL}(\delta_2)\big)E\big[\rho(V^n, X'^n)|(V^n, L^n) \notin \mathcal{T}^n_{VL}(\delta_2)\big] \\
&\quad + \sum_{\mathcal{T}^n_{\tilde{V}\tilde{L}}\subset\mathcal{T}^n_{VL}(\delta_2)} \Pr\big((V^n, L^n) \in \mathcal{T}^n_{\tilde{V}\tilde{L}}\big)E\big[\rho(V^n, X'^n)|(V^n, L^n) \in \mathcal{T}^n_{\tilde{V}\tilde{L}}\big] \\
&< D_1, \qquad (89)
\end{aligned}$$

for sufficiently large n.

2. The Condition of Nearly Uniformity

By the definition of function F in the step 4(i) of the coding scheme,

$\Pr\{F = e\} \le \Pr\big\{(V^n, L^n) \notin \mathcal{T}^n_{VL}(\delta_2)\big\} = 1 - P^n_{VL}(\mathcal{T}^n_{VL}(\delta_2)\big\}$, and hence by (63),

$$|\Pr\{F = e\} - |\mathcal{A}|^{-1}| \le \max\{2^{-n\zeta}, |\mathcal{A}|^{-1}\} \longrightarrow 0 \ (n \to \infty). \qquad (90)$$

Next fix an $\ell^n \in \mathcal{T}^n_L(\delta_1)$ with ED $P_{\tilde{L}}$ (say), let $P_{\tilde{L}\tilde{U}}$ be the joint ED generated by $P_{\tilde{L}}$, and let $\mathcal{Q}(\tilde{L}\tilde{U})$ be the set of joint ED's $P_{\tilde{V}\tilde{L}\tilde{U}}$ with marginal distribution $P_{\tilde{L}\tilde{U}}$ and generated by the joint ED of some (δ_1, δ_2)-typical sequence. Then $\Pr\{U'^n = u^n|L^n = \ell^n\} > 0$, only if $u^n \in \mathcal{U}(\ell^n) = \big\{\tilde{u}^n_m(\ell^n) : m = 1, 2, \ldots, M\big\}$.

Moreover, for a (δ_1, δ_2)-typical sequence (v^n, ℓ^n) with joint ED $P_{\tilde{V}\tilde{L}}$, $\tilde{u}^n_m(\ell^n) \in \mathcal{U}(\ell^n)$, by the coding scheme

$$\begin{aligned}
&\Pr\big\{V^n = v^n, U'^n = u^n_m(\ell^n)|L = \ell^n\big\} \\
&= \begin{cases} P^n_{V|L}(V^n = v^n|\ell^n)|\mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)|^{-1} & \text{if } u^n_m(\ell^n) \in \mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n) \\ 0 & \text{else.} \end{cases} \qquad (91)
\end{aligned}$$

Recalling (78), then we have that for all $\ell^n \in \mathcal{T}^n_{\tilde{L}} \subset \mathcal{T}^n_L(\delta_1)$, $\tilde{u}^n_m(\ell^n) \in \mathcal{U}(\ell^n)$

$$\begin{aligned}
&\Pr\left\{U'^n = \tilde{u}^n_m(\ell^n)|L = \ell^n\right\} \\
&= \sum_{P_{\tilde{V}\tilde{L}\tilde{U}} \in \mathcal{Q}(\tilde{L}\tilde{U})} \sum_{v^n \in \mathcal{T}^n_{\tilde{V}|\tilde{L}\tilde{U}}(\ell^n \tilde{u}^n_m(\ell^n))} P^n_{V|L}(v^n|\ell^n)|\mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)|^{-1}. \qquad (92)
\end{aligned}$$

By (77) we have that

$$\left[M(1+\varepsilon)\right]^{-1}\frac{|\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)|}{|\mathcal{T}^n_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)|} < |\mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)|^{-1} < \left[M(1-\varepsilon)\right]^{-1}\frac{|\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)|}{|\mathcal{T}^n_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)|}. \qquad (93)$$

On the other hand,

$$\begin{aligned}
&\sum_{P_{\tilde{V}\tilde{L}\tilde{U}} \in \mathcal{Q}(\tilde{L}\tilde{U})} \sum_{v^n \in \mathcal{T}^n_{\tilde{V}|\tilde{L}\tilde{U}}(\ell^n \tilde{u}^n_m(\ell^n))} P^n_{V|L}(v^n|\ell^n)\frac{|\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)|}{|\mathcal{T}^n_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)|} \\
&= \sum_{P_{\tilde{V}\tilde{L}\tilde{U}} \in \mathcal{Q}(\tilde{L}\tilde{U})} \sum_{v^n \in \mathcal{T}^n_{\tilde{V}|\tilde{L}\tilde{U}}(\ell^n \tilde{u}^n_m(\ell^n))} P^n_{V^n|L}(\mathcal{T}^n_{\tilde{V}|\tilde{L}}(\ell^n)|\ell^n)\frac{|\mathcal{T}^n_{\tilde{U}|\tilde{L}}(\ell^n)|}{|\mathcal{T}^n_{\tilde{V}|\tilde{L}}(\ell^n)||\mathcal{T}^n_{\tilde{U}|\tilde{V}\tilde{L}}(v^n,\ell^n)|} \\
&= \sum_{P_{\tilde{V}\tilde{L}\tilde{U}} \in \mathcal{Q}(\tilde{L}\tilde{U})} P^n_{V|L}\left(\mathcal{T}^n_{\tilde{V}|\tilde{L}}(\ell^n)|\ell^n\right) \\
&= \Pr\left\{(V^n, \ell^n) \in \mathcal{T}^n(\delta_1, \delta_2)|\ell^n\right\}, \qquad (94)
\end{aligned}$$

where the first equality holds because the value of $P^n_{V|L}(v^n|\ell^n)$ for given ℓ^n depends on v^n through the conditional ED; the second equality holds by the fact that $\frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{V}|\tilde{L}}t_{\tilde{U}|\tilde{V}\tilde{L}}} = \frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{V}\tilde{U}|\tilde{L}}} = \frac{1}{t_{\tilde{V}|\tilde{L}\tilde{U}}}$; and the last equality holds because $P_{\tilde{V}\tilde{L}\tilde{U}}$ is generated by $P_{\tilde{V}\tilde{L}}$ uniquely (see its definition in condition (iii)).

Thus by combining (62), (92)–(94), we obtain for an $\eta > 0$ with $\eta \to 0$ as $n \to \infty, \varepsilon \to 0$

$$(1-\eta)M^{-1} < \Pr\left\{U'^n = \tilde{u}^n_m(\ell^n)|L = \ell^n\right\} < (1+\eta)M^{-1}, \qquad (95)$$

for $\ell^n \in \mathcal{T}^n_{L(\delta_1)}$, $\tilde{u}^n_m(\ell^n) \in \mathcal{U}(\ell^n)$.

So for $m \in \{1, 2, \ldots, M\}$, $j \in \{1, 2, \ldots, J\}$,

$$\begin{aligned}\Pr\left\{F = (m, j)\right\} &= \Pr\left\{U'^n = \tilde{u}_m^n(L^n), L^n \in \mathcal{L}_j\right\} \\ &= \sum_{\ell^n \in \mathcal{L}_j} P_L^n(\ell^n) \Pr\left\{U'^n = \tilde{u}_m^n(\ell^n) | L = \ell^n\right\} \\ &< (1+\eta) M^{-1} P_L^n(\mathcal{L}_j). \end{aligned} \tag{96}$$

Since $|\mathcal{T}_{\tilde{L}}^n| > 2^{n\left(H(\tilde{L})+\frac{\varepsilon_1}{2}\right)}$ for sufficiently large n, by (67) and (83), we have that $\frac{|\mathcal{T}_{\tilde{L}}^n|}{J} > 2^{\frac{n}{2}\varepsilon_1}$ and hence by (85)

$$|\mathcal{L}_j \cap \mathcal{T}_{\tilde{L}}^n| \le \left\lceil \frac{|\mathcal{T}_{\tilde{L}}^n|}{J} \right\rceil < \frac{|\mathcal{T}_{\tilde{L}}^n|}{J} + 1 < \frac{|\mathcal{T}_{\tilde{L}}^n|}{J}\left(1 + 2^{-\frac{n}{2}\varepsilon_1}\right).$$

Because the value of $P_L^n(\ell^n)$ depends on ℓ^n through its ED, this means that

$$P_L^n(\mathcal{L}_j \cap \mathcal{T}_{\tilde{L}}^n) < J^{-1} P_L^n(\mathcal{T}_{\tilde{L}}^n)\left(1 + 2^{-\frac{n}{2}\varepsilon}\right)$$

and consequently

$$P_L^n(\mathcal{L}_k) < P_L^n\left(\mathcal{T}_L^n(\delta_1)\right) J^{-1}\left(1 + 2^{-\frac{n}{2}\varepsilon_1}\right) \tag{97}$$

which with (96) is followed by

$$\Pr\left\{F = (m, j)\right\} < M^{-1} J^{-1}(1+\eta)\left(1 + 2^{-\frac{n}{2}\varepsilon_1}\right) P_L^n\left(\mathcal{T}_L^n(\delta_1)\right). \tag{98}$$

Similarly we have that

$$\Pr\left\{F = (m, j)\right\} > M^{-1} J^{-1}(1-\eta)\left(1 - 2^{-\frac{n}{2}\varepsilon_1}\right) P_L^n\left(\mathcal{T}_L^n(\delta_1)\right). \tag{99}$$

Now (61), (98) and (99) together imply that for an $\eta' > 0$ with $\eta' \to 0$ as $n \to \infty$, $\eta \to 0$,

$$\sum_{(m,j)} \left| \Pr\left\{F = (m, j)\right\} - |\mathcal{A}|^{-1} \right| < \eta', \tag{100}$$

which with (90) completes the proof of condition of nearly uniformity.

3. The Rate:
   In (70) one can choose

$$\alpha > I\big(U; Y(\mathcal{W})|L\big) - \varepsilon' \text{ for all } \varepsilon' \text{ with } \varepsilon_2 < \varepsilon' < \frac{1}{2}\xi.$$

   Then by (63), (83), (84), (100), we know that for an $\eta'' > 0$ with $\eta'' \to 0$ as $n \to \infty$, $\eta' \to 0$

$$\begin{aligned}\frac{1}{n}H(F) &> \frac{1}{n}\log|\mathcal{A}| - \eta'' > I\big(U; Y(\mathcal{W})|L\big) - \varepsilon' + H(L) - 2\varepsilon_1 - \eta' \\ &= I\big(U; Y(\mathcal{W})|L\big) + I(U; L) + H(L|U) - \varepsilon' - 2\varepsilon_1 - \eta' \\ &= I\big(U; L, Y(\mathcal{W})\big) + H(L|U) - \varepsilon' - 2\varepsilon_1 - \eta', \end{aligned} \tag{101}$$

   for sufficiently large n.
4. Estimation of Probability of Error:
   In and only in the following three cases an error occurs.
   Case 1. The source outputs a non-(δ_1, δ_2)-typical sequence whose probability is less than $2^{-n\zeta}$ by (63).
   Now we assume that a (δ_1, δ_2)-typical sequence (v^n, ℓ^n) with joint ED $P_{\tilde{V}\tilde{L}}$ is output. So the sender first chooses a $\tilde{u}_m^n(\ell^n) \in \mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n, \ell^n)$, then an $x^n \in \mathcal{X}^n$ according to his private randomness and sends x^n through the channel. Consequently a $y^n \in \mathcal{Y}^n$ is output by the channel. Then in the following two cases an error occurs.
   Case 2. A codeword $\tilde{u}_m(\ell^n) \in \mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n) \subset \mathcal{U}^n(\ell^n)$ is chosen and an output sequence

$$y^n \notin \mathcal{Y}_m(\ell^n) = \bigcup_{P_{\overline{Y}|\tilde{L}\tilde{U}} \in \mathcal{Q}_{\mathcal{W}}(\ell^n, \tilde{u}_m(\ell^n))} \mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}\big(\ell^n \tilde{u}_m(\ell^n), \tau\big)$$

   is output of the channel. Suppose now $W \in \mathcal{W}$ governs the channel. Then by (64), and (65) the probability that $y^n \in \mathcal{Y}^n$ is output of the channel under the condition that $(V^n, L^n) = (v^n, \ell^n) \in \mathcal{T}^n(\delta_1, \delta_2)$ is output of the correlated source and $U'^n = \tilde{u}_m^n(\ell^n) \in \mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n, \ell^n)$ is chosen is

$$\begin{aligned}&\Pr\big\{Y'^n = y^n | (V^n, L^n) = (v^n, \ell^n), U'^n = \tilde{u}_m^n(\ell^n)\big\} \\ &= \sum_{x^n \in \mathcal{X}^n} P^n_{X|VLU}(x^n|v^n, \ell^n, \tilde{u}_m^n(\ell^n)) W^n(y^n|x^n) \\ &= P^n_{\tilde{Y}(\mathcal{W})|\tilde{V}\tilde{L}\tilde{U}}\big(y^n|v^n, \ell^n, \tilde{u}_m(\ell^n)\big). \end{aligned} \tag{102}$$

On the other hand

$$\mathcal{T}^n_{\tilde{Y}(W)|\tilde{V}\tilde{L}\tilde{U}}\big(v^n\ell^n u^n_m(\ell^n),\tau\big)\subset\mathcal{T}^n_{\tilde{Y}(W)|\tilde{L}\tilde{U}}\big(\ell^n u^n_m(\ell^n),\tau\big)\subset\mathcal{Y}_m.$$

So the probability that such an error occurs vanishes exponentially as n grows.

Case 3. A codeword $\tilde{u}^n_m(\ell^n)$ is chosen and a $y^n\in\mathcal{Y}_m\cap\Big[\bigcup_{m'\neq m}\mathcal{Y}_{m'}\Big]$ is output of the channel.

Now by (91), (93), (95), and simple calculation, we obtain that

$$\begin{aligned}&\big[(1-\eta)(1-\varepsilon)\big]^{-1}P^n_{V|L}(v^n|\ell^n)\frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{U}|\tilde{V}\tilde{L}}}\\&<\Pr\big\{V^n=v^n|L^n=\ell^n,U'^n=\tilde{u}^n_m(\ell^n)\big\}\\&<\big[(1+\eta)(1+\varepsilon)\big]^{-1}P^n_{V|L}(v^n|\ell^n)\frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{U}|\tilde{V}\tilde{L}}}\end{aligned}\tag{103}$$

for (δ_1,δ_2)-typical sequences (v^n,ℓ^n) with joint ED $P_{\tilde{V}\tilde{L}}$ and $\tilde{u}_m(\ell^n)\in\mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n,\ell^n)$, where $P_{\tilde{V}\tilde{L}\tilde{U}}$ is the ED generated by $P_{\tilde{V}\tilde{L}}$.

Moreover, since $t_{\tilde{U}|\tilde{V}\tilde{L}}=\frac{t_{\tilde{V}\tilde{U}|\tilde{L}}}{t_{\tilde{V}|\tilde{L}}}$, $t_{\tilde{U}\tilde{V}|\tilde{L}}=t_{\tilde{U}|\tilde{L}}t_{\tilde{V}|\tilde{L}\tilde{U}}$, and since for given ℓ^n, the value of $P^n_{V|L}(v^n|\ell^n)$ depends on v^n through the conditional ED,

$$P^n_{V|L}(v^n|\ell^n)\frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{U}|\tilde{V}\tilde{L}}}=P^n_{V|L}(v^n|\ell^n)\frac{t_{\tilde{V}|\tilde{L}}}{t_{\tilde{V}|\tilde{L}\tilde{U}}}=P^n_{V|L}\big(\mathcal{T}^n_{\tilde{V}|\tilde{L}}(\ell^n)|\ell^n\big)\frac{1}{t_{\tilde{V}|\tilde{L}\tilde{U}}}.\tag{104}$$

Further it is well known that for all $(v^n,\ell^n,u^n)\in\mathcal{T}^n_{\tilde{V}\tilde{L}\tilde{U}}$, $\lim\limits_{n\to\infty}\frac{1}{n}\Big(\log P^n_{\tilde{V}|\tilde{L}\tilde{U}}(v^n|\ell^n,u^n)-\log\frac{1}{t_{\tilde{V}|\tilde{L}\tilde{U}}}\Big)=0$.

So by (103) and (104), we have that

$$\begin{aligned}&\Pr\big\{V=v^n|L^n=\ell^n,U'^n=\tilde{u}^n_m(\ell^n)\big\}\\&<2^{n\theta}P^n_{V|L}\big(\mathcal{T}^n_{\tilde{V}|\tilde{L}}(\ell^n)|\ell^n\big)P^n_{\tilde{V}|\tilde{L}\tilde{U}}\big(v^n|\ell^n,\tilde{u}^n_m(\ell^n)\big)\\&\le2^{n\theta}P^n_{\tilde{V}|\tilde{L}\tilde{U}}\big(v^n|\ell^n,\tilde{u}^n_m(\ell^n)\big)\end{aligned}\tag{105}$$

for (δ_1,δ_2)-typical sequences (v^n,ℓ^n) with ED $P_{\tilde{V}\tilde{L}}$, $\tilde{u}^n_m\in\mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(\ell^n)\subset\mathcal{U}(\ell^n)$ and sufficiently large n, and a $\theta\to0$ as $n\to\infty$.

We choose $\theta<\frac{1}{20}\varepsilon_2$.

Since $\Pr\left\{(V^n, L^n) = (v^n, \ell^n), U'^n = u^n\right\} > 0$ only if (v^n, ℓ^n) is (δ_1, δ_2) typical and $u^n \in \mathcal{U}_{\tilde{U}|\tilde{V}\tilde{L}}(v^n, \ell^n)$, by (102) and (105) we have that

$$\begin{aligned}
&\Pr\left\{Y'^n = y^n | L^n = \ell^n, U'^n = \tilde{u}_m^n(\ell^n)\right\} \\
&= \sum_{v^n \in \mathcal{V}^n} \Pr\left\{V^n = v^n | L^n = \ell^n, U'^n = \tilde{u}_m(\ell^n)\right\} \\
&\qquad \Pr\left\{Y'^n = y^n | (V^n, L^n) = (v^n, \ell^n), U'^n = \tilde{u}_m^n(\ell^n)\right\} \\
&\le \sum_{v^n \in \mathcal{V}^n} 2^{n\theta} P^n_{\tilde{V}|\tilde{L}\tilde{U}}(v^n | \ell^n, u_m^n(\ell^n)) P^n_{\tilde{Y}(W)|\tilde{V}\tilde{L}\tilde{U}}(y^n | v^n, \ell^n, u^n) \\
&\le 2^{n\theta} P^n_{\tilde{Y}(W)|\tilde{L}\tilde{U}}\left(y^n | \ell^n, u_m^n(\ell^n)\right) \qquad (106)
\end{aligned}$$

for $\ell^n \in \mathcal{T}_L^n(\delta_1)$, $\tilde{u}_m(\ell^n) \in \mathcal{U}(\ell^n)$ and $y^n \in \mathcal{Y}^n$ if $W \in \mathcal{W}$ governs the channel. Now we obtain an upper bound in terms of a product probability distribution $P^n_{\tilde{Y}(\mathcal{W})|\tilde{L}\tilde{U}}\left(y^n | \ell^n, u_m^n(\ell^n)\right)$ whose value depends on y^n through the conditional ED. Consequently by (80) and (106) we have that for all $\ell^n \in \mathcal{T}_L^n(\delta_1)$, $\tilde{u}_m(\ell^n) \in \mathcal{U}(\ell^n)$ with joint ED $P_{\tilde{L}\tilde{U}}$, $P_{\tilde{Y}|\tilde{L}\tilde{U}} \in \mathcal{Q}_{\mathcal{W}}\left(\ell^n, \tilde{u}_m(\ell^n), \tau\right)$

$$\begin{aligned}
&M^{-1} \sum_{m=1}^{M} \Pr\left\{ Y'^n \in \mathcal{T}^n_{\tilde{Y}|\tilde{L}\tilde{U}}\left(\ell^n, u_m^n(\ell^n)\right) \cap \left[\bigcup_{m' \neq m} \mathcal{Y}_{m'}(\ell^n) \right] | \ell^n, \tilde{u}_m^n(\ell^n) \right\} \\
&\le 2^{n\theta} M^{-1} \sum_{m=1}^{M} P^n_{\tilde{Y}(W)|\tilde{L}\tilde{U}} \left\{ \mathcal{T}^n_{\tilde{Y}|\tilde{L}\tilde{U}}\left(\ell^n u_m^n(\ell^n)\right) \cap \left[\bigcup_{m' \neq m} \mathcal{Y}_{m'}(\ell^n) \right] | \ell^n, u_m^n(\ell^n) \right\} \\
&\le 2^{n\theta} \cdot 2^{-\frac{n}{9}\varepsilon_2} P^n_{\tilde{Y}(W)|\tilde{L}\tilde{U}} \left\{ \mathcal{T}^n_{\tilde{Y}|\tilde{L}\tilde{U}}\left(\ell^n, u_m^n(\ell^n)\right) | \ell^n, \tilde{u}_m^n(\ell^n) \right\} \\
&\le 2^{-n\left(\frac{1}{9}\varepsilon_2 - \theta\right)} < 2^{-\frac{n}{20}\varepsilon_2}, \qquad (107)
\end{aligned}$$

where the last inequality holds by our choice $\theta < \frac{\varepsilon_2}{20}$. Recalling

$$\mathcal{Y}_m(\ell^n) = \bigcup_{P_{\tilde{Y}|\tilde{L}\tilde{U}} \in \mathcal{Q}_{\mathcal{W}}(\ell^n \tilde{u}_m(\ell^n), \tau)} \mathcal{T}^n_{\tilde{Y}|\tilde{L}\tilde{U}}\left(\ell^n u_m^n(\ell^n)\right),$$

by the union bound and (107) we obtain that

$$M^{-1}\sum_{m=1}^{M}\Pr\left\{Y'^n\in\mathcal{Y}_m(\ell^n)\cap\left[\bigcup_{m'\neq m}\mathcal{Y}_m(\ell^n)\right]|L^n=\ell^n,U'^n=\tilde{u}_m(\ell^n)\right\}$$

$$<(n+1)^{|\mathcal{L}||\tilde{U}||\mathcal{Y}|}2^{-\frac{n}{20}\varepsilon_2}$$

$$<2^{-\frac{n}{21}\varepsilon_2} \tag{108}$$

for $\ell^n\in\mathcal{T}_L^n(\delta_1)$, $\tilde{u}_m^n(\ell^n)\in\mathcal{U}(\ell^n)$ and sufficiently large n. Finally by (95) and (108) we obtain an upper bound to the probability that on error of this ED occurs, under the condition $L^n=\ell^n\in\mathcal{T}_L^n(\delta_1)$.

$$\sum_{m=1}^{M}\Pr\left\{U'^n=\tilde{u}_m(\ell^n)|L^n=\ell^n\right\}$$

$$\Pr\left\{Y'^n\in\mathcal{Y}_m(\ell^n)\cap\left[\bigcup_{m'\neq m}\mathcal{Y}_{m'}(\ell^n)\right]|L^n=\ell^n,U'^n=\tilde{u}_m(\ell^n)\right\}$$

$$<(1+\eta)\sum_{m=1}^{M}M^{-1}$$

$$\Pr\left\{Y'^n\in\mathcal{Y}_m(\ell^n)\cap\left[\bigcup_{m'\neq m}\mathcal{Y}_{m'}(\ell^n)\right]|L^n=\ell^n,U'^n=\tilde{u}_m(\ell^n)\right\}$$

$$<(1+\eta)2^{-\frac{n}{21}\varepsilon_2}, \tag{109}$$

which completes the proof because by definition $\sum_{m=1}^{M}\Pr\left\{U'^n=\tilde{u}_m(\ell^n)|L^n=\ell^n\right\}=1$ for all $\ell^n\in\mathcal{T}_L^n(\delta_1)$.

□

Remark Our model of identification becomes that in [27] if L takes a constant value with probability one. So our proof of the lemma above provides a new proof of [27, Theorem 4] (as special case) without using the Gelfand-Pinsker Theorem in [20].

Corollary 168 (Direct part of Theorem 156) *For all single channels W*

$$C_{\mathrm{CRI}}\big((V,L),W,D_1\big)\geq\max_{(V,L,U,X,Y)\in\mathcal{Q}((V,L),W,D_1)}\big[I(U;L,Y)+H(L|U)\big].$$

Lemma 169 (Direct part of Theorem 159) *For all compound channels* $\mathcal{W}$

$$C_{\mathrm{CRII}}\big((V,L),W,R_{\mathrm{K}},D_1\big) \geq \sup_{(V,L,U,X)\in\mathcal{Q}_1^*((V,L),\mathcal{W},R_{\mathrm{K}},D_1)} \big[I\big(U;L,Y\big)+H(L|U)\big]+R_{\mathrm{K}}. \tag{110}$$

Proof By the same reason as in the proof of the previous lemma, it is sufficient for us to show the achievability of $I\big(U;L,Y(\mathcal{W})\big)+H(L|U)+R_{\mathrm{K}}$ for (V,L,U,X) with $E\rho(V,X)<D_1$ and for some $\xi>0$

$$I\big(U;Y(\mathcal{W})|L\big)+R_{\mathrm{K}}-I(U;V|L)>\xi. \tag{111}$$

In the case $I\big(U;Y(\mathcal{W})|L\big) > I(U;V|L)$, by the previous lemma $I\big(U;LY(\mathcal{W})\big)+H(L|U)$ is achievable even if the noiseless channel is absent. So sender and receiver may generate $n\big(I(U;LY(\mathcal{W}))+H(L|U)\big)$ bits of common randomness and at the same time the sender sends R_{K} bits of his private randomness via the noiseless channel to the receiver to make additionally nR_{K} bits of common randomness. That is, the rate $I\big(U;L,Y(\mathcal{W})\big)+H(L|U)+R_{\mathrm{K}}$ is achievable.

So, next we may assume that $I\big(U;Y(\mathcal{W})|L\big)\leq I(U;V;|L)$. Moreover we can assume $I\big(U;Y(\mathcal{W})|L\big)>0$, because otherwise $I\big(U;L,Y(\mathcal{W})\big)+H(L|U)+R_{\mathrm{K}}=I(U;L)+H(L|U)+R_{\mathrm{K}}=H(L)+R_{\mathrm{K}}$ is achievable as follows. We partition $\mathcal{T}_L^n(\delta_1)$ into $\mathcal{L}_j$, $j=1,2,\ldots,J$ as in the step 4) of the coding scheme in the proof of the previous lemma to get $n\big(H(L)-2\varepsilon_1\big)$ bits of common randomness and get other nR_{K} bits of common randomness by using the noiseless channel. Thus it is sufficient for us to assume that

$$0<I\big(U;Y(\mathcal{W})|L\big)\leq I(U;V|L)<I\big(U;Y(\mathcal{W})|L\big)+R_{\mathrm{K}}-\xi, \tag{112}$$

for a ξ with $0<\xi<R_{\mathrm{K}}$.

We shall use (δ_1,δ_2)-typical sequences, the joint ED's $P_{\tilde{L}\tilde{U}}$ and $P_{\tilde{V}\tilde{L}\tilde{U}}$ generated by the ED's $P_{\tilde{L}}$ and $P_{\tilde{U}\tilde{L}}$ respectively, and the RV's $(\tilde{V},\tilde{L},\tilde{U},\tilde{X})$ and $\big(\tilde{V},\tilde{L},\tilde{U},\tilde{X},\tilde{Y}(\mathcal{W})\big)$ in (64) and (65) satisfying (66)–(69), which are defined in the conditions (i)–(iv) in the proof of the previous lemma.

Instead of the choice α in (70) we now choose $\beta_1,\beta_2>0$ and $\beta_3\geq 0$ for arbitrarily small but fixed ε_2 with $0<\varepsilon_2<\frac{1}{2}\xi$ such that

$$I\big(U;Y(\mathcal{W})|L\big)-\frac{3}{2}\varepsilon_2<\beta_1<I\big(U;Y(\mathcal{W})|L\big)-\varepsilon_2, \tag{113}$$

$$I(U;V|L)-I\big(U;Y(\mathcal{W})|L\big)+\xi\leq\beta_2\leq R_{\mathrm{K}} \tag{114}$$

and

$$0\leq\beta_3=R_{\mathrm{K}}-\beta_2. \tag{115}$$

Notice that the existence and positivity of β_2 are guaranteed by (112).

By adding both sides of the first inequalities in (113) and (114), we obtain that

$$\beta_1 + \beta_2 > I(U; V|L) + \left(\xi - \frac{3}{2}\varepsilon_2\right), \tag{116}$$

and by the first inequality in (113) and the equality in (115) we have that

$$\beta_1 + \beta_2 + \beta_3 > I\big(U; Y(\mathcal{W})|L\big) + R_{\mathrm{K}} - \frac{3}{2}\varepsilon_2. \tag{117}$$

Let $\xi - \frac{3}{2}\varepsilon_2 = 2\eta$ and rewrite (116) as

$$\beta_1 + \beta_2 > I(U; V|L) + 2\eta. \tag{118}$$

Then $\eta > \frac{\xi}{8} > 0$ by our choice $\varepsilon_2 < \frac{1}{2}\xi$.

Next as in the proof to the previous lemma we fix an (arbitrary small) positive ε_2, η, choose ε_1 (and consequently δ_1, δ_2) sufficiently small so that $\varepsilon_1 < \min\left(\frac{1}{2}\varepsilon_2, \frac{1}{2}\eta\right)$. Then by (69) and the second inequality in (115) we have that

$$\beta_1 < I\big(\tilde{U}; \tilde{Y}(\mathcal{W})|\tilde{L}\big) - \frac{\varepsilon_2}{2}, \tag{119}$$

and by (70) and (118) we have that

$$\beta_1 + \beta_2 > I(\tilde{U}; \tilde{V}|\tilde{L}) + \frac{3}{2}\eta. \tag{120}$$

Without loss of generality we assume that $2^{n\beta_1}$, $2^{n\beta_2}$ and $2^{n\beta_3}$ are integers and denote by $M_1 = 2^{n\beta_1}$, $I = 2^{n\beta_2}$ and $K' = 2^{n\beta_3}$.

Then similarly as in the proof of the previous lemma, we have that for sufficiently large n, sufficiently small τ, all joint ED's $P_{\tilde{V}\tilde{L}\tilde{U}}$ generated by ED's of (δ_1, δ_2)-typical sequences and $\mathcal{Q}_{\mathcal{W}}(\ell^n u^n, \tau)$ in the proof of the previous lemma,

$$2^{n\eta}\frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{U}|\tilde{V}\tilde{L}}} < M_1 I \tag{121}$$

and

$$M_1 < 2^{-\frac{n}{3}\varepsilon_2}\frac{t_{\tilde{U}|\tilde{L}}}{t_{\tilde{U}|\tilde{L}\overline{Y}}}, \tag{122}$$

for all $P_{\overline{Y}|\tilde{L}\tilde{U}} \in \mathcal{Q}_{\mathcal{W}}(\ell^n u^n, \tau)$.

Coding scheme.

1. Choosing the Codebook:
We choose a codebook for all $\ell^n \in \mathcal{T}_L^n(\delta_1)$ in a similar way as in the step 1 of the coding scheme in the proof of the previous lemma. But we now use Lemma 164 for $\alpha = \beta_1 + \beta_2$ and Lemma 166 for $\gamma = \frac{\varepsilon_2}{3}$ instead of Lemmas 164 and 165. Thus by random choice we obtain subsets of $\mathcal{T}_U^n$ $\mathcal{U}^i(\ell^n) = \left\{\tilde{u}_{m,i}^n(\ell^n) : m = 1, 2, \ldots, M_1\right\}$ for $i = 1, 2, \ldots, I$ for all $\ell^n \in \mathcal{T}_L^n(\delta_1)$ such that for

$$\mathcal{U}^*(\ell^n) = \bigcup_{i=1}^{I} \mathcal{U}^i(\ell^n), \tag{123}$$

and $\mathcal{U}^*_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n) = \mathcal{U}^*(\ell^n) \cap \mathcal{T}^n_{\tilde{U}|\tilde{V}\tilde{L}}(v^n, \ell^n)$, where $P_{\tilde{V}\tilde{L}\tilde{U}}$ is the ED generated by the joint ED $P_{\tilde{V}\tilde{L}}$ of (δ_1, δ_2)-sequences (v^n, ℓ^n) as before, and with an abuse of notation in the union in (123): counting it twice and labeling it as different elements $\tilde{u}_{m,i}^n(\ell^n)$ and $\tilde{u}_{m',i'}^n(\ell^n)$ if a codeword appears twice in it, the following holds.

$$\left|\mathcal{U}^*_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n) - M_1 I \frac{t_{\tilde{U}|\tilde{V}\tilde{L}}}{t_{\tilde{U}|\tilde{V}\tilde{L}}}\right| < M_1 I \frac{t_{\tilde{U}|\tilde{V}\tilde{L}}}{t_{\tilde{U}|\tilde{V}\tilde{L}}}\varepsilon, \tag{124}$$

and for $\mathcal{Q}_{\mathcal{W}}(\ell^n v^n, \tau)$ in the proof of the previous lemmas and any conditional ED $P_{\tilde{Y}|\tilde{L}\tilde{U}}$,

$$I^{-1}\sum_{i=1}^{I} M_1^{-1} \sum_{m=1}^{M_1} \left| \mathcal{T}^n_{\tilde{Y}|\tilde{L}\tilde{U}}\left(\ell^n u_{m,i}^n(\ell^n)\right) \cap \left[\bigcup_{m' \neq m} \bigcup_{\substack{P_{\overline{Y}|\tilde{L}\tilde{U}} \in \\ \mathcal{Q}_{\mathcal{W}}(\ell^n v^n)}} \mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}\left(\ell^n u_{m',i}^n(\ell^n)\right) \right] \right|$$

$$< 2^{-\frac{n}{7}\varepsilon_2} t_{\tilde{Y}|\tilde{L}\tilde{V}} \tag{125}$$

here (124) and (125) are analogous to (77) and (80) respectively, and are shown in an analogous way.

2. Choosing Inputs of the Channels:
In the current model, we have an additional noiseless channel with rate R_{K} except for the noisy channel which exists in the Model I. The sender chooses the inputs of the two channels as follows.

2(i) Choosing the input sequence of the noisy channel:

— In the case that the source outputs a (δ_1, δ_2)-typical sequence (v^n, ℓ^n) with joint ED $P_{\tilde{V}\tilde{L}}$, by (124) for the ED $P_{\tilde{V}\tilde{L}\tilde{U}}$ generated by $P_{\tilde{V}\tilde{L}}$, $\mathcal{U}^*_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n) \neq \emptyset$. Then similarly to the Step 2 of the coding scheme in the proof of the previous lemma, the sender randomly and uniformly chooses

a member of $\mathcal{U}^*_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)$, say $\tilde{u}^n_{m,i}(\ell^n)$, and according to the probability $P_{X|VLU}\left(x^n|v^n, \ell^n, \tilde{u}^n_{m,i}(\ell^n)\right)$ chooses an input sequence x^n of the channel $\mathcal{W}$ and sends x^n through the channel.

— In the case that the output of the source is non-(δ_1, δ_2)-typical, the sender sends an arbitrary fixed sequence x^n_e through the channel.

2(ii) Choosing the Input of the Noiseless Channel:

— In the case that a (δ_1, δ_1)-typical sequence (v^n, ℓ^n) with joint ED $P_{\tilde{V}\tilde{L}}$ is output of the correlated channel, the sender first spends $\log I = n\beta_2$ bits to send the index $i \in \{1, 2, \ldots, I\}$ to the receiver via the noiseless channel if a codeword $\tilde{u}^n_{m,i}(\ell^n) \in \mathcal{U}^i(\ell^n) \subset \mathcal{U}^*(\ell^n)$ is chosen in the substep 2(i) in the current coding scheme, then he randomly and uniformly chooses a $k' \in \{1, 2, \ldots, K'\}$ independent of the output of the source and sends it through the noiseless channel by using the rest of $nR_{\mathrm{K}} - n\beta_2 = n\beta_3 = \log K'$ bits.

— In the case that a non-(δ_1, δ_2)-typical sequence is output, the sender sends a constant message through the noiseless channel.

3. Choosing the common range $\mathcal{A}$ of functions F and G:
Let J be as in (83) and

$$\mathcal{A} = \left[\{1, 2, \ldots, M_1\} \times \{1, 2, \ldots, I\} \times \{1, 2, \ldots, K'\} \times \{1, 2, \ldots, J\}\right] \cup \{e\}. \tag{126}$$

4. Defining the functions G and G:
Partition $\mathcal{T}^n_L(\delta_1)$ into $\mathcal{L}_j$, $j = 1, 2, \ldots, J$ as in the step 4 of the coding scheme in the proof of the previous lemma and let $\mathcal{K}_n = \{1, 2, \ldots, I\} \times \{1, 2, \ldots, K'\}$.

4(i) Defining function F:
The sender decides on the value of function F according to the output of the correlated source and his private randomness as follows.

— In the case that a (δ_1, δ_2)-typical sequence (v^n, ℓ^n) is output, F takes value (m, i, k', j) if $\ell^n \in \mathcal{L}_j$, $\tilde{u}^n_{m,i}(\ell^n) \in \mathcal{U}j(\ell^n) \cap \mathcal{U}^*_{\tilde{U}|\tilde{V}\tilde{L}}(v^n\ell^n)$ is chosen in step 2) of the current coding scheme, and k' is chosen for sending it via the noiseless channel in the last $n\beta_3$ bits (that means (i, k') is sent through the noiseless channel).

— In the other case $F = e$.

4(ii) Defining function G:
The receiver decides on the value of the function G according to the output $(i, k') \in \mathcal{K}_n$ of the noiseless channel, the output ℓ^n of the component L^n of the correlated source, and the output $y^n \in \mathcal{Y}^n$ of the noisy compound channel $\mathcal{W}$ as follows.

Let

$$\mathcal{Y}_{m,i}(\ell^n) = \bigcup_{P_{\overline{Y}|\tilde{L}\tilde{U}} \in \mathcal{Q}_{\mathcal{W}}(\ell^n \tilde{u}^n_{m,i}(\ell^n),\tau)} \mathcal{T}^n_{\overline{Y}|\tilde{L}\tilde{U}}\big(\ell^n u^n_m, i(\ell^n)\big)$$

for $m = 1, 2, \ldots, M_1$, $i = 1, 2, \ldots, I$, and the ED $P_{\tilde{L}\tilde{U}}$ generated by the ED $P_{\tilde{L}}$ of $\ell^n \in \mathcal{L}_j \subset \mathcal{T}^n_L(\delta_1)$.

— In the case that (i, k') is output of the noiseless channel, $\ell^n \in \mathcal{T}^n_L(\delta_1)$ is output of the source, and there exists an $m \in \{1, 2, \ldots, M_1\}$ such that the output of the noisy compound channel $\mathcal{W}$, $y^n \in \mathcal{Y}_{m,i}(\ell^n) \setminus \left\{ \bigcup_{m' \neq m} \mathcal{Y}_{m',i}(\ell^n) \right\}$, G takes value (m, i, k', j) if $\ell^n \in \mathcal{L}_j$.

— In the other case $G = e$.

Analysis.

1. – 3. Distortion criterion, the nearly uniformity condition, and the rate.

One can verify the distortion criterion, the nearly uniformity condition and the rate

$$\frac{1}{n} \log H(F) > \beta_1 + \beta_2 + \beta_3 + o(1) = I\big(U; Y(\mathcal{W})|L\big) + R_{\mathrm{K}} + o(1)$$

(c.f. (117)), and obtain analogous inequalities

$$(1 - \eta)(M_1 I)^{-1} < \Pr\left\{U'^n = u^n_{m,i}(\ell^n)|L = \ell^n\right\} < (1 + \eta)(M_1 I)^{-1} \tag{127}$$

to the inequalities in (95) for $\ell^n \in \mathcal{T}^n_L(\delta_1)$, $u^n_{m,i}(\ell^n) \in \mathcal{U}^*(\ell^n)$ and RV U'^n chosen by the sender in step 2 of the coding scheme in the same way as in parts 1 – 3 of the Analysis in the proof of the previous lemma except that the roles of $\mathcal{U}(\ell^n)$ and (77) there are played by $\mathcal{U}^*(\ell^n) = \bigcup_{i=1}^{I} \mathcal{U}^i(\ell^n)$ and (124). Notice that in those parts of the proof of the previous lemma (80) is not used, neither is (125) here correspondingly.

4. Estimation of probability of error:

By the same reason as in the proof of the previous lemma, the probabilities of errors of the first two types, the error caused by that a non-(δ_1, δ_2)-typical sequence is output and the error caused by that $\tilde{u}_{m,i}(\ell^n)$ is chosen and $y^n \notin \mathcal{Y}_{m,i}(\ell^n)$ is output of the noisy compound channel exponentially vanish as n grows.

Next by replacing $\mathcal{U}(\ell^n)$ and (80) by $\mathcal{U}^i(\ell^n)$ and (125), in the same way as in the proof of the previous lemma we now obtain

$$align(M_1 I)^{-1} \sum_{i=1}^{I} \sum_{m=1}^{M_1} \Pr\left\{ Y'^n \in \mathcal{Y}_{m,i}(\ell^n) \cap \left[\bigcup_{m' \neq m} \mathcal{Y}_{m',i}(\ell^n) \right] | \ell^n, u_{m,i}^n(\ell^n) \right\}$$

$$align < 2^{-\frac{n}{21}\varepsilon_2} \tag{128}$$

instead of (108).

Finally analogously to in the way to obtain (109) in the proof of the previous lemma from (95) and (108), we finish the proof by combining (127) and (128). □

Corollary 170 (Direct part of Theorem 157) *For all single channels W*

$$C_{\mathrm{CRII}}\big((V, L), W, R_{\mathrm{K}}, D_1\big) \geq \max_{\substack{(V,L,U,X,Y)\in \\ \mathcal{Q}^*((V,L),W,R_{\mathrm{K}},D_1)}} \big(I(U; L, Y) + H(L|U)\big) + R_{\mathrm{K}}.$$

6 The Converse Theorems for Common Randomness

To obtain single letter characterizations for the converse parts of coding theorems for common randomness, we need a useful identity which appears in [18] (on page 314).

Lemma 171 (Csiszár and Körner) *Let (A^n, B^n) be an arbitrary pair of random sequences and let C be an arbitrary RV. Then*

$$H(A^n|C) - H(B^n|C) = \sum_{t=1}^{n} [H(A_t|A_{t+1}, A_{t+2}, \ldots, A_n, B^{t-1}, C) - H(B_t|A_{t+1}, A_{t+2}, \ldots, A_n, B^{t-1}, C)]. \tag{129}$$

Proof

$$H(A^n|C) - (B^n|C)$$

$$= \sum_{t=0}^{n-1} H(A_{t+1}, A_{t+2}, \ldots, A_n, B^t|C) - \sum_{t=1}^{n} H(A_{t+1}, A_{t+2}, \ldots, A_n, B^t|C)$$

$$= \sum_{t=1}^{n} H(A_t, A_{t+1}, \ldots, A_n, B^{t-1}|C) - \sum_{t=1}^{n} H(A_{t+1}, A_{t+2}, \ldots, A_n, B^t|C)$$

$$= \sum_{t=1}^{n} [H(A_t, A_{t+1}, \ldots, A_n, B^{t-1}|C) - H(A_{t+1}, \ldots, A_n, B^{t-1}|C)]$$

$$- \sum_{t=1}^{n} [H(A_{t+1}, A_{t+2}, \ldots, A_n, B^{t}|C) - H(A_{t+1}, \ldots, A_n, B^{t-1}|C)]$$

$$= \sum_{t=1}^{n} [H(A_t|A_{t+1}, A_{t+2}, \ldots, A_n, B^{t-1}, C)$$

$$- H(B_t|A_{t+1}, A_{t+2}, \ldots, A_n, B^{t-1}, C)], \tag{130}$$

where $(A_{t+1}, A_{t+2}, \ldots, A_n, B^t)$ to be understood as A^n and B^n when $t = 0$ and $t = n$ respectively. □

Lemma 172 (Converse part of Theorem 156) *For single channel W,*

$$C_{\mathrm{CRI}}((V, L), W, D_1) \leq max_{(V,L,U,X,Y)\in\mathcal{Q}((V,L),W,D_1)}[I(U; LY) + H(L|U)]. \tag{131}$$

Proof Assume that for a source output of length n there are functions F and K such that for the channel W^n and the distortion measure (10)–(16) hold. Denote by X^n and Y^n the random input and output of the channel generated by the correlated source (V^n, L^n), sender's private randomness M, and the channel.

Then (10) can be rewritten in terms of (V^n, X^n) as

$$\frac{1}{n}\mathbb{E}\rho(V^n, X^n) \leq D_1 \tag{132}$$

Further by Fano inequality (Lemma 48), (11)–(14), we have that

$$\begin{aligned} H(F) &\leq H(F) - H(F|G) + n\lambda \log \kappa + h(\lambda) \\ &= I(F; G) + n\lambda \log \kappa + h(\lambda) \\ &\leq I(F; L^n, Y^n) + n\lambda \log \kappa + h(\lambda) \\ &= I(F; Y^n|L^n) + I(F; L^n) + n\lambda \log \kappa + h(\lambda) \\ &\leq I(F; Y^n|L^n) + H(L^n) + n\lambda \log \kappa + h(\lambda) \\ &= I(F; Y^n|L^n) + \sum_{t=1}^{n} H(L_t) + n\lambda \log \kappa + h(\lambda) \\ &= \sum_{t=1}^{n} I(F; Y_t|L^n, Y^{t-1}) + \sum_{t=1}^{n} H(L_t) + n\lambda \log \kappa + h(\lambda), \end{aligned} \tag{133}$$

where $h(z) = -z\log z - (1-z)\log(1-z)$ for $z \in [0,1]$ is the binary entropy. Here the first inequality follows from the Fano inequality (Lemma 48), (11), (12) and (14); the second inequality holds by (13); and the third equality holds because the source is memoryless. Since $I(F; V^n, L^n) \le H(F)$, the first four lines in (133) is followed by

$$\begin{aligned}
0 &\le I(F; L^n, Y^n) - I(F; V^n, L^n) + n\lambda\log\kappa + h(\lambda)\\
&\le [I(F; Y^n|L^n) + I(F; L^n)] - [I(F; V^n|L^n) + I(F; L^n)] + n\lambda\log\kappa + h(\lambda)\\
&= I(F; Y^n|L^n) - I(F; V^n|L^n) + n\lambda\log\kappa + h(\lambda)\\
&= [H(Y^n|L^n) - H(Y^n|L^n, F)]\\
&\quad - [H(V^n|L^n) - H(V^n|L^n, F)] + n\lambda\log\kappa + h(\lambda)\\
&= [H(Y^n|L^n) - H(V^n|L^n)]\\
&\quad + [H(V^n|L^n, F) - H(Y^n|L^n, F)] + n\lambda\log\kappa + h(\lambda). \qquad (134)
\end{aligned}$$

To obtain a single letter characterization we substitute A^n, B^n and C in (129) by V^n, Y^n and (L^n, F) respectively and so

$$\begin{aligned}
&H(V^n|L^n F) - H(Y^n|L^n F)\\
&= \sum_{t=1}^{n} [H(V_t|V_{t+1}, V_{t+2}, \ldots, V_n, L^n, Y^{t-1}, F)\\
&\qquad - H(Y_t|V_{t+1}, V_{t+2}, \ldots, V_n, L^n, Y^{t-1}, F)]. \qquad (135)
\end{aligned}$$

Moreover because the source is memoryless, we have

$$H(V^n|L^n) = \sum_{t=1}^{n} H(V_t|L_t). \qquad (136)$$

We now substitute (134), (135); (136) and $H(Y^n|L^n) = \sum_{t=1}^{n} H(Y_t|L^n, Y^{t-1})$ into (133) and continue it,

$$\begin{aligned}
0 &\le \sum_{t=1}^{n} [H(Y_t|L^n, Y^{t-1}) - H(V_t|L_t)] + \sum_{t=1}^{n} [H(V_t|V_{t+1}, V_{t+2}, \ldots, V_n, L^n, Y^{t-1}, F)\\
&\quad - H(Y_t|V_{t+1}, V_{t+2}, \ldots, V_n, L^n, Y^{t-1}, F)] + n\lambda\log\kappa + h(\lambda)\\
&= \sum_{t=1}^{n} [H(Y_t|L^n, Y^{t-1}) - H(Y_t|V_{t+1}, V_{t+2}, \ldots, V_n, L^n, Y^{t-1}, F)]
\end{aligned}$$

$$
\begin{aligned}
&-\sum_{t=1}^{n}[H(V_t|L_t) - H(V_t|V_{t+1}, V_{t+2}, \ldots, V_n, L^n, Y^{t-1}, F)] + n\lambda \log \kappa + h(\lambda) \\
&= \sum_{t=1}^{n} I(Y_t; V_{t+1}, V_{t+2}, \ldots, V_n, F|L^n, Y^{t-1}) \\
&- \sum_{t=1}^{n} I(V_t; V_{t+1}, V_{t+2}, \ldots, V_n, L_1, L_2 \ldots, L_{t-1}, L_{t+1}, \ldots, L_n, Y^{t-1}, F|L_t) \\
&+ n\lambda \log \kappa + h(\lambda) \\
&\leq \sum_{t=1}^{n}[I(Y_t; V_{t+1}, V_{t+2}, \ldots, V_n L_1, L_2 \ldots, L_{t-1}, L_{t+1}, \ldots, L_n, Y^{t-1}, F|L_t)] \\
&- \sum_{t=1}^{n} I(V_t; V_{t+1}, V_{t+2}, \ldots, V_n, L_1, L_2 \ldots, L_{t-1}, L_{t+1}, \ldots, L_n, Y^{t-1}, F|L_t)] \\
&+ n\lambda \log \kappa + h(\lambda). \qquad (137)
\end{aligned}
$$

Let J be the RV taking values in $\{1, 2, \ldots, n\}$ uniformly, and

$$U_J = (V_{J+1}, V_{J+2}, \ldots, V_n, L_1, L_2 \ldots, L_{J-1}, L_{J+1}, \ldots, L_n, Y^{J-1}, F). \qquad (138)$$

Then J and (V_J, L_J) are independent i. e., $I(J; V_J, L_J) = 0$. Thus (137) is rewritten and continued in the following a few lines.

$$
\begin{aligned}
0 &\leq nI(U_J; Y_J|L_J, J) - nI(U_J; V_J|L_J, J) + n\lambda \log \kappa + h(\lambda) \\
&= n[I(U_J; L_J, Y_J|J) - I(U_J; L_J|J)] - [I(U_J; V_J, L_J|J) - I(U_J; L_J|J) \\
&+ n\lambda \log \kappa + h(\lambda) \\
&= nI(U_J; L_J, Y_J|J) - nI(U_J; V_J, L_J|J) + n\lambda \log \kappa + h(\lambda) \\
&\leq nI(U_J, J; L_J, Y_J) - n[I(U_J, J; V_J, L_J) - I(J; V_J, L_J)] + n\lambda \log \kappa + h(\lambda) \\
&= nI(U_J, J; L_J, Y_J) - nI(U_J, J; V_J, L_J) + n\lambda \log \kappa + h(\lambda). \qquad (139)
\end{aligned}
$$

Next we denote by

$$(V'', L'', U'', X'', Y'') = (V_J, L_J, U_J J, X_J, Y_J) \qquad (140)$$

for the uniformly distributed J and U_J in (138). Then, obviously (V'', L'') has the same probability distribution with the generic (V, L) of the correlated source, the conditional probability distribution $P_{Y''|X''} = W$, and $(V''L''U'', X'', Y'')$ forms a Markov Chain. Namely, the joint distribution of $(V'', L'', U'', X'', Y'')$ is $P_{V''L''U''X''Y''} = P_{VL}P_{U''X''|V''L''}W$. With the defined random variables, (132) is rewritten as

$$\mathbb{E}\rho(V'', X'') = \mathbb{E}[\mathbb{E}\rho(V'', X'')|J] = \mathbb{E}[\mathbb{E}\rho(V_J, X_J)|J] = \frac{1}{n}\mathbb{E}\rho(V^n, X^n) \le D_1. \tag{141}$$

Moreover, by substituting (140) in (139) and then dividing both sides of resulting inequality by n, we obtain that

$$0 \le I(U''; L'', Y'') - I(U''; V'', L'') + o(1), \tag{142}$$

(as $\lambda \to 0$).

Because the set $\{P_{V,L,U,X,Y} : (V, L, U, X, Y) \in \mathcal{Q}((V, L), W, D_1)\}$ is a closed set, by (141) and (142) is sufficient for us to complete the proof to show that

$$\frac{1}{n}H(F) \le I(U''; L'', Y'') + H(L''|U'') + o(1)$$

for $\lambda \to 0$. This is done by dividing both sides of (133) by n and continuing it by the following few lines.

$$\begin{aligned}
\frac{1}{n}H(F) &\le \frac{1}{n}\sum_{t=1}^{n} I(F; Y_t|L^n, Y^{t-1}) + \frac{1}{n}\sum_{t=1}^{n} H(L_t) + \lambda \log \kappa + \frac{1}{n}h(\lambda), \\
&\le \frac{1}{n}\sum_{t=1}^{n} I(V_{t+1}, V_{t+2}, \ldots, V_n, F; Y_t|L^n, Y^{t-1}) + \frac{1}{n}\sum_{t=1}^{n} H(L_t) \\
&\quad + \lambda \log \kappa + \frac{1}{n}h(\lambda), \\
&\le \frac{1}{n}\sum_{t=1}^{n} I(V_{t+1}, \ldots, V_n, L_1, \ldots, L_{t-1}, L_{t+1}, \ldots, L_n, Y^{t-1}, F; Y_t|L_t) \\
&\quad + \frac{1}{n}\sum_{t=1}^{n} H(L_t) + \lambda \log \kappa + \frac{1}{n}h(\lambda) \\
&= I(U_J; Y_J|L_J, J) + H(L_J|J) + \lambda \log \kappa + \frac{1}{n}h(\lambda)
\end{aligned}$$

$$\leq I(U_J, J; Y_J|L_J) + H(L_J|J) + \lambda \log \kappa + \frac{1}{n}h(\lambda)$$

$$= I(U_J, J; Y_J|L_J) + H(L_J) + \lambda \log \kappa + \frac{1}{n}h(\lambda)$$

$$= I(U_J, J; Y_J|L_J) + I(U_J; L_J) + H(L_J|U_J) + \lambda \log \kappa + \frac{1}{n}h(\lambda)$$

$$\leq I(U_J, J; Y_J|L_J) + I(U_J, J; L_J) + H(L_J|U_J) + \lambda \log \kappa + \frac{1}{n}h(\lambda)$$

$$= I(U_J, J; L_J, Y_J) + H(L_J|U_J) + \lambda \log \kappa + \frac{1}{n}h(\lambda)$$

$$= I(U''; L'', Y'') + H(L''|U'') + \lambda \log \kappa + \frac{1}{n}h(\lambda), \tag{143}$$

where the second equality holds because U_J is independent of J. Finally the upper bound to the size of $\mathcal{U}$ follows from the Support Lemma in [11] (as well on page 310 in the book [18]). □

Lemma 173 (Converse part of Theorem 157) *For a single channel W,*

$$C_{\mathrm{CRI}}((V, L), W, R_{\mathrm{K}}, D_1) \leq \max_{\substack{(V,L,U,X,Y)\in \\ \mathcal{Q}^*((V,L),W,R_{\mathrm{K}},D_1)}} [I(U; L, Y) + H(L|U)] + R_{\mathrm{K}}. \tag{144}$$

Proof Let $\{(V^n, L^n)\}_{n=1}^{\infty}$ be a correlated source with generic (V, L), W be a noisy channel, and R_{K} and D_1 be the key rate and the distortion criterion in the Model II of common randomness respectively. Let F and G be functions satisfying (10)–(12), (17), and (14)–(16) in the Model II of common randomness (for output sequence of source of length n). Denote by X^n and K_n inputs of noisy channel W^n and the noiseless channel chosen by the sender according to the output of the correlated source and his/her private randomness. Then (132) holds and similarly to (133) by Fano inequality (Lemma 48), we have that

$$H(F) \leq I(F; G) + n\lambda \log \kappa + h(\lambda)$$

$$\leq I(F; Y^n, L^n, K_n) + n\lambda \log \kappa + h(\lambda)$$

$$= I(F; Y^n, L^n) + I(F; K_n|Y^n, L^n) + n\lambda \log \kappa + h(\lambda)$$

$$= I(F; Y^n|L^n) + I(F; L^n) + I(F; K_n|Y^n, L^n) + n\lambda \log \kappa + h(\lambda)$$

$$\leq I(F; Y^n|L^n) + H(L^n) + H(K_n|Y^n, L^n) + n\lambda \log \kappa + h(\lambda)$$

$$\begin{aligned}
&\le I(F;Y^n|L^n)+H(L^n)+H(K_n)+n\lambda\log\kappa+h(\lambda)\\
&\le I(F;Y^n|L^n)+H(L^n)+nR_{\mathrm{K}}+n\lambda\log\kappa+h(\lambda)\\
&=\sum_{t=1}^{n} I(F;Y_t|L^n,Y^{t-1})+\sum_{t=1}^{n} H(L_t)+nR_{\mathrm{K}}+n\lambda\log\kappa+h(\lambda),
\end{aligned} \tag{145}$$

where the second inequality holds by (17). Analogously to (134) we have

$$\begin{aligned}
0&\le I(F;Y^n,L^n,K_n)-I(F;V^n,L^n)+n\lambda\log\kappa+h(\lambda)\\
&=I(F;Y^n,L^n)-I(F;V^n,L^n)+I(F;K_n|Y^n,L^n)+n\lambda\log\kappa+h(\lambda)\\
&\le I(F;Y^n,L^n)-I(F;V^n,L^n)+H(K_n|Y^n,L^n)+n\lambda\log\kappa+h(\lambda)\\
&\le I(F;Y^n,L^n)-I(F;V^n,L^n)+nR_{\mathrm{K}}+n\lambda\log\kappa+h(\lambda).
\end{aligned} \tag{146}$$

Note that we only used the basic properties of Shannon information measures, Lemma 171, and the assumption that the correlated source is memoryless in the estimation of $I(F;Y^n,L^n)-I(F;V^n,L^n)$ in the part of (134)–(137) and all these are available here. So we have the same estimation here i.e.,

$$\begin{aligned}
&I(F;Y^n,L^n)-I(F;V^nL^n)\\
&\le\sum_{t=1}^{n} I(Y_t;V_{t+1},V_{t+2},\ldots,V_n,L_1,L_2\ldots,L_{t-1},L_{t+1},\ldots,L_n,Y^{t-1},F|L_t)\\
&\quad-\sum_{t=1}^{n} I(V_t;V_{t+1},V_{t+2},\ldots,V_n,L_1,L_2\ldots,L_{t-1},L_{t+1},\ldots,L_n,Y^{t-1},F|L_t)\\
&\quad+n\lambda\log\kappa+h(\lambda).
\end{aligned} \tag{147}$$

Let U_J and J be defined as in (138). Then (147) is rewritten as

$$\begin{aligned}
I(F;Y^n,L^n)-I(F;V^nL^n)\le{}&nI(U_J,J;L_J,Y_J)-nI(U_J,J;V_J,L_J)\\
&+n\lambda\log\kappa+h(\lambda).
\end{aligned} \tag{148}$$

Let (V'',L'',U'',X'',Y'') is defined as in the previous lemma. Then (141) and $P_{V''L''U''X''Y''}=P_{VL}P_{U''X''|V''L''}W$ are certainly fulfilled. But now (146)–(148) lead us to

$$0\le I(U'';L'',Y'')-I(U'';V'',L'')+R_{\mathrm{K}}+o(1). \tag{149}$$

In the same way as (143) we can show

$$\sum_{t=1}^{n} I(F; Y_t|L^n, Y^{t-1}) + \sum_{t=1}^{n} H(L_t) + nR_K + n\lambda \log \kappa + h(\lambda)$$

$$\leq nI(U''; L'', Y'') + nH(U''|L'') + n\lambda \log \kappa + h(\lambda) \tag{150}$$

which with (145) yields

$$\frac{1}{n}H(F) \leq I(U''; L''Y'') + H(U''|L'') + R_K + \lambda \log \kappa + \frac{1}{n}h(\lambda).$$

Again $|\mathcal{U}|$ is bounded by the Support Lemma. Thus our proof is finished. □

Finally it immediately follows from Lemmas 172 and 173 that

Corollary 174 *For compound channel $\mathcal{W}$,*

(i) *(The converse part of Theorem 158:)*

$$C_{\text{CRI}}((V, L), \mathcal{W}, D_1) \leq \inf_{W \in \mathcal{W}} \max_{\substack{(V,L,U,X,Y)\in \\ \mathcal{Q}((V,L),W,D_1)}} [I(U; L, Y) + H(L|U)] \tag{151}$$

and

(ii) *(The converse part of Theorem 159:)*

$$C_{\text{CRII}}((V, L), \mathcal{W}, R_K, D_1)$$

$$\leq \inf_{W \in \mathcal{W}} \max_{\substack{(V,L,U,X,Y)\in \\ \mathcal{Q}^*((V,L),W,R_K,D_1)}} [I(U; L, Y) + H(L|U)] + R_K. \tag{152}$$

7 Construction of Watermarking Identification Codes from Common Randomness

Ahlswede and Dueck found in [10] (chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas") that an identification code with the same rate can be always obtained from the common randomness between a sender and receiver under the condition

$$\text{The sender can send a massage with arbitrarily small}$$

$$\text{but positive rate (in the exponential sense)} \tag{153}$$

Thus under the condition (153) the capacity of identification is not smaller than that of common randomness. Note that the sets $\mathcal{Q}((V, L), W, D_1)$, $\mathcal{Q}^{**}(V, W, R_k, D_1)$, $\mathcal{Q}_1((V, L), \mathcal{W}, D_1)$, and $\mathcal{Q}_1^{**}(V, \mathcal{W}, R_k, D_1)$ are not empty implies the condition (153) in the Theorems 160, 161, 162, and 163 respectively. Consequently Theorems 160, 161, 162, and 163 follows from Theorems 156, 157, 158, and 159 respectively.

8 A Converse Theorem of a Watermarking Coding Theorem Due to Steinberg-Merhav

In order to construct identification codes in [27], Y. Steinberg and N. Merhav introduced the following code to build common randomness between sender and receiver and obtained an inner bound of the capacity region. This inner bound is sufficient for their goal. We shall show that it is as well tight. This would support their conjecture that the lower bound in their Theorem 4 ([27]) is tight although it does not imply it.

Let $\{V^n\}_{n=1}^{\infty}$ be a memoryless source with alphabet $\mathcal{V}$ and generic V and W be a noisy channel with input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ respectively. A pair of functions (f, g) is called an $(n, M, J, \delta, \lambda, D)$ watermarking transmission code with a common experiment, distortion measure ρ, distortion level D and covertext P_V if the followings are true.

— f is a function from $\mathcal{V}^n \times \{1, 2, \ldots, M\}$ to $\{1, 2, \ldots, J\} \times \mathcal{X}^n$.
— g is a function from $\mathcal{Y}^n$ to $\{1, 2, \ldots, J\} \times \{1, 2, \ldots, M\}$.

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{v^n\in\mathcal{V}^n} P_V^n(v^n)W^n(\{y : g(y^n) = (f_J(v^n, m), m)\}|f_X(v^n, m)) \geq 1 - \lambda, \tag{154}$$

where f_X and f_J are projections of f to $\mathcal{X}^n$ and $\{1, 2, \ldots, J\}$ respectively.

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{v^n\in\mathcal{V}^n} P_V^n(v^n)\rho(v^n, f_X(v^n, m)) \leq D. \tag{155}$$

For $m = 1, 2, \ldots, M$, there exists a subset $\mathcal{B}^{(m)} \subset \{1, 2, \ldots, J\}$ of cardinality $|\mathcal{B}^{(m)}| \geq J2^{-n\delta}$ such that

$$J^{-1}2^{-n\delta} \leq P_V^n\{f_J(V^n, m) = j\} \leq J^{-1}2^{n\delta} \tag{156}$$

for all j and

$$\sum_{j\in\mathcal{B}^{(m)}} P_V^n\{f_J(V^n, m) = j\} \geq 1 - \lambda. \tag{157}$$

g serves as a decoding function here. (156) and (157) play the same role as nearly uniform condition in construction of identification codes from common randomness. In fact one can find the nearly uniform condition (16) is stronger but for the purpose to construct identification codes the conditions (156) and (157) are strong enough.

A pair (R_1, R_2) is called achievable with distortion D if for all positive reals δ, λ, and ϵ there is an $(n, M, J, \delta, \lambda, D)$ code defined as above such that

$$\frac{1}{n}\log M > R_1 - \epsilon \tag{158}$$

and

$$\frac{1}{n}\log J > R_2 - \epsilon. \tag{159}$$

The set of achievable pair of rates is called capacity region and denoted by $\mathcal{R}$. Denote by $\mathcal{R}^{(*)}$ the subset of pair of real numbers such that there exist RV's (V, U, X, Y) taking values in $\mathcal{V}\times\mathcal{U}\times\mathcal{X}\times\mathcal{Y}$ such that $|\mathcal{U}| \leq |\mathcal{Y}| + |\mathcal{X}|$, for all $v \in \mathcal{V}, u \in \mathcal{U}, x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$P_{VUXY}(v, u, x, y) = P_V(v)P_{UX|V}(u, x|v)W(y|x),$$

$$\mathbb{E}\rho(V, X) \leq D,$$

$$0 \leq R_1 \leq I(U; Y) - I(U; V), \tag{160}$$

and

$$0 \leq R_2 \leq I(U; V). \tag{161}$$

Theorem 175 (Steinberg and Merhav 2001 [27])

$$\mathcal{R}^* \subset \mathcal{R}. \tag{162}$$

We now show the opposite contained relation holds i.e.,

Theorem 176

$$\mathcal{R} \subset \mathcal{R}^*. \tag{163}$$

Proof Let (f, g) be a pair of functions satisfying (154)–(159) for sufficiently large n (, which is specified later,) and Z_n be a RV with uniform distribution over $\{1, 2, \ldots, M\}$. Further let $f(V^n, Z_n) = (B_n, X^n)$, where B_n and X^n have ranges $\{1, 2, \ldots, J\}$ and $\mathcal{X}^n$ respectively and Y^n be the random output of the channel W^n when X^n is input.

Then (156) and (157) are rewritten as

$$J^{-1}2^{-n\delta} \le P_{B_n|Z_n}(j|m) \le J^{-1}2^{n\delta} \tag{164}$$

for all $j \in \mathcal{B}^{(m)}$ and

$$P_{B_n|Z_n}(B_n \in \mathcal{B}^{(m)}|m) \ge 1 - \lambda \tag{165}$$

respectively. So,

$$\begin{aligned} H(B_n|Z_n) &= \sum_{m=1}^{M} P_{Z_n}(m) H(B_n|Z_n = m) \\ &\ge -\sum_{m=1}^{M} P_{Z_n}(m) \sum_{j\in\mathcal{B}^{(m)}} P_{B_n|Z_n}(j|m) \log P_{B_n|Z_n}(j|m) \\ &\ge -\sum_{m=1}^{M} P_{Z_n}(m) \sum_{j\in\mathcal{B}^{(m)}} P_{B_n|Z_n}(j|m) \log J^{-1}2^{n\delta} \\ &= (\log J - n\delta) \sum_{m=1}^{M} P_{Z_n}(m) P_{B_n|Z_n}(B_n \in \mathcal{B}^{(m)}|m) \\ &\ge (\log J - n\delta)(1 - \lambda) \end{aligned} \tag{166}$$

where the second inequality holds by (164) and the last inequality follows from (165). Or equivalently

$$\frac{1}{n} \log J \le \frac{\frac{1}{n} H(B_n|Z_n)}{1 - \lambda} + \delta. \tag{167}$$

Since $H(B_n) \le \log J$, (167) implies that for a function θ such that $\theta(\delta, \lambda) \to 0$ as $\delta, \lambda \to 0$,

$$\frac{1}{n} \log J - \theta(\delta, \lambda) < \frac{1}{n} H(B_n|Z_n) \le \frac{1}{n} H(B_n) \le \frac{1}{n} \log J. \tag{168}$$

which says that B_n and Z_n are "nearly independent". Moreover because Z_n is independent of V^n, by Fano's inequality (Lemma 48),

$$\begin{aligned} R_1 - \varepsilon < \frac{1}{n}\log M &= \frac{1}{n}H(Z_n) \\ &= \frac{1}{n}H(Z_n|V^n) \\ &\le \frac{1}{n}H(B_n, Z_n|V^n) \\ &\le \frac{1}{n}[H(B_n, Z_n|V^n) - H(B_n, Z_n|Y^n)] + \lambda \log JM + \frac{1}{n}h(\lambda) \\ &= \frac{1}{n}[I(B_n, Z_n; Y^n) - I(B_n, Z_n; V^n)] + \lambda\frac{1}{n}\log JM + \frac{1}{n}h(\lambda) \end{aligned} \tag{169}$$

where the second inequality follows from Fano's inequality (Lemma 48). Since B_n is a function of V^n and Z_n, we have also

$$H(B_n, Z_n|V^n) \le H(V^n, Z_n|V^n) = H(Z_n), \tag{170}$$

which and (168) are followed by

$$\begin{aligned} R_2 - \varepsilon < \frac{1}{n}\log J &< \frac{1}{n}H(B_n|Z_n) + \theta(\delta, \lambda) \\ &= \frac{1}{n}[H(B_n, Z_n) - H(Z_n)] + \theta(\delta, \lambda) \\ &\le \frac{1}{n}[H(B_n, Z_n) - H(B_n, Z_n|V^n)] + \theta(\delta, \lambda) \\ &= \frac{1}{n}I(B_n, Z_n; V^n) + \theta(\delta, \lambda). \end{aligned} \tag{171}$$

So far we have had a non-single-letter characterization of the capacity region (169) and (171). In the remaining part of the proof we shall reduce it to a single letter one.

First we substitute A^n, B^n, and C in (129) by V^n, Y^n, and (B_n, Z_n) respectively and obtain that

$$\begin{aligned}
&H(V^n|B_n, Z_n) - H(Y^n|B_n, Z_n) \\
&= \sum_{t=1}^{n} [H(V_t|V_{t+1}, V_{t+2}, \ldots, V_n, Y^{t-1}, B_n, Z_n) \\
&\quad - H(Y_t|V_{t+1}, V_{t+2}, \ldots, V_n, Y^{t-1}, B_n, Z_n)]. \qquad (172)
\end{aligned}$$

Next we note that $H(V^n) = \sum_{t=1}^{n} H(V_t)$ because the source is memoryless and $H(Y^n) = \sum_{t=1} H(Y_t|Y^{t-1})$. Therefore, we have

$$\begin{aligned}
&I(B_n, Z_n; Y^n) - I(B_n, Z_n; V^n) \\
&= H(Y^n) - H(V^n) + [H(V^n|B_n, Z_n) - H(Y^n|B_n, Z_n)] \\
&= \sum_{t=1}^{n} H(Y_t|Y^{t-1}) - \sum_{t=1}^{n} H(V_t) + \sum_{t=1}^{n} [H(V_t|V_{t+1}, V_{t+2}, \ldots, V_n, Y^{t-1}, B_n, Z_n) \\
&\quad - H(Y_t|V_{t+1}, V_{t+2}, \ldots, V_n, Y^{t-1}, B_n, Z_n)] \\
&= \sum_{t=1}^{n} [H(Y_t|Y^{t-1}) - H(Y_t|V_{t+1}, V_{t+2}, \ldots, V_n, Y^{t-1}, B_n, Z_n)] \\
&\quad - \sum_{t=1}^{n} [H(V_t) - H(V_t|V_{t+1}, V_{t+2}, \ldots, V_n, Y^{t-1}, B_n, Z_n)] \\
&= \sum_{t=1}^{n} I(V_{t+1}, V_{t+2}, \ldots, V_n, B_n, Z_n; Y_t|Y^{t-1}) \\
&\quad - \sum_{t=1}^{n} I(V_{t+1}, V_{t+2}, \ldots, V_n, Y^{t-1}, B_n, Z_n; V_t) \\
&\le \sum_{t=1}^{n} I(V_{t+1}, V_{t+2}, \ldots, V_n, Y^{t-1}, B_n, Z_n; Y_t) \\
&\quad - \sum_{t=1}^{n} I(V_{t+1}, V_{t+2}, \ldots, V_n, Y^{t-1}, B_n, Z_n; V_t). \qquad (173)
\end{aligned}$$

Moreover,

$$I(B_n, Z_n; V^n) = \sum_{t=1}^{n} I(B_n, Z_n; V_t | V_{t+1}, V_{t+2}, \ldots, V_n)$$

$$\leq \sum_{t=1}^{n} I(V_{t+1}, V_{t+2}, \ldots, V_n, B_n, Z_n; V_t)$$

$$\leq \sum_{t=1}^{n} I(V_{t+1}, V_{t+2}, \ldots, V_n, Y^{t-1}, B_n, Z_n; V_t). \tag{174}$$

So we may let I be a RV taking values in $\{1, 2, \ldots, n\}$ uniformly and $U' = (V_{I+1}, V_{I+2}, \ldots, V_n, Y^{I-1}, B_n, Z_n)$ and conclude by (171), (172), (173), (174)

$$R_1 - \varepsilon \leq I(U'; Y_I | I) - I(U'; V_I | I) + \lambda \log JM + \frac{1}{n} h(\lambda)$$

$$\leq I(U', I; Y_J) - I(U', I; V_I) + I(I; V_I) + \lambda \log JM + \frac{1}{n} h(\lambda), \tag{175}$$

and

$$R_2 - \varepsilon \leq I(U'; V_I | I) \leq I(U', I; V_I) + \theta(\delta, \lambda). \tag{176}$$

Let $U = (U', I)$, $V' = V_I$, $X = X_I$ and $Y = Y_I$. Then $P_{V'} = P_V$, $(V'U, X, Y)$ forms a Markov chain and (176) can be re-written as

$$R_2 \leq I(U; V') + \theta(\delta, \lambda),$$

and

$$E\, P(v', x') < D.$$

Further that $I(I; V_I) = 0$ (as the source is stationary) and (175) are followed by

$$R_1 \leq I(U; Y) - I(U; V') + \lambda \log JM + \frac{1}{n} h(\lambda) + \varepsilon.$$

Finally $|\mathcal{U}|$ is bounded by the support Lemma in the standard way. □

References

1. R. Ahlswede, Channels with arbitrarily varying channel probability functions in the presence of noiseless feedback, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **25**, 239-252, (1973)
2. R. Ahlswede, Elimination of correlation in random codes for arbitrarily varying channels. Zeitschrift Wahrscheinlichkeitstheorie und verw. Geb. **33**, 159–175 (1978)
3. R. Ahlswede, General theory of information trannsfer, Preprint 97-118, SFB 343 Diskrete Strukturen in der Mathematik, 1997, this volume pages
4. R. Ahlswede, V.B. Balakirsky, Identification under random processes, Preprint 95–098, SFB 343, "Diskrete Strukturen in der Mathematik", Universität Bielefeld, Problemy peredachii informatsii (special issue devoted to M.S. Pinsker), vol. 32, no. 1 (1996), pp. 144–160
5. R. Ahlswede, N. Cai, The AVC with noiseless feedback and maximal error probability: a capacity formula with a trichotomy, in *Numbers, Information, and Complexity* (Festschrift for Rudolf Ahlswede), ed. by I. Althöfer, N. Cai, G. Dueck, L. Khachatrian, M. Pinsker, A. Sarkozy, I. Wegener, Z. Zhang (Kluwer, 2000), pp. 151–176
6. R. Ahlswede, N. Cai, Watermarking identification codes with related topics in common randomness, in *General Theory of Information Transfer and Combinatorics*, Lecture Notes in Computer Science, vol. 4123 (Springer, 2006), pp. 107–153
7. R. Ahlswede, I. Csiszár, Common randomness in information theory and cryptography, Part I: Secret sharing. IEEE Trans. Inform. Theory **39**, 1121–1132 (1993)
8. R. Ahlswede, I. Csiszár, Common randomness in information theory and cryptography, Part II: CR capacity. IEEE Trans. Inform. Theory **44**(1), 225 (1998)
9. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inform. Theory **35**, 15–29 (1989)
10. R. Ahlswede, G. Dueck, Identification in the presence of feedback — a discovery of new capacity formulas. IEEE Trans. Inf. Theory **35**, 30–39 (1989)
11. R. Ahlswede, J. Körner, Source coding with side information and a converse for degraded broadcast channels. IEEE Trans. Inf. Theory **IT-21**(6), 629–637 (1975)
12. R. Ahlswede, C. Kleinewächter, Pathological examples with different common randomness and (second order) identification capacities, this volume pages
13. 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. Theory **41**(4), 1040–1050 (1995)
14. M. Barni, F. Bartolini, A. De Rosa, A. Piva, Capactiy of the watermark channel: how many bits can be hidden within a digital image? in, *Proceedings of SPIE*, SPIE, vol. 3657, pp. 437–448, San Jose, Jan (1999)
15. M. Burnashav, On identification capacity of infinite alphabets or continuous time channel. IEEE Trans. Inf. Theory **IT–46**, 2407–2414 (2000)
16. N. Cai, K.Y. Lam, On identification secret sharing schemes. Inf. Comput. **184**(2), 298–310 (2003)
17. I.J. Cox, M.L. Miller, A. Mckellips, Watermarking as communications with side information. Proc. IEEE **87**(7), 1127–1141 (1999)
18. I. Csiszár, J. K"orner, *Information Theory: Coding Theorem for Discrete Memoryless Systems* (Academic, New York, 1982)
19. I. Csiszár, P. Narayan, Common randomness and secret key generation with a helper. IEEE Trans. Inf. Theory **46**(2), 344–366 (2000)
20. S.I. Gelfand, M.S. Pinsker, Coding for channels with random parameters. Probl. Control Inform. Theory **9**, 19–31 (1980)
21. U.M. Maurer, Secret key agreement by public discussion from common information. IEEE Trans. Inform. Theory. **39**(3), 733–742 (1993)
22. N. Merhav, On random coding error exponents of watermarking codes. IEEE Trans. Inform. Theory **46**(2), 420–430 (2000)

23. P. Moulin, J.A. O'Sulliovan, Information-theoretic analysis of information hiding, preprint (1999)
24. J.A. O'Sullivan, P. Moulin, J.M. Ettinger, Information theoretic analysis of steganography, in *Proceedings of ISIT '98* 297 (1998)
25. S.D. Servetto, C.I. Podilchuk, K. Ramchandran, Capacity issues in digital image watermarking, in *Proceedings of ICIP '98* (1998)
26. Y. Steinberg, New converses in the theory of identification via channels. IEEE Trans. Inform. Theory **44**, 984–998 (1998)
27. Y. Steinberg, N. Merhav, Indentification in the presence of side infromation with application to watermarking. IEEE Trans. Inform. Theory **47**, 1410–1422 (2001)
28. S. Venkatesan, V. Anantharam, The common randomness capacity of a pair of independent dicrete memoryless channels. IEEE Trans. Inform. Theory **44**, 215–224 (1998)
29. S. Venkatesan, V. Anantharam, The common randomness capacity of network of discret mamoryless channels. IEEE Trans. Inform. Theory **IT-46**, 367–387 (2000)
30. R.W. Yeung, *A First Course in Information Theory* (Kluwer Academic, 2002)

Transmission, Identification and Common Randomness Capacities for Wire-Tap Channels with Secure Feedback from the Decoder

We analyze wire-tap channels with secure feedback from the legitimate receiver. We present a lower bound on the transmission capacity (Theorem 178), which we conjecture to be tight and which is proved to be tight (Corollary 180) for Wyner's original (degraded) wire-tap channel and also for the reversely degraded wire-tap channel for which the legitimate receiver gets a degraded version from the enemy (Corollary 181).

Somewhat surprisingly we completely determine the capacities of secure common randomness (Theorem 183) and secure identification (Theorem 184 and Corollary 185). Unlike for the DMC, these quantities are different here, because identification is linked to non-secure common randomness.

1 Introduction

The main results are mentioned in the abstract.

After giving standard concepts in Sect. 2, known results and techniques for the wire-tap channel in Sect. 3, we state and prove Theorem 178 in Sect. 4. Our code construction relies upon a lemma for balanced coloring from [1], which has already been proved useful for secrecy problems in [2] and [3] (see chapters "Perspectives" and "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints").

The transmission capacities for the two kinds of degraded wire-tap channels are derived in Sect. 5. Particularly interesting is an example of a reversely degraded channel, where the channel $W_1' : \mathcal{X} \to \mathcal{Z}$ for the wiretapper is noiseless (for instance with binary alphabets) and the channel $W_2' : \mathcal{Z} \to \mathcal{Y}$ for the legal receiver is a noisy binary symmetric channel with crossover probability $p \in (0, 1/2)$. Here the wiretapper is in a better position than the legal user and therefore the capacity

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_17

is zero, if there is no feedback. However, by our corollary the capacity is positive, because the feedback serves as a secure key shared by sender and receiver.

In Sect. 6 a discussion based on the construction for transmission in Sect. 4, known results and constructions for identification [8, 9, 15], common randomness [7, 9] and all other references builds up the intuition for our solutions of the capacity problems for common randomness and identification in Sects. 7 and 8.

2 Notation and Definitions

Throughout the lecture $\mathcal{U}$, $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are finite sets and their elements are written as corresponding lower letters e.g. u, x, y, and z. The letters U, X, Y, Z etc. will be used for RV's with values in the corresponding sets, $\mathcal{U}, \ldots$. $\mathcal{T}^n_X$, $\mathcal{T}^n_{Y|X}(x^n)$, $\mathcal{T}^n_{XYZ}$, etc. are sets of n-typical, conditional typical and joint typical sequences, and sets of δ-typical, conditional typical and joint typical sequences are written as $\mathcal{T}^n_{X,\delta}$, $\mathcal{T}^n_{Y|X,\delta}(x^n)$, $\mathcal{T}^n_{XYZ,\delta}$, etc.

Then a (discrete memoryless) wire-tap channel is specified by a stochastic matrix $W : \mathcal{X} \to \mathcal{Y} \times \mathcal{Z}$, where $\mathcal{X}$ serves as input alphabet, $\mathcal{Y}$ as output alphabet of the legal receiver and $\mathcal{Z}$ as output alphabet of a wiretapper. The channel works as follows: the legal receiver receives an output sequence y^n and the wiretapper receives an output sequence z^n with probability

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

In the case of transmission the sender's goal is to send to the receiver a message U uniformly distributed on a large set of messages with vanishing probability of error such that the wiretapper almost knows nothing about the message. Randomization at the sender side is allowed. The wiretapper, who knows the coding scheme but not the message, tries to learn about the message as much as possible.

For given $\lambda, \mu > 0$, a (λ, μ)-code of length n with a set of messages $\mathcal{M}$ is a system $\{(Q_m : \mathcal{D}_m) : m \in \mathcal{M}\}$, where the Q_m's for $m \in \mathcal{M}$ are probability distributions on $\mathcal{X}^n$, and the $\mathcal{D}_m$'s are pairwise disjoint subsets of $\mathcal{Y}^n$, such that

$$|\mathcal{M}|^{-1} \sum_{m \in \mathcal{M}} \sum_{x^n \in \mathcal{X}^n} Q_m(x^n) \sum_{z^n \in \mathcal{Z}^n} W^n(\mathcal{D}_m, z^n | x^n) > 1 - \lambda, \tag{1}$$

and

$$\frac{1}{n} I(U; Z^n) < \mu, \tag{2}$$

if Z^n is the random output sequence generated by the message U through the channel. The transmission capacity of the wire-tap channel is the maximal non-

negative number C_{wt} such that for $\mathcal{M}, \lambda, \mu, \varepsilon > 0$ and all sufficiently large length n, there exists a (λ, μ)-code with rate $\frac{1}{n} \log |\mathcal{M}| > C_{\mathrm{wt}} - \varepsilon$. The security criterion (2) is strengthened in [11] to

$$I(U; Z) < \mu. \tag{3}$$

In the current lecture we assume the output y_t at time t is completely and immediately feedback to the sender via a secure noiseless channel such that the wiretapper has no knowledge about the feedback (except his own output z^n). Then for $\lambda, \mu > 0$, a (λ, μ)-code of length n for the wire-tap channel with secure feedback is a system $\{(Q, \mathcal{D}_m) : m \in \mathcal{M}\}$ where $\mathcal{D}_m$, $m \in \mathcal{M}$, are pairwise disjoint subsets of $\mathcal{Y}^n$ as before and Q is a stochastic matrix $Q : \mathcal{M} \times \mathcal{Y}^{n-1} \to \mathcal{X}^n$ with

$$Q(x^n | m, y^{n-1}) = \prod_{t=1}^{n} Q(x_t | m, y^{t-1})$$

for $x^n \in \mathcal{X}$, $y^{n-1} \in \mathcal{Y}^{n-1}$, and $m \in \mathcal{M}$, such that

$$|\mathcal{M}|^{-1} \sum_{m \in \mathcal{M}} \sum_{x^n \in \mathcal{X}} \sum_{z^n \in \mathcal{Z}^n} \sum_{y^n \in \mathcal{D}_m} Q(x^n | m, y^{n-1}) W^n(y^n, z^n | x^n) > 1 - \lambda \tag{4}$$

and (2) holds. The transmission capacity is defined analogously and denoted by C_{wtf}. In Theorem 178 in Sect. 4 we shall prove our (direct) coding theorem with the stronger security criterion (3).

3 Previous and Auxiliary Results

Our code construction is based on a coding lemma and a code for wire-tap channel without feedback. A balanced coloring lemma originally was introduced by Ahlswede [1] and we need its following variation.

Lemma 177 *For all $\delta, \eta > 0$, sufficiently large n, all n-ED P_{XY} and all $x^n \in \mathcal{T}^n_X$, there exists a γ-coloring $c : \mathcal{T}^n_{Y|X}(x^n) \to \{0, 1, 2, \ldots, \gamma - 1\}$ of $\mathcal{T}^n_{Y|X}(x^n)$ such that for all joint n-ED P_{XYZ} with marginal distribution P_{XY} and $\gamma^{-1} |\mathcal{T}^n_{Y|XZ}(x^n, z^n)| > 2n\eta$, $x^n, z^n \in \mathcal{T}^n_{XZ}$,*

$$|c^{-1}(k)| \leq \gamma^{-1} |\mathcal{T}^n_{Y|XZ}(x^n, z^n)|(1 + \delta), \tag{5}$$

for $k = 0, 1, \ldots, \gamma - 1$, where c^{-1} is the inverse image of c.

Proof Let us randomly and independently color $y^n \in \mathcal{T}^n_{Y|X}(x^n)$ with γ colors and uniform distribution over $\mathcal{T}^n_{Y|X}(x^n)$. Let for $k = 0, 1, \ldots, \gamma - 1$

$$S_k(y^n) = \begin{cases} 1 & \text{if } y^n \text{ is colored by } k \\ 0 & \text{else.} \end{cases} \tag{6}$$

Then for a joint ED P_{XZY} and $z^n \in \mathcal{T}^n_{Z|X}(x^n)$, by Chernoff bound,

$$\begin{aligned}
&\Pr\left\{\sum_{y^n \in \mathcal{T}^n_{Y|XZ}(x^n,z^n)} S_k(y^n) > \gamma^{-1}|\mathcal{T}^n_{Y|XZ}(x^n, y^n)|(1+\delta)\right\} \\
&\le e^{-\frac{\delta}{2}\gamma^{-1}|\mathcal{T}^n_{Y|XZ}(x^n,z^n)|(1+\delta)} \prod_{y^n \in \mathcal{T}^n_{Y|XZ}(x^n,z^n)} E\, e^{\frac{\delta}{2}S_k(y^n)} \\
&= c^{-\frac{\delta}{2}\gamma^{-1}|\mathcal{T}^n_{Y|XZ}(x^n,z^n)|(1+\delta)} \left[(1-\gamma^{-1}) + \gamma^{-1}e^{\frac{\delta}{2}}\right]^{|\mathcal{T}^n_{Y|XZ}(x^n,z^n)|} \\
&= e^{-\frac{\delta}{2}\gamma^{-1}|\mathcal{T}^n_{Y|XZ}(x^n,z^n)|(1+\delta)} \left[1 + (e^{\frac{\delta}{2}} - 1)\gamma^{-1}\right]^{|\mathcal{T}^n_{Y|XZ}(x^n,z^n)|} \\
&\le e^{-\frac{\delta}{2}\gamma^{-1}|\mathcal{T}^n_{Y|XZ}(x^n,z^n)|(1+\delta)} \left[1 + \gamma^{-1}(\frac{\delta}{2} + \frac{\delta^2}{8}e)\right]^{|\mathcal{T}^n_{Y|XZ}(x^n,z^n)|} \\
&\le \exp_e\left\{-\frac{\delta}{2}\gamma^{-1}|\mathcal{T}^n_{Y|XZ}(x^n, z^n)|(1+\delta) + \gamma^{-1}\left(\frac{\delta}{2} + \frac{\delta^2}{8}e\right)|\mathcal{T}^n_{Y|XZ}(x^n, z^n)|\right\} \\
&= \exp_e\left\{-\frac{\delta}{2}\gamma^{-1}|\mathcal{T}^n_{Y|XZ}(x^n, z^n)|\left(1 - \frac{e}{4}\right)\delta\right\} \\
&\le e^{-\frac{e\delta^2}{24}\gamma^{-1}|\mathcal{T}^n_{Y|XZ}(x^n,z^n))|} \\
&\le e^{-\frac{e\delta^2}{24}2^{n\eta}},
\end{aligned} \tag{7}$$

if $\gamma^{-1}|\mathcal{T}^n_{Y|XZ}(x^n, z^n)| > 2^{n\eta}$ and $\frac{\delta}{2} \le 1$.

Here, to obtain the 2nd and 3rd inequalities, we use for $x \in [0, 1]$ the inequalities $e^x \le 1 + x + \frac{e}{2}x^2$ and $1 + x \le e^x$ respectively.

Equation (5) follows from (7) because the numbers of sequences z^n and n-EDs increase exponentially and polynomially respectively as the length increases. □

To prove the (direct part of) coding theorem for wire-tap channel (without feedback) [11] Csiszár and Körner used a special code and we shall use its following improvement [10].

For a given wire-tap channel such that for an input RV X and its output RV's Y and Z for the legal user and wiretapper respectively

$$I(X;Y) - I(X;Z) > 0 \tag{8}$$

all $\lambda', \mu' > 0$ $0 < \varepsilon' < I(X;Y) - I(X;Z)$ and sufficiently large n, there exists a set of codewords

$$\{u_{m,\ell} : m = 0, 1, 2, \ldots, M-1,\ \ell = 0, 1, 2, \ldots, L-1\}$$

in $\mathcal{T}_X^n$ having the following properties.

$$I(X;Y) - I(X;Z) - \varepsilon' < \frac{1}{n}\log M \le I(X;Y) - I(X;Z) - \frac{\varepsilon'}{2} \tag{9}$$

$$I(X;Z) + \frac{\varepsilon'}{8} \le \frac{1}{n}\log L < I(X;Z) + \frac{\varepsilon'}{4}. \tag{10}$$

For a set of properly chosen decoding sets $\{\mathcal{D}_{m,\ell}\}$,

$$\{(u_{m,\ell}, \mathcal{D}_{m,\ell}) : m = 0, 1, 2, \ldots, M-1,\ \ell = 0, 1, 2, \ldots, L-1\}$$

is a λ-code for the legal user.

Let $V, \tilde{Z}$ be RV's taking values in $\mathcal{M} \times \mathcal{Z}^n$, where $\mathcal{M} = \{0, 1, \ldots, M-1\}$, with probability for $(m, z^n) \in \mathcal{M} \times \mathcal{Z}^n$

$$\Pr\{V, \tilde{Z}) = (m, z^n)) = \sum_{\ell=0}^{L-1} L^{-1} P_{Z|X}^n(z^n|u_{m,\ell}).$$

Then

$$I(V;\tilde{Z}) < \mu'. \tag{11}$$

4 The Coding Theorem for Transmission and Its Proof

Let $\mathcal{Q}$ be the set of quadruples of RV's (U, X, Y, Z) taking values in $\mathcal{U} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ for a finite set $\mathcal{U}$ with probability

$$\Pr((U, X, Y, Z) = (u, x, y, z)) = P_{UX}(ux)W(yz|x) \tag{12}$$

for $(u, x, y, z) \in \mathcal{U} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$.

Then

Theorem 178 *The capacity of a wire-tap channel with feedback satisfies*

$$C_{\mathrm{wtf}} \geq \max_{(U,X,Y,Z)\in\mathcal{Q}} \min[|I(U;Y) - I(U;Z)|^{+} + H(Y|U,Z), I(U;Y)]. \tag{13}$$

Proof For a $(U, X, Y, Z) \in \mathcal{Q}$, to show the achievability, one may introduce an auxiliary channel $P_{X|U}$ and construct a code for the channel

$$W'(y,z|u) = \sum_{x} P_{X|U}(x|u)W(y,z|x).$$

Then it is sufficient to show that $|I(X;Y) - I(X;Z)|^{+} + H(Y|XZ)$ is achievable. Let us fix $\lambda, \mu, \varepsilon > 0$ and construct a (λ, μ)-code with rate

$$|I(X;Y) - I(X,Z)|^{+} + H(Y|XZ) - \varepsilon. \tag{14}$$

To this end, let $\lambda', \mu', \varepsilon'$ be positive small real numbers specified later.

Let $\mathcal{U} = \{u_{m,\ell} : m = 0, 1, 2, \ldots, M-1, \ell = 0, 1, 2, \ldots, L-1\}$ be the codebook if in the previous section for a sufficiently large n (8) holds i.e., $I(X;Y)-I(X;Z) > 0$.

In the case that (8) does not hold we choose $M = 1$ and take a codebook of an arbitrary λ'-code for the legal user, with rate $I(X;Y) - \varepsilon' < R \triangleq \frac{1}{n}\log L \leq I(X;Y) - \frac{\varepsilon'}{2}$ as our codebook:

$$\mathcal{U} = \{u_{0,\ell} : \ell = 0, 1, 2, \ldots, L-1\}.$$

Our code consists of N blocks of length n and sends a message

$$(U'_1,\ U'_2U''_2,\ U'_3U''_3,\ \ldots,\ U'_NU''_N$$

uniformly distributed on $\mathcal{M}' \times (\mathcal{M}' \times \mathcal{M}'')^{N-1}$, where

$$\mathcal{M}' = \{0, 1, 2, \ldots, M-1\}, \qquad \mathcal{M}'' = \{0, 1, \ldots, L''-1\}, \tag{15}$$

and $L'' = \min\{L, 2^{n(H(Y|XZ)-\frac{3}{4})}\}$.

In particular $M = 1$, $\mathcal{M}'$ is a dummy message set. Then the rate of the messages is

$$R^* = \frac{1}{n}\log M + \frac{1}{n}\log L'' - \frac{1}{nN}\log L'' \geq \frac{1}{n}\log M + \frac{1}{n}\log L'' - \frac{1}{N}\log|\mathcal{Y}|.$$

That is by (9), and (10)

$$R^* \geq \begin{cases} I(X;Y) - I(X;Z) - \varepsilon' + \min\left[I(X;Z) + \frac{\varepsilon'}{8}, H(Y|XZ) - \frac{\varepsilon}{4}\right] \\ \quad -\frac{1}{N}\log|\mathcal{Y}| \qquad \text{if } I(X;Y) - I(X;Z) > 0 \\ \min\left[I(X;Y) - \frac{\varepsilon'}{2}, H(Y|XZ) - \frac{\varepsilon}{4}\right] - \frac{1}{N}\log|\mathcal{Y}| \qquad \text{else.} \end{cases} \tag{16}$$

By choosing $\varepsilon' < \frac{\varepsilon}{2}$ and $N > 2\varepsilon^{-1}\log|\mathcal{Y}|$ in (4.5) we have

$$R^* > \min[|I(X;Y) - I(X;Z)|^+ + H(Y|XZ), I(X;Y)] - \varepsilon \tag{17}$$

our desired rate.

In each block, we use a codebook

$$\mathcal{U} = \{u_{m,\ell} : m = 0, 1, \ldots, M-1, \ell = 0, 1, 2, \ldots, L-1\}$$

defined as above. Suppose the sender wants to send $(m'_1,\ m'_2m''_2,\ \ldots,\ m'_Nm''_N)$ to the receiver. Then our code consists of the following components.

1. In the first block the sender randomly chooses a $u_{m'_1,\ell}$ from the codebook with uniform distribution on $\{u_{m'_1}, j : j = 0, 1, \ldots, L-1\}$ and sends the codeword to the receiver. Then by choosing a proper decoder the receiver can decode $u_{m'_1,\ell}$ and therefore m'_1 correctly with probability $1 - \lambda'$.
2. From the first to the $N-1$st blocks, for all $u_{m,\ell} \in \mathcal{U}$, color all $\mathcal{T}^n_{\bar{Y}|\bar{X}}(u_{m,\ell}) \subset \mathcal{T}^n_{Y|X,\delta_1}(u_{m,\ell})$ with L'' colors such that for a suitably small $\delta_2 > o$ all n-joint ED $P_{\bar{X}\bar{Y}Z}$ with $P_{\bar{X}} = P_X$ and

$$\sum_{yz} |P_{\bar{Y}\bar{Z}\bar{X}}(y,z|x) - P_{YZ|X}(yz|x)| < \delta_2. \tag{18}$$

   $\mathcal{T}^n_{\bar{Y}|\bar{X}\bar{Z}}(u_{m,\ell}, z^n)$ is properly colored in the sense of Lemma 177.
3. For $j = 1, 2, \ldots, N-1$ after the sender receives output y^n of the jth block, he gives up if $y^n \notin \mathcal{T}^n_{Y|X,\delta_1}(u(j))$, where $u(j)$ is the input sequence in $\mathcal{X}^n$ sent by the sender in the jth block. Then the probability for giving up at the jth block is exponentially small in n. In the case $y^n \in \mathcal{T}^n_{Y|X,\delta_1}(u(j))$, y^n receives a coloring $c_{u(j)}(y^n) \in \{0, 1, \ldots, L''-1\}$ in the coloring for $\mathcal{T}^n_{\bar{Y}|\bar{X}}(u(j))$, where $P_{\bar{X}\bar{Y}}$ is the joint ED of $(u(j), Y^n)$.

   3.1. In the case $L \leq 2^{[H(Y|XZ)-\frac{3}{4}]}$ i.e., $L'' = L$, the sender sends $U_{m'_{j+1}m''_{j+1}} \oplus c_{m(j)}(y^n) \triangleq u(j+1)$ in the codebook $\mathcal{U}$ in the $j+1$st block, where $\oplus$ is the addition modulo L''.

3.2. In the case $L > 2^{n[H(Y|XZ)-\frac{3}{4}]}$, without loss of generality, we assume $L''|L$. Then the sender partitions $\{0, 1, \ldots, L-1\}$ into L'' segments of equal size. He randomly chooses an ℓ''_{j+1} in the $m''_{j+1} \oplus c_{u(j)}(y^n)$ segment with equal probability and sends $u_{m'_{j+1},\ell''_{j+1}}$ in the codebook in the $(j+1)$-th block.

4. For $j = 1, 2, \ldots, N$ in the jth block the receiver decode separately by using a proper decoder and obtains a $\bar{u}(j)$ in the jth block. Thus $\bar{u}(j) = u(j)$ with probability λ' for a given j. Let $\lambda' < M^{-1}\lambda$, then $\bar{u}(j) = u(j)$ with probability larger than $1-\lambda$ for all j. The receiver declares $m'_1 = \bar{m}'_1$ if $\bar{u}(1) = u_{\bar{m}'_1,\ell}$ for some ℓ. The receiver declares $m'_j m''_j = \bar{m}'_j \bar{m}''_j$ for $\bar{m}''_j = \ell_j \ominus c_{\bar{u}(j-1)}(y^n)$ if in the $(j-1)$-th block he receives y^n and $\bar{u}(j) = u_{\bar{m}'_j \ell_j}$ in the case $L'' = L$ and $\bar{u}(j) = u_{\bar{m}'_j,\ell'_j}$ for an ℓ'_j in the ℓ_jth segment in the case $L'' < L$, for $j = 2, 3, \ldots, N$. Obviously

$$(\bar{m}'_1, \bar{m}'_2\bar{m}''_2, \ldots, \bar{m}'_N\bar{m}''_N) = (m'_1 m m'_2 m''_2, \ldots, m'_N m''_N)$$

if $\bar{u}(j) = u(j)$ for all j.

We have seen that the probability of error is smaller than λ and it is sufficient for us to verify the security criterion.

Denote by $\tilde{X}_j$, $\tilde{Y}_j$ and $\tilde{Z}_j$, the random input and outputs in the jth block generated by the code and the random message, $(U'_1, U'_2U''_2, \ldots, U'_NU''_N)$ respectively, for $j = 1, 2, \ldots, N$. Notice here $\tilde{X}_j$, $\tilde{Y}_j$, and $\tilde{Z}_j$ are random sequences of length n. Let K_j be the coloring of the random output sequences of the legal receiver in the jth block. Write $U'^N = (U'_1, U'_2, \ldots, U'_N)$, $U''^N = (U''_1, U''_2, \ldots, U''_N)$ (where U'_1 is a dummy constant), $\tilde{X}^N = (\tilde{X}_1, \ldots, \tilde{X}_N)$, $\tilde{Y}^N = (\tilde{Y}_1, \ldots, \tilde{Y}_N)$ and $\tilde{Z}^N = (\tilde{Z}_1, \ldots, \tilde{Z}_N)$. Then we are concerned about an upper bound to $I(U'^N U''^N; \tilde{Z}^N)$. At first we bound $I(U'^N; \tilde{Z}^N)$ with (11). Denote $\tilde{Z}^{\bar{j}} = (\tilde{Z}_1, \tilde{Z}_2, \ldots, \tilde{Z}_{j-1}, \tilde{Z}_{j+1}, \ldots, \tilde{Z}_N)$.

Then by symmetry, independent of $\tilde{Z}^{\bar{j}}$ and U'^{j-1}, given $U'_j = m$, the input of the channel in the jth block is uniformly distributed on the sub-codebook $\{u_{m,\ell} : \ell = 0, 1, \ldots, L-1\}$. For $j = 1$ it immediately follows from the step 1 of the coding scheme. For $j > 1$, it is sufficient for us to show that $P_{U''_j \oplus K_{j-1}|U'^{j-1}\tilde{Z}^{\bar{j}}}$ is uniform. Indeed, for all ℓ, u'^{j-1}, and $z^{\bar{j}}$

$$\Pr\{U''_j \oplus K_{j-1} = \ell | U'^{j-1} = u'^{j-1}, \tilde{Z}^{\bar{j}} = z^{\bar{j}}\}$$

$$= \sum_{m''=0}^{L''-1} L''^{-1} \Pr\{K_{j-1} = \ell \ominus m'' | U^{j-1} = u'^{j-1}, \tilde{Z}^{\bar{j}} = Z^{\bar{j}}\} = L''^{-1}.$$

This means that for all j and $(V, \tilde{Z})$ in (11) we have

$$H(U'_j|U'^{j-1}\tilde{Z}^N) = H(U'_j|\tilde{Z}_j, U'^{j-1}Z^{\bar{j}}) = H(U|\tilde{Z})$$

and therefore by (11)

$$I(U'_j; U'^{j-1}\tilde{Z}^N) < \mu'$$

since U'_j and V have the same distribution.

Consequently

$$I(U'^N; Z^N) = \sum_{j=1}^{N} I(U_j; Z^N|U^{j-1}) \leq \sum_{j=1}^{N} I(U_j; U'^{j-1}Z^N) \leq N\mu'. \tag{19}$$

Next we bound $I(U''_j; \tilde{Z}^N|U'^N U''^{j-1})$. At first we observe that by our coding scheme U''_j is independent of $U'^N U''^{j-1}\tilde{Z}^i$ for all $i < j$ and therefore

$$I(U''_j; \tilde{Z}_i|U'^N U''^{j-1}\tilde{Z}^{i-1}) = 0,$$

or

$$\begin{aligned} &I(U''_j; \tilde{Z}^N|U'^N U''^{j-1}) \\ &= \sum_{i=1}^{j-1}[I(U''_j; \tilde{Z}_i|U'^N U''^{j-1}\tilde{Z}^{i-1}) + I(U''_j; \tilde{Z}_j|U'^N U''^{j-1}\tilde{Z}^{j-1}) \qquad (20) \\ &\qquad + I(U''_j; \tilde{Z}^N_{j+1}|U'^N U''^{j-1}\tilde{Z}^j)] \\ &= I(U''_j; \tilde{Z}_j|U'^N U''^{j-1}\tilde{Z}^{j-1}) + I(U''_j; \tilde{Z}^N_{j+1}|U'^N U''^{j-1}\tilde{Z}^j), \qquad (21) \end{aligned}$$

where $\tilde{Z}^N_{j+1} = (\tilde{Z}_{j+1}, \ldots, \tilde{Z}_N)$.

Moreover by our coding scheme under the condition given $U'^N U''^{j-1}\tilde{Z}^{j-1}$

$$U''_j \Leftrightarrow U''_j \oplus K_{j-1} \Leftrightarrow \tilde{Z}_j$$

form a Markov chain i.e., by the data processing inequality (Lemma 49).

$$\begin{aligned} I(U''_j; \tilde{Z}_j|U''^N U''^{j-1}Z^{j-1}) &\leq I(U''_j; U''_j \oplus K_{j-1}|U'^N U''^{j-1}Z^{j-1}) \\ &= I(U''_j; K_{j-1}|U'^N U''^{j-1}Z^{j-1}) \\ &\leq I(U'^N U''^j Z^{j-1}; K_{j-1}). \qquad (22) \end{aligned}$$

However, because

$$U'^N U''^j \tilde{Z}^{j-1} \Leftrightarrow \tilde{X}_{j-1}\tilde{Z}_{j-1} \Leftrightarrow K_{j-1}$$

forms a Markov chain, (22) implies

$$I(U''_j; \tilde{Z}_j | U'^N U''^j Z^{j-1}) \leq I(\tilde{X}_{j-1}\tilde{Z}_{j-1}; K_{j-1}). \tag{23}$$

For $j-1$

$$W_{j-1} = \begin{cases} 0 & \text{if } \tilde{Y}_{j-1} \in \mathcal{T}^n_{Y|X,\delta_1}(\tilde{X}_{j-1}) \\ 1 & \text{else,} \end{cases}$$

then recalling that the output of legal user is colored by Lemma 177 in the $j-1$st block, by AEP we have

$$\Pr\{K_{j-1} = k | \tilde{X}_{j-1} = x^n, \tilde{Z}_{j-1} = j^n W_{j-1} = 0\} \leq L''^{-1}(1+\delta).$$

Thus

$$\begin{aligned} H(K_{j-1}|\tilde{X}_{j-1}\tilde{Z}_{j-1}) &\geq (1-2^{-n\theta}) H(K_{j-1}|\tilde{X}_{j-1}\tilde{Z}_{j-1}W_{j-1}=0) \\ &\geq (1-2^{-n\theta})[\log L'' - \log(1+\delta)], \end{aligned}$$

for a $\theta > 0$ as $\Pr(W_j = 0) > 1 - 2^{-n\theta}$.

Thus for a $\mu'' > 0$ with $\mu'' \to 0$ as $\delta \to 0$,

$$I(\tilde{X}_{j-1}\tilde{Z}_{j-1}; K_{j-1}) = H(K_{j-1}) - \log L'' + \mu'' \leq \mu'', \tag{24}$$

for sufficiently large n.

Similarly by the coding scheme under the condition given U'^N

$$U''^j Z^j \Leftrightarrow K_j \Leftrightarrow Z^N_{j+1}$$

forms a Markov chain and therefore

$$\begin{aligned} I(U''_j; Z^N_{j+1} | U''^N U''^{j-1}) &\leq I(U''^j Z^j; \tilde{Z}^N_{j+1} | U'^N) \\ &\leq I(U''^j Z^j; K_j | U'^N) \leq I(U'^N U''^j \tilde{Z}^j; K_j). \end{aligned} \tag{25}$$

However, by the coding scheme

$$U'^N U''^j \tilde{Z}^j \Leftrightarrow \tilde{X}_j \tilde{Z}_j \Leftrightarrow K_j$$

forms a Markov chain and so we can continue to bound (25) as

$$I(U_j''; Z_{j+1}^N | U'^N U''^{j-1} Z^j) \le I(\tilde{X}_j \tilde{Z}_j; K_j). \tag{26}$$

By replacing $j-1$ by j in (24) and applying it to (26) we have

$$I(U_j''; Z_{j+1}^N | U'^N U''^{j-1} Z^j) \le \mu''. \tag{27}$$

Finally, we combine (19), (21), (22), (23), and (27), to obtain

$$\begin{aligned}
&I(U'^N U''^N; \tilde{Z}^N) \\
&\quad = I(U'^N; \tilde{Z}^N) + I(U''^N; \tilde{Z}^N | U'^N) \\
&\quad \le N\mu' + \sum_{j=2}^{N} I(U_j''; \tilde{Z}^N | U'^N U''^{j-1}) \\
&\quad = N\mu' + \sum_{j=2}^{N} [I(U_j''; \tilde{Z}_j | U'^N U''^{j-1} \tilde{Z}^{j-1}) + I(U_j''; \tilde{Z}_{j+1}^N | U'^N U''^{j-1} \tilde{Z}_j)] \\
&\quad \le N\mu' + \sum_{j=2}^{N} [I(\tilde{X}_{j-1} \tilde{Z}_{j-1}; K_{j-1}) + I(U_j''; \tilde{Z}_{j+1}^N | U'^N U''^{j-1} \tilde{Z}_j)] \\
&\quad \le N\mu' + 2(N-1)\mu'' < \mu,
\end{aligned}$$

for sufficiently small μ' and μ''.

This completes our proof. □

5 Capacity of Two Special Families of Wire-Tap Channels

In this section we apply Theorem 178 to show the following upper bound of capacity, which is believed not to be tight in general, but is tight for wire-tap channels with certain Markovities.

Let $\mathcal{Q}'$ be the set of triples of RV's (X, Y, Z) with joint distribution

$$P_{XYZ}(x, y, z) = P_X(x) W(y, z|x)$$

for $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathcal{Z}$.

Then

Lemma 179 *For all wire-tap channels*

$$C_{\text{wtf}} \leq \max_{(X,Y,Z)\in\mathcal{Q}'} \min[H(Y|Z), I(X;Y)]. \tag{28}$$

Proof For a given (λ, μ)-code for the wire-tap channel, let X^n, Y^n, Z^n be the input and outputs generated by uniformly distributed messages U through the code. Then in the same way to show the converse coding theorem of a (two terminal) noisy channel with feedback, one obtains that

$$C_{\text{wtf}} \leq \frac{1}{n}\sum_{t=1}^{n} I(X_t; Y_t) + \varepsilon' \tag{29}$$

where $\varepsilon' \to 0$ as $\lambda \to 0$.

On the other hand, by the security condition and Fano's inequality (Lemma 48) we have

$$\begin{aligned}
C_{\text{wtf}} = \frac{1}{n}H(U) &\leq \frac{1}{n}H(U|Z^n) + \mu \\
&\leq \frac{1}{n}H(U|Z^n) - \frac{1}{n}H(H|Y^n) + \lambda \log|\mathcal{X}| + \frac{1}{n}h(\lambda) + \mu \\
&\leq \frac{1}{n}H(U|Z^n) - \frac{1}{n}H(U|Y^nZ^n) + \lambda \log|\mathcal{X}| + \frac{1}{n}h(\lambda) + \mu \\
&= \frac{1}{n}I(U; Y^n|Z^n) + \varepsilon'' \leq \frac{1}{n}H(Y^n|Z^n) + \varepsilon'' \\
&= \frac{1}{n}\sum_{t=1}^{n} H(Y_t|Z^nY^{t-1}) + \varepsilon'' \leq \frac{1}{n}\sum_{t=1}^{n} H(Y_t|Z_t) + \varepsilon'',
\end{aligned} \tag{30}$$

where $h(\lambda) = -\lambda\log\lambda - (1-\lambda)\log(1-\lambda)$ and $\varepsilon'' = \lambda\log|\mathcal{X}| + \frac{1}{n}h(\lambda) + \mu \to 0$ as $\lambda, \mu \to 0$.

Let $(UXYZ)$ be a quadruple of RV's with distribution

$$P_{UXYZ}(t,z,y,z) = \frac{1}{n}\sum_{t=1}^{n} P_{X_tY_tZ_t}(x,y,z)$$

for $t \in \{1,2,\ldots,n\}$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $z \in \mathcal{Z}$.

Then $(XYZ) \in \mathcal{Q}'$ and by (29) and (30) for $\varepsilon = \max(\varepsilon', \varepsilon'')$

$$C_{\text{wtf}} \leq \min[H(Y|ZU), I(X;Y|U)] + \varepsilon \leq \min[H(Y|Z), I(X;Y)] + \varepsilon,$$

where $\varepsilon \to 0$ as $\lambda, \mu \to 0$.

That is, (28). □

Corollary 180 *For a wire-tap channel W such that there exist $W_1 : \mathcal{X} \to \mathcal{Y}$, and $W_2 : \mathcal{Y} \to \mathcal{Z}$ with*

$$W(y, z|x) = W_1(y|x)W_2(z|y), \tag{31}$$

for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathcal{Z}$

$$C_{\text{wtf}} = \max_{(X,Y,Z)\in\mathcal{Q}'}, \min[H(Y|Z), I(X; Y)].$$

Proof By Markov condition (31), we have that for all $(X, Y, Z) \in \mathcal{Q}'$

$$I(X; Y) - I(X, Z) \geq 0 \tag{32}$$

and

$$I(X; Z|Y) = 0. \tag{33}$$

Thus

$$\begin{aligned} |I(X; Y) - I(X; Z)|^+ + H(Y|XZ) &= H(X|Z) - H(X|Y) + H(Y|XZ) \\ &= H(XY|Z) - H(X|Y) \\ &= H(Y|Z) + H(X|YZ) - H(X|Y) \\ &= H(Y|Z) + I(X; Z|Y) \\ &= H(Y|Z). \end{aligned}$$

Then corollary follows from Theorem 178 and Lemma 179. □

Corollary 181 *For a wire-tap channel such that there exist $W_1' : \mathcal{X} \to \mathcal{Z}$ and $W_2' : \mathcal{Z} \to \mathcal{Y}$ with*

$$W(y, z|x) = W_1'(z|x)W_2'(y|z) \tag{34}$$

for $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathcal{Z}$

$$C_{\text{wtf}} = \max_{(X,Y,Z)\in\mathcal{Q}'}, \min[H(Y|Z), I(X; Y)].$$

Proof The Markov condition (34) implies that

$$I(X; Y) - I(X; Z) \leq 0 \tag{35}$$

and

$$H(Y|XZ) = H(Y|Z), \tag{36}$$

which yield

$$|I(X;Y) - I(X;Z)|^+ + H(Y|XZ) = H(Y|XZ) = H(Y|Z). \tag{37}$$

Thus the corollary follows from Theorem 178 and Lemma 179. □

Example An interesting example is a special channel for which W_1' is a noiseless channel and W_2' is a noisy channel in Corollary 181 e.g., W_1 is a noiseless binary channel, W_2'' is a binary symmetric channel with crossover probability $p \in \left(0, \frac{1}{2}\right)$. For this channel the wiretapper is in a better position than the legal user. So the capacity is zero without feedback. The feedback makes the capacity positive by our Corollary 181 as it serves as a secure key shared by sender and receiver. ▲

6 Discussion: Transmission, Building Common Randomness and Identification

As goals of communications are considered transmission i.e., sending a given message from a set of messages, building common randomness i.e., to provide a random resource shared by users, and identification i.e., identifying whether an event of interest to a particular user occurs ([2–5, 13]).

Roughly saying in a given communication system, the capacity of transmission is upper bounded by the capacity of common randomness, since common randomness shared by a sender and receiver can be built by transmission whereas the capacity of identification is lower bounded by capacity of common randomness, if the former is positive, which is shown by a scheme in [5] (chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas" in Part I) to build identification codes by common randomness. That is,

$$\text{capacity of transmission} \leq \text{capacity of common randomness} \tag{38}$$

$$\leq \text{capacity of identification.} \tag{39}$$

However, in different communication systems equalities in (39) may or may not hold. In this section we illustrate the variety in two-terminal channels and wire-tap channels. More examples in more complicated communication systems can be found e.g., in [2, 3, 12, 15].

First of all, obviously the first inequality in (39) is always an equality for a two terminal channel without feedback, because all information obtained by the receiver is from the transmission via the channel. Moreover, it has been shown in

[4] (chapter "Identification via Channels" in Part I) that the second inequality is an equality and therefore the three quantities in (39) are actually the same if the channel is discrete memoryless. A channel with rapidly increasing alphabet (as the length of codes grows) for which the capacity of identification is strictly larger than capacity of common randomness was described in [6]. It was shown in [8] that under a certain condition the capacity of common randomness (which is equal to the capacity of transmission) for Gaussian channels is finite whereas the capacity of identification is infinite in the same communication system. We notice that Gaussian channels have continuous, or infinite alphabets. It is natural to expect that for a discrete channel whose input alphabet "reasonably" increases the last two quantities, or consequently the three quantities in (39) are equal. This was shown in [14] for all channels whose input alphabets exponentially increase as the lengths of codes linearly increase.

The situation of two terminal channels is different when feedback is present. In this case the capacity of identification, which is equal to the capacity of common randomness, is strictly larger than the capacity of transmission for simplest channels, namely discrete memoryless channels (chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas". The reason is clear. On one hand, it is well known, feedback does not increase the capacity of transmission for discrete memoryless channels. On the other hand, the feedback provides a random resource, shared by sender and receiver, the random output, whose rate, roughly speaking, is input entropy. Obviously it increases common randomness between sender and receiver and therefore capacity of identification.

Next we turn to wire-tap channels without feedback. More precisely, we mean secure common randomness shared by sender and receiver, about which the wire-tapper has (almost) no knowledge. By the same reason as for two terminal channels without feedback, the capacity of (secure) common randomness is not larger than the capacity of transmission over the wire-tap channel. In fact it is shown in [2] and [3] (chapters "Perspectives" and "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints"), that it may not be larger than the capacity of transmission even in the case where a public forward channel with unbounded capacity is available to the sender and receiver. This intuitively is not surprising. Ahlswede and Zhang observed in [7] (see chapter "Identification via Channels with Noisy Feedback" in Part I) that to keep the message to be identified in secret a secure common randomness with positive rate is sufficient and the *major part of common randomness between the legitimate communicator applied in the identification code in* [5] *(chapter* "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas"*) can be publicly sent.*

Based on this observation they show that the capacity of identification is strictly larger than the capacity of secure common randomness. A more detailed analysis in [9] shows that the amount of secure common randomness needed only depends on the probability of second error and security criterion and is independent of the rate of messages. For fixed criterion of error and security, a constant amount – or zero-rate – of secure common randomness is sufficient, if provided with sufficiently large public common randomness.

Ahlswede gave an example of a non stationary memoryless channel with double exponentially growing input alphabet with identification capacity 1 and common randomness capacity 0. The structure of this channel has some similarities to the structure of ID-codes used in most of the achievability proofs for ID-coding theorems, thus it can be viewed as a channel with "built–in" ID-encoder. Kleinewächter gave a counter example for the other direction. For given real numbers C_{ID} and C_{CR} with $0 < C_{\mathrm{ID}} < C_{\mathrm{CR}}$, he explicitly construct a discrete channel with memory and noiseless passive feedback with identification capacity C_{ID} and common randomness capacity C_{CR}. This channel is constructed in such away that it can be used in two ways. In one respect, the channel is good for the generation of common randomness, in the other it is suitable for identification. It is quite reasonable to consider channels with memory. One may think for example of a system where data are transmitted by different voltage levels at high frequency. Because of the electrical capacity of the system it can be difficult to switch from a low voltage level to a high one and vice versa. There are also certain types of magnetic recording devices have problems with long sequences of the same letter. These examples for instance lead to the notion of run length limited codes. A third example are systems requiring the use of binary codewords which have approximately the same number of zeroes and ones. This limitation arises if the system can only transmit an unbiased alternating current, therefore these codes are called DC-free.

Let us return to our main topic wire-tap channels with secure feedback and investigate (39) in this communication system. We immediately find that the observation about wire-tap channels without feedback is still valid when feedback is present, because there is nothing in the observation which links to the existence of feedback. This means that the capacity of identification must be the capacity of "public" common randomness between sender and receiver i.e., the maximum rate of common randomness shared by the sender and the receiver, neglecting whether or how much the wiretapper knows about it once a positive amount of secure common randomness is provided. But now the public common randomness is the maximum output entropy for the channel $W_1 : \mathcal{X} \to \mathcal{Y}$ defined by

$$W_1(y|x) = \sum_{z \in \mathcal{Z}} W(y, z|x) \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y}, \tag{40}$$

or in other words $\max_{(X,Y,Z)\in\mathcal{Q}'} H(Y)$, for $\mathcal{Q}'$ as defined in Sect. 5. So we conclude that in this case the capacity of identification is either zero or $\max_{(X,Y,Z)\in\mathcal{Q}'}, H(Y)$. The only problem left is to find suitable conditions for the positivity of the capacity. We shall discuss this later.

To see the relation of the first pair of quantities in (39), we take a look at our main result

Theorem 182 *The information theoretical meaning of mutual information in* (13) *is obvious. The capacity of transmission with security criterion can not exceed that*

without it. So we expect this term could be removed in the formula of capacity of common randomness. To investigate the remaining term in (13)*, let us recall our coding scheme in Sect.* 4.

From the first block to the second last block, the transmission in each block has two tasks, sending a secret message m'_j (in the jth block) with a rate $\sim |I(U;Y) - I(U;Z)|$; and generating a secure common randomness with a rate $\sim H(Y|UZ)$, which will be used as a private key to send message m''_{j+1} in the next block. This gives us a secure common randomness with rate $\sim H(Y|UZ)$. The reason for the fact that U occurs in the "condition" is that the key for the $(j+1)$-th block has to be independent of the message sent in the jth block. For secure common randomness itself this is not necessary. So we expect that the capacity of common randomness is $\max_{(X,Y,Z)\in\mathcal{Q}'} H(Y|Z)$, which actually is shown in the next section.

But before this we have a remaining problem, namely the positivity of the capacity of identification, which should be discussed. First we notice that to have positive capacity of identification, the capacity of the channel W_1 in (40), where we do not count wiretapper's role, has to be positive. By counting wiretapper's role, we look for an input RV X, the conditional entropy $H(Y|Z)$ for output RV Y and Z has to be positive, because otherwise the wiretapper would know everything known by the legal receiver. *We shall show that the two necessary conditions together are sufficient for the positivity*.

7 The Secure Common Randomness Capacity in the Presence of Secure Feedback

Let $\mathcal{J}_n = \{0, 1, \ldots, J_n - 1\}$ be a finite set (whose size depends on n), $\lambda, \mu > 0$. An (n, J_n, λ, μ)-common randomness for the wire-tap channel with secure feedback is a pair of random variables (K_n, L_n) defined on the same domain $\mathcal{J}_n$ with the following properties.

There exists a RV U taking value in a finite set $\mathcal{U}$ and three functions $\theta^n : \mathcal{U} \times \mathcal{Y}^{n-1} \to \mathcal{X}^n$, $\varphi : \mathcal{U} \times \mathcal{Y}^n \to \mathcal{J}_n$, and $\Psi : \mathcal{Y}^n \to \mathcal{J}_n$ such that for all $u \in \mathcal{U}$ and $y^{n-1} \in \mathcal{Y}^{n-1}$

$$\theta^n(u, y^{n-1}) = (\theta_1(u), \theta_2(u, y_1), \ldots, \theta_n(u, y^{n-1})), \tag{41}$$

$$K_n = \varphi(U, Y^n), \tag{42}$$

$$L_n = \Psi(Y^n), \tag{43}$$

where Y^n and Z^n are output RV's for the legal receiver and the wiretapper, respectively, generated by random variable U, encoding function θ^n, and the channel W.

I.e.

$$\Pr((Y^n, Z^n) = (y^n, z^n)) = \sum_{u \in \mathcal{U}} \Pr(U = u) W(y_1, z_1 | \theta_1(u)) \prod_{t=2}^{n} W(y_t, z_t | \theta_t(u, y^{t-1})). \tag{44}$$

$$\Pr(K_n \neq L_n) < \lambda, \tag{45}$$

$$\frac{1}{n} H(K_n | Z^n) > \frac{1}{n} \log J_n - \mu. \tag{46}$$

$\frac{1}{n} \log J_n$ is called rate of the code and the capacity of the (secure) common randomness, denoted by C_{wtfCR}, is defined as the maximum achievable rate in the standard way.

Theorem 183

$$C_{wtfCR} = \max_{(X,Y,Z) \in \mathcal{Q}'} H(Y|Z), \tag{47}$$

in particular, the RHS of (7.7) is achievable if (7.6) is replaced by a stronger condition

$$H(K_n | Z^n) > \log J_n - \mu. \tag{48}$$

Proof The proofs to both, direct and converse parts, are straightforward. They immediately follow from the proofs for Theorem 178 and Lemma 179, respectively.

Let $(X', Y, Z) \in \mathcal{Q}'$ achieve the maximum at RHS (47). Apply Lemma 177 to color sets of typical remaining sequences $\mathcal{T}^n_{Y'} \subset \mathcal{T}^n_{Y,\delta}$,[1] then it follows from the proof of Theorem 178 (the part to show (23)) that for any fixed $\mu > 0$ and sufficiently large n

$$H(\tilde{K} | Z^n) > \log J_n - \mu,$$

where $\tilde{K}$ is the random J_n-coloring obtained from Lemma 177.

Choose $K_n = L_n = \tilde{K}$, then the proof of the direct part is done. To show the converse part we apply Fano's inequality (Lemma 48) to (45). Then

$$\begin{aligned} \frac{1}{n} \log J_n &\leq \frac{1}{n} H(K_n | Z^n) + \mu \\ &\leq \frac{1}{n} H(K_n | Z^n) - \frac{1}{n} H(K_n | Y^n) + \mu + \frac{1}{n} \lambda \log J_n + \frac{1}{n} h(\lambda) \end{aligned}$$

[1]More precisely, let $\mathcal{X}_0 = \{x_0\}$, $x^n = (x_0, x_0, \ldots, x_0)$, and (X, X', Y, Z) be RV's with joint distribution $\Pr((X, X', Y, Z) = (x^n, x'^n, y^n, z^n)) = P_{X'YZ}(x'^n, y^n, z^n)$ for all x'^n, y^n, z^n and coloring for the "conditional" typical sequences $\mathcal{T}^n_{Y|X}(x^n) = \mathcal{T}^n_{Y'}$.

$$\leq \frac{1}{n}H(K_n|Z^n) - \frac{1}{n}H(K_n|Y^n, Z^n) + \mu + \frac{1}{n}\lambda \log J_n + \frac{1}{n}h(\lambda)$$

$$\leq \frac{1}{n}I(K_n; Y^n|Z^n) + \mu + \frac{1}{n}\lambda \log J_n + \frac{1}{n}h(\lambda)$$

$$\leq \frac{1}{n}H(Y^n|Z^n) + \mu + \frac{1}{n}\lambda \log J_n + \frac{1}{n}h(\lambda).$$

Now the converse follows as in the proof for Lemma 179. □

8 The Secure Identification Capacity in the Presence of Secure Feedback

In this section let us take a look at the coding theorem for identification codes. First we have to formally define the codes and capacity. An $(n, |\mathcal{M}|, \lambda_1, \lambda_2, \mu)$-identification code for a wire-tap channel with secure feedback is a system $\{Q, \mathcal{D}_m : m \in \mathcal{M}\}$ such that $Q : \mathcal{M} \times Y^{n-1} \to \mathcal{X}^n$ is a stochastic matrix with

$$Q(x^n|m, y^{n-1}) = Q_1(x_1|m)\prod_{t=2}^{n} Q_t(x_t|m, y^{t-1})$$

for $m \in \mathcal{M}$, $y^{n-1} \in \mathcal{Y}^{n-1}$, for all $m \in \mathcal{M}$

$$\sum_{x^n \in \mathcal{X}^n} \sum_{y^n \in \mathcal{D}_m} Q_n(x_1|m)\prod_{t=2}^{n} Q_t(x_t|m, y^{t-1})W_1(y_t|x_t) > 1 - \lambda_1,$$

for $m, m' \in \mathcal{M}$ with $m \neq m'$

$$\sum_{x^n \in \mathcal{X}^n} \sum_{y^n \in \mathcal{D}'_m} Q_1(x_1|m)\prod_{t=2}^{n} Q_t(x_t|m, y^{t-1})W_1(y_t|x_t) < \lambda_2,$$

and for all $m, m' \in \mathcal{M}$, $m \neq m'$ and $\mathcal{V} \subset Z^n$

$$\sum_{x^n \in \mathcal{X}^n} \sum_{y^n \in \mathcal{Y}^n} Q_1(x_1|m')\prod_{t=2}^{n} Q_t(x_t|m', y^{t-1})W(y^n, \mathcal{V}|x^n)$$

$$+ \sum_{x^n \in \mathcal{X}^n} \sum_{y^n \in \mathcal{Y}^n} Q_1(x_1|m)\prod_{t=2}^{n} Q_t(x_t|m, y^{t-1})W(y^n, \mathcal{V}^c|x^n) > 1 - \mu.$$

Then capacity of identification is defined in the standard way and denoted by C_{wtfID}.

C_{wtfID} is upper bounded by the RHS of (49), follows from the converse of the coding theorem of identification with feedback for channel W_1 (chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas"). In the case that II holds, one can construct a code achieving $H(Y)$ asymptotically from the code in chapter "Identification via Channels with Noisy Feedback" by replacing the ordinary code for W_1 by a uniform partition of output sequences for the legal receiver and a code for the wire-tap channel without feedback by a code for the same channel but with feedback.

Furthermore the two conditions in III

Theorem 184 *The following statements are equivalent.*

(i)

$$C_{wtfID} = \max_{(X,Y,Z)\in\mathcal{Q}'} H(Y) \tag{49}$$

(ii) $C_{wtf} > 0$

(iii) *There exists an* $(X, Y, Z) \in \mathcal{Q}'$ *such that*

$$H(Y|Z) > 0$$

and the channel W_1 *has positive capacity.*

Proof The converse of the coding theorem i.e., C_{wtfID} is upper bounded by the right hand side of (49) follows from the converse of coding theorem of identification with feedback for channel W_1 ([4], chapter "Identification via Channels"). In the case that (ii) holds, one can construct a code achieving $H(Y)$ asymptotically from the code in theorem 2 (see chapter "Identification via Channels") by replacing the ordinary code for W_1 by a uniform partition of output sequences for the legal receiver and a code for the wiretap channel without feedback by a code for the same channel but with feedback.

Furthermore the two conditions in (iii) obviously are necessary for positivity of C_{wtfID}. The only thing left to be proved is that (iii) implies (ii). Let $(X_i, Y_i, Z_i) \in \mathcal{Q}'$ for $i = 0, 1$ such that $H(Y_0|Z_0) > 0$ and $I(X_1, Y_1) > 0$. By Theorem 178, it is sufficient for us to find $(U, X, Y, Z) \in \mathcal{Q}$ such that $I(U; Y) > 0$ and $H(Y|UZ) > 0$. Obviously we are done, if $I(X_0; Y_0) > 0$ or $H(Y_1|U_1, Z_1) > 0$. Otherwise we have to construct a quadruple of RV's $(U, X, Y, Z) \in \mathcal{Q}$ from (X_0, Y_0, Z_0) and (X_1, Y_1, Z_1) such that $H(Y|UZ) > 0$ and $I(U; Y) > 0$. To this end, let $\mathcal{U} = \mathcal{X} \cup \{u_0\}$, (where u_0 is a special letter not in $\mathcal{X}$), and for all $u \in \mathcal{U}$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$, let (U, X, Y, Z) be a quadruple of RV's such that

$$P_{UXYZ}(u, x, y, z) = \begin{cases} \frac{1}{2} P_{X_0Y_0Z_0}(x, y, z) & \text{if } u = u_0 \\ \frac{1}{2} P_{X_1Y_1Z_1}(x, y, z) & \text{if } u \in \mathcal{X} \text{ and } u = x \\ 0 & \text{otherwise.} \end{cases}$$

Then $(U, X, Y, Z) \in \mathcal{Q}$, $P_{YZ|U}(y|u_0) = P_{Y_0Z_0}(yz)$ for all $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$. $P_0(u_0) = \frac{1}{2}$ and therefore

$$H(Y|UZ) = \sum_{u \in \mathcal{U}} P_U(u) H(Y|U = uZ) \geq \frac{1}{2} H(Y|U = u_0 Z) = \frac{1}{2} H(Y_0|Z_0) > 0.$$

On the other hand for

$$S = \begin{cases} 0 & \text{if } U = u_0 \\ 1 & \text{otherwise,} \end{cases}$$

for all $u \in \mathcal{X}$, $y \in \mathcal{Y}$

$$P_{UY|S}(u, y|S = 1) = P_{X_1Y_1}(u, y)$$

and $P_s(1) = \frac{1}{2}$ and consequently

$$I(U; Y) = I(US; Y) \geq I(U; Y|S) \geq P_s(1) I(U; Y|S = 1) = \frac{1}{2} I(X_1; Y_1) > 0.$$

That is, (U, X, Y, Z) is as desired. □

We conclude with the

Corollary 185

$$C_{wtfID} = \begin{cases} \max\limits_{(X,Y,Z) \in \mathcal{Q}'} H(Y|Z) \\ 0 \end{cases}$$

and $C_{wtfID} = 0$ iff for all $(X, Y, Z) \in \mathcal{Q}'$ $H(Y|Z) = 0$ or the capacity of W_1 is zero.

Proof That for all $(X, Y, Z) \in \mathcal{Q}'$, $H(Y|Z) = 0$ implies that the wiretapper knows what the receiver receives with probability one no matter how the sender chooses the input and that the capacity of W_1 is zero means the sender may not change the output distributions at the terminal for the legal receiver. So in both cases $C_{\text{wtfID}} = 0$. Thus the corollary follows from Theorem 184. □

References

1. R. Ahlswede, Coloring hypergraphs: a new approach to multiuser source coding. J. Comb. Inf. Syst. Sci. **1**, 76–115 (1979) and **2**, 220–268 (1980)
2. R. Ahlswede, I. Csiszár, Common randomness in information theory and cryptography, Part I: Secret sharing. IEEE Trans. Inform. Theory **39**, 1121–1132 (1993)

3. R. Ahlswede, I. Csiszár, Common randomness in information theory and cryptography, Part II: CR capacity. IEEE Trans. Inform. Theory **44**(1), 225–240 (1998)
4. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inform. Theory **35**, 15–29 (1989)
5. R. Ahlswede, G. Dueck, Identification in the presence of feedback — a discovery of new capacity formulas. IEEE Trans. Inform. Theory **35**, 30–39 (1989)
6. R. Ahlswede, C. Kleinewächter, Pathological examples with different common randomness and (second order) identification capacities, this volume pages
7. 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. Theory **41**(4), 1040–1050 (1995)
8. M. Burnashav, On identification capacity of infinite alphabets or continuous time channel. IEEE Trans. Inf. Theory **IT–46**, 2407–2414 (2000)
9. N. Cai, K.Y. Lam, On identification secret sharing schemes. Inf. Comput. **184**(2), 298–310 (2003)
10. I. Csiszár, Almost independence and secrecy capacity, Probl. Inform. Trans. **32**, 40–47 (1996)
11. I. Csiszár, J. Körner, Broadcast channels with confidential messages. IEEE Trans. Inf. Theory **24**, 339–348 (1978)
12. I. Csiszár, P. Narayan, The secret key capacity for multiple terminals, in *Information Theory Workshop, Proceedings of the 2002*. IEEE (2002)
13. C.E. Shannon, A mathematical theory of communication. Bell. Syst. Tech. J. **27**, 379–423 (1948)
14. Y. Steinberg, New converses in the theory of identification via channels. IEEE Trans. Inform. Theory **44**, 984–998 (1998)
15. S. Venkatesan, V. Anantharam, The common randomness capacity of network of discret mamoryless channels. IEEE Trans. Inf. Theory **46**, 367–387 (2000)

Secrecy Systems for Identification Via Channels with Additive-Like Instantaneous Block Encipherer

In this lecture we propose a model of secrecy systems for identification via channels with ALIB encipherers and find the smallest asymptotic key rate of the ALIB encipherers needed for the requirement of security.

1 Introduction

Attention: This is the only lecture in the book which works with the *optimistic capacity*, which is the optimal rate achievable with arbitrary small error probability again and again as the blocklength goes to infinity.

The criticism of this concept made in [B34] has been supplemented by a new aspect: *in cryptology enemies strongest time in wire-tapping must be taken into consideration!*

The model of identification via channels was introduced by Ahlswede and Dueck. [2] (see chapter "Identification via Channels", Part I) based on the following cases. The receivers of channels only are interested in whether a specified message was sent but not in which message was sent and the senders do not know in which message the receivers are interested. Sometimes the sender requires that the message sent can be identified only by legitimate receivers of the channel but not by any one else (e.g., wiretapper). For example, a company produces N kinds of products which are labeled by $j = 1, 2, \cdots, N$. The company wants to sell a kind of products only to the members of the company's association. For other customers it even does not want them to know what it is going to sell. In this case the company can use a secrecy system for identification via channels with additive-like instantaneous block (ALIB) encipherers, i.e., the sender encrypts the message (identification code) with a private key sending it via the channel and sends the same key only to the members of the company's association through a secure channel. The secrecy system with ALIB encipherers was investigated by Ahlswede and Dueck [1], but their model needs to

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_18

be adapted to satisfy the requirement of identification via channels. In this lecture we consider the model of secrecy systems for identification via channels with ALIB encipherers and investigate the smallest asymptotic key rate of the ALIB encipherers needed for the requirement of security.

In Sect. 2, we review the necessary background of identification via channels. Our model is described in Sect. 3. Our result for symmetric channels is proved in Sect. 4.

2 Background

Let $\mathcal{X}$, $\mathcal{K}$, $\mathcal{Y}$ and $\mathcal{Z}$ be finite sets. For simplicity, we assume that $\mathcal{X} = \mathcal{K} = \mathcal{Y} = \mathcal{Z} = GF(q)$ with $q \geq 2$. Let $W = \{W^n\}_{n=1}^{\infty}$ be a memoryless channel with transmission matrix $(w(z|x); x \in \mathcal{X}, z \in \mathcal{Z})$.

Definition 186 A randomized $(n, N_n, \mu_n, \lambda_n)$ identification (ID) code for the channel W^n is a system $\{(Q_i, D_i); 1 \leq i \leq N_n\}$, where Q_i is a probability distribution (PD) of the random codeword $X^n(i)$ generated by a randomized encoder $\varphi_n(i)$, i.e. $Q_i(x^n) = \Pr\{X^n(i) = x^n\}$, $x^n \in \mathcal{X}^n$, $D_i \subset \mathcal{Z}^n$ is a decoding set.

Denote by $Z^n(i)$ the output of W^n when the input is $X^n(i)$ and $Q_i W^n$ the PD of $Z^n(i)$. Set $\mu_n^{(i)} = Q_i W^n(D_i^c) = \Pr\{Z^n(i) \in \mathcal{Z}^n - D_i\}$ and $\lambda_n^{(j,i)} = Q_j W^n(D_i) = \Pr\{Z^n(j) \in D_i\}(j \neq i)$. $\mu_n = \max\limits_{1 \leq i \leq N_n} \mu_n^{(i)}$ and $\lambda_n = \max\limits_{1 \leq j, i \leq N_n, j \neq i} \lambda_n^{(j,i)}$ are called the error probability of the first and second kind for the ID code, respectively, $\frac{1}{n} \log \log N_n = r_n$ is called the rate of the ID code.

Definition 187 A rate R is a (μ, λ)-achievable rate if there exists a sequence of $(n, N_n, \mu_n, \lambda_n)$ ID codes for the channel W^n $(1 \leq n < \infty)$ satisfying the following conditions.

(i) $\limsup\limits_{n \to \infty} \mu_n \leq \mu$,
(ii) $\limsup\limits_{n \to \infty} \lambda_n \leq \lambda$,
(iii) $\liminf\limits_{n \to \infty} r_n \geq R$.

The (μ, λ)-Id capacity for the channel W is defined by

$$D(\mu, \lambda|W) = \sup(R|R \text{ is } (\mu, \lambda) \text{ -achievable}).$$

Theorem 188 (Ahlswede and Dueck 1989 [2]) *Let $W = \{W^n\}_{n=1}^{\infty}$ be an arbitrary channel. If there exists a number ε satisfying $0 \leq \varepsilon \leq \mu$ and $0 \leq \varepsilon \leq \lambda$, then it holds that $D(\mu, \lambda|W) \geq C(\varepsilon|W)$, where $C(\varepsilon|W)$ denotes the ε-channel capacity of the channel W which is defined as follows.*

Definition 189 Rate R is ε-achievable if there exists a sequence of (n, M_n, ε_n) codes for the channel $W^n (1 \leq n \leq \infty)$ satisfying the following conditions.

(i) $\limsup_{n\to\infty} \varepsilon_n \leq \varepsilon$,
(ii) $\liminf_{n\to\infty} \frac{1}{n} \log M_n \geq R$.

The ε-channel capacity for the channel W is defined by

$$C(\varepsilon|W) = \sup(R|R \text{ is } \varepsilon\text{-achievable}).$$

Theorem 188 is proved by using the following lemma.

Lemma 190 (Ahlswede and Dueck 1989 [2]) *Let $\mathcal{M}$ be an arbitrary finite set of size $M =| \mathcal{M} |$. Choose constants τ and κ satisfying $0 < \tau \leq \frac{1}{3}$ and $0 < \kappa < 1$ and $\kappa \log(\frac{1}{\tau} - 1) \geq \log 2 + 1$, where the natural logarithms are used. Define $N = \lfloor e^{\tau M}/Me \rfloor$. Then, there exist N subsets $A_1, A_2, \cdots, A_N$ of $\mathcal{M}$ satisfying $| A_i | = \lfloor \tau M \rfloor$ $(1 \leq i \leq N)$ and $| A_i \cap A_j | < \kappa \lfloor \tau M \rfloor (i \neq j)$.*

Using Lemma 190 the ID-code for proving Theorem 188 can be constructed as follows.

Let $\gamma > 0$ be an arbitrarily small constant and set $R = C(\varepsilon|W) - \gamma$. By Definition 189 R is ε-achievable as a rate of the transmission code. Therefore, there exists a sequence of (n, M_n, ε_n) codes for the channel $W^n (1 \leq n < \infty)$ satisfying the following conditions:

(i) $\limsup_{n\to\infty} \varepsilon_n \leq \varepsilon$,
(ii) $\liminf_{n\to} \frac{1}{n} \log M_n \geq R$,

where ε_n denotes the maximum decoding error probability of the code. Denote the (n, M_n, ε_n) code by $\mathcal{C}_n = \{c_1, c_2, \cdots, c_{M_n}\}$
$(c_i \in \mathcal{X}^n)$ and let E_i be the decoding region corresponding to $c_i (1 \leq i \leq M_n)$.

Now we apply Lemma 190 by setting $\mathcal{M} = \{1, 2, \cdots, M_n\}, M = M_n, \tau = \tau_n = \frac{1}{(n+3)}, \kappa = \kappa_n = \frac{2}{\log(n+2)}$ and $N = N_n = \lfloor e^{\tau_n M_n}/M_n e \rfloor$. Since all conditions of Lemma 190 are satisfied, there exist N_n subsets $A_1, A_2, \cdots, A_{N_n}$ of $\mathcal{M}$ satisfying $|A_j| = \lfloor \tau_n M_n \rfloor (1 \leq j \leq N_n)$ and $| A_j \cap A_k | < \kappa_n \lfloor \tau_n M_n \rfloor (j \neq k)$. Define the subsets S_j $(1 \leq j \leq N_n)$ of $\mathcal{C}_n$ by $S_j = \bigcup_{i \in A_j} \{c_i\}$ and let Q_j denote the uniform distribution over S_j. Define $D_j = \bigcup_{i \in A_j} E_i$ as the decoding set corresponding to Q_j.
It is shown that the constructed ID code $\{(Q_j, D_j); 1 \leq j \leq N_n\}$ can be used to prove Theorem 188.

Theorem 188 gives the direct theorem on the ID coding problem. We need the converse theorem also. Since the converse theorem is essentially related to the channel resolvability problem, we can introduce the channel resolvability instead.

Let $W = \{W^n\}_{n=1}^{\infty}$ be an arbitrary channel with input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ respectively. Let $Y = \{Y^n\}_{n=1}^{\infty}$ be the output from the channel W corresponding

to a given input $X = \{X^n\}_{n=1}^{\infty}$. We transform the uniform random number U_{M_n} of size M_n into another input $\widetilde{X} = \{\widetilde{X}^n\}_{n=1}^{\infty}$. That is, $\widetilde{X}^n = f_n(U_{M_n})$, $f_n : \{1, 2, \cdots, M_n\} \to \mathcal{X}^n$.

Denote by $\widetilde{Y} = \{\widetilde{Y}^n\}_{n=1}^{\infty}$ the output from the channel W with an input $\widetilde{X}$. The problem of how we can choose the size M_n of the uniform random number U_{M_n} and the transform f_n such that the variational distance between $Y = \{Y^n\}_{n=1}^{\infty}$ and $\widetilde{Y} = \{\widetilde{Y}^n\}_{n=1}^{\infty}$ satisfies $\lim_{n\to\infty} d(Y^n, \widetilde{Y}^n) = 0$ is sometimes called the channel resolvability problem. In this problem, the criterion of approximation can be slightly generalized to $\limsup_{n\to\infty} d(Y^n, \widetilde{Y}^n) \le \delta$, where δ is an arbitrary constant satisfying $0 \le \delta < 2$.

Definition 191 Rate R is δ-achievable for an input $X = \{X^n\}_{n=1}^{\infty}$ if there exists a sequence of transforms $\widetilde{X}^n = f_n(U_{M_n})(1 \le n < \infty)$ satisfying

$$\limsup_{n\to\infty} d(Y^n, \widetilde{Y}^n) \le \delta \qquad \text{and} \qquad \limsup_{n\to\infty} \frac{1}{n} \log M_n \le R,$$

where Y^n and $\widetilde{Y}^n$ denote the channel outputs corresponding to X^n and $\widetilde{X}^n$, respectively. The channel δ-resolvability for an input X is defined by

$$S_X(\delta|W) = \inf(R|R \text{ is } \delta\text{-achievable for an input } X).$$

Theorem 192 (Han 2003 [4]) *Let W be an arbitrary channel with time structure and X an arbitrary input variable. Then, it holds that $S_X(\delta|W) \le \overline{I}(X;Y)$ for all $\delta \ge 0$, where Y denotes the channel output variable corresponding to X and $\overline{I}(X;Y)$ represents the sup-mutual information rate defined by*

$$\overline{I}(X;Y) = p - \limsup_{n\to\infty} \frac{1}{n} \log \frac{W^n(Y^n|X^n)}{P_{Y^n}(Y^n)} \tag{1}$$

$$= \inf\left(\alpha \left| \lim_{n\to\infty} \Pr_{X^nY^n}\left\{\frac{1}{n}\log\frac{W^n(Y^n|X^n)}{P_{Y^n}(Y^n)} > \alpha\right\} = 0\right.\right). \tag{2}$$

3 Model

In this section we propose a model of the secrecy systems for identification via channels with ALIB encipherers. We keep the notations and assumptions given in Sect. 2 for reviewing the background of identification via channels.

Let $\{(Q_i, D_i) : 1 \le i \le N_n\}$ be the $(n, N_n, \mu_n, \lambda_n)$ ID code constructed as in the proof of Theorem 188 for the channel W. Recall that an (n, R) ALIB encipherer is a subset $C \subset \mathcal{K}^n$ with $\mid C \mid < e^{nR}$. Let $f : \mathcal{X} \times \mathcal{K} \to \mathcal{Y}$ be a function, where $f(x, \cdot)$ is bijective for each $x \in \mathcal{X}$ and $f(\cdot, k)$ is bijective for each $k \in \mathcal{K}$. $f^n : \mathcal{X}^n \times \mathcal{K}^n \to \mathcal{Y}^n$ denotes the n-fold product of f. Given a pair (f, C) we define a secrecy system which works as follows. If the sender wants to send a message

$i(1 \le i \le N_n)$, he sends the random codeword $X^n(i)$ generated by the randomized encoder $\varphi_n(i)$. Before he transmits $X^n(i)$ he uses a random key generator K^n to generate k^n according to the uniform distribution on C. Then the sender encrypts $X^n(i)$ into the random cryptogram $Y^n(i) = f^n(X^n(i), K^n)$ and sends it to the receiver over the channel W^n. Suppose that $X^n(i)$ and K^n are mutually independent. The used key k^n is sent to the receiver over a secure channel. Denote by $\widetilde{Z}^n(i)$ the output of the channel W^n when the input is the cryptogram $Y^n(i)$. In general, the receiver cannot use the same key k^n to recover the received codeword $Z^n(i)$ from the received cryptogram $\widetilde{Z}^n(i)$ since the channel W^n is noisy. In order to solve this problem, we assume that $f(x, k) = x + k$, where $+$ operates in $GF(q)$. Then we have $Y^n(i) = X^n(i) + K^n$. Further, we need to assume that the channel W^n is memoryless with *symmetric transmission matrix*, more specifically, the output and input of the channel W^n have the following relation: $\widetilde{Z}^n(i) = Y^n(i) + E^n$, where $E^n = (E_1, E_2, \cdots, E_n)$ is a sequence of independent RV's with the same PD on $GF(q)$. Combining the two assumptions, we obtain $\widetilde{Z}^n(i) = X^n(i) + K^n + E^n = Z^n(i) + K^n$ or $Z^n(i) = \widetilde{Z}^n(i) - K^n$. Hence the receiver can get $Z^n(i)$ from $\widetilde{Z}^n(i)$ by using the same key k^n and decides that the message $i(1 \le i \le N_n)$ is sent if $Z^n(i) \in D_i$. Since the PD of $Z^n(i)$ is $Q_i W^n$ and $Q_i W^n(D_i^c) \le \mu_n$, $Q_j W^n(D_i) \le \lambda_n (j \ne i)$, the receiver can identify the message i with error probabilities of the first kind and second kind not greater than μ_n and λ_n, respectively. Another customer intercepts the channel output $\widetilde{Z}^n(i)$ and attempts to identify a message $j(1 \le j \le N_n)$ being sent. Since the customer does not know the actual key k^n being used, he has to use $\widetilde{Z}^n(i)$ and his knowledge of the system for deciding that the message j is sent. We need a security condition under which the customer can not decide for any fixed message $j(1 \le j \le N_n)$ being sent with small error probability. Such a condition was given by Ahlswede and Zhang [3] (see chapter "Identification via Channels with Noisy Feedback", Part I) for investigating the problem of identification via a wiretap channel. This condition is also suitable for our model. The condition is stated as follows.

Security Condition. For any pair of messages $(i, j)(1 \le i \ne j \le N_n)$ and $D \subset \mathcal{Z}^n$, it holds that $\widetilde{Q}_i W^n(D^c) + \widetilde{Q}_j W^n(D) > 1 - \delta_n$ and $\lim_{n\to\infty} \delta_n = 0$, where $\widetilde{Q}_i$ and $\widetilde{Q}_i W^n$ denote the PD of $Y^n(i)$ and $\widetilde{Z}^n(i)$ respectively.

From the identity $\widetilde{Q}_i W^n(D^c) + \widetilde{Q}_i W^n(D) = 1$ for any $i(1 \le i \le N_n)$ and any $D \subset \mathcal{Z}$ and the Security Condition, we obtain $\widetilde{Q}_j W^n(D) > 1 - \widetilde{Q}_i W^n(D^c) - \delta_n = \widetilde{Q}_i W^n(D) - \delta_n$ for any pair $(i, j)(1 \le i \ne j \le N_n)$. Therefore, the Security Condition means that $\widetilde{Q}_i W^n$ and $\widetilde{Q}_j W^n$ are almost the same for any pair (i, j) with $i \ne j$. Hence the customer can not decide on any fixed message $j(1 \le j \le N_n)$ being sent with small error probability.

We are interested in the following problem. What is the largest rate R of the ALIB encipherer C so that the distributions $\widetilde{Q}_i W^n (i = 1, 2, \cdots, N_n)$ satisfy the Security Condition.

4 Main Result

For the model of a secrecy system described in Sect. 3 we obtain the following main result.

Theorem 193

1. *Assume for the alphabets* $\mathcal{X} = \mathcal{K} = \mathcal{Y} = \mathcal{Z} = GF(q)(q \geq 2)$ *and that* $W = \{W^n\}_{n=1}^{\infty}$ *is a memoryless symmetric channel with the transmission matrix* $(w(z|x) > 0; x \in \mathcal{X}, z \in \mathcal{Z})$.
2. *Assume that the function* $f(x, k) = x + k$, *where* $+$ *operates in the finite field* $GF(q)$.
3. *Suppose that the random key* K^n *has uniform distribution on the ALIB encipherer* $C \subset \mathcal{K}^n$ *and is mutually independent with each random codeword* $X^n(i)(1 \leq i \leq N_n)$.

Then, the secrecy system for identification via the channel W *with ALIB encipherers possesses the following properties.*

(i) *The secrecy system can transmit* N_n *messages* $i = 1, 2, \cdots, N_n$ *with*

$$\liminf_{n\to\infty} \frac{1}{n} \log\log N_n \geq \log q + \sum_{z\in\mathcal{Z}} w(z|x) \log w(z|x) - \gamma,$$

where $\gamma > 0$ *is an arbitrarily small number and* $x \in \mathcal{X}$ *is fixed, the legitimate receiver can identify the message* $i(1 \leq i \leq N_n)$ *with arbitrarily small error probability.*

(ii) *The smallest asymptotic key rate* R *of the ALIB encipherer* C *is* $R = -\sum_{z\in\mathcal{Z}} w(z|x) \log w(z|x)$ *(*$x \in \mathcal{X}$ *is fixed) for the distributions* $\widetilde{Q}_i W^n (i = 1, 2, \cdots, N_n)$ *satisfying the Security Condition. Hence, the other customer can not judge any fixed message* $j(1 \leq j \leq N_n)$ *being sent from* $\widetilde{Q}_i W^n$ *with small error probability.*

Proof

(i) By assumption 1, the transmission capacity of the channel W is $C(W) = C(0|W) = \log q + \sum_{z\in\mathcal{Z}} w(z|x) \log w(z|x)$. Using Theorem 188 with $\varepsilon = 0$, we obtain that the (μ, λ)-ID capacity of the channel W, $D(\mu, \lambda|W) \geq C(W)$ for $\mu \geq 0, \lambda \geq 0$. Hence, there exists a sequence of $(n, N_n, \mu_n, \lambda_n)$ ID codes for the channel $W^n(1 \leq n < \infty)$ satisfying the conditions: 1: $\lim_{n\to\infty} \mu_n = 0$; 2: $\lim_{n\to\infty} \lambda_n = 0$; 3: $\liminf_{n\to\infty} r_n \geq C(W) - \gamma$. Using the Id codes in the secrecy system, the property (1) holds.

(ii) By assumption 2, the random cryptogram $Y^n(i) = X^n(i) + K^n$, where the random key K^n has uniform distribution on an ALIB encipherer $C \subset \mathcal{K}^n$. Ahlswede and Dueck [1] have pointed out that $Y^n(i)$ and K^n can be regarded

as the output and input of the channel denoted by $V = \{V^n\}_{n=1}^{\infty}$. In the case of identification, the channel V is a general channel rather than a memoryless channel. By assumption 3, the transmission probability of the channel V^n can be defined as $V^n_{y^n|k^n} = \sum_{x^n} Q_i(x^n)\delta(y^n, x^n + k^n)$, where

$$\delta(y^n, x^n + k^n) = \begin{cases} 1, & \text{if } y^n = x^n + k^n, \\ 0, & \text{otherwise.} \end{cases}$$

In order to prove property (ii), we want to apply Theorem 192 for the general channel V. First, we consider the input U^n of the channel V^n which has uniform distribution on the ALIB encipherer $C = \mathcal{K}^n$. It is evident that the PD of the output $Y^n(i)$ corresponding to the input U^n is the uniform distribution on $\mathcal{Y}^n$, i.e., $\widetilde{Q}_i(y^n) = \Pr\{Y^n(i) = y^n\} = q^{-n}$ for any $y^n \in \mathcal{Y}$ and any $i(1 \leq i \leq N_n)$. By the assumption 1, it is also evident that the PD of the output $\widetilde{Z}^n(i)$ of the channel W^n corresponding to the input $Y^n(i)$ is the uniform distribution on $\mathcal{Z}^n$, i.e., $\widetilde{Q}_i W^n(z^n) = q^{-n}$ for any $z^n \in \mathcal{Z}^n$ and any $i(1 \leq i \leq N_n)$. Hence $\widetilde{Q}_i W^n(1 = 1, 2, \cdots, N_n)$ satisfy the Security Condition. But the key rate of $C = \mathcal{K}^n$ equals $\log q$, it can be reduced. Then, applying Theorem 192 for the input $U = \{U^n\}_{n=1}^{\infty}$ and $\delta = 0$, we obtain $S_U(0|V) \leq \overline{I}(U, Y(i))$, where $Y(i) = \{Y^n(i)\}_{n=1}^{\infty}$. We use formula (2) to compute $\overline{I}(U, Y(i))$. We have seen that $\Pr\{Y^n(i) = y^n\} = P_{Y^n(i)}(y^n) = q^{-n}$ for any $y^n \in \mathcal{Y}^n$ and $V^n_{y^n|k^n} = \sum_{x^n} Q_i(x^n)\delta(y^n, x^n + k^n) = \sum_{x^n \in S_i} \mid S_i \mid^{-1} \delta(y^n, x^n + k^n)$ for $k^n \in \mathcal{K}^n$, where $\mid S_i \mid = \tau_n M_n$, $\tau_n = \frac{1}{(n+3)}$, $\liminf_{n\to\infty} \frac{1}{n} \log M_n \geq C(W) - \gamma$. Then, the joint distribution of U^n and $Y^n(i)$

$$\Pr\{U^n = k^n,\ Y^n(i) = y^n\} = \begin{cases} q^{-n} \mid S_i \mid^{-1}, & \text{for } y^n \in S_i + k^n = \{x^n + k^n\} \\ & \text{for } x^n \in S_i \\ 0, & \text{else.} \end{cases}$$

Hence,

$$\frac{1}{n} \log \frac{V^n(Y^n(i)|U^n)}{P_{Y^n(i)}(Y^n(i))} = \frac{1}{n} \log \frac{|S_i|^{-1}}{q^{-n}} = \log q - \frac{1}{n} \log |S_i|$$

$$= \log q - \frac{1}{n} \log M_n + \frac{1}{n} \log(n + 3)$$

with probability one. Therefore, by formula (1):

$$\overline{I}(U; Y(i)) \leq \log q - C(W) + \gamma = -\sum_{z \in \mathcal{Z}} w(z|x) \log w(z|x) + \gamma.$$

Since γ is an arbitrarily small number, so $\overline{I}(U, Y(i)) = H(\{w(z|x); z \in \mathcal{Z}\})$, where $H(\cdot)$ is the entropy function. Then, we obtain $S_U(0|V) \leq H(\{w(z|x); z \in \mathcal{Z}\})$. By the Definition 191, there exists a sequence of transforms $K^n = f_n(U_{M_n})(1 \leq n < \infty)$ satisfying $\lim_{n\to\infty} d(Y^n(i), \widetilde{Y}^n(i)) = 0$ and $\liminf_{n\to\infty} \frac{1}{n} \log M_n \leq H(\{w(z|x); z \in \mathcal{Z}\}) + \gamma$, where $Y^n(i)$ and $\widetilde{Y}^n(i)$ denote the outputs of channel V corresponding to the inputs U^n and K^n respectively.
In other words, there exists a sequence of (n, R) ALIB encipherers C with $R \leq H(\{w(z|x); z \in \mathcal{Z}\}) + \gamma$, such that if the random key K^n generates the key k^n according to the uniform distribution on C, then the random cryptogram $\widetilde{Y}^n(i) = X^n(i) + K^n$ satisfies $\lim_{n\to\infty} d(Y^n(i), \widetilde{Y}^n(i)) = 0$.

In the following, in order to avoid confusion, the PDs of $Y^n(i)$ and $\widetilde{Y}^n(i)$ are denoted by $Q_{Y^n(i)}$ and $\widetilde{Q}_i$, respectively, denote $\widetilde{Z}^n(i)$ the output of the channel W^n corresponding to the input $\widetilde{Y}^n(i)$. Now, we prove that the PD of $\widetilde{Z}^n(i)$, $\widetilde{Q}_i W^n (i = 1, 2, \cdots, N_n)$ satisfies the Security Condition. In fact, $Q_{Y^n(i)} W^n$ is the uniform distribution on $\mathcal{Z}^n$ and $Q_{Y^n(i)} W^n(D) + Q_{Y^n(i)} W^n(D^c) = 1$ for any $D \subset \mathcal{Z}^n$. On the other hand,

$$\begin{aligned} d(Q_{Y^n(i)} W^n, \widetilde{Q}_i W^n) &= \sum_{z^n \in \mathcal{Z}^n} |Q_{Y^n(i)} W^n(z^n) - \widetilde{Q}_i W^n(z^n)| \\ &\leq \sum_{z^n \in \mathcal{Z}^n} \sum_{y^n \in \mathcal{Y}^n} |Q_{Y^n(i)}(y^n) - \widetilde{Q}_i(y^n)| W^n_{z^n|y^n} \\ &= d(Q_{Y^n(i)}, \widetilde{Q}_i). \end{aligned}$$

Consequently, $\lim_{n\to\infty} d(Q_{Y^n(i)} W^n, \widetilde{Q}_i W^n) = 0$. Evidently, for any $i (1 \leq i \leq N_n)$,

$$| Q_{Y^n(i)} W^n(D^c) - \widetilde{Q}_i W^n(D^c) | \leq d(Q_{Y^n(i)} W^n, \widetilde{Q}_i W^n),$$

then,

$$\widetilde{Q}_i W^n(D^c) \geq Q_{Y^n(i)} W^n(D^c) - d(Q_{Y^n(i)} W^n, \widetilde{Q}_i W^n).$$

Similarly, for any $j (j \neq i)$,

$$Q_j W^n(D) \geq Q_{Y^n(j)} W^n(D) - d(Q_{Y^n(j)} W^n, \widetilde{Q}_j W^n).$$

Combine these two inequalities and set

$$\delta_n = 2[d(Q_{Y^n(i)} W^n, \widetilde{Q}_i W^n) + d(Q_{Y^n(j)} W^n, \widetilde{Q}_j W^n)].$$

We obtain $\widetilde{Q}_i W^n(D^c) + \widetilde{Q}_j W^n(D) > 1 - \delta_n$ and $\lim_{n\to\infty} \delta_n = 0$. Our proof is complete. □

References

1. R. Ahlswede, G. Dueck, Bad codes are good ciphers. Prob. Cont. Inf. Theory **11**, 337–351 (1982)
2. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**, 15–29 (1989)
3. 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. Theory **41**(4), 1040–1050 (1995)
4. T.S. Han, *Information-Spectrum Methods in Information Theory* (Springer, Berlin, 2003)

Part IV
Identification for Sources, Identification Entropy, and Hypothesis Testing

Identification for Sources

1 Introduction

1.1 Pioneering Model

The classical transmission problem deals with the question how many possible messages can we transmit over a noisy channel? Transmission means there is an answer to the question "What is the actual message?" In the identification problem we deal with the question how many possible messages the receiver of a noisy channel can identify? Identification means there is an answer to the question "Is the actual message u?" Here u can be any member of the set of possible messages.

Allowing randomized encoding the optimal code size grows double exponentially in the blocklength and somewhat surprisingly the second order capacity equals Shannon's first order transmission capacity (see [3], chapter "Identification via Channels" in Part I).

Thus Shannon's Channel Coding Theorem for Transmission is paralleled by a Channel Coding Theorem for Identification. It seems natural to look for such a parallel for sources, in particular for noiseless coding. This was suggested by Ahlswede in [1].

Let $(\mathcal{U}, P)$ be a source, where $\mathcal{U} = \{1, 2, \ldots, N\}$, $P = (P_1, \ldots, P_N)$, and let $\mathcal{C} = \{c_1, \ldots, c_N\}$ be a binary prefix code (PC) for this source with $\|c_u\|$ as length of c_u. Introduce the RV U with $\Pr(U = u) = p_u$ for $u = 1, 2, \ldots, N$ and the RV C with $C = c_u = (c_{u_1}, c_{u_2}, \ldots, c_{u\|c_u\|})$ if $U = u$.

We use the PC for noiseless identification, that is user u wants to know whether the source output equals u, that is, whether C equals c_u or not. He iteratively checks whether $C = (C_1, C_2, \ldots)$ coincides with c_u in the first, second, etc. letter and stops when the first different letter occurs or when $C = c_u$.

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_19

What is the expected number $L_{\mathcal{C}}(P,u)$ of checkings?

In order to calculate this quantity we introduce for the binary tree $T_{\mathcal{C}}$, whose leaves are the codewords $c_1,\ldots,c_N$, the sets of leaves $\mathcal{C}_{ik}(1 \le i \le N; 1 \le k)$, where $\mathcal{C}_{ik} = \{c \in \mathcal{C} : c$ coincides with c_i exactly until the k'th letter of $c_i\}$. If C takes a value in $\mathcal{C}_{uk}, 0 \le k \le \|c_u\| - 1$, the answers are k times "Yes" and 1 time "No". For $C = c_u$ the answers are $\|c_u\|$ times "Yes". Thus[1]

$$L_{\mathcal{C}}(P,u) = \sum_{k=0}^{\|c_u\|-1} P(C \in \mathcal{C}_{uk})(k+1) + \|c_u\| P_u.$$

For code $\mathcal{C}$ $L_{\mathcal{C}}(P) = \max_{1\le u\le N} L_{\mathcal{C}}(P,u)$ is the expected number of checkings in the worst case and $L(P) = \min_{\mathcal{C}} L_{\mathcal{C}}(P)$ is this number for a best code.

Analogously, if $\tilde{\mathcal{C}}$ is a randomized coding, $L_{\tilde{\mathcal{C}}}(P,u)$, $L_{\tilde{\mathcal{C}}}(P)$ and $\tilde{L}(P)$ were also introduced in [1].

What are the properties of $L(P)$ and $\tilde{L}(P)$? In analogy to the role of entropy $H(P)$ in Shannon's Noiseless Source Coding Theorem they can be viewed as approximations to a kind of "identification entropy" functional H_I.

Their investigation is left to future research. We quickly report now two simpler pioneering questions and partial answers from [1]. They shed some light on the idea that in contrast to classical entropy H, which takes values between 0 and ∞, the right functional H_I shall have 2 as maximal value.

Let us start with $P_N = \left(\frac{1}{N},\ldots,\frac{1}{N}\right)$ and set $f(N) = L(P_N)$.

1. What is $\sup_N f(N)$ or $\lim_{N\to\infty} f(N)$?

   Starting with an identification code for $N = 2^{k-1}$ a new one for 2^k users is constructed by adding for half of all users a 1 as prefix to the codewords and a 0 for the other half. Obviously we are getting an identification code with twice as many codewords in this way. Now user u has to read the first bit. With probability $\frac{1}{2}$ he then stops and with probability $\frac{1}{2}$ he needs only an expected number of $f(2^{k-1})$ many further checkings. Now an optimal identification code is at least as good as the constructed one and we get the recursion

   $$f(2^k) \le 1 + \frac{1}{2} f(2^{k-1}),\ f(2) = 1$$

   and therefore

   $$f(2^k) \le 2 - 2^{-(k-1)}.$$

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

[1] Probability distributions and codes depend on N, but are mostly written without an index N.

2. Is $\tilde{L}(P) \leq 2$?
   This is the case under the stronger assumption that encoder and decoder have access to a random experiment with unlimited capacity of common randomness (see [2], chapter "The Role of Common Randomness in Information Theory and Cryptography: Secrecy Constraints" in Part III).
   For $P = (P_1, \ldots, P_N)$, $N \leq 2^n$ write $P^{(n)} = (P_1, \ldots, P_N, 0, \ldots, 0)$ with 2^n components. Use a binary regular tree of depth n with leaves $1, 2, \ldots, 2^n$ represented in binary expansions.
   The common random experiment with 2^n outcomes can be used to use 2^n cyclic permutations of $1, 2, \ldots, 2^n$ for 2^n deterministic codes. For each u we get equally often 0 and 1 in its representation and an expected word length $\leq 2 - \frac{1}{2^{n-1}} \leq 2$. The error probability is 0.

Remark Note that the *same* tree $T_{\mathcal{C}}$ can be used by *all* users in order to answer their question ("Is it me or not?").

1.1.1 Further Models and Definitions

The model of identification for sources described can be extended (as for channels in the spirit of [1]) to *generalized identification* (GI) as follows.

There is now a set of users $\mathcal{V}$ (not necessarily equal to $\mathcal{U}$), where user $v \in \mathcal{V}$ has a set $\mathcal{U}_v \subset \mathcal{U}$ of source outputs of his interest, that is, he wants to know whether the source output u is in $\mathcal{U}_v$ or not.

Furthermore we speak of *generalized identification with decoding* (GID), if user v not only finds out whether the output is in $\mathcal{U}_v$, but also identifies it if it is in $\mathcal{U}_v$.

Obviously the two models coincide if $|\mathcal{U}_v| = 1$ for $v \in \mathcal{V}$. Also, they specialize to the original model in *1.1*, if $\mathcal{V} = \mathcal{U}$ and $\mathcal{U}_v = \{v\}$ for $v \in \mathcal{U}$.

For our analysis we use the following definition. We denote by $D(x)$ the set of all proper prefixes of $x \in \{0, 1\}^*$, i.e.

$$D(x) \triangleq \{y \in \{0, 1\}^* : y \text{ is prefix of } x \text{ and } \|y\| < \|x\|\}. \tag{1}$$

e stands for the empty word in $\{0, 1\}^*$. For a set $A \subset \{0, 1\}^*$ we extend this notion to

$$D(A) \triangleq \bigcup_{x \in A} D(x). \tag{2}$$

$\{0, 1\}^*$ can be viewed as a binary, regular infinite tree with root e. A code $\mathcal{C}$ corresponds to the subtree $T_{\mathcal{C}}$ with root e and leaves $c_1, \ldots, c_N$.

In the sequel we use a specific example of a code for illustrations of concepts and ideas.

Example Let $\mathcal{C}$ be the set of all words of length 3. Notice that $D(010) = \{e, 0, 01\}$ and $D(\{001, 010\}) = \{e, 0, 00, 01\}$. ▲

The set $\mathcal{C}_v = \{c_u : u \in \mathcal{U}_v\}$ is a code for user v. For GID its codewords have to be uniquely decodable by user v in order to identify the source output. For this he uses the set of stop sequences

$$\mathcal{S}_v = \big\{y_1 \dots y_k : y_1 \dots y_{k-1} \in D(\mathcal{C}_v) \text{ and } y_1 \dots y_k \notin D(\mathcal{C}_v)\big\}. \tag{3}$$

By definition of D, $\mathcal{C}_v$ is contained in $\mathcal{S}_v$. We can also write

$$\mathcal{S}_v = \big\{xy : x \in \{0,1\}^*, y \in \{0,1\} \text{ with } x \in D(\mathcal{C}_v) \text{ and } xy \notin D(\mathcal{C}_v)\big\}. \tag{4}$$

(For $k = 1$, $y_1 \dots y_{k-1}$ describes the empty word e or the root of the code tree which is element of each set $D(\mathcal{C}_v)$.)

Example For the code of the previous example for $\mathcal{C}_v = \{010\}$ we have $\mathcal{S}_v = \{1, 00, 011, 010\}$ and for $\mathcal{C}_v = \{001, 010\}$ we have $\mathcal{S}_v = \{1, 000, 001, 010, 011\}$. ▲

With the families of sets of stop sequences $\mathcal{S}_v$ we derive first in Sect. 2 general lower bounds on the number of checkings for both models. In Sect. 3 we consider a uniform source and show that $\lim_{N\to\infty} f(N) = 2$. Then, in Sect. 4, we derive bounds on the maximal individual (average) identification length, which is introduced in item C of Sect. 2.

Finally, in Sect. 5, we introduce an *average identification* length for the case $\mathcal{V} = \mathcal{U}, \mathcal{U}_v = \{v\}$ for $v \in \mathcal{V}$ and derive asymptotic results.

2 A Probabilistic Tool for Generalized Identification

General supposition: We consider here prefix codes $\mathcal{C}$, which satisfy the Kraft inequality with equality, that is,

$$\sum_{u\in\mathcal{U}} 2^{-\|c_u\|} = 1. \tag{5}$$

We call them saturated, because they cannot be enlarged.

A. GID

For all $x \in \{0,1\}^*$ let

$$q_{\mathcal{C}}(P,x) = \begin{cases} 0, & \text{if } x \notin D(\mathcal{C}) \cup \mathcal{C} \\ P_u, & \text{if } x = c_u \\ q_{\mathcal{C}}(P,x0) + q_{\mathcal{C}}(P,x1), & \text{if } x \in D(\mathcal{C}). \end{cases}$$

The general supposition implies that for any set of stopping sequences $\mathcal{S}_v$ we have $\mathcal{S}_v \subset D(\mathcal{C}) \cup \mathcal{C}$ and the probability for user v to stop in $x \in \mathcal{S}_v$ equals

$q_c(P, x)$. After stopping in x user v has read $\|x\|$ many bits. Therefore the average identification length of user v is

$$L_{\mathcal{C}}(P, v) = \sum_{x \in \mathcal{S}_v} q_{\mathcal{C}}(P, x)\|x\|. \tag{6}$$

By definition of $q_{\mathcal{C}}$ we get

$$L_{\mathcal{C}}(P, v) = \sum_{x \in D(\mathcal{C}_v)} q_{\mathcal{C}}(P, x). \tag{7}$$

By construction $\mathcal{S}_v$ forms a prefix code. Each codeword has to be uniquely decoded by user v. Furthermore the probabilities $q_{\mathcal{C}}(P, x)$, $x \in \mathcal{S}_v$, define a probability distribution on $\mathcal{S}_v$ by

$$P_{\mathcal{C},v}(x) \triangleq q_{\mathcal{C}}(P, x) \text{ for all } x \in \mathcal{S}_v. \tag{8}$$

By the Noiseless Coding Theorem $L_{\mathcal{C}}(P, v)$ can be lower bounded by the entropy $H(P_{\mathcal{C},v})$. More directly, using the grouping axiom we get

$$H(P_{\mathcal{C},v}) = \sum_{x \in D(\mathcal{C}_v)} q_{\mathcal{C}}(P, x) h\left(\frac{q_{\mathcal{C}}(P, x1)}{q_{\mathcal{C}}(P, x)}\right), \tag{9}$$

where h is the binary entropy function, and thus

$$L_{\mathcal{C}}(P, v) - H(P_{\mathcal{C},v}) = \sum_{x \in D(\mathcal{C}_v)} q_{\mathcal{C}}(P, x)\left(1 - h\left(\frac{q_{\mathcal{C}}(P, x1)}{q_{\mathcal{C}}(P, x)}\right)\right). \tag{10}$$

Suppose $P_u > 0$ for all $1 \le u \le N$, then

$$q_{\mathcal{C}}(P, x) > 0 \text{ and with } \left(\frac{q_{\mathcal{C}}(P, x1)}{q_{\mathcal{C}}(P, x)}\right) \le 1 \text{ for all } x \in D(\mathcal{C})$$

it follows under the general supposition (2.1) for every user $v \in \mathcal{V}$ the average identification length satisfies

Theorem 194

$$L_{\mathcal{C}}(P, v) \ge H(P_{\mathcal{C},v}) \textit{ with "=" iff } \frac{q_{\mathcal{C}}(P, x1)}{q_{\mathcal{C}}(P, x)} = \frac{1}{2} \textit{ for all } x \in D(\mathcal{C}_v). \tag{11}$$

Since P is fixed we write now $L_{\mathcal{C}}(v)$ for $L_{\mathcal{C}}(P, v)$.

B. GI

Suppose we have a node x and a user v with the properties

(i) all codewords having x as prefix are all elements of $\mathcal{C}_v$ or
(ii) they are all not in $\mathcal{C}_v$.

In this case user v can stop in x and decide whether v occurred or not. By construction of the stop sequences $\mathcal{S}_v$ in (3) only case (i) can occur. Therefore we have to start the following algorithm to generate modified sets $\mathcal{S}_v$.

1. If $\mathcal{C}_v$ contains two codewords different only in the last position, say $x_1 \dots x_k 0$ and $x_1 \dots x_k 1$ then
   (i) remove these two codewords from $\mathcal{C}_v$ and insert $x_1 \dots x_k$. This new codeword has the probability $q_{\mathcal{C}}(P, x_1 \dots x_k)$.
   (ii) repeat step 1. Else continue with 2.
2. With the modified sets $\mathcal{C}_v$ construct the sets $\mathcal{S}_v$ as defined in (3).

The definition of $L_{\mathcal{C}}(P, v)$, $P_{\mathcal{C},v}$ and $H(P_{\mathcal{C},v})$ are as in (6), (8) and (9). Also the formulas (10) and (11) hold.

Example Let $\mathcal{C}_v = \{000, 001, 010\}$. After step 1 of the algorithm we get $\mathcal{C}_v = \{00, 010\}$. With step 2 we define $D(\mathcal{C}_v) = \{\emptyset, 0, 01\}$ and $\mathcal{S}_v = \{1, 00, 010, 011\}$. ▲

C. Maximal Individual (Expected) Identification Length L(P)

For a given probability distribution P and a given code $\mathcal{C}$ user v has uniquely to decode the codewords in $\mathcal{C}_v$.

Using (11) we can lower bound $L(P)$ as follows:

(i) Take the set of pairs $\mathcal{M} = \{(\mathcal{C}_v, v) : L(P) = L_{\mathcal{C}}(P, v)\}$.
(ii) Define

$$H_{\max}(P) = \max_{(\mathcal{C}_v, v) \in \mathcal{M}} H(P_{\mathcal{C},v}).$$

Then

$$L(P) \geq H_{\max}(P).$$

Remark Note that

1.

$$\sum_{x \in D(\mathcal{C})} q_{\mathcal{C}}(P, x) = \sum_{u=1}^{N} P_u \|c_u\|.$$

2. Using the grouping axiom it holds

$$\sum_{x\in D(\mathcal{C})} q_{\mathcal{C}}(P,x)h\left(\frac{q_{\mathcal{C}}(P,x1)}{q_{\mathcal{C}}(P,x)}\right) = H(P)$$

for all codes $\mathcal{C}$.

3. If for each code $\mathcal{C}$ there exists a set $\mathcal{C}_v$ (in case B after modification) such that $D(\mathcal{C}_v) = D(\mathcal{C})$, then $L(P) = \sum_{u=1}^{N} P_u \|c_u\|$ where the code $\mathcal{C}$ is the Huffman-code for the probability distribution P.

Example Suppose that $|\mathcal{V}| = \binom{N}{K}$, $K \geq \frac{N}{2}$, and $\{\mathcal{U}_v : v \in \mathcal{V}\} = \binom{[N]}{K}$.

1. In case A there exists for each code $\mathcal{C}$ a set $\mathcal{C}_v$ such that $D(\mathcal{C}_v) = D(\mathcal{C})$.
2. In case B with $K = \frac{N}{2}$ there exists for each code $\mathcal{C}$ a set $\mathcal{C}_v$ such that $D(\mathcal{C}_v) = D(\mathcal{C})$.
3. In case B if $K = N$ and thus $\mathcal{V} = \{v_1\}, \mathcal{U}_{v_1} = [N]$, then after modifying $\mathcal{C}_{v_1}$ the set $D(\mathcal{C}_{v_1})$ contains only the root of the tree which means the user v_1 has to read nothing from the received codeword (because he knows already the answer). ▲

Remark The example above is motivated by K-identification for channels!

3 The Uniform Distribution

Now we return to the original model of the first subsection of Sect. 1 with $\mathcal{V} = \mathcal{U}$ and $\mathcal{C}_v = \{c_v\}$ for each $v \in \mathcal{V}$. Let $P = (\frac{1}{N}, \ldots, \frac{1}{N})$. We construct a prefix code $\mathcal{C}$ in the following way. In each node (starting at the root) we split the number of remaining codewords in proportion as close as possible to $(\frac{1}{2}, \frac{1}{2})$.

1. Suppose $N = 2^k$. By construction our code $\mathcal{C}$ contains all binary sequences of length k. It follows that

$$q_{\mathcal{C}}(P,x) = \frac{1}{N}\frac{N}{2^{\|x\|}} = 2^{-\|x\|} \tag{12}$$

and by (7)

$$L_{\mathcal{C}}(P) = \sum_{x\in D(\mathcal{C}_v)} q_{\mathcal{C}}(P,x) = \sum_{i=0}^{k-1} 2^{-i} = 2 - 2^{-k+1} = 2 - \frac{2}{N}. \tag{13}$$

2. Suppose $2^{k-1} < N < 2^k$. By construction the remaining code contains only the codeword lengths $k-1$ and k.

By (7) we add the weights ($q_{\mathcal{C}}(P,x)$) of all nodes of a path from the root to a codeword (leave). Therefore in the worst case, N is odd and we have to add the larger weight.
At the root we split ($\frac{N-1}{2}$, $\frac{N-1}{2}+1$). Now we split again the larger one and in the worst case this number is again odd. It follows in general that

$$q_{\mathcal{C}}(P,x) \leq \frac{1}{N}\left(\frac{N-1}{2^{\|x\|}}+1\right). \tag{14}$$

Therefore

$$L_{\mathcal{C}}(P) \leq \sum_{i=0}^{k-1}\frac{1}{N}\left(\frac{N-1}{2^i}+1\right) = \sum_{i=0}^{k-1}2^{-i} - \frac{1}{N}\sum_{i=0}^{k-1}2^{-i} + \frac{1}{N}\sum_{i=0}^{k-1}1$$

$$= 2 - \frac{1}{N} - \frac{2}{N} + \frac{2}{N^2} + \frac{k}{N} = 2 + \frac{k-3}{N} + \frac{2}{N^2}. \tag{15}$$

With $k = \lceil log_2(N)\rceil$ it follows

Theorem 195 *For* $P = \left(\frac{1}{N},\ldots,\frac{1}{N}\right)$

$$\lim_{N\to\infty} L_{\mathcal{C}}(P) = 2 \tag{16}$$

4 Bounds on L(P) for General P = ($P_1, \ldots, P_N$)

4.1 An Upper Bound

We will now give an inductive construction for identification codes to derive an upper bound on $L(P)$. Let $P = (P_1,\ldots,P_N)$ be the probability distribution. W.l.o.g. we can assume that $P_i \geq P_j$ for all $i < j$. For $N = 2$ of course we assign 0 and 1 as codewords. Now let $N > 2$. We have to consider two cases:

1. $P_1 \geq 1/2$. In this case we assign 0 as codeword to message 1. We set $P_i'' = \frac{P_i}{\sum_{j=2}^{N} P_j}$ for $i = 2,\ldots,N$. By induction we can construct a code for the probability distribution $P'' = (P_2'',\ldots,P_N'')$ and messages 2 to N get the corresponding codewords for P'' but prefixed with a 1.
2. $P_1 < 1/2$. Choose ℓ such that $\delta_\ell = |\frac{1}{2} - \sum_{i=1}^{\ell} P_i|$ is minimal. Set $P_i' = \frac{P_i}{\sum_{j=1}^{\ell} P_j}$ for $i = 1,\ldots,\ell$ and $P_i'' = \frac{P_i}{\sum_{j=\ell+1}^{N} P_j}$ for $i = \ell+1,\ldots,N$. Analogous to the first case we construct codes for the distributions $P' = (P_1',\ldots,P_\ell')$ (called the *left side*) and $P'' = (P_{\ell+1}'',\ldots,P_N'')$ (called the *right side*). We get the code for

P by prefixing the codewords from the left side with 0 and the codewords from the right side with 1.

Trivially this procedure yields a prefix code.

Theorem 196 *Let $N \in \mathbb{N}$ and let $P = (P_1, \ldots, P_N)$. The previous construction yields a prefix code with $L(P) < 3$.*

Proof The case $N = 2$ is trivial. Now let $N \geq 3$.

Case 1. $P_1 \geq 1/2$: In this case we have $L(P) = 1 + \max\left\{P_1, L(P'')\sum_{i=2}^{N} P_i\right\}$, where $L(P'')$ denotes the corresponding maximal identification length for probability distribution P''. If the maximum is assumed for P_1 we have $L(P) \leq 2$, otherwise we get by induction $L(P) < 1 + 3 \cdot 1/2 < 3$.

Case 2. $P_1 < 1/2$ for $i = 1, \ldots, N$: In this case we have

$$L(P) = 1 + \max\left\{L(P') \cdot \sum_{i=1}^{\ell} P_i, \quad L(P'') \cdot \sum_{i=\ell+1}^{N} P_i\right\}.$$

Choose ℓ' such that $\sum_{i=1}^{\ell'} P_i \leq 1/2 < \sum_{i=1}^{\ell'+1} P_i$. Obviously either $\ell = \ell'$ or $\ell = \ell' + 1$.

Subcase. $\ell = \ell'$. Suppose the maximum is assumed on the left side. Then without changing the maximal identification length we can construct a new probability distribution $P''' = (P'''_1, \ldots, P'''_{\ell+1})$ by $P'''_1 = \sum_{i=\ell+1}^{N} P_i$ and $P'''_i = P_{i-1}$ for $2 \leq i \leq \ell + 1$. Since $P'''_1 \geq 1/2$ we are back in case 1. If the maximum is assumed on the right side then let $P'''_1 = \sum_{i=1}^{\ell} P_i$ and $P'''_i = P_{i+\ell-1}$ for all $2 \leq i \leq n - \ell + 1$. Notice that in this case $P'''_1 \geq 1/3$ (because $P'''_1 \geq 1/2 - P'''_2/2 \geq 1/2 - P'''_1/2$). Thus by induction $L(P''') \leq 1 + 3 \cdot 2/3 \leq 3$.

Subcase. $\ell = \ell' + 1$. If the maximum is on the right side we set $P'''_1 = \sum_{i=1}^{\ell} P_i \geq 1/2$, $P'''_i = P_{i+\ell-1}$ for $2 \leq i \leq n - \ell + 1$ and we are again back in case 1. Now suppose the maximum is taken on the left side. Since $\sum_{i=1}^{\ell} P_i - 1/2 \leq 1/2 - \sum_{i=1}^{\ell'} P_i$ it follows that $\delta_\ell \leq P_\ell/2$. Because $P_{\ell'} \leq (2\ell')^{-1}$ we have $\delta_\ell \leq (4\ell')^{-1} = (4(\ell - 1))^{-1}$. Also note that $\ell \geq 2$. The case $\ell = 2$ is again trivial. Now let $\ell > 2$. Then $L(P) < 3 \cdot (1/2 + \frac{1}{4(\ell-1)}) \leq 3 \cdot (1/2 + 1/8) < 3$. □

5 An Average Identification Length

We consider here the case where not only the source outputs but also the users occur at random. Thus in addition to the source $(\mathcal{U}, P)$ and RV U, we are given $(\mathcal{V}, Q)$, $\mathcal{V} \equiv \mathcal{U}$, with RV V independent of U and defined by $\Pr(V = v) = Q_v$ for $v \in \mathcal{V}$. The source encoder knows the value u of U, but not that of V, which chooses the user v with probability Q_v. Again let $\mathcal{C} = \{c_1, \ldots, c_N\}$ be a binary prefix code and let $L_{\mathcal{C}}(P, u)$ be the expected number of checkings on code $\mathcal{C}$ for user u. Instead of

$L_{\mathcal{C}}(P) = \max_{u\in\mathcal{U}} L_{\mathcal{C}}(P, u)$, the maximal number of expected checkings for a user, we consider now the average number of expected checkings

$$L_{\mathcal{C}}(P, Q) = \sum_{v\in\mathcal{V}} Q_v L_{\mathcal{C}}(P, v) \tag{17}$$

and the average number of expected checkings for a best code

$$L(P, Q) = \min_{\mathcal{C}} L_{\mathcal{C}}(P, Q). \tag{18}$$

(The models GI and GID can also be considered.)

We also call $L(P, Q)$ the average identification length. $L_{\mathcal{C}}(P, Q)$ can be calculated by the formula

$$L_{\mathcal{C}}(P, Q) = \sum_{x\in D(\mathcal{C})} q_{\mathcal{C}}(Q, x) q_{\mathcal{C}}(P, x). \tag{19}$$

In the same way as (19) we get the conditional entropy

$$H_{\mathcal{C}}(P\|Q) = \sum_{x\in D(\mathcal{C})} q_{\mathcal{C}}(Q, x) q_{\mathcal{C}}(P, x) h\left(\frac{q_{\mathcal{C}}(P, x1)}{q_{\mathcal{C}}(P, x)}\right). \tag{20}$$

5.1 Q *is the Uniform Distribution on* $\mathcal{V} = \mathcal{U}$

We begin with $|\mathcal{U}| = N = 2^k$, choose $\mathcal{C} = \{0, 1\}^k$ and note that

$$\sum_{x\in D(\mathcal{C})\|x\|=i} q_{\mathcal{C}}(P, x) = 1 \text{ for all } 0 \le i \le k. \tag{21}$$

By (12) for all $x \in \{0, 1\}^*$ with $\|x\| \le k$

$$q_{\mathcal{C}}(Q, x) = 2^{-\|x\|} \tag{22}$$

and thus by (19) and then by (21)

$$L_{\mathcal{C}}(P, Q) = \sum_{i=0}^{k-1} \sum_{\substack{x\in D(\mathcal{C})\\ \|x\|=i}} 2^{-i} q_{\mathcal{C}}(P, x) \tag{23}$$

$$= \sum_{i=0}^{k-1} 2^{-i} = 2 - 2^{-k+1} = 2 - \frac{2}{N}. \tag{24}$$

We continue with the case $2^{k-1} < N < 2^k$ and construct the code $\mathcal{C}$ again as in Sect. 3. By (14)

$$q_{\mathcal{C}}(Q, x) \leq \frac{1}{N}\left(\frac{N-1}{2^{\|x\|}} + 1\right). \tag{25}$$

Therefore

$$\begin{aligned} L_{\mathcal{C}}(P, Q) &= \sum_{x \in D(\mathcal{C})} q_{\mathcal{C}}(Q, x) q_{\mathcal{C}}(P, x) \leq \frac{1}{N} \sum_{x \in D(\mathcal{C})} (\frac{N-1}{2^{\|x\|}} + 1) q_{\mathcal{C}}(P, x) \\ &= \frac{1}{N} \sum_{i=0}^{k-1} (\frac{N-1}{2^i} + 1) \sum_{\substack{x \in D(\mathcal{C}) \\ \|x\| = i}} q_{\mathcal{C}}(P, x) \leq \frac{1}{N} \sum_{i=0}^{k-1} (\frac{N-1}{2^i} + 1) \cdot 1 \\ &= 2 + \frac{k-3}{N} + \frac{2}{N^2} \qquad \text{(see (15))}. \end{aligned} \tag{26}$$

With $k = \lceil log_2(N) \rceil$ it follows that

Theorem 197 *Let $N \in \mathbb{N}$ and $P = (P_1, \ldots, P_N)$, then for $Q = \left(\frac{1}{N}, \ldots, \frac{1}{N}\right)$*

$$\lim_{N \to \infty} L_{\mathcal{C}}(P, Q) = 2. \tag{27}$$

5.2 *The Example Above in Model GID with Average Identification Length for a Uniform* Q*

We get now

$$L_{\mathcal{C}}(P, Q) = \sum_{x \in D(\mathcal{C})} \frac{|\{v : x \in D(\mathcal{C}_v)\}|}{|\mathcal{V}|} q_{\mathcal{C}}(P, x) \tag{28}$$

and for the entropy in (20)

$$H_{\mathcal{C}}(P \| Q^*) = \sum_{x \in D(\mathcal{C})} \frac{|\{v : x \in D(\mathcal{C}_v)\}|}{|\mathcal{V}|} q_{\mathcal{C}}(P, x) h\left(\frac{q_{\mathcal{C}}(P, x1)}{q_{\mathcal{C}}(P, x)}\right). \tag{29}$$

Furthermore let $\mathcal{C}_0$ be the set of all codes $\mathcal{C}$ with $L_{\mathcal{C}}(P, Q^*) = L(P, Q^*)$. We define

$$H(P\|Q^*) = \max_{\mathcal{C}\in\mathcal{C}_0} H_{\mathcal{C}}(P\|Q^*). \tag{30}$$

Then

$$L(P, Q) \geq H(P\|Q^*). \tag{31}$$

Case $N = 2^n$. We choose $\mathcal{C} = \{0, 1\}^n$ and calculate $\frac{|\{v:x\in D(\mathcal{C}_v)\}|}{|\mathcal{V}|}$. Notice that for any $x \in D(\mathcal{C})$ we have $2^{n-\|x\|}$ many codewords with x as prefix.

Order this set. There are $\binom{N-1}{K-1}$ $(K-1)$-element subsets of $\mathcal{C}$ containing the first codeword in this set. Now we take the second codeword and $K-1$ others, but not the first. In this case we get $\binom{N-2}{K-1}$ further sets and so on.

Therefore

$$|\{v : x \in D(\mathcal{C}_v)\}| = \sum_{j=1}^{2^{n-\|x\|}} \binom{2^n - j}{K-1} \tag{32}$$

and (30) yields

$$\begin{aligned} L_{\mathcal{C}}(P, Q^*) &= \frac{1}{\binom{N}{K}} \sum_{x\in D(\mathcal{C})} \sum_{j=1}^{2^{n-\|x\|}} \binom{2^n - j}{K-1} q_{\mathcal{C}}(P, x) \\ &= \frac{1}{\binom{2^n}{K}} \sum_{i=0}^{n-1} \left(\sum_{j=1}^{2^{n-i}} \binom{2^n - j}{K-1} \right) \left(\sum_{\substack{x\in D(\mathcal{C}) \\ \|x\|=i}} q_{\mathcal{C}}(P, x) \right) \\ &= \frac{1}{\binom{2^n}{K}} \sum_{i=0}^{n-1} \left(\sum_{j=1}^{2^{n-i}} \binom{2^n - j}{K-1} \right) \qquad \text{(by (21))}. \end{aligned} \tag{33}$$

Lets abbreviate this quantity as $g(n, K)$. Its asymptotic behaviour remains to be analyzed.

The exact values are

$$g(n, 1) = 2 - \frac{2}{2^n}$$

$$g(n, 2) = \frac{2}{3} \frac{5 \cdot 2^{-n} - 9 + 4 \cdot 2^n}{2^n - 1}$$

$$g(n,3) = -\frac{2}{7}\frac{49\cdot 2^n - 70 + 32\cdot 2^{-n} - 11\cdot 4^n}{(2^n-1)(2^n-2)}$$

$$g(n,4) = \frac{4}{105}\frac{-2220 + 908\cdot 2^{-n} - 705\cdot 4^n + 1925\cdot 2^n + 92\cdot 8^n}{(2^n-1)(2^n-2)(2^n-3)}$$

We calculated the limits ($n \to \infty$)

K	1	2	3	4	5	6	7	8	9
$\lim_{n\to\infty} g(n,K)$	2	$\frac{8}{3}$	$\frac{22}{7}$	$\frac{368}{105}$	$\frac{2470}{651}$	$\frac{7880}{1953}$	$\frac{150266}{35433}$	$\frac{13315424}{3011805}$	$\frac{2350261538}{513010785}$

This indicates that $\sup_K \lim_{n\to\infty} g(n,K) = \infty$.

References

1. R. Ahlswede, General theory of information transfer, Preprint 97–118, SFB 343 "Diskrete Strukturen in der Mathematik", Universität Bielefeld, General Theory of Information Transfer and Combinatorics, Report on a Research Project at the ZIF (Center of interdisciplinary studies) in Bielefeld Oct. 1, 2002–August 31, 2003, edit R. Ahlswede with the assistance of L. Bäumer and N. Cai, also Special issue of Discrete Mathematics
2. 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. Inf. Theory **44**(1), 55–62 (1998)
3. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**(1), 15–29 (1989)

Identification Entropy

Shannon [5] has shown that a source $(\mathcal{U}, P, U)$ with output satisfying Prob $(U = u) = P_u$, can be encoded in a prefix code $\mathcal{C} = \{c_u : u \in \mathcal{U}\} \subset \{0, 1\}^*$ such that for the entropy

$$H(P) = \sum_{u \in \mathcal{U}} -p_u \log p_u \leq \sum p_u \text{ length}(c_u) \leq H(P) + 1.$$

We use a prefix code $\mathcal{C}$ for another purpose, namely noiseless identification, that is every user who wants to know whether a u $(u \in \mathcal{U})$ of his interest is the actual source output or not can consider the RV C with $C = c_u = (c_{u_1}, \ldots, c_{u\|c_u\|})$ and check whether $C = (C_1, C_2, \ldots)$ coincides with c_u in the first, second etc. letter and stop when the first different letter occurs or when $C = c_u$. Let $L_{\mathcal{C}}(P, u)$ be the expected number of checkings, if code $\mathcal{C}$ is used.

Our discovery is an identification entropy, namely the function

$$H_I(P) = 2\left(1 - \sum_{u \in \mathcal{U}} P_u^2\right).$$

We prove that $L_{\mathcal{C}}(P, P) = \sum_{u \in \mathcal{U}} P_u \, L_{\mathcal{C}}(P, u) \geq H_I(P)$ and thus also that

$$L(P) = \min_{\mathcal{C}} \max_{u \in \mathcal{U}} L_{\mathcal{C}}(P, u) \geq H_I(P)$$

and related upper bounds, which demonstrate the operational significance of identification entropy in noiseless source coding similar as Shannon entropy does in noiseless data compression.

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_20

Also other averages such as $L_{\mathcal{C}}(P) = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} L_{\mathcal{C}}(P,u)$ are discussed in particular for Huffman codes where classically equivalent Huffman codes may now be different.

We also show that prefix codes, where the codewords correspond to the leaves in a regular binary tree, are universally good for this average.

1 Introduction

Shannon's Channel Coding Theorem for Transmission [5] is paralleled by a Channel Coding Theorem for Identification [3] (see Lecture 1 in Part I). In [1] we introduced noiseless source coding for identification and suggested the study of several performance measures.

Interesting observations were made already for uniform sources $P^N = \left(\frac{1}{N},\ldots,\frac{1}{N}\right)$, for which the worst case expected number of checkings $L(P^N)$ is approximately 2. Actually in [4] (see Lecture 18) it is shown that $\lim_{N\to\infty} L(P^N) = 2$.

Recall that in channel coding going from transmission to identification leads from an exponentially growing number of manageable messages to double exponentially many. Now in source coding roughly speaking the range of average code lengths for data compression is the interval $[0,\infty)$ and it is $[0,2)$ for an average expected length of optimal identification procedures. Note that no randomization has to be used here.

A discovery of the paper presented in this lecture [2] is an identification entropy, namely the functional

$$H_I(P) = 2\left(1-\sum_{u=1}^{N} P_u^2\right) \tag{1}$$

for the source $(\mathcal{U},P)$, where $\mathcal{U} = \{1,2,\ldots,N\}$ and $P = (P_1,\ldots,P_N)$ is a probability distribution.

Its operational significance in identification source coding is similar to that of classical entropy $H(P)$ in noiseless coding of data: it serves as a good lower bound.

Beyond being continuous in P it has three basic properties.

I. Concavity

For $p = (p_1,\ldots,p_N), q = (q_1,\ldots,q_N)$ and $0 \le \alpha \le 1$

$$H_I(\alpha p + (1-\alpha)q) \ge \alpha H_I(p) + (1-\alpha)H_I(q).$$

This is equivalent with

$$\sum_{i=1}^{N}(\alpha p_i+(1-\alpha)q_i)^2 = \sum_{i=1}^{N}\alpha^2 p_i^2+(1-\alpha)^2 q_i^2+\sum_{i\neq j}\alpha(1-\alpha)p_i q_j \leq \sum_{i=1}^{N}\alpha p_i^2+(1-\alpha)q_i^2$$

or with

$$\alpha(1-\alpha)\sum_{i=1}^{N} p_i^2 + q_i^2 \geq \alpha(1-\alpha)\sum_{i\neq j} p_i q_j,$$

which holds, because $\sum\limits_{i=1}^{N}(p_i - q_i)^2 \geq 0$.

II. Symmetry

For a permutation $\Pi : \{1, 2, \ldots, N\} \to \{1, 2, \ldots, N\}$ and $\Pi P = (P_{1\Pi}, \ldots, P_{N\Pi})$

$$H_I(P) = H_I(\Pi P).$$

III. Grouping Identity

For a partition $(\mathcal{U}_1, \mathcal{U}_2)$ of $\mathcal{U} = \{1, 2, \ldots, N\}$, $Q_i = \sum_{u\in\mathcal{U}_i} P_u$ and $P_u^{(i)} = \frac{P_u}{Q_i}$ for $u \in \mathcal{U}_i$ $(i = 1, 2)$

$$H_I(P) = Q_1^2 H_I(P^{(1)}) + Q_2^2 H_I(P^{(2)}) + H_I(Q), \text{ where } Q = (Q_1, Q_2).$$

Indeed,

$$Q_1^2 2\left(1 - \sum_{j\in\mathcal{U}_1}\frac{P_j^2}{Q_1^2}\right) + Q_2^2 2\left(1 - \sum_{j\in\mathcal{U}_2}\frac{P_j^2}{Q_2^2}\right) + 2(1 - Q_1^2 - Q_2^2)$$

$$= 2Q_1^2 - 2\sum_{j\in\mathcal{U}_1} P_j^2 + 2Q_2^2 - 2\sum_{j\in\mathcal{U}_2} P_j^2 + 2 - 2Q_1^2 - 2Q_2^2$$

$$= 2\left(1 - \sum_{j=1}^{N} P_j^2\right).$$

Obviously, $0 \leq H_I(P)$ with equality exactly if $P_i = 1$ for some i and by concavity $H_I(P) \leq 2\left(1 - \frac{1}{N}\right)$ with equality for the uniform distribution.

Remark Another important property of $H_I(P)$ is Schur convexity.

2 Noiseless Identification for Sources and Basic Concept of Performance

For the source $(\mathcal{U}, P)$ let $\mathcal{C} = \{c_1, \ldots, c_N\}$ be a binary prefix code (PC) with $\|c_u\|$ as length of c_u. Introduce the RV U with $\text{Prob}(U = u) = P_u$ for $u \in \mathcal{U}$ and the RV C with $C = c_u = (c_{u1}, c_{u2}, \ldots, c_{u\|c_u\|})$ if $U = u$. We use the PC for noiseless identification, that is a user interested in u wants to know whether the source output equals u, that is, whether C equals c_u or not. He iteratively checks whether $C = (C_1, C_2, \ldots)$ coincides with c_u in the first, second etc. letter and stops when the first different letter occurs or when $C = c_u$. What is the expected number $L_{\mathcal{C}}(P, u)$ of checkings?

Related quantities are

$$L_{\mathcal{C}}(P) = \max_{1 \leq u \leq N} L_{\mathcal{C}}(P, u), \tag{2}$$

that is, the expected number of ckeckings for a person in the worst case, if code $\mathcal{C}$ is used,

$$L(P) = \min_{\mathcal{C}} L_{\mathcal{C}}(P), \tag{3}$$

the expected number of checkings in the worst case for the best code, and finally, if users are chosen by a RV V independent of U and defined by $\text{Prob}(V = v) = Q_v$ for $v \in \mathcal{V} = \mathcal{U}$, (see Lecture 18, 18.5) we consider

$$L_{\mathcal{C}}(P, Q) = \sum_{v \in \mathcal{U}} Q_v L_{\mathcal{C}}(P, v) \tag{4}$$

the average number of expected checkings, if code $\mathcal{C}$ is used, and also

$$L(P, Q) = \min_{\mathcal{C}} L_{\mathcal{C}}(P, Q) \tag{5}$$

the average number of expected checkings for a best code.

A natural special case is the mean number of expected checkings

$$\bar{L}_{\mathcal{C}}(P) = \sum_{u=1}^{N} \frac{1}{N} L_{\mathcal{C}}(P, u), \tag{6}$$

which equals $L_{\mathcal{C}}(P, Q)$ for $Q = \left(\frac{1}{N}, \ldots, \frac{1}{N}\right)$, and

$$\bar{L}(P) = \min_{\mathcal{C}} \bar{L}_{\mathcal{C}}(P). \tag{7}$$

Another special case of some "intuitive appeal" is the case $Q = P$. Here we write

$$L(P, P) = \min_{\mathcal{C}} L_{\mathcal{C}}(P, P). \tag{8}$$

It is known that Huffman codes minimize the expected code length for PC.

This is not the case for $L(P)$ and the other quantities in identification (see the last example of the next section). It was noticed already in [1, 4] that a construction of code trees balancing probabilities like in the Shannon-Fano code is often better. In fact Theorem 3 of [4] (Theorem 196 here) establishes that $L(P) < 3$ for every $P = (P_1, \ldots, P_N)$!

Still it is also interesting to see how well Huffman codes do with respect to identification, because of their classical optimality property. This can be put into the following

Problem Determine the region of simultaneously achievable pairs

$$(L_{\mathcal{C}}(P), \sum_{u} P_u \| c_u \|)$$

for (classical) transmission and identification coding, where the $\mathcal{C}$'s are PC. In particular, what are extremal pairs? We begin here with first observations.

3 Examples for Huffman Codes

We start with the uniform distribution

$$P^N = (P_1, \ldots, P_N) = \left(\frac{1}{N}, \ldots, \frac{1}{N}\right), 2^n \leq N < 2^{n+1}.$$

Then $2^{n+1} - N$ codewords have the length n and the other $2N - 2^{n+1}$ codewords have the length $n + 1$ in any Huffman code. We call the $N - 2^n$ nodes of length n of the code tree, which are extended up to the length $n + 1$ *extended nodes*.

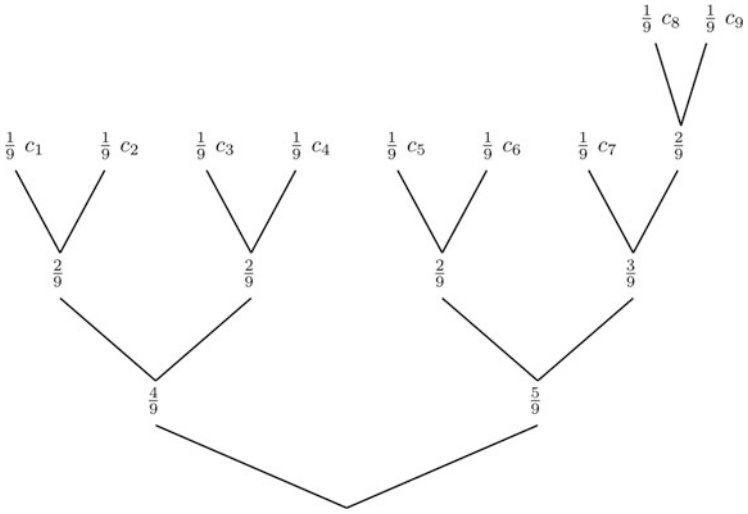
All Huffman codes for this uniform distribution differ only by the positions of the $N - 2^n$ extended nodes in the set of 2^n nodes of length n.

The average codeword length (for data compression) does not depend on the choice of the extended nodes.

However, the choice influences the performance criteria for identification!

Clearly there are $\binom{2^n}{N-2^n}$ Huffman codes for our source.

Example $N = 9$, $\mathcal{U} = \{1, 2, \ldots, 9\}$, $P_1 = \cdots = P_9 = \frac{1}{9}$. ▲



Here $L_{\mathcal{C}}(P) \approx 2.111$, $L_{\mathcal{C}}(P, P) \approx 1.815$ because

$$L_{\mathcal{C}}(P) = L_{\mathcal{C}}(c_8) = \frac{4}{9} \cdot 1 + \frac{2}{9} \cdot 2 + \frac{1}{9} \cdot 3 + \frac{2}{9} \cdot 4 = 2\frac{1}{9}$$

$$L_{\mathcal{C}}(c_9) = L_{\mathcal{C}}(c_8),\ L_{\mathcal{C}}(c_7) = 1\frac{8}{9},\ L_{\mathcal{C}}(c_5) = L_{\mathcal{C}}(c_6) = 1\frac{7}{9},$$

$$L_{\mathcal{C}}(c_1) = L_{\mathcal{C}}(c_2) = L_{\mathcal{C}}(c_3) = L_{\mathcal{C}}(c_4) = 1\frac{6}{9}$$

and therefore

$$L_{\mathcal{C}}(P, P) = \frac{1}{9}\left[1\frac{6}{9}\cdot 4 + 1\frac{7}{9}\cdot 2 + 1\frac{8}{9}\cdot 1 + 2\frac{1}{9}\cdot 2\right] = 1\frac{22}{27} = \bar{L}_{\mathcal{C}},$$

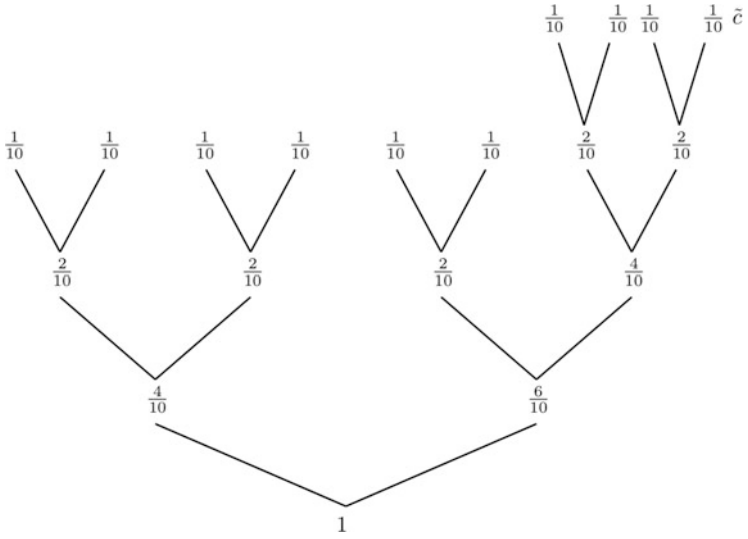
because P is uniform and the $\binom{2^3}{9-2^3} = 8$ Huffman codes are equivalent for identification.

Remark Notice that Shannon's data compression gives

$$H(P) + 1 = \log 9 + 1 > \sum_{u=1}^{9} P_u \|c_u\| = \frac{1}{9}3\cdot 7 + \frac{1}{9}4\cdot 2 = 3\frac{2}{9} \geq H(P) = \log 9.$$

Example $N = 10$. There are $\binom{2^3}{10-2^3} = 28$ Huffman codes.

The four worst Huffman codes are maximally unbalanced.

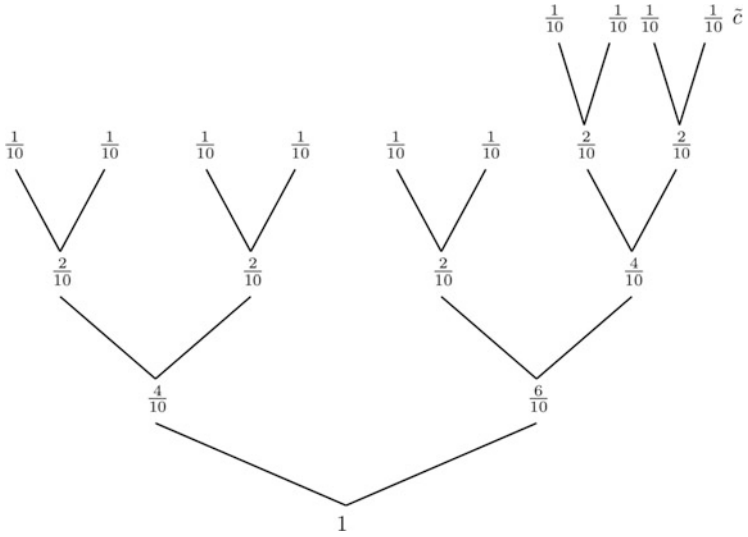


Here $L_{\mathcal{C}}(P) = 2.2$ and $L_{\mathcal{C}}(P, P) = 1.880$, because

$$L_{\mathcal{C}}(P) = 1 + 0.6 + 0.4 + 0.2 = 2.2$$

$$L_{\mathcal{C}}(P, P) = \frac{1}{10}[1.6\cdot 4 + 1.8\cdot 2 + 2.2\cdot 4] = 1.880.$$

One of the 16 best Huffman codes



Here $L_{\mathcal{C}}(P) = 2.0$ and $L_{\mathcal{C}}(P, P) = 1.840$ because

$$L_{\mathcal{C}}(P) = L_{\mathcal{C}}(\tilde{c}) = 1 + 0.5 + 0.3 + 0.2 = 2.000$$

$$L_{\mathcal{C}}(P, P) = \frac{1}{5}(1.7 \cdot 2 + 1.8 \cdot 1 + 2.0 \cdot 2) = 1.840$$

▲

The best identification performances of Huffman codes for the uniform distribution

N	8	9	10	11	12	13	14	15
$L_{\mathcal{C}}(P)$	1.750	2.111	2.000	2.000	1.917	2.000	1.929	1.933
$L_{\mathcal{C}}(P, P)$	1.750	1.815	1.840	1.860	1.861	1.876	1.878	1.880

Actually $\lim_{N\to\infty} L_{\mathcal{C}}(P^N) = 2$, but bad values occur for $N = 2^k + 1$ like $N = 9$ (see chapter "Identification for Sources").

One should prove that a best Huffman code for identification for the uniform distribution is best for the worst case and also for the mean.

However, for non-uniform sources generally Huffman codes are not best.

Example Let $N = 4$, $P(1) = 0.49$, $P(2) = 0.25$, $P(3) = 0.25$, $P(4) = 0.01$. Then for the Huffman code $\|c_1\| = 1$, $\|c_2\| = 2$, $\|c_3\| = \|c_4\| = 3$ and thus $L_{\mathcal{C}}(P) =$

$1 + 0.51 + 0.26 = 1.77$, $L_{\mathcal{C}}(P, P) = 0.49 \cdot 1 + 0.25 \cdot 1.51 + 0.26 \cdot 1.77 = 1.3277$, and $\bar{L}_{\mathcal{C}}(P) = \frac{1}{4}(1 + 1.51 + 2 \cdot 1.77) = 1.5125$.

However, if we use $\mathcal{C}' = \{00, 10, 11, 01\}$ for $\{1, \ldots, 4\}$ (4 is on the branch together with 1), then $L_{\mathcal{C}'}(P, u) = 1.5$ for $u = 1, 2, \ldots, 4$ and all three criteria give the same value 1.500 better than $L_{\mathcal{C}}(P) = 1.77$ and $\bar{L}_{\mathcal{C}}(P) = 1.5125$.

But notice that $L_{\mathcal{C}}(P, P) < L_{\mathcal{C}'}(P, P)$! ▲

4 An Identification Code Universally Good for all P on $\mathcal{U} = \{1, 2, \ldots, N\}$

Theorem 198 *Let $P = (P_1, \ldots, P_N)$ and let $k = \min\{\ell : 2^\ell \geq N\}$, then the regular binary tree of depth k defines a PC $\{c_1, \ldots, c_{2^k}\}$, where the codewords correspond to the leaves. To this code $\mathcal{C}_k$ corresponds the subcode $\mathcal{C}_N = \{c_i : c_i \in \mathcal{C}_k, 1 \leq i \leq N\}$ with*

$$2\left(1 - \frac{1}{N}\right) \leq 2\left(1 - \frac{1}{2^k}\right) \leq \bar{L}_{\mathcal{C}_N}(P) \leq 2\left(2 - \frac{1}{N}\right) \tag{9}$$

and equality holds for $N = 2^k$ on the left sides.

Proof By definition,

$$\bar{L}_{\mathcal{C}_N}(P) = \frac{1}{N} \sum_{u=1}^{N} L_{\mathcal{C}_N}(P, u) \tag{10}$$

and abbreviating $L_{\mathcal{C}_N}(P, u)$ as $L(u)$ for $u = 1, \ldots, N$ and setting $L(u) = 0$ for $u = N + 1, \ldots, 2^k$ we calculate with $P_u \triangleq 0$ for $u = N + 1, \ldots, 2^k$

$$\begin{aligned}
\sum_{u=1}^{2^k} L(u) = {} & \left[(P_1 + \cdots + P_{2^k})2^k\right] \\
& + \left[(P_1 + \cdots + P_{2^{k-1}})2^{k-1} + (P_{2^{k-1}+1} + \cdots + P_{2^k})2^{k-1}\right] \\
& + \left[(P_1 + \cdots + P_{2^{k-2}})2^{k-2} + (P_{2^{k-2}+1} + \cdots + P_{2^{k-1}})2^{k-2}\right. \\
& + (P_{2^{k-1}+1} + \cdots + P_{2^{k-1}+2^{k-2}})2^{k-2}
\end{aligned}$$

$$
\begin{aligned}
&+ (P_{2^{k-1}+2^{k-2}+1} + \cdots + P_{2^k})2^{k-2}\big] \\
&+ \quad \cdots \\
&\vdots \\
&+ \big[(P_1 + P_2)2 + (P_3 + P_4)2 + \cdots + (P_{2^k-1} + P_{2^k})2\big] \\
&= 2^k + 2^{k-1} + \cdots + 2 = 2(2^k - 1)
\end{aligned}
$$

and therefore

$$
\sum_{u=1}^{2^k} \frac{1}{2^k} L(u) = 2\left(1 - \frac{1}{2^k}\right). \tag{11}
$$

Now $2\left(1 - \frac{1}{N}\right) \le 2\left(1 - \frac{1}{2^k}\right) = \sum_{u=1}^{2^k} \frac{1}{2^k} L(u) \le \sum_{u=1}^{N} \frac{1}{N} L(u) = \frac{2^k}{N} \sum_{u=1}^{2^k} \frac{1}{2^k} L(u) = \frac{2^k}{N} 2\left(1 - \frac{1}{2^k}\right) \le 2\left(2 - \frac{1}{N}\right)$, which gives the result by (10).

Notice that for $N = 2^k$, a power of 2, by (11)

$$
\bar{L}_{\mathcal{C}_N}(P) = 2\left(1 - \frac{1}{N}\right). \tag{12}
$$

□

Remark 199 The upper bound in (9) is rough and can be improved significantly.

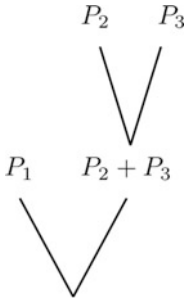
5 Identification Entropy $H_I(P)$ and Its Role as Lower Bound

Recall from the Introduction that

$$
H_I(P) = 2\left(1 - \sum_{u=1}^{N} P_u^2\right) \text{ for } P = (P_1 \dots P_N). \tag{13}
$$

We begin with a small source.

Example Let $N = 3$. W.l.o.g. an optimal code $\mathcal{C}$ has the structure ▲



Proposition 200

$$\bar{L}_{\mathcal{C}}(P) = \sum_{u=1}^{3} \frac{1}{3} L_{\mathcal{C}}(P, u) \geq 2\left(1 - \sum_{u=1}^{3} P_u^2\right) = H_I(P).$$

Proof Set $L(u) = L_{\mathcal{C}}(P, u)$. $\sum\limits_{u=1}^{3} L(u) = 3(P_1 + P_2 + P_3) + 2(P_2 + P_3)$.

This is smallest, if $P_1 \geq P_2 \geq P_3$ and thus $L(1) \leq L(2) = L(3)$. Therefore $\sum\limits_{u=1}^{3} P_u L(u) \leq \frac{1}{3} \sum\limits_{u=1}^{3} L(u)$. Clearly $L(1) = 1$, $L(2) = L(3) = 1 + P_2 + P_3$ and $\sum\limits_{u=1}^{3} P_u L(u) = P_1 + P_2 + P_3 + (P_2 + P_3)^2$.

This does not change if $P_2 + P_3$ is constant. So we can assume $P = P_2 = P_3$ and $1 - 2P = P_1$ and obtain

$$\sum_{u=1}^{3} P_u L(u) = 1 + 4P^2.$$

On the other hand

$$2\left(1 - \sum_{u=1}^{3} P_u^2\right) \leq 2\left(1 - P_1^2 - 2\left(\frac{P_2 + P_3}{2}\right)^2\right), \tag{14}$$

because $P_2^2 + P_3^2 \geq \frac{(P_2+P_3)^2}{2}$.

Therefore it suffices to show that

$$1 + 4P^2 \geq 2\big(1 - (1-2P)^2 - 2P^2\big)$$

$$= 2(4P - 4P^2 - 2P^2)$$

$$= 2(4P - 6P^2) = 8P - 12P^2$$

or that $1 + 16P^2 - 8P = (1-4P)^2 \geq 0$. □

We are now prepared for the first main result for $L(P, P)$.

Central in our derivations are proofs by induction based on *decomposition formulas for trees*.

Starting from the root a binary tree $\mathcal{T}$ goes via 0 to the subtree $\mathcal{T}_0$ and via 1 to the subtree $\mathcal{T}_1$ with sets of leaves $\mathcal{U}_0$ and $\mathcal{U}_1$, respectively. A code $\mathcal{C}$ for $(\mathcal{U}, P)$ can be viewed as a tree $\mathcal{T}$, where $\mathcal{U}_i$ corresponds to the set of codewords $\mathcal{C}_i, \mathcal{U}_0 \cup \mathcal{U}_1 = \mathcal{U}$.

The leaves are labeled so that $\mathcal{U}_0 = \{1, 2, \ldots, N_0\}$ and $\mathcal{U}_1 = \{N_0 + 1, \ldots, N_0 + N_1\}$, $N_0 + N_1 = N$. Using probabilities

$$Q_i = \sum_{u \in \mathcal{U}_i} P_u, \qquad i = 0, 1$$

we can give the decomposition in

Lemma 201 *For a code $\mathcal{C}$ for $(\mathcal{U}, P^N)$*

$$L_{\mathcal{C}}((P_1, \ldots, P_N), (P_1, \ldots, P_N))$$
$$= 1 + L_{\mathcal{C}_0}\left(\left(\frac{P_1}{Q_0}, \ldots, \frac{P_{N_0}}{Q_0}\right), \left(\frac{P_1}{Q_0}, \ldots, \frac{P_{N_0}}{Q_0}\right)\right) Q_0^2$$
$$+ L_{\mathcal{C}_1}\left(\left(\frac{P_{N_0+1}}{Q_1}, \ldots, \frac{P_{N_0+N_1}}{Q_1}\right), \left(\frac{P_{N_0+1}}{Q_1}, \ldots, \frac{P_{N_0+N_1}}{Q_1}\right)\right) Q_1^2.$$

This readily yields

Theorem 202 *For every source $(\mathcal{U}, P^N)$*

$$L(P^N) \geq L(P^N, P^N) \geq H_I(P^N).$$

Proof For $N = 2$ and any $\mathcal{C}$ $L_{\mathcal{C}}(P^2, P^2) \geq P_1 + P_2 = 1$, but

$$H_I(P^2) = 2(1 - P_1^2 - (1-P_1)^2) = 2(2P_1 - 2P_1^2) = 4P_1(1-P_1) \leq 1. \qquad (15)$$

This is the induction beginning.

For the induction step use for any code $\mathcal{C}$ the decomposition formula in Lemma 201 and of course the desired inequality for N_0 and N_1 as induction

hypothesis.

$$
\begin{aligned}
&L_{\mathcal{C}}((P_1, \ldots, P_N), (P_1, \ldots, P_N)) \\
&\geq 1 + 2\left(1 - \sum_{u \in \mathcal{U}_0} \left(\frac{P_u}{Q_0}\right)^2\right) Q_0^2 + 2\left(1 - \sum_{u \in \mathcal{U}_1} \left(\frac{P_u}{Q_1}\right)^2\right) Q_1^2 \\
&\geq H_I(Q) + Q_0^2 H_I(P^{(0)}) + Q_1^2 H_I(P^{(1)}) = H_I(P^N),
\end{aligned}
$$

where $Q = (Q_0, Q_1)$, $1 \geq H(Q)$, $P^{(i)} = \left(\frac{P_u}{Q_i}\right)_{u \in \mathcal{U}_i}$, and the grouping identity is used for the equality. This holds for every $\mathcal{C}$ and therefore also for $\min_{\mathcal{C}} L_{\mathcal{C}}(P^N)$. □

6 On Properties of $\bar{\mathbf{L}}(\mathbf{P^N})$

Clearly for $P^N = \left(\frac{1}{N}, \ldots, \frac{1}{N}\right)$ $\bar{L}(P^N) = L(P^N, P^N)$ and Theorem 202 gives therefore also the lower bound

$$
\bar{L}(P^N) \geq H_I(P^N) = 2\left(1 - \frac{1}{N}\right), \tag{16}
$$

which holds by Theorem 198 only for the Huffman code, but then for all distributions.

We shall see later that $H_I(P^N)$ is not a lower bound for general distributions P^N! Here we mean non-pathological cases, that is, not those where the inequality fails because $\bar{L}(P)$ (and also $L(P, P)$) is not continuous in P, but $H_I(P)$ is, like in the following case.

Example Let $N = 2^k + 1$, $P(1) = 1 - \varepsilon$, $P(u) = \frac{\varepsilon}{2^k}$ for $u \neq 1$, $P^{(\varepsilon)} = \left(1 - \varepsilon, \frac{\varepsilon}{2^k}, \ldots, \frac{\varepsilon}{2^k}\right)$, then

$$
\bar{L}(P^{(\varepsilon)}) = 1 + \varepsilon 2\left(1 - \frac{1}{2^k}\right) \tag{17}
$$

and $\lim_{\varepsilon \to 0} \bar{L}(P^{(\varepsilon)}) = 1$ whereas

$$
\lim_{\varepsilon \to 0} H_I(P^{(\varepsilon)}) = \lim_{\varepsilon \to 0} \left(2\left(1 - (1 - \varepsilon)^2 - \left(\frac{\varepsilon}{2^k}\right)^2 2^k\right)\right) = 0.
$$

▲

However, such a discontinuity occurs also in noiseless coding by Shannon.

The same discontinuity occurs for $L(P^{(\varepsilon)}, P^{(\varepsilon)})$.

Furthermore, for $N = 2 \quad P^{(\varepsilon)} = (1-\varepsilon, \varepsilon)$, $\bar{L}(P^{(\varepsilon)}) = 1 \quad L(P^{(\varepsilon)}, P^{(\varepsilon)}) = 1$ and $H_I(P^{(\varepsilon)}) = 2(1 - \varepsilon^2 - (1-\varepsilon)^2) = 0$ for $\varepsilon = 0$.

However, $\max\limits_{\varepsilon} H_I(P^{(\varepsilon)}) = \max\limits_{\varepsilon} 2(-2\varepsilon^2 + 2\varepsilon) = 1$ (for $\varepsilon = \frac{1}{2}$). Does this have any significance?

There is a second decomposition formula, which gives useful lower bounds on $\bar{L}_{\mathcal{C}}(P^N)$ for codes $\mathcal{C}$ with corresponding subcodes $\mathcal{C}_0, \mathcal{C}_1$ with uniform distributions.

Lemma 203 *For a code $\mathcal{C}$ for $(\mathcal{U}, P^N)$ and corresponding tree $\mathcal{T}$ let $T_{\mathcal{T}}(P^N) = \sum_{u \in \mathcal{U}} L(u)$. Then (in analogous notation)*

$$T_{\mathcal{T}}(P^N) = N_0 + N_1 + T_{\mathcal{T}_0}(P^{(0)})Q_0 + T_{\mathcal{T}_1}(P^{(1)})Q_1.$$

However, identification entropy is not a lower bound for $\bar{L}(P^N)$. We strive now for the worst deviation by using Lemma 203 and by starting with $\mathcal{C}$, whose parts $\mathcal{C}_0, \mathcal{C}_1$ satisfy the entropy inequality.

Then inductively

$$T_{\mathcal{T}}(P^N) \geq N + 2\left(1 - \sum_{u \in \mathcal{U}_0} \left(\frac{P_u}{Q_0}\right)^2\right) N_0 Q_0 + 2\left(1 - \sum_{u \in \mathcal{U}_1} \left(\frac{P_u}{Q_1}\right)^2\right) N_1 Q_1 \tag{18}$$

and

$$\frac{T_{\mathcal{T}}(P^N)}{N} \geq 1 + \sum_{i=0}^{1} 2\left(1 - \sum_{u \in \mathcal{U}_i} \left(\frac{P_u}{Q_i}\right)^2\right) \frac{N_i Q_i}{N} \triangleq A, \text{ say.}$$

We want to show that for

$$2\left(1 - \sum_{u \in \mathcal{U}} P_u^2\right) \triangleq B, \text{ say,}$$

$$A - B \geq 0. \tag{19}$$

We write

$$A - B = \left[-1 + 2\sum_{i=0}^{1} \frac{N_i Q_i}{N}\right] + 2\left[\sum_{u \in \mathcal{U}} P_u^2 - \sum_{i=0}^{1} \sum_{u \in \mathcal{U}_i} \left(\frac{P_u}{Q_i}\right)^2 \frac{N_i Q_i}{N}\right]$$

$$= C + D, \quad \text{say.} \tag{20}$$

C and D are functions of P^N and the partition $(\mathcal{U}_0, \mathcal{U}_1)$, which determine the Q_i's and N_i's. The minimum of this function can be analysed without reference to codes. Therefore we write here the partitions as $(\mathcal{U}_1, \mathcal{U}_2)$, $C = C(P^N, \mathcal{U}_1, \mathcal{U}_2)$ and $D = D(P^N, \mathcal{U}_1, \mathcal{U}_2)$. We want to show that

$$\min_{P^N, (\mathcal{U}_1, \mathcal{U}_2)} C(P^N, \mathcal{U}_1, \mathcal{U}_2) + D(P^N, \mathcal{U}_1, \mathcal{U}_2) \geq 0. \tag{21}$$

6.1 A First Idea

Recall that the proof of (15) used

$$2Q_0^2 + 2Q_1^2 - 1 \geq 0. \tag{22}$$

Now if $Q_i = \frac{N_i}{N}$ $(i = 0, 1)$, then by (22)

$$A - B = \left[-1 + 2 \sum_{i=0}^{1} \frac{N_i^2}{N^2} \right] + 2 \left[\sum_{u \in \mathcal{U}} P_u^2 - \sum_{u \in \mathcal{U}} P_u^2 \right] \geq 0.$$

A goal could be now to achieve $Q_i \sim \frac{N_i}{N}$ by rearrangement not increasing $A - B$, because in case of equality $Q_i = \frac{N_i}{N}$ that does it.

This leads to a nice *problem of balancing a partition* $(\mathcal{U}_1, \mathcal{U}_2)$ of $\mathcal{U}$. More precisely for $P^N = (P_1, \ldots, P_N)$

$$\varepsilon(P^N) = \min_{\phi \neq \mathcal{U}_1 \subset \mathcal{U}} \left| \sum_{u \in \mathcal{U}_1} P_u - \frac{|\mathcal{U}_1|}{N} \right|.$$

Then clearly for an optimal $\mathcal{U}_1$

$$Q_1 = \frac{|\mathcal{U}_1|}{N} \pm \varepsilon(P^N) \quad \text{and} \quad Q_2 = \frac{N - |\mathcal{U}_1|}{N} \mp \varepsilon(P^N).$$

Furthermore, one comes to a question of some independent interest. What is

$$\max_{P^N} \varepsilon(P^N) = \max_{P^N} \min_{\phi \neq \mathcal{U}_1 \subset \mathcal{U}} \left| \sum_{u \in \mathcal{U}_1} P_u - \frac{|\mathcal{U}_1|}{N} \right| ?$$

One can also go from sets $\mathcal{U}_1$ to distributions $\mathcal{R}$ on $\mathcal{U}$ and get, perhaps, a smoother problem in the spirit of game theory.

However, we follow another approach here.

6.2 A Rearrangement

We have seen that for $Q_i = \frac{N_i}{N}$ $D = 0$ and $C \geq 0$ by (22). Also, there is "air" up to 1 in C, if $\frac{N_i}{N}$ is away from $\frac{1}{2}$. Actually, we have

$$C = -\left(\frac{N_1}{N} + \frac{N_2}{N}\right)^2 + 2\left(\frac{N_1}{N}\right)^2 + 2\left(\frac{N_2}{N}\right)^2 = \left(\frac{N_1}{N} - \frac{N_2}{N}\right)^2. \tag{23}$$

Now if we choose for $N = 2m$ even $N_1 = N_2 = m$, then the air is out here, $C = 0$, but it should enter the second term D in (20).

Let us check this case first. Label the probabilities $P_1 \geq P_2 \geq \cdots \geq P_N$ and define $\mathcal{U}_1 = \left\{1, 2, \ldots, \frac{N}{2}\right\}, \mathcal{U}_2 = \left\{\frac{N}{2} + 1, \ldots, N\right\}$. Thus obviously

$$Q_1 = \sum_{u \in \mathcal{U}_1} P_u \geq Q_2 = \sum_{u \in \mathcal{U}_2} P_u$$

and

$$D = 2\left(\sum_{u \in \mathcal{U}} P_u^2 - \sum_{i=1}^{2} \frac{1}{2Q_i} \sum_{u \in \mathcal{U}_i} P_u^2\right).$$

Write $Q = Q_1, 1 - Q = Q_2$. We have to show

$$\sum_{u \in \mathcal{U}_1} P_u^2 \left(1 - \frac{1}{(2Q)^2}\right) \geq \sum_{u \in \mathcal{U}_2} P_u^2 \left(\frac{1}{(2Q_2)^2} - 1\right)$$

or

$$\sum_{u \in \mathcal{U}_1} P_u^2 \frac{(2Q)^2 - 1}{(2Q)^2} \geq \sum_{u \in \mathcal{U}_2} P_u^2 \left(\frac{1 - (2(1-Q))^2}{(2(1-Q))^2}\right). \tag{24}$$

At first we decrease the left hand side by replacing $P_1, \ldots, P_{\frac{N}{2}}$ all by $\frac{2Q}{N}$. This works because $\sum P_i^2$ is Schur-concave and $P_1 \geq \cdots \geq P_{\frac{N}{2}}$, $\frac{2Q}{N} = \frac{2(P_1 + \cdots + P_{\frac{N}{2}})}{N} \geq P_{\frac{N}{2}+1}$, because $\frac{2Q}{N} \geq P_{\frac{N}{2}} \geq P_{\frac{N}{2}+1}$. Thus it suffices to show that

$$\frac{N}{2}\left(\frac{2Q}{N}\right)^2 \frac{(2Q)^2 - 1}{(2Q)^2} \geq \sum_{u \in \mathcal{U}_2} P_u^2 \frac{1 - (2(1-Q))^2}{(2(1-Q))^2} \tag{25}$$

or that

$$\frac{1}{2N} \geq \sum_{u \in \mathcal{U}_2} P_u^2 \frac{1-(2(1-Q))^2}{(2(1-Q))^2((2Q)^2-1)}. \tag{26}$$

Secondly we increase now the right hand side by replacing $P_{\frac{N}{2}+1}, \ldots, P_N$ all by their maximal possible values $\left(\frac{2Q}{N}, \frac{2Q}{N}, \ldots, \frac{2Q}{N}, q\right) = (q_1, q_2, \ldots, q_t, q_{t+1})$, where $q_i = \frac{2Q}{N}$ for $i = 1, \ldots, t, q_{t+1} = q$ and $t \cdot \frac{2Q}{N} + q = 1 - Q, t = \left\lfloor \frac{(1-Q)N}{2Q} \right\rfloor$, $q < \frac{2Q}{N}$.

Thus it suffices to show that

$$\frac{1}{2N} \geq \left(\left\lfloor \frac{(1-Q)N}{2Q} \right\rfloor \cdot \left(\frac{2Q}{N}\right)^2 + q^2 \right) \frac{1-(2(1-Q))^2}{(2(1-Q))^2((2Q)^2-1)}. \tag{27}$$

Now we inspect the easier case $q = 0$. Thus we have $N = 2m$ and equal probabilities $P_i = \frac{1}{m+t}$ for $i = 1, \ldots, m + t = m$, say for which (27) goes wrong! We arrived at a very simple counterexample.

Example In fact, simply for $P_M^N = \left(\frac{1}{M}, \ldots, \frac{1}{M}, 0, 0, 0\right)$ we have $\lim_{N\to\infty} \bar{L}(P_M^N) = 0$, whereas

$$H_I(P_M^N) = 2\left(1 - \frac{1}{M}\right) \text{ for } N \geq M.$$

▲

Notice that here

$$\sup_{N,M} |\bar{L}(P_M^N) - H_I(P_M^N)| = 2. \tag{28}$$

This leads to the following problem solved in the next section.

Problem 204 Is $\sup_P |\bar{L}(P) - H_I(P)| = 2$?

7 Upper Bounds on $\bar{\mathbf{L}}(\mathbf{P^N})$

We know from Theorem 198 that

$$\bar{L}(P^{2^k}) \leq 2\left(1 - \frac{1}{2^k}\right) \tag{29}$$

and come to the following problem.

Problem 205 Is $\bar{L}(P^N) \leq 2\left(1-\frac{1}{2^k}\right)$ for $N \leq 2^k$?

This is the case, if the answer to the next question is positive.

Problem 206 Is $\bar{L}\left(\left(\frac{1}{N},\ldots,\frac{1}{N}\right)\right)$ monotone increasing in N?

In case the inequality in Problem 205 does not hold then it should with a very small deviation. Presently we have the following result, which together with (28) settles Problem 204.

Theorem 207 *For* $P^N = (P_1,\ldots,P_N)$

$$\bar{L}(P^N) \leq 2\left(1-\frac{1}{N^2}\right).$$

Proof (The Induction Beginning $\bar{L}(P^2) = 1 \leq 2\left(1-\frac{1}{4}\right)$ ***Holds)*** Define now $\mathcal{U}_1 = \left\{1,2,\ldots,\left\lfloor\frac{N}{2}\right\rfloor\right\}, \mathcal{U}_2 = \left\{\left\lfloor\frac{N}{2}\right\rfloor+1,\ldots,N\right\}$ and Q_1, Q_2 as before. Again by the decomposition formula of Lemma 2 and induction hypothesis

$$T(P^N) \leq N + 2\left(1-\frac{1}{\left\lfloor\frac{N}{2}\right\rfloor^2}\right) Q_1 \left\lfloor\frac{N}{2}\right\rfloor + 2\left(1-\frac{1}{\left\lceil\frac{N}{2}\right\rceil^2}\right) Q_2 \cdot \left\lceil\frac{N}{2}\right\rceil$$

and

$$\bar{L}(P^N) = \frac{1}{N}T(P^N) \leq 1 + \frac{2\left\lfloor\frac{N}{2}\right\rfloor Q_1 + 2\left\lceil\frac{N}{2}\right\rceil Q_2}{N} - \frac{2}{\left\lfloor\frac{N}{2}\right\rfloor}\cdot\frac{Q_1}{N} - \frac{2Q_2}{\left\lceil\frac{N}{2}\right\rceil N} \tag{30}$$

Case N Even $\bar{L}(P^N) \leq 1 + Q_1 + Q_2 - \left(\frac{4}{N^2}Q_1 + \frac{4}{N^2}Q_2\right) = 2 - \frac{4}{N^2} = 2\left(1-\frac{2}{N^2}\right) \leq 2\left(1-\frac{1}{N^2}\right)$

Case N Odd $\bar{L}(P^N) \leq 1 + \frac{N-1}{N}Q_1 + \frac{N+1}{N}Q_2 - 4\left(\frac{Q_1}{(N-1)N} + \frac{Q_2}{(N+1)N}\right) \leq 1 + 1 + \frac{Q_2-Q_1}{N} - \frac{4}{(N+1)N}$

Choosing the $\left\lceil\frac{N}{2}\right\rceil$ smallest probabilities in $\mathcal{U}_2$ (after proper labeling) we get for $N \geq 3$

$$\bar{L}(P^N) \leq 1+1+\frac{1}{N\cdot N} - \frac{4}{(N+1)N} = 2 + \frac{1-3N}{(N+1)N^2} \leq 2 - \frac{2}{N^2} = 2\left(1-\frac{1}{N^2}\right),$$

because $1-3N \leq -2N-2$ for $N \geq 3$. □

8 The Skeleton

Assume that all individual probabilities are powers of $\frac{1}{2}$

$$P_u = \frac{1}{2^{\ell_u}}, \quad u \in \mathcal{U}. \tag{31}$$

Define then $k = k(P^N) = \max_{u \in \mathcal{U}} \ell_u$.

Since $\sum_{u \in \mathcal{U}} \frac{1}{2^{\ell_u}} = 1$ by Kraft's theorem there is a PC with codeword lengths

$$\|c_u\| = \ell_u. \tag{32}$$

Notice that we can put the probability $\frac{1}{2^k}$ at all leaves in the binary regular tree and that therefore

$$L(u) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{2^3} 3 + \cdots + \frac{1}{2^t} t + \cdots + \frac{2}{2^{\ell_u}}. \tag{33}$$

For the calculation we use

Lemma 208 *Consider the polynomials* $G(x) = \sum_{t=1}^{r} t \cdot x^t + r x^r$ *and* $f(x) = \sum_{t=1}^{r} x^t$, *then*

$$G(x) = x\ f'(x) + r\ x^r = \frac{(r+1)x^{r+1}(x-1) - x^{r+2} + x}{(x-1)^2} + r\ x^r.$$

Proof Using the summation formula for a geometric series

$$f(x) = \frac{x^{r+1} - 1}{x - 1} - 1$$

$$f'(x) = \sum_{t=1}^{r} t\ x^{t-1} = \frac{(r+1)x^r(x-1) - x^{r+1} + 1}{(x-1)^2}.$$

This gives the formula for G. □

Therefore for $x = \frac{1}{2}$

$$G\left(\frac{1}{2}\right) = -(r+1)\left(\frac{1}{2}\right)^r - \left(\frac{1}{2}\right)^r + 2 + r\left(\frac{1}{2}\right)^r$$

$$= -\frac{1}{2^{r-1}} + 2$$

and since $L(u) = G\left(\frac{1}{2}\right)$ for $r = \ell_u$

$$L(u) = 2\left(1 - \frac{1}{2^{\ell_u}}\right) = 2\left(1 - \frac{1}{2^{\log \frac{1}{P_u}}}\right)$$

$$= 2(1 - P_u). \tag{34}$$

Therefore

$$L(P^N, P^N) \leq \sum_u P_u(2(1 - P_u)) = H_I(P^N) \tag{35}$$

and by Theorem 202

$$L(P^N, P^N) = H_I(P^N). \tag{36}$$

Theorem 209 *For* $P^N = (2^{-\ell_1}, \ldots, 2^{-\ell_N})$ *with 2-powers as probabilities*

$$L(P^N, P^N) = H_I(P).$$

This result shows that identification entropy is a right measure for identification source coding. For Shannon's data compression we get for this source $\sum_u p_u \|c_u\| = \sum_u p_u \ell_u = -\sum_u p_u \log p_u = H(P^N)$, again an identity.

For general sources the minimal average length deviates there from $H(P^N)$, but by not more than 1.

Presently we also have to accept some deviation from the identity.

We give now a first (crude) approximation. Let

$$2^{k-1} < N \leq 2^k \tag{37}$$

and that the probabilities are sums of powers of $\frac{1}{2}$ with exponents not exceeding k

$$P_u = \sum_{j=1}^{\alpha(u)} \frac{1}{2^{\ell_{uj}}}, \ell_{u1} \leq \ell_{u2} \leq \cdots \leq \ell_{u\alpha(u)} \leq k. \tag{38}$$

We now use the *idea of splitting object* $\boldsymbol{u}$ into *objects*

$$u1, \ldots, u\alpha(u). \tag{39}$$

Since

$$\sum_{u,j} \frac{1}{2^{\ell_{uj}}} = 1 \tag{40}$$

again we have a PC with codewords c_{uj} $(u \in \mathcal{U}, j = 1, \ldots, \alpha(u))$ and a regular tree of depth k with probabilities $\frac{1}{2^k}$ on all leaves.

Person u can find out whether u occured, he can do this (and more) by finding out whether $u1$ occured, then whether $u2$ occured, etc. until $u\alpha(u)$. Here

$$L(us) = 2\left(1 - \frac{1}{2^{\ell_{us}}}\right) \tag{41}$$

and

$$\sum_{u,s} L(us) P_{us} = 2\left(1 - \sum_{u,s} \frac{1}{2^{\ell_{us}}} \cdot \frac{1}{2^{\ell_{us}}}\right) = 2\left(1 - \sum_{u}\left(\sum_{s=1}^{\alpha(u)} P_{us}^2\right)\right). \tag{42}$$

On the other hand, being interested only in the original objects this is to be compared with $H_I(P^N) = 2\left(1 - \sum_u \left(\sum_s P_{us}\right)^2\right)$, which is smaller.

However, we get

$$\left(\sum_s P_{us}\right)^2 = \sum_s P_{us}^2 + \sum_{s \neq s'} P_{us} P_{us'} \leq 2 \sum_s P_{us}^2$$

and therefore

$$L(P^N, P^N) \leq 2\left(1 - \sum_u \left(\sum_{s=1}^{\alpha(u)} P_{us}^2\right)\right) \leq 2\left(1 - \frac{1}{2}\sum_u P_u^2\right). \tag{43}$$

For $P_u = \frac{1}{N} (u \in \mathcal{U})$ this gives the upper bound $2\left(1 - \frac{1}{2N}\right)$, which is better than the bound in Theorem 207 for uniform distributions.

9 Directions for Research

1. Study

$$L(P, R) \text{ for } P_1 \geq P_2 \geq \cdots \geq P_N \text{ and } R_1 \geq R_1 \geq \cdots \geq R_N.$$

2. Our results can be extended to q-ary alphabets, for which then identification entropy has the form

$$H_{I,q}(P) = \frac{q}{q-1}\left(1 - \sum_{i=1}^{N} P_i^2\right).$$

3. So far we have considered prefix-free codes. One also can study
   (a) fix-free codes
   (b) uniquely decipherable codes
4. Instead of the number of checkings one can consider other cost measures like the αth power of the number of checkings and look for corresponding entropy measures.
5. The analysis on universal coding can be refined.
6. In [4] (see Lecture 18) first steps were taken towards source coding for K-identification. This should be continued with a reflection on entropy and also towards GTIT.
7. **Grand ideas:** Other data structures
   (a) Identification source coding with *parallelism*: there are N identical code-trees, each person uses his own, but informs others
   (b) Identification source coding with *simultaneity*: $m (m = 1, 2, \ldots, N)$ persons use simultaneously the same tree.
8. We know that with probability at least $\lambda \in (0, 1)$ there is a subset $\mathcal{U}$ of cardinality $\exp\{H(P)/(1-\lambda)\}$.
   Is there such a result for $H_I(P)$?
   It is very remarkable that in our world of source coding the classical range of entropy $[0, \infty)$ is replaced by $[0, 2)$—singular, dual, plural—there is some appeal to this range.

References

1. R. Ahlswede, General theory of information transfer, Preprint 97–118, SFB 343 "Diskrete Strukturen in der Mathematik", Universität Bielefeld, General Theory of Information Transfer and Combinatorics, Report on a Research Project at the ZIF (Center of interdisciplinary studies) in Bielefeld Oct. 1, 2002–August 31, 2003, edit R. Ahlswede with the assistance of L. Bäumer and N. Cai, also Special issue of Discrete Mathematics
2. R. Ahlswede, Identification entropy, in *General Theory of Information Transfer and Combinatorics*, Lecture Notes in Computer Science, vol. 4123 (Springer, Berlin, 2006), pp. 595–613
3. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**(1), 15–29 (1989)

4. R. Ahlswede, B. Balkenhol, C. Kleinewächter, Identification for sources, in *General Theory of Information Transfer and Combinatorics*, ed. by R. Ahlswede et al. Lecture Notes in Computer Science, vol. 4123 (Springer, Berlin, 2006)
5. C.E. Shannon, A mathematical theory of communication, Bell Syst. Techn. J. **27**, 379–423, 623–656 (1948)

An Interpretation of Identification Entropy

After Ahlswede introduced identification for source coding he discovered identification entropy and demonstrated that it plays a role analogously to classical entropy in Shannon's noiseless source coding. We give now even more insight into this functional interpreting its two factors.

1 Introduction

1.1 Terminology

Identification in source coding started in [3]. Then identification entropy was discovered and its operational significance in noiseless source coding was demonstrated in [4] (see chapter "Identification Entropy").

Familiarity with that lecture is helpful, but not necessary here. As far as possible we also use its notation.

Differences come from the fact that we use now a q-ary coding alphabet $\mathcal{X} = \{0, 1, \ldots, q-1\}$, whereas earlier only the case $q = 2$ was considered and it was remarked only that all results generalize to arbitrary q. In particular the identification entropy, abbreviated as ID-entropy, for the source $(\mathcal{U}, P, U)$ has the form

$$H_{I,q}(P) = \frac{q}{q-1}\left(1 - \sum_{u\in\mathcal{U}} P_u^2\right). \tag{1}$$

Shannon (in 1948) has shown that a source $(\mathcal{U}, P, U)$ with output U satisfying $\Pr(U = u) = P_u$, can be encoded in a prefix code $\mathcal{C} = \{c_u : u \in \mathcal{U}\} \subset$

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_21

$\{0, 1, \ldots, q-1\}^*$ such that for the q-ary entropy

$$H_q(P) = \sum_{u \in \mathcal{U}} = -P_u \log_q P_u \leq \sum_{u \in \mathcal{U}} P_u ||c_u|| \leq H_q(P) + 1$$

where $||c_u||$ is the length of c_u.

We use a prefix code, abbreviated as PC, $\mathcal{C}$ for another purpose, namely, noiseless identification, that is, every user who wants to know whether a v, $v \in \mathcal{U}$ of his interest is the actual source output or cannot consider the RV C with $C = c_u = (c_{u_1}, \ldots, c_{u_{||c_u||}})$ if $U = u$ and check whether $C = (C_1, C_2, \ldots)$ coincides with c_v in the first, second, etc., letter and stop when the first different letter occurs or when $C = c_u$. Let $L_{\mathcal{C}}(P, u)$ be the expected number of checkings, if code $\mathcal{C}$ is used.

Related quantities are

$$L_{\mathcal{C}}(P) = \max_{v \in \mathcal{U}} L_{\mathcal{C}}(P, u) \tag{2}$$

that is, the expected number of checkings for a person in the worst case, if code $\mathcal{C}$ is used

$$L(P) = \min_{\mathcal{C}} L_{\mathcal{C}}(P), \tag{3}$$

the expected number of checkings in the worst case for a best code, and finally, if v's are chosen by a RV V independent of U and defined by $\Pr(V = v) = Q_v$ for $v \in \mathcal{V} = \mathcal{U}$, we consider

$$L_{\mathcal{C}}(P, Q) = \sum_{v \in \mathcal{U}} Q_v L_{\mathcal{C}}(P, v), \tag{4}$$

the average number of expected checkings, if code $\mathcal{C}$ is used, and also

$$L(P, Q) = \min_{\mathcal{C}} L_{\mathcal{C}}(P, Q) \tag{5}$$

the average number of expected checkings for a best code.

A natural special case is the mean number of expected checkings

$$\bar{L}_{\mathcal{C}}(P) = \sum_{u=1}^{N} \frac{1}{N} L_{\mathcal{C}}(P, u) \qquad \text{if } \mathcal{U} = [N] \tag{6}$$

which equals $L_{\mathcal{C}}(P, Q)$ for $Q = \left(\frac{1}{N}, \ldots, \frac{1}{N}\right)$, and

$$\bar{L}(P) = \min_{\mathcal{C}} \bar{L}_{\mathcal{C}}(P). \tag{7}$$

Another special case of some "intuitive appeal" is the case $Q = P$. Here we write

$$L(P, P) = \min_{\mathcal{C}} L_{\mathcal{C}}(P, P). \tag{8}$$

It is known that Huffman codes minimize the expected code length for a PC.
This is not always the case for $L(P)$ and the other quantities in identification.
In this lecture an important incentive comes from Theorem 4 of [1].

For $P^N = (2^{-l_1}, \ldots, 2^{-l_N})$, that is with 2-powers as probabilities $L(P^N, P^N) = H_l(P^N)$. Here the assumption means that there is a *complete* prefix code (i.e., equality holds in Kraft's equality).

1.2 A New Terminology Involving Proper Common Prefices

The quantity $L_{\mathcal{C}}(P, Q)$ is defined below also for the case of not necessarily independent U and V. It is conveniently described in a terminology involving proper common prefices

Definition 210 For an encoding $c : \mathcal{U} \to \mathcal{X}^*$, we define for two words $w, w' \in \mathcal{X}^*$ $cp(w, w')$ as the number of proper common prefices including the empty word, which equals the length of the maximal proper common prefix plus 1.

For example $cp(11\ 000) = 1$, $cp(0110\ 0100) = 3$, and $cp(1001\ 1001) = 4$ (since the proper common prefices are $\emptyset, 01, 100$).

Now with encoding c for PC $\mathcal{C}$ and RV's U and V, $cp(c_U, c_V)$ measures the time steps it takes to decide whether U and V are equal, that is, the checking time or waiting time, which we denote by

$$W_{\mathcal{C}}(U, V) = cp(c_U, c_V). \tag{9}$$

Clearly, we can write the expected waiting time as

$$\mathbb{E}W_{\mathcal{C}}(U, V) = \mathbb{E}cp(c_U, c_V) \tag{10}$$

It is readily verified that for independent U, V, that is, $\Pr(U = u, V = v) = P_u Q_v$

$$\mathbb{E}W_{\mathcal{C}}(U, V) = L_{\mathcal{C}}(P, Q) = \mathbb{E}cp(c_U, c_V). \tag{11}$$

We give now another description for $\mathbb{E}W_{\mathcal{C}}(U, V)$. For a word $w \in \mathcal{X}^*$ and a code $\mathcal{C}$ define as subset of $\mathcal{U}$

$$\mathcal{U}(\mathcal{C}, w) = \{u \in \mathcal{U} : c_u \quad \text{has proper prefix} \quad w\} \tag{12}$$

and its indicator function $1_{\mathcal{U}(\mathcal{C},w)}$. Now

$$\begin{aligned}\mathbb{E}\, cp(c_U, c_V) &= \sum_{u,v\in\mathcal{U}} \Pr(U=u, V=v) cp(c_u, c_v)\\ &= \sum_{u,v\in\mathcal{U}} \Pr(U=u, V=v) \sum_{w} 1_{\mathcal{U}(\mathcal{C},w)}(u) 1_{\mathcal{U}(\mathcal{C},w)}(v)\\ &= \sum_{w} \Pr\left(U\in\mathcal{U}(\mathcal{C},w), V\in\mathcal{U}(\mathcal{C},w)\right)\end{aligned}$$

and by (11)

$$\mathbb{E}\, W_{\mathcal{C}}(U,V) = \sum_{w} \Pr\left(U\in\mathcal{U}(\mathcal{C},w), w\in\mathcal{U}(\mathcal{C},w)\right). \tag{13}$$

1.3 Matrix Notation

Next we look at the double infinite matrix

$$\Lambda = \left(cp(w,w')\right)_{w\in\mathcal{X}^*, w'\in\mathcal{X}^*} \tag{14}$$

and its minor $\Lambda^{(L)}$ labeled by sequences in $\mathcal{X}^{\leq L}$.

Henceforth we assume that U and V are independent and have distributions P and Q. We can then use (11). For a prefix code $\mathcal{C}$ P induces the distribution $P_{\mathcal{C}}$ and Q induces the distribution $Q_{\mathcal{C}}$, when for $u, v \in \mathcal{U}$

$$P_{\mathcal{C}}(c_u) = P_u,\ Q_{\mathcal{C}}(c_v) = Q_v \tag{15}$$

and

$$P_{\mathcal{C}}(x) = Q_{\mathcal{C}}(x) = 0 \quad \text{for} \quad x\in\mathcal{X}^*\setminus\mathcal{C}. \tag{16}$$

Viewing both, $P_{\mathcal{C}}$ and $Q_{\mathcal{C}}$ as row vectors, for the corresponding column vector $Q_{\mathcal{C}}^T$ (11) can be written in the form

$$L_{\mathcal{C}}(P,Q) = P_{\mathcal{C}}\Lambda Q_{\mathcal{C}}^T. \tag{17}$$

It is clear from (10) that a non-complete prefix code, that is one for which the Kraft sum is smaller than 1, can be improved for identification by shortening a suitable

codeword. Hence *an optimal ID source code is necessarily complete*. In such a code

$$\max_{u\in\mathcal{U}} \|c_u\| \le |\mathcal{U}| - 1 \tag{18}$$

and one can replace Λ by its submatrix $\Lambda^{(L)}$ for $L = |\mathcal{U}| - 1$. This implies

$$L_{\mathcal{C}}(P, Q) = P_{\mathcal{C}}^{(L)} \Lambda^{(L)} (Q_{\mathcal{C}}^{(L)})^T, \tag{19}$$

where $P_{\mathcal{C}}^{(L)}$, and $Q_{\mathcal{C}}^{(L)}$ are row vectors obtained by deleting the components $y \notin \mathcal{X}^{\le L}$.

Sometimes the expressions (17) or (18) are more convenient for the investigation of $L_{\mathcal{C}}(P, Q)$. For example it is easy to see that Λ and therefore also $\Lambda^{(L)}$ are positive semidefinite. Indeed, let Δ (resp. $\Delta^{(L)}$) be a matrix whose rows are labeled by sequences in $\mathcal{X}^*$ (resp. $\mathcal{X}^{\le L}$) and whose columns are labeled by sequences in $\mathcal{X}^*$ (resp. $\mathcal{X}^{\le L-1} \cup \{\text{empty sequence}\}$) such that its (x, y)-entry is

$$\delta_y^*(x) = \begin{cases} 1 & \text{if } y \text{ is a proper prefix of } x \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\Delta\, \Delta^T = \Lambda \quad \text{and} \quad \Delta^{(L)} (\Delta^{(L)})^T = \Lambda^{(L)} \tag{20}$$

and hence Λ and $\Lambda^{(L)}$ are positive semidefinite. Therefore by (18) $L_{\mathcal{C}}(P, P)$ *is* $(\cup)$*-convex in* P.

Furthermore, for sources $(\mathcal{U}, P)$ with $|\mathcal{U}| = 2^k$ and for block code $\mathcal{C} = \{0, 1\}^k$, the uniform distribution on $\mathcal{U}$ achieves $\min_P L_{\mathcal{C}}(P, P)$.[1]

Another interesting observation on (19) is that as the wth component of $P_{\mathcal{C}}^{(L)} \Delta^{(L)}$ (resp. $Q_{\mathcal{C}}^{(L)} \Delta^{(L)}$) is $P\big(\mathcal{U}(\mathcal{C}, w)\big)$ (resp. $Q\big(\mathcal{U}(\mathcal{C}, w)\big)$), application of the Cauchy-Schwarz inequality to (19) yields

$$\big[P_{\mathcal{C}}^{(L)} \Lambda^{(L)} (Q_{\mathcal{C}}^{(L)})^T\big]^2 \le \big[P_{\mathcal{C}}^{(L)} \Delta^{(L)} (P_{\mathcal{C}}^{(L)})^T\big] \cdot \big[Q_{\mathcal{C}}^{(L)} \Delta^{(L)} (Q_{\mathcal{C}}^{(L)})^T\big] \tag{21}$$

and equality holds iff for all w

$$P\big(\mathcal{U}(\mathcal{C}, w)\big) = Q\big(\mathcal{U}(\mathcal{C}, w)\big).$$

We state this in equivalent form as follows.

[1] A proof is given in chapter "L-Identification for Sources".

Lemma 211

$$L_{\mathcal{C}}(P, Q)^2 \leq L_{\mathcal{C}}(P, P)L_{\mathcal{C}}(Q, Q) \tag{22}$$

and equality holds iff for all w

$$P\big(\mathcal{U}(\mathcal{C}, w)\big) = Q\big(\mathcal{U}(\mathcal{C}, w)\big),$$

which implies $L_{\mathcal{C}}(P, Q) = L_{\mathcal{C}}(P, P) = L_{\mathcal{C}}(Q, Q)$.

This suggests to introduce

$$\mu_{\mathcal{C}}(P, Q) = \frac{L_{\mathcal{C}}(P, Q)^2}{L_{\mathcal{C}}(P, P)L_{\mathcal{C}}(Q, Q)} \leq 1$$

as a measure of similarly of sources P and Q with respect to the code $\mathcal{C}$.

Intuitively we feel that for a good code for source P and Q as user distribution P and Q should be very dissimilar, because then the user waits less time until he knows that the output of U is not what he wants.

This idea will be used later for code construction. Actually it is clear even in the general case where U and V are not necessarily independent.

To simplify the discussion we assume here that the alphabet $\mathcal{X}$ is binary, i.e. $q = 2$.

Then the first bit of a codeword partitions the source $\mathcal{U}$ into two parts $\bar{\mathcal{U}}(i_1)$; $i_1 = 0, 1$; where $\bar{\mathcal{U}}(i_1) = \{u \in \mathcal{U} : c_{u1} = i_1\}$. By (13), to minimize $\mathbb{E}\, W_{\mathcal{C}}(U, V)$ one has to choose a partition such that $\Pr\big(U \in \bar{\mathcal{U}}(i_1), V \in \bar{\mathcal{U}}(i_1)\big)$'s are small simultaneously for $i_1 = 0, 1$. To construct a good code one can continue this line: partition $\bar{\mathcal{U}}(i_1)$ to $\bar{\mathcal{U}}(i_1, i_2)$'s such that

$$\Pr\big(U \in \bar{\mathcal{U}}(i_1, i_2), V \in \bar{\mathcal{U}}(i_1, i_2) \mid U \in \bar{\mathcal{U}}(i_1), V \in \bar{\mathcal{U}}(i_1)\big)\text{'s}$$

are as small as possible for $i_1, i_2 = 0, 1$ and so on.

When U and V are independent the requirement for a good code is that the difference between $P\big(\bar{\mathcal{U}}(i_1, \ldots, i_k)\big)$ and $Q\big(\bar{\mathcal{U}}(i_1, \ldots, i_k)\big)$ is large.

We call this the *Local Unbalance Principle* in contrast to the *Global Balance Principle* below.

Another extremal case is that U and V are equal with probability one and in this case one may never use the unbalance principle. However in this case the identification for the source makes no sense: The user knows that his output definitely comes! But still we can investigate the problem by assuming that with high probability $U = V$. More specifically, we consider the limit of $\mathbb{E}\, W_{\mathcal{C}}(U_k, V_k)$ for a sequence of RV's $(U_k, V_k)_{k=1}^{\infty}$ such that U_k converges to V_k in probability.

Then it follows from (11) that $\mathbb{E}\ W_{\mathcal{C}}(U_k, V_k)$ converges to the average length of codewords, the *classical object in source coding*! In this sense identification for sources is a generalization of source coding (data compression).

One of the discoveries of [4] is that ID-entropy is a lower bound to $L_{\mathcal{C}}(P, P)$. In (2) we repeat the original proof and we give in (3) another proof of this fact via two basic tools, Lemmas 214 and 215 for $L_{\mathcal{C}}(P^n, P^n)$, where P^n is the distribution of a memoryless source. It provides a clear information-theoretical meaning of the two factors $\frac{q}{q-1}$ and $\left(1 - \sum_{u \in \mathcal{U}} P_u^2\right)$ of ID-entropy. Next we consider in (4) sufficient and necessary conditions for a prefix code $\mathcal{C}$ to achieve the ID-entropy lower bound for $L_{\mathcal{C}}(P, P)$. *Quite surprisingly it turns out that the ID-entropy bound for ID-time is achieved by a variable length code iff the Shannon entropy bound for the average length of codewords is achieved by the same code* (Theorem 213).

Finally we end the lecture in (5) with a global balance principle to find good codes (Theorem 219).

2 An Operational Justification of ID-Entropy as Lower Bound for $\mathbf{L}_{\mathcal{C}}(\mathbf{P}, \mathbf{P})$

Recall from the introduction that for $q = 2$

$$H_I(P) = 2\left(1 - \sum_{u=1}^{N} P_u^2\right), \qquad \text{for } P = (P_1, \ldots, P_N).$$

We repeat the first main result for $L(P, P)$ from [4] (see chapter "Identification Entropy").

Central in our derivation is a proof by induction based on a *decomposition formula for trees*.

Starting from the root a binary tree $\mathbb{T}$ goes via 0 to the subtree $\mathbb{T}_0$ and via 1 to the subtree $\mathbb{T}_1$ with sets of leaves $\mathcal{U}_0$ and $\mathcal{U}_1$, respectively. A code $\mathcal{C}$ for $(\mathcal{U}, P)$ can be viewed as a tree $\mathbb{T}$, where $\mathcal{U}_i$ corresponds to the set of codewords $\mathcal{C}_i$, $\mathcal{U}_0 \cup \mathcal{U}_1 = \mathcal{U}$.

The leaves are labeled so that $\mathcal{U}_0 = \{1, 2, \ldots, N_0\}$ and $\mathcal{U}_1 = \{N_0 + 1, \ldots, N_0 + N_1\}$, $N_0 + N_1 = N$. Using probabilities

$$Q_i = \sum_{u \in \mathcal{U}_i} P_u, \qquad i = 0, 1$$

we can give the decomposition in the following

Lemma 212 *For a code $\mathcal{C}$ for $(\mathcal{U}, P^N)$*

$$\begin{aligned} &L_{\mathcal{C}}((P_1, \ldots, P_N), (P_1, \ldots, P_N)) \\ &= 1 + L_{\mathcal{C}_0}\left(\left(\frac{P_1}{Q_0}, \ldots, \frac{P_{N_0}}{Q_0}\right), \left(\frac{P_1}{Q_0}, \ldots, \frac{P_{N_0}}{Q_0}\right)\right) Q_0^2 \\ &\quad + L_{\mathcal{C}_1}\left(\left(\frac{P_{N_0+1}}{Q_1}, \ldots, \frac{P_{N_0+N_1}}{Q_1}\right), \left(\frac{P_{N_0+1}}{Q_1}, \ldots, \frac{P_{N_0+N_1}}{Q_1}\right)\right) Q_1^2. \end{aligned}$$

This readily yields the following result

Theorem 213 *For every source $(\mathcal{U}, P^N)$*

$$L(P^N) \geq L(P^N, P^N) \geq H_I(P^N).$$

Proof We proceed by induction on N. The base case $N = 2$ is established as follows. For $N = 2$ and any $\mathcal{C}$ $L_{\mathcal{C}}(P^2, P^2) \geq P_1 + P_2 = 1$, but

$$H_I(P^2) = 2(1 - P_1^2 - (1 - P_1)^2) = 2(2P_1 - 2P_1^2) = 4P_1(1 - P_1) \leq 1.$$

For the induction step, for any code $\mathcal{C}$ the decomposition formula in Lemma 212 and, of course, the desired inequality for N_0 and N_1 as induction hypothesis.

$$\begin{aligned} &L_{\mathcal{C}}((P_1, \ldots, P_N), (P_1, \ldots, P_N)) \\ &\geq 1 + 2\left(1 - \sum_{u \in \mathcal{U}_0} \left(\frac{P_u}{Q_0}\right)^2\right) Q_0^2 + 2\left(1 - \sum_{u \in \mathcal{U}_1} \left(\frac{P_u}{Q_1}\right)^2\right) Q_1^2 \\ &\geq H_I(Q) + Q_0^2 H_I(P^{(0)}) + Q_1^2 H_I(P^{(1)}) = H_I(P^N), \end{aligned}$$

where $Q = (Q_0, Q_1)$, $1 \geq H(Q)$, $P^{(i)} = \left(\frac{P_u}{Q_i}\right)_{u \in \mathcal{U}_i}$, and the grouping identity is used for the equality. This holds for every $\mathcal{C}$ and therefore also for $\min_{\mathcal{C}} L_{\mathcal{C}}(P^N)$. □

The approach readily extends also to the q-ary case.

3 An Alternative Proof of the ID-Entropy Lower Bound for $L_{\mathcal{C}}(P, P)$

First we establish Lemma 214 below, which holds for the more general case $\mathbb{E}\, W_{\mathcal{C}}(U, V)$. Let $\left((U^n, V^n)\right)_{n=1}^{\infty}$ be a discrete memoryless correlated source with generic pair of variables (U, V). Again U^n serves as (random) source and V^n serves

as random user. For a given code $\mathcal{C}$ for (U, V) let $\mathcal{C}^n$ be the code obtained by encoding the components of sequence $u^n \in \mathcal{U}^n$ iteratively. That is, for all $u^n \in \mathcal{U}^n$

$$c^n_{u^n} = (c_{u_1}, c_{u_2}, \ldots, c_{u_n}). \tag{23}$$

Lemma 214

$$\mathbb{E}\, W_{\mathcal{C}^n}(U^n, V^n) = \mathbb{E}\, W_{\mathcal{C}}(U, V) \left(1 + \sum_{t=1}^{n-1} \Pr(U^t = V^t)\right) \tag{24}$$

and therefore,

$$\lim_{n\to\infty} \mathbb{E}\, E_{\mathcal{C}^n}(U^n, V^n) = \frac{\mathbb{E}\, W_{\mathcal{C}}(U, V)}{1 - \Pr(U = V)}. \tag{25}$$

Proof Since

$$\Pr(U^n = V^n) = \prod_{t=1}^{n} \Pr(U_t = V_t) = \overset{n}{\Pr}(U = V)$$

Equation (25) follows from (24) immediately by the summation formula for geometric series.

To show (24) we define first for all $t \geq 2$ RV's

$$Z_t = \begin{cases} 0 & \text{if } U^{t-1} \neq V^{t-1} \\ 1 & \text{otherwise.} \end{cases} \tag{26}$$

and for $t = 1$ we let Z_1 be a constant for convenience of notation. Further we let W_t be the waiting time for the random user V^n in the t-th block.

Conditional on $Z_t = 1$ it is defined like $W_{\mathcal{C}}(U, V)$ in (9) and conditional on $Z_t = 0$ obviously $\Pr(W_t = 0 \mid Z_t = 0) = 1$, because the random user has made his decision before the t-th step. Moreover by the definition of $\mathcal{C}^n$

$$\mathbb{E}[W_t \mid Z_t = 1] = \mathbb{E}\, W_{\mathcal{C}}(U, V) \tag{27}$$

and consequently,

$$\mathbb{E}[\mathbb{E}(W_t \mid Z_t)] = \begin{cases} \Pr(U^{t-1} = V^{t-1}) \mathbb{E}\, W_{\mathcal{C}}(U, V) & \text{for } t = 2, 3, \ldots, n \\ \mathbb{E}\, W_{\mathcal{C}}(U, V) & \text{for } t = 1 \end{cases} \tag{28}$$

where (28) holds in case $t = 1$, because the random user has to wait for the first outcome. Therefore it follows that

$$\mathbb{E}\, W_{\mathcal{C}^n}(U^n, V^n) = \mathbb{E}\, W^n = \sum_{t=1}^{n} \mathbb{E}\, W_t = \sum_{t=1}^{n} \mathbb{E}[\mathbb{E}(W_t \mid Z_t)]$$

$$= \mathbb{E}\, W_{\mathcal{C}}(U, V) + \sum_{t=1}^{n-1} \Pr(U^t, V^t)\mathbb{E}\, W_{\mathcal{C}}(U, V)$$

as we wanted to show. □

Next we consider the case where U and V are i.i.d. with distribution P so that

$$\Pr(U^n = u^n, V^n = v^n) = \prod_{t=1}^{n} P_{u_t} \cdot P_{v_t}. \tag{29}$$

More specifically we are looking for a lower bound on $L_{\mathcal{C}}(P^n, P^n)$ for *all* prefix codes $\mathcal{C}$ over $\mathcal{U}^n$.

Lemma 215 *For all $\varepsilon > 0$ there exists an $\eta > 0$ such that for sufficiently large n and all positive integers*

$$L_n \leq n\big(H(P) - \varepsilon\big)(\log q)^{-1} \tag{30}$$

for all prefix codes $\mathcal{C}$ over $\mathcal{U}^n$

$$L_{\mathcal{C}}(P^n, P^n) > (1 - 2^{-n\eta}) \sum_{t=0}^{L_n - 1} q^{-t}. \tag{31}$$

Proof For given $\varepsilon > 0$ we choose $\delta > 0$ such that for a $\tau > 0$ and sufficiently large n for familiar typical sequences

$$P^n(\mathcal{T}^n_{P,\delta}) > 1 - 2^{-n\tau}$$

and for all $u^n \in \mathcal{T}^n_{P,\delta}$

$$P(u^n) < 2^{-n(H(P) - \frac{\varepsilon}{2})}.$$

Since for a prefix code $\mathcal{C}$

$$|\{u^n \in \mathcal{U}^n : \|c_{u^n}\| \le L_n\}| \le q^{L_n} \tag{32}$$

$$\begin{aligned}\Pr(\|c_{u^n}\| \le L_n) &= \Pr(\|c_{v^n}\| \le L_n)\\ &\le \Pr(V^n \notin \mathcal{T}^n_{P,\delta}) + \Pr(V^n \in \mathcal{T}^n_{P,\delta}, \|c_{v^n}\| \le L_n)\\ &< 2^{-n\tau} + |\{u^n : \|c_{u^n}\| < L_n\}| \cdot 2^{-n(H(P)-\frac{\varepsilon}{2})}\\ &\le 2^{-n\tau} + q^{L_n} 2^{-n(H(P)-\frac{\varepsilon}{2})}. \end{aligned} \tag{33}$$

However, (30) implies that

$$q^{L_n} \le 2^{n(H(P)-\varepsilon)}.$$

This together with (33) yields

$$\Pr(\|c_{U^n}\| \le L_n) < 2^{-n\tau} + 2^{-n\frac{\varepsilon}{2}} < 2^{-n\delta} \tag{34}$$

for $\delta \triangleq \min\left(\frac{\tau}{2}, \frac{\varepsilon}{4}\right)$.

Next, for the distribution P and the code $\mathcal{C}$ over $\mathcal{U}^n$ we construct a related source $(\tilde{U}, \tilde{P})$ and a code $\tilde{\mathcal{C}}$ over $\tilde{\mathcal{U}}$ as follows.

The new set $\tilde{\mathcal{U}}$ contains $\{u^n \in \mathcal{U}^n : \|c_{u^n}\| \le L_n\}$ and for its elements $\tilde{P}(u^n) = P^n(u^n)$ and the new $\sim$-coding is $\tilde{c}_{u^n} = c_{u^n}$.

Now we define the additional elements in $\tilde{\mathcal{U}}$ with its $\tilde{P}$ and $\tilde{c}$.

We partition $\{u^n \in \mathcal{U}^n : \|c_{u^n}\| > L_n\}$ into subsets $S_j (1 \le j \le J)$ according to the L_n-th prefix and use letter g_j to represent S_j and put the set $\tilde{\tilde{\mathcal{U}}} = \{g_j : 1 \le j \le J\}$ into $\tilde{\mathcal{U}}$ so that

$$\tilde{\mathcal{U}} = \{u^n \in \mathcal{U}^n : \|c_{u^n}\| \le L_n\} \cup \tilde{\tilde{\mathcal{U}}}.$$

Then we define $\tilde{P}(g_j) = \sum_{u^n \in S_j} P(u^n)$ and let $\tilde{c}_{g_j}$ be the common L_nth prefix of the c_{u^n}'s for the u^n's in S_j. That is, we consider all u^n sharing the same L_nth prefix in c_{u^n} as a single element. Obviously,

$$L_{\mathcal{C}}(P^n, P^n) \ge L_{\tilde{\mathcal{C}}}(\tilde{P}, \tilde{P}). \tag{35}$$

Finally let $\tilde{U}_n$ and $\tilde{V}_n$ be RV's for the new source and new random user with distribution $\tilde{P}$ and let Z be a RV such that

$$Z = \begin{cases} 0 & \text{if both } \|c_{U^n}\| \text{ and } \|c_{V_n}\| \text{ are larger than } L_n \\ 1 & \text{otherwise.} \end{cases}$$

Then

$$L_{\tilde{\mathcal{C}}}(\tilde{P}, \tilde{P}) = \mathbb{E}W = \mathbb{E}(W \mid Z) \geq \Pr(Z = 0)\mathbb{E}(W \mid Z = 0)$$

$$= \Pr(\|c_{U^n}\| \geq L_n) \Pr(\|c_{V^n}\| \geq L_n) \cdot L_{\tilde{\tilde{\mathcal{C}}}}(\tilde{\tilde{P}}, \tilde{\tilde{P}}) \tag{36}$$

where W is the random waiting time, $\tilde{\tilde{P}}$ is the common conditional distribution of $\tilde{U}_n$ given $\tilde{U}_n \in \tilde{\tilde{\mathcal{U}}}$, and $\tilde{V}_n$ given $\tilde{V}_n \in \tilde{\tilde{\mathcal{U}}}$, i.e.,

$$\tilde{\tilde{P}}(g_j) = \frac{\tilde{P}(g)}{\tilde{P}(\tilde{\tilde{\mathcal{U}}})} \qquad \text{for } g_j \in \tilde{\tilde{\mathcal{U}}}$$

and $\tilde{\tilde{\mathcal{C}}}$ is the restriction of $\tilde{\mathcal{C}}$ to $\tilde{\tilde{\mathcal{U}}}$.

Notice that $\tilde{\tilde{\mathcal{C}}}$ is a block code of length L_n. In order to bound $L_{\tilde{\tilde{\mathcal{C}}}}(\tilde{\tilde{P}}, \tilde{\tilde{P}})$ we extend $\tilde{\tilde{\mathcal{U}}}$ to a set of cardinality q^{L_n} in the case of necessity and assign zero probabilities and a codeword of length L_n not in $\tilde{\tilde{\mathcal{C}}}$. This little modification obviously does not change the value of $L_{\tilde{\tilde{\mathcal{C}}}}(\tilde{\tilde{P}}, \tilde{\tilde{P}})$. Thus, if we denote the uniform distribution over the extended set $\tilde{\tilde{\mathcal{U}}}$ by $\bar{P}$, we have

$$L_{\tilde{\tilde{\mathcal{C}}}}(\tilde{\tilde{P}}, \tilde{\tilde{P}}) \geq L_{\tilde{\tilde{\mathcal{C}}}}(\bar{P}, \bar{P}) \tag{37}$$

where $\tilde{\tilde{\mathcal{C}}}$ is a bijective block code $\tilde{\tilde{\mathcal{U}}} \to \mathcal{X}^{L_n}$.

It is clear that $\mathcal{U}(\tilde{\mathcal{C}}, \omega) \neq \emptyset$ iff the length of ω is smaller than $L_n - 1$ and

$$\mathcal{U}(\tilde{\mathcal{C}}, \omega) = \mathcal{X}^{L_* - 1}, \text{ if } \|\omega\| = \ell \leq L_n - 1.$$

Then it follows from (13) that

$$L_{\tilde{\tilde{\mathcal{C}}}}(\bar{P}, \bar{P}) = \sum_{t=0}^{L_n - 1} q^t [q^{L_n - t} \cdot q^{-L_n}]^2 = \sum_{t=0}^{L_n} q^{-t}. \tag{38}$$

Finally we combine (34)–(38) and Lemma 215 follows. □

An immediate consequence is the following corollary.

Corollary 216

$$\varliminf_{n \to \infty} L(P^n, P^n) \geq \sum_{t=0}^{\infty} q^{-t} = \frac{q}{q-1}. \tag{39}$$

Furthermore for i.i.d. RV's U, V with distribution P we have

$$\Pr(U = V) = \sum_{u \in \mathcal{U}} P_u^2$$

and from (25) and (39) follows the ID-entropy bound.

Corollary 217 *(See chapter "Identification Entropy")*

$$L_{\mathcal{C}}(P, P) \geq \frac{q}{q-1} \left(1 - \sum_{u \in \mathcal{U}} P_u^2\right). \tag{40}$$

This derivation provides a clear information-theoretical meaning to the two factors in ID-entropy: $\frac{q}{q-1}$ is a universal lower bound on the ID-waiting time for a discrete memoryless source with an independent user having the same distribution P. $\frac{1}{1-\sum\limits_{u \in \mathcal{U}} P_u^2}$ is the cost paid for coding the source component-wise and leaving time for the random user in the following sense.

Let us imagine the following procedure:

At a unit of time the random source U^n outputs a symbol U_t and the random user V^n, who wants to know whether $U^n = V^n$, checks whether U_t coincides with his own symbol V_t. He will end if not. Then the waiting time for him is ℓ with probability

$$\Pr(U^{\ell-1} = V^{\ell-1}) \Pr(U_\ell \neq V_\ell) = \Pr(U = V)^{\ell-1}\big(1 - \Pr(U = V)\big) \qquad \text{for } \ell \leq n.$$

Letting $n \to \infty$ we obtain a geometric distribution.

The expected waiting time is

$$\begin{aligned} \mathbb{E}W &= \sum_{\ell=0}^{\infty} \ell \Pr(U = V)^{\ell-1}\big(1 - \Pr(U = V)\big) \\ &= \sum_{\ell=0}^{\infty} (\ell + 1) \Pr(U = V)^{\ell} - \sum_{\ell=0}^{\infty} \Pr(U = V)^{\ell} \\ &= \sum_{\ell=0}^{\infty} \Pr(U = V)^{\ell} = \frac{1}{1 - \Pr(U = V)} \end{aligned} \tag{41}$$

which equals $\frac{1}{1-\sum\limits_{u} P_u^2}$ in the case of i.i.d. RV's.

(Actually (24) holds for all stationary sources and we choose a memoryless source for simplicity.) In general (25) has the form

$$\lim_{n\to\infty} \mathbb{E}W_{\mathcal{C}^n}(U^n, V^n) = \mathbb{E}W_{\mathcal{C}}(U, V) \cdot \lim_{n\to\infty} \left(1 + \sum_{t=1}^{n-1} \Pr(U^t = V^t)\right). \tag{42}$$

By monotonicity the limit at the RHS and therefore also at the LHS exists and equals a positive finite or infinite value.

When it is finite one may replace $\Pr(U = V)^{t-1}$, $\Pr(U = V)$ and $\Pr(U = V)^t$ in the first lines of (41) by $\Pr(U^{t-1} = V^{t-1})$, $\Pr(U_t = V_t \mid U^{t-1} = V^{t-1})$ and $\Pr(U^t = V^t)$, respectively, and obtain

$$\lim_{n\to\infty} \left(1 + \sum_{t=1}^{n-1} \Pr(U^t = V^t)\right)$$

$$= \sum_{t=0}^{\infty} t \Pr(U^{t-1} = V^{t-1}) \cdot \Pr(U_t \neq V_t \mid U^{t-1} = V^{t-1}) = \mathbb{E}L, \tag{43}$$

the expectation of random leaving time L for a stationary source.

Thus (42) is rewritten as

$$\lim_{n\to\infty} W_{\mathcal{C}^n}(U^n, V^n) = W_{\mathcal{C}}(U, V)\mathbb{E}L. \tag{44}$$

Now the information-theoretical meaning of (44) is quite clear. One encodes a source $(U^n, V^n)_{n=1}^{\infty}$ with alphabet $\mathcal{U}$ component by component by a variable length code $\mathcal{C}$. The first term at the right hand side of (44) is the expected waiting time in a block and the second term is the expected waiting time for different U_t and V_t.

4 Sufficient and Necessary Conditions for a Prefix Code $\mathcal{C}$ to Achieve the ID-Entropy Lower Bound of $L_{\mathcal{C}}(P, P)$

Quite surprisingly the ID-entropy bound to ID-waiting time is achieved by a variable length code iff the Shannon entropy bound to the average lengths of codewords is achieved by the same code.

For the proof we use a simple consequence of the Cauchy-Schwarz inequality, which states for two sequences of real numbers $(a_1, a_2, \ldots, a_k)$ and $(b_1, b_2, \ldots, b_k)$ that

$$\left(\sum_{i=1}^{k} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{k} a_i^2\right)\left(\sum_{i=1}^{k} b_i^2\right) \tag{45}$$

with equality iff for some constant, say γ, $a_i = \gamma\ b_i$ for all i or $b_i = c\ a_i$ for all i.

Choosing $b_i = 1$ for all i one has

$$\left(\sum_{i=1}^{k} a_i\right)^2 \leq k \sum_{i=1}^{k} a_i^2 \tag{46}$$

with equality iff $a_1 = a_2 = \cdots = a_k$.

Theorem 218 *Let $\mathcal{C}$ be a prefix code. Then the following statements are equivalent*

(i) $\sum_{u \in U} P_u \|c_u\| = H(P)$

(ii) For all $\omega \in \mathcal{X}^*$ with $\mathcal{U}(\mathcal{C}, \omega) \neq \emptyset$

$$P\big(\mathcal{U}(\mathcal{C}, \omega)\big) = q^{-\|\omega\|} \tag{47}$$

and for all $u, u' \in \mathcal{U}$ $\|c_u\| = \|c_{u'}\|$ and that c_u and $c_{u'}$ share the same prefix of length $\|c_u\| - 1$ implies

$$P_u = P_{u'}. \tag{48}$$

(iii)

$$L_{\mathcal{C}}(P, P) = \frac{q}{q-1}\left(1 - \sum_{u \in \mathcal{U}} P_u^2\right). \tag{49}$$

Proof It is well-known that (i) is equivalent to

(i') For all $u \in \mathcal{U}$

$$\|c_u\| = -[\log q]^{-1} \log P_u \text{ or } P_u = q^{-\|c_u\|}. \tag{50}$$

Notice that for (i) the code $\mathcal{C}$ is necessarily complete. We shall show that

$$(i') \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (i').$$

Ad (i')⇒(ii) For all ω with $\mathcal{U}(\mathcal{C}, \omega) \neq \emptyset$ the code $\mathcal{C}_\omega$ obtained by deleting the common prefix ω from all the codewords c_u, $u \in \mathcal{U}(\mathcal{C}, \omega)$, is a complete code on $\mathcal{U}(\mathcal{C}, \omega)$, because $\mathcal{C}$ is a complete code. That is,

$$\sum_{u \in \mathcal{U}(\mathcal{C}, \omega)} q^{-[\|c_u\| - \|\omega\|]} = 1$$

and, consequently, by (50)

$$P\big(\mathcal{U}(\mathcal{C},\omega)\big) = \sum_{u\in\mathcal{U}(\mathcal{C},\omega)} P_u = \sum_{u\in\mathcal{U}(\mathcal{C},\omega)} q^{-\|c_u\|} = q^{-\|\omega\|} \sum_{u\in\mathcal{U}(\mathcal{C},\omega)} q^{(\|c_u\|-\|\omega\|)} = q^{-\|\omega\|}.$$

Ad (ii)⇒(iii) Suppose (47) holds for all ω and we prove (iii) by induction on $\ell_{\max}(\mathcal{C}) = \max_{u\in\mathcal{U}} \|c_u\|$.

In case $\ell_{\max}(\mathcal{C}) = 1$ both sides of (49) are one. Assume (iii) holds for all codes $\mathcal{C}'$ with $\ell_{\max}(\mathcal{C}') \le L-1$ and let $\ell_{\max}(\mathcal{C}) = L$. Let $\mathcal{U}_1(\mathcal{C})$ and $\mathcal{U}_{(\alpha)}(\mathcal{C})$, be as in the proof of (11) and let $\mathcal{C}_{(\alpha)}$ be the prefix code for the source with alphabet $\mathcal{U}_{(\alpha)}(\mathcal{C})$ and distribution $P_{(\alpha)}$ such that for all $u \in \mathcal{U}_{(\alpha)}(\mathcal{C})$ and $\mathcal{X}' = \{c_u : u \in \mathcal{U}_1(\mathcal{C})\}$

$$P_{(\alpha)}(u) = P^{-1}\big(\mathcal{U}_{(\alpha)}(\mathcal{C})\big) P_u.$$

Then (47) and (48) imply that (ii) holds for all $\mathcal{C}_{(\alpha)}$, $\alpha \in \mathcal{U}_1(\mathcal{C})$ and for all $\beta \in \mathcal{U}_1(\mathcal{C})$

$$P_\beta = |\mathcal{U}_1(\mathcal{C})|^{-1} P\big(\mathcal{U}_1(\mathcal{C})\big). \tag{51}$$

Next we apply (47) to all ω with $\mathcal{U}(\mathcal{C},\omega)$ and $\|\omega\| = 1$ and obtain

$$\Pr\big(U \notin \mathcal{U}_1(\mathcal{C})\big) = \big(q - |\mathcal{U}_1(\mathcal{C})|\big) q^{-1}, \tag{52}$$

which with (51) yields for all $\beta \in \mathcal{U}_1(\mathcal{C})$

$$P_\beta = q^{-1}. \tag{53}$$

Moreover, by induction hypothesis for all $\mathcal{C}_{(\alpha)}$ and $P_{(\alpha)}$, $\alpha \in \mathcal{U}_1(\mathcal{C})$

$$L_{\mathcal{C}_{(\alpha)}}(P_{(\alpha)}, P_{(\alpha)}) = \frac{q}{q-1}\left(1 - q^2 \sum_{u\in\mathcal{U}_{(\alpha)}(\mathcal{C})} P_u^2\right) \tag{54}$$

as by (47)

$$P\big(\mathcal{U}_{(\alpha)}(\mathcal{C})\big) = q^{-1} \tag{55}$$

for all $\alpha \in \mathcal{X}^\Delta = \mathcal{X} \setminus \{c_u : u \in \mathcal{U}_1(\mathcal{C})\}$ (say).

Finally, like in the proof of (11) we have

$$L_{\mathcal{C}}(P,P) = 1 + \sum_{\alpha\in\mathcal{X}^\Delta} P^2\big(\mathcal{U}_{(\alpha)}(\mathcal{C})\big) L_{\mathcal{C}_{(\alpha)}}(P_{(\alpha)}, P_{(\alpha)}) \tag{56}$$

$$= 1 + \sum_{\alpha\in\mathcal{X}^\Delta} \frac{1}{q(q-1)}\left[1 - q^2 \sum_{u\in\mathcal{U}_{(\alpha)}(\mathcal{C})} P_u^2\right]$$

$$= 1 + \frac{|\mathcal{X}^{\Delta}|}{q(q-1)} - \frac{q}{q-1} \sum_{u \notin \mathcal{U}_1(\mathcal{C})} P_u^2 \tag{57}$$

$$= 1 + \frac{q - |\mathcal{U}_1(\mathcal{C})|}{q(q-1)} - \frac{q}{q-1} \sum_{u \in \mathcal{U}} P_u^2 + \frac{q}{q-1} |\mathcal{U}_1(\mathcal{C})| q^{-2}$$

$$= \frac{q}{q-1} \left(1 - \sum_{u \in \mathcal{U}} P_u^2 \right), \qquad \text{that is (49)}, \tag{58}$$

where the second equality holds by (54), the third equality holds, because $\{\mathcal{U}_1(\mathcal{C}), \mathcal{U}_{(\alpha)}(\mathcal{C}), \alpha \in \mathcal{X}'\}$ is a partition of $\mathcal{U}$, and the fourth equality follows from (53) and the definition of $\mathcal{X}^{\Delta}$.

Ad (iii)⇒(i') Again we proceed by induction on the maximum length of codewords.

Suppose first that for a code $\mathcal{C}$ $\quad \ell_{\max}(\mathcal{C}) = 1$. Then $L_{\mathcal{C}}(P, P) = 1$ and $|\mathcal{U}| \leq q$. Applying (46) to the ID-entropy we get

$$\frac{q}{q-1} \left(1 - \sum_{u \in \mathcal{U}} P_u^2 \right) \leq \frac{q}{q-1} (1 - |\mathcal{U}|^{-1})$$

with equality iff P is the uniform distribution. On the other hand, since $|\mathcal{U}| \leq q$, $\frac{q}{q-1}(1 - |\mathcal{U}|^{-1}) \leq \frac{q}{q-1} \left(1 - \frac{1}{q}\right) = 1$ and the equality holds iff $|\mathcal{U}| = q$. Then (49) holds iff P is uniform and $|\mathcal{U}| = q$, i.e., (50).

Assume now that the implication (iii) ⇒ (i') holds for all codes with maximum lengths $\leq L - 1$ and that $\mathcal{C}$ is a prefix code of maximum length $\ell_{\max}(\mathcal{C}) = L$.

Without loss of generality we can assume that $\mathcal{C}$ is complete, because otherwise we can add "dummy" symbols with 0 probability to $\mathcal{U}$ and assign to them suitable codewords so that the Kraft sum equals 1, but this does not change equality (49).

Having completeness we can assume that for $(ak) \leq q^{L-1}$ there are kq symbols $u(i, j)$ $(1 \leq i \leq k, 0 \leq j \leq q-1)$ in $\mathcal{U}$ with $\|c_{u(i,j)}\| = L$ and such that $c_{u(i,0)}, c_{u(i,1)}, \ldots, c_{u(i,q-1)}$ share a prefix ω_i of length $L-1$ for $i = 1, 2, \ldots, k$.

Let $u(1), \ldots, u(k)$ be k "new symbols" not in the original $\mathcal{U}$ and consider

$$\mathcal{U}' = \left[\mathcal{U}\{u(i, j) : 1 \leq i \leq k, 0 \leq j \leq q-1\}\right] \cup \{u(i) : 1 \leq i \leq k\}$$

and the probability distribution P' defined by

$$P'_{u'} = \begin{cases} P_{u'} & \text{if } u' \in \mathcal{U} \cap \mathcal{U}' \\ \sum_{j=0}^{q-1} P_{u(i,j)} & \text{if } u' = u(i) \text{ for some } i. \end{cases} \tag{59}$$

Next we define a prefix code $\mathcal{C}'$ for the source $(\mathcal{U}', P')$ by using $\mathcal{C}$ as follows:

$$c'_{u'} = \begin{cases} c_{u'} & \text{if } u' \in \mathcal{U} \cap \mathcal{U}' \\ \omega_i & \text{if } u' = u(i) \text{ for some } i. \end{cases} \tag{60}$$

Then for $u' \in \mathcal{U} \cap \mathcal{U}'$ $\|c'_{u'}\| = \|c_{u'}\|$ and $\|c'_{u(1)}\| = \|c'_{u(2)}\| = \cdots = \|c'_{u(k)}\| = L-1$.

Therefore by induction hypothesis

$$L_{\mathcal{C}'}(P', P') \geq \frac{q}{q-1}\left(1 - \sum_{u' \in \mathcal{U}'} P'^2_{u'}\right) \tag{61}$$

and equality holds iff $P_u = q^{-\|c_u\|}$ for $u \in \mathcal{U} \cap \mathcal{U}'$ and

$$\sum_{j=0}^{q-1} P_{u(i,j)} = P'_{u(i)} = q^{-(L-1)} \qquad \text{for } i = 1, 2, \ldots, k.$$

Furthermore, it follows from (46) and the definition of $L_{\mathcal{C}}(P, P)$ and $L_{\mathcal{C}'}(P', P')$ that

$$\begin{aligned} L_{\mathcal{C}}(P, P) &= L_{\mathcal{C}'}(P', P') + \sum_{i=1}^{k}\left(\sum_{j=0}^{q-1} P_{u(i,j)}\right)^2 \\ &= L_{\mathcal{C}'}(P', P') + \sum_{i=1}^{k} P'^2_{u(i)} \\ &\geq \frac{q}{q-1}\left(1 - \sum_{u' \in \mathcal{U}'} P'^2_{u'}\right) + \sum_{i=1}^{k} P'^2_{u(i)} \\ &= \frac{q}{q-1}\left(1 - \sum_{u \in \mathcal{U} \cap \mathcal{U}'} P_u^2\right) + \sum_{i=1}^{k}\left(1 - \frac{q}{q-1}\right) P'^2_{u(i)} \\ &= \frac{q}{q-1}\left[1 - \sum_{u \in \mathcal{U} \cap \mathcal{U}'} P_u^2 - \sum_{i=1}^{k} q^{-1}\left(\sum_{j=0}^{q-1} P_{u(i,j)}\right)^2\right] \\ &\geq \frac{q}{q-1}\left[1 - \sum_{u \in \mathcal{U}} P_u^2\right]. \end{aligned} \tag{62}$$

By (60) the first inequality holds iff $P_u = q^{-\|c_u\|}$ for $u \in \mathcal{U} \cap \mathcal{U}'$ and $\sum_{j=0}^{q-1} P_{u(i,j)} = q^{-(L-1)}$ for $i = 1, 2, \ldots, k$; it follows from (46) that the last inequality holds and with equality iff

$$P_{u(i,0)} = P_{u(i,1)} = \cdots = P_{u(i,q-1)} \text{ for } i = 1, 2, \ldots, k.$$

In order to have

$$L_{\mathcal{C}}(P, P) = \frac{q}{q-1}\left[1 - \sum_{u \in \mathcal{U}} P_u^2\right]$$

the two inequalities in (62) must be equalities. However, this is equivalent with (50), i.e. (i'). □

5 A Global Balance Principle to Find Good Codes

In case U and V are i.i.d. there is no gain in using the local unbalance principle (LUP). But in this case Corollary 216 and (46) provide a way to find a good code. We first rewrite Corollary 216 as

$$\mathbb{E}W_{\mathcal{C}}(U, V) = \sum_{u} \sum_{\omega \in \mathcal{X}^u} \Pr\big(U \in \mathcal{U}(\mathcal{C}, \omega), V \in \mathcal{U}(\mathcal{C}, \omega)\big).$$

By the assumptions on U and V with their distribution P

$$L_{\mathcal{C}}(P, P) = \sum_{u} \sum_{\omega \in \mathcal{X}^u} P^2\big(\mathcal{U}(\mathcal{C}, \omega)\big). \tag{63}$$

Notice that in case

$$P_{n,\mathcal{C}} \triangleq \sum_{\omega \in \mathcal{X}^n} P\big(\mathcal{U}(\mathcal{C}, \omega)\big)$$

is a constant $\sum_{\omega \in \mathcal{X}^n} P^2\big(\mathcal{U}(\mathcal{C}, \omega)\big)$ is minimized by choosing the $P\big(\mathcal{U}(\mathcal{C}, \omega)\big)$'s uniformly. This gives us a global balance principle (GBT) for finding good codes.

We shall see the roles of both, the LUP and the GBP in the proof of the following coding theorem for discrete memoryless sources (DMS's).

Theorem 219 *For a DMS* $(U^n, V^n)_{n=1}^{\infty}$ *with generic distribution* $P_{UV} = PQ$*, i.e. the generic RV's* U *and* V *are independent and* $P_U = P$*,* $P_V = Q$

$$\lim_{n\to\infty} L(P^n, Q^n) = \begin{cases} 1 & \text{if } P \neq Q \\ \frac{q}{q-1} & \text{if } P = Q. \end{cases} \tag{64}$$

Proof Trivially $L_{\mathcal{C}}(P, Q) \geq 1$ and by Corollary 217, $\frac{q}{q-1}$ is a lower bound to $\lim_{n\to\infty} L(P^n, P^n)$. Hence we only have to construct codes to achieve asymptotically the bounds in (64).

Case $P \neq Q$ We choose a $\delta > 0$ so that for sufficiently large n

$$\mathcal{T}^n_{P,\delta} \cap \mathcal{T}^n_{Q,\delta} = \emptyset \tag{65}$$

and for a $\theta > 0$

$$P(\mathcal{T}^n_{P,\delta}) > 1 - 2^{n\theta} \text{ and } Q(\mathcal{T}^n_{Q,\delta}) > 1 - 2^{n\theta}. \tag{66}$$

Partition $\mathcal{U}^n$ into two parts $\mathcal{U}_0$ and $\mathcal{U}_1$ such that $\mathcal{U}_0 \supset \mathcal{T}^n_{P,\delta}$ and $\mathcal{U}_1 \supset \mathcal{T}^n_{Q,\delta}$.

To simplify matters we assume $q = 2$. This does not loose generality since enlarging the alphabet cannot make things worse.

Let $\ell_i = \lceil \log |\mathcal{U}_i| \rceil$ and $\psi_i : \mathcal{U}_i \to 2^{[\ell_i]}$ for $i = 1, 2$. Then we define a code $\mathcal{C}$ by $c_{u^n} = \big(i, \psi_i(u^n)\big)$ if $u^n \in \mathcal{U}_i$ and show that $L_{\mathcal{C}}(P^n, Q^n)$ is arbitrarily close to one if n is sufficiently large. Actually it immediately follows from (11)

$$\begin{aligned}
&L_{\mathcal{C}}(P^n, Q^n) \\
&\quad = \sum_{u^n, u'^n \in \mathcal{U}^n} P^n(c_{u^n}) Q^n(c_{u'^n}) cp(c_{u^n}, c_{u'^n}) \\
&\quad = \sum_{u^n \in \mathcal{U}_0} \sum_{u'^n \in \mathcal{U}_0} P^n(c_{u^n}) Q^n(c_{u'^n}) cp(c_{u^n}, c_{u'^n}) \\
&\quad + \sum_{u^n \in \mathcal{U}_0} \sum_{u'^n \in \mathcal{U}_1} P^n(c_{u^n}) Q^n(c_{u'^n}) cp(c_{u^n}, c_{u'^n}) \\
&\quad + \sum_{u^n \in \mathcal{U}_1} \sum_{u'^n \in \mathcal{U}_0} P^n(c_{u^n}) Q^n(c_{u'^n}) cp(c_{u^n}, c_{u'^n}) \\
&\quad + \sum_{u^n \in \mathcal{U}_1} \sum_{u'^n \in \mathcal{U}_1} P^n(c_{u^n}) Q^n(c_{u'^n}) cp(c_{u^n}, c_{u'^n}) \\
&\quad < \ell_0 \sum_{u^n \in \mathcal{U}_0} P^n(c_{u^n}) \sum_{u'^n \in \mathcal{U}_0} Q^n(c_{u'^n}) + \sum_{u^n \in \mathcal{U}_0} P^n(c_{u^n}) \sum_{u'^n \in \mathcal{U}_1} Q^n(c_{u'_n})
\end{aligned}$$

$$+\sum_{u^n\in\mathcal{U}_1}P^n(c_{u^n})\sum_{u'^n\in\mathcal{U}_0}Q^n(c_{u'^n})+\ell_1\sum_{u^n\in\mathcal{U}_1}P^n(c_{u^n})\sum_{u'^n\in\mathcal{U}_1}Q^n(c_{u'^n})$$

$$\le\left[\sum_{u^n\in\mathcal{U}_0}P^n(c_{u^n})\sum_{u'^n\in\mathcal{U}_1}Q^n(c_{u'^n})+\sum_{u^n\in\mathcal{U}_1}P^n(c_{u^n})\sum_{u'^n\in\mathcal{U}_0}Q^n(c_{u'^n})\right]$$

$$+\lceil n\log|\mathcal{U}|\rceil\left[\sum_{u^n\in\mathcal{U}_0}(c_{u^n})\sum_{u'^n\in\mathcal{U}_0}Q^n(c_{u'^n})+\sum_{u^n\in\mathcal{U}_1}P^n(c_{u^n})\sum_{u'^n\in\mathcal{U}_0}Q^n(c_{u'^n})\right]$$

$$\le 1+\lceil n\log|\mathcal{U}|\rceil\left[\sum_{u'^n\in\mathcal{U}_0}Q^n(c_{u'^n})+\sum_{u^n\in\mathcal{U}_1}P^n(c_{u^n})\right]$$

and therefore,

$$L_{\mathcal{C}}(P^n,Q^n)<1+\lceil n\log|\mathcal{U}|\rceil 2^{-n\theta+1}\to 1\qquad\text{as } n\to\infty, \tag{67}$$

where the second inequality holds because

$$\ell_i=\lceil\log|\mathcal{U}_i|\rceil\le\lceil\log|\mathcal{U}^n|\rceil\qquad\text{for } i=0,1$$

and the last inequality follows from (66).

Case $P=Q$ Now we let $P=Q$. For $0<\alpha<H(P)$ let $\mathcal{P}_n(>\alpha)$ be the set of n-ED's (n-empirical distributions) $\tilde{P}$ on $\mathcal{U}$ with $|\mathcal{T}^n_{\tilde{P}}|>2^{n\alpha}$. Then there is a positive θ such that the empirical distribution of the output U^n (resp. V^n) is in $\mathcal{P}_n(>\alpha)$ with probability larger than $1-2^{n\theta}$.

Next we choose an integer ℓ_n such that for

$$\beta\triangleq\frac{1}{4}\min(\theta,\alpha)\quad 2^{\frac{n}{2}\beta}<q^{\ell_n}\le 2^{n\beta}. \tag{68}$$

Label sequences in $\mathcal{T}^n_{\tilde{P}}$ for $\tilde{P}\in\mathcal{P}_n(>\alpha)$ by $0,1,\ldots,|\mathcal{T}^n_{\tilde{P}}|-1$ and let Ψ_1 be a mapping from $\mathcal{U}^n$ to $\mathcal{X}^{\ell_n}$, where $\mathcal{X}=\{0,1,\ldots,q-1\}$ as follows.

If u^n has ED $\tilde{P}$ in $\mathcal{P}_n(>\alpha)$ and got an index $ind(u^n)$ with q-ary representation $(x_k,x_{k-1},\ldots,x_2,x_1)$ i.e., $ind(u^n)=\sum_{i=0}^{k}x_iq^{i-1}$ for $0\le x_i\le q-1$, $k=\lceil\log|\mathcal{T}^n_{\tilde{P}}|\rceil$, then let

$$\Psi_1(u^n)=(x_1,x_2,\ldots,x_{\ell_n}). \tag{69}$$

If the ED of u^n is not in $\mathcal{P}_n(> \alpha)$, we arbitrarily choose a sequence in $\mathcal{X}^{\ell_n}$ as $\psi_1(u^n)$.

For any fixed $t \leq \ell_n$, $\tilde{P} \in P_n(> \alpha)$, and $x^t \in \mathcal{X}^t$ let $\mathcal{U}(\tilde{P}, x^t)$ be the set of sequences in $\mathcal{T}^n_{\tilde{P}}$ such that x^t is a prefix of $\psi_1(u^n)$. Then it is not hard to see that for all x^t, x'^t with $t \leq \ell_n$

$$\big||\mathcal{U}(\tilde{P}, x^t)| - |\mathcal{U}(\tilde{P}, x'^t)|\big| \leq 1.$$

More specifically for all $t \leq \ell_n$ and $x^t \in \mathcal{X}^t$

$$|\mathcal{U}(\tilde{P}, x^t)| = \sum_{j=t+1}^{k} a_j q^{j-1-t} \text{ or } \sum_{j=t+1}^{k} a_j q^{j-1-t} + 1,$$

if $|\mathcal{T}^n_{\tilde{P}}| = \sum_{j=1}^{k} a_j q^{j-1}$ with $a_k \neq 0$, $0 \leq a_j \leq q-1$ for $j = 1, 2, \ldots, k-1$.

Let $\mathcal{U}(x^t) = \bigcup_{\text{all } \tilde{P}} \mathcal{U}(\tilde{P}, x^t)$ (here it does not matter whether $\tilde{P} \in \mathcal{P}_n(> \alpha)$ or not).

Thus we partition $\mathcal{U}^n$ into q^t parts as $\{\mathcal{U}(x^t) : x^t \in \mathcal{X}^t\}$ for $t \leq \ell_n$.

By the asymptotic equipartition property (AEP), the difference of the conditional probability of the event that the output of U^n is in $\mathcal{U}(x^t)$ given that the ED of U^n is in $\mathcal{P}_n(> \alpha)$ and q^{-1} is not larger than

$$\min_{\tilde{P} \in \mathcal{P}_n(> \alpha)} |\mathcal{T}_{\tilde{P}}|^{-1} < 2^{-n\alpha}.$$

Recalling that with probability $1 - 2^{-n\theta}$ U^n has ED in $\mathcal{P}_n(> \alpha)$ and the assumption that V^n has the same distribution as U^n, we obtain that

$$\Pr\big(U^n \in \mathcal{U}(x^t)\big) = \Pr\big(V^n \in \mathcal{U}(x^t)\big) = P^n\big(\mathcal{U}(x^t)\big)$$

and for all $x^t \in \mathcal{X}^t$

$$(1 - 2^{-n\theta})(q^{-t} - 2^{-n\alpha}) \leq P^n\big(\mathcal{U}(x^t)\big) \leq (1 - 2^{-n\theta})(q^{-t} + 2^{-n\alpha}) + 2^{-n\theta},$$

which implies that for all $x^t \in \mathcal{X}^t$

$$|P^n\big(\mathcal{U}(x^t)\big) - q^{-t}| \leq 2^{-n\theta} + 2^{-n\alpha} < 2^{-2n\beta}, \tag{70}$$

when $\beta \triangleq \frac{1}{4}\min(\theta, \alpha)$.

Recall that Ψ_1 is a function from $\mathcal{U}^n$ to $\mathcal{X}^{\ell_n}$ and that the definition of $\mathcal{U}(x^t)$, $\mathcal{U}(x^{\ell_n})$ is actually the inverse image of $\mathcal{X}^{\ell_n}$ under Ψ_1, i.e. $\mathcal{U}(\mathcal{X}^{\ell_n}) = \Psi_1^{-1}(\mathcal{X}^{\ell_n})$.

Let furthermore $\ell^*(x^{\ell_n}) \triangleq \left\lceil \frac{\log |\mathcal{U}(x^n)|}{\log q} \right\rceil$ and let Ψ_2 be a function on $\mathcal{U}^n$ such that its restriction on $\mathcal{U}(x^{\ell_n})$ is an injection into $\mathcal{X}^{\ell^*(x^{\ell_n})}$ for all x^{ℓ_n}. Then our decoding function is defined as

$$c = (\Psi_1, \Psi_2). \tag{71}$$

To estimate $L_{\mathcal{C}}(P^n, P^n)$ we introduce an auxiliary source with alphabet $\mathcal{X}^{\ell_n}$ and probability distribution P^* such that for all $x^{\ell_n} \in \mathcal{X}^{\ell_n}$

$$P^*(x^{\ell_n}) = P^n\big(\mathcal{U}(x^{\ell_n})\big).$$

We divide the waiting time for identification with code $\mathcal{C}$ into two parts according to the two components Ψ_1 and Ψ_2 in (71), and we let W_1 and W_2 be the random waiting times of the two parts, respectively. Now let Z be a binary RV such that

$$Z = \begin{cases} 0 & \text{if } \Psi_1(U^n) \neq \Psi_1(V^n) \\ 1 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} &L_{\mathcal{C}}(P^n, P^n) \\ &\quad = \mathbb{E}(W_1 + W_2) = \mathbb{E}\, W_1 + \mathbb{E}\big(\mathbb{E}(W_2 \mid Z)\big) \\ &\quad = \mathbb{E}\, W_1 + \Pr(Z = 1)\mathbb{E}(W_2 \mid Z = 1) \\ &\quad = \mathbb{E}\, W_1 + \left[\sum_{x^{\ell_n}} P^n\big(\Psi_1(U^n) = x^{\ell_1}\big) P^n\big(\Psi_1(V^n) = x^{\ell_n}\big)\right] \cdot \mathbb{E}(W_2 \mid Z = 1) \\ &\quad = \mathbb{E}\, W_1 + \left\{\sum_{x^{\ell_n}} \big[P^n\big(\mathcal{U}(x^{\ell_n})\big)\big]^2\right\} \mathbb{E}(W_2 \mid Z = 1). \end{aligned} \tag{72}$$

Let $\mathcal{C}^*$ be the code for the auxiliary source with encoding function $c^* = \Psi_1$. Then we have that

$$\mathbb{E}\, W_1 = L_{\mathcal{C}^*}(P^*, P^*) \tag{73}$$

and with the notation in Corollary 216 $\mathcal{U}(\mathcal{C}^*, x^t) = \mathcal{U}(x^t)$ and $P^*\big(\mathcal{U}(\mathcal{C}^*, x^t)\big) = P^n\big(\mathcal{U}(x^t)\big)$ for $x^t \in \mathcal{X}^t$ with $t \leq \ell_n$. For all $x^t \in \mathcal{X}^t$, $t \leq \ell_n$, we denote

$$\delta(x^t) = q^{-t} - p^n\big(\mathcal{U}(x^t)\big).$$

Then we have for all $t \leq \ell_n$ $\sum\limits_{x^t \in \mathcal{X}^t} \delta(x^t) = 0$ and by (70) $\delta(x^t) < 2^{-2n\beta}$.

Now we apply Corollary 216 to estimate

$$
\begin{aligned}
L_{\mathcal{C}^*}(P^*, P^*) &= \sum_{t=0}^{\ell_n} \sum_{x^t \in \mathcal{X}^t} \left[P^*\big(\mathcal{U}(\mathcal{C}^*, x^t)\big)\right]^2 \\
&= \sum_{t=0}^{\ell_n} \sum_{x^t \in \mathcal{X}^t} \left[P^n\big(\mathcal{U}(x^t)\big)\right]^2 = \sum_{t=0}^{\ell_n} \sum_{x^t \in \mathcal{X}^t} \big(q^{-t} - \delta(x^t)\big)^2 \\
&= \sum_{t=0}^{\ell_n} \left[q^t \cdot q^{-2t} - 2q^{-t} \sum_{x^t \in \mathcal{X}^t} \delta(x^t) + \sum_{x^t \in \mathcal{X}^t} \delta(x^t)^2 \right] \\
&\leq \sum_{t=0}^{\ell_n} q^{-t} + \sum_{t=0}^{\ell_n} q^t \cdot 2^{-4n\beta} \\
&< \sum_{t=0}^{\infty} q^{-t} + \frac{q^{\ell_n+1} - 1}{q-1} 2^{-4n\beta} \\
&< \frac{q}{q-1} + \frac{1}{q-1} q^{\ell_n+1} 2^{-4n\beta}. \qquad (74)
\end{aligned}
$$

Moreover, by definition of Ψ_2 and W_2

$$
\mathbb{E}\big(W_2 \mid Z = 1)\big) \leq \left\lceil \frac{n \log |\mathcal{U}|}{\log q} \right\rceil
$$

and in (74) we have shown that

$$
\sum_{x^{\ell_n}} \left[P^n\big(\mathcal{U}(x^{\ell_n})\big)\right]^2 \leq q^{-\ell_n} + q^{\ell_n} \cdot 2^{-4n\beta}.
$$

Consequently

$$
\left\{ \sum_{x^{\ell_n} \in \mathcal{X}^{\ell_n}} \left[P^n\big(\mathcal{U}(x^t)\big)\right]^2 \right\} \mathbb{E}(W_2 \mid Z = 1) \leq [q^{-\ell_n} + q^{\ell_n} 2^{-4n\beta}] \left\lceil \frac{n \log |\mathcal{U}|}{\log q} \right\rceil. \qquad (75)
$$

Finally, by combining (72)–(75) with the choice of β in (68) we have that

$$
\overline{\lim_{n\to\infty}} L_{\mathcal{C}}(P^n, P^n) < \frac{q}{q-1},
$$

the desired inequality. □

It is interesting that the limits of the waiting time of ID-codes in the left hand side of (64) are independent of the generic distributions P and Q and only depend on whether they are equal.

In the case that they are not equal it is even independent of the alphabet size. In particular in case $P \neq Q$, we have seen in the proof that the key step is how to distribute the first symbol and the local unbalance principle (LUP) is applied in the second step. Moreover for a good code the random user with exponentially vanishing probability needs to wait for the second symbol. So the remaining parts of codewords are not so important.

Similarly in the case $P = Q$, where we use instead of the LUP the GBP, the key parts of codewords is a relatively small prefix (in the proof it is the ℓ_n-th prefix) and after that the user with exponentially small probability has to wait. Thus again the remaining part of codewords is less important.

6 Comments on Generalized Entropies

After the discovery of ID-entropies in [4] work of Tsallis [11] and also [12] was brought to our attention. The equalities (1) and (2) in [12] are here (76) and (77). The letter q used there corresponds to our letter α, because for us q gives the alphabet size. The generalisation of Boltzmann's entropy

$$H(P) = -k \sum P_u ln P_u$$

is

$$S_\alpha(P) = k \frac{1}{\alpha - 1} \left(1 - \sum_{u=1}^{N} P_u^\alpha \right) \tag{76}$$

for any real $\alpha \neq 1$. Notice that $\lim_{\alpha \to 1} S_\alpha(P) = H(P)$, which can be named $S_1(P)$.

One readily verifies that for product-distributions $P \times Q$ for independent RV's

$$S_\alpha(P \times Q) = S_\alpha(P) + S_\alpha(Q) - \frac{(\alpha - 1)}{k} S_\alpha(P) S_\alpha(Q) \tag{77}$$

Since in all cases $S_\alpha \geq 0$, $\alpha < 1$, $\alpha = 1$ and $\alpha > 1$ respectively correspond to superadditivity, additivity and subadditivity (also called for the purposes in statistical physics superextensitivity, extensitivity, and subextensitivity).

We recall the grouping identity of [4] (chapter "Identification Entropy").

For a partition $(\mathcal{U}_1, \mathcal{U}_2)$ of $\mathcal{U} = \{1, 2, \ldots, N\}$, $Q_i = \sum_{u \in \mathcal{U}_i} P_u$, and $P_u^{(i)} = \frac{P_u}{Q_i}$ for $u \in \mathcal{U}_i$, $i = 1, 2$

$$H_{I,q}(P) = H_{I,q}(Q) + \sum_i Q_i^2 H_{I,q}\left(\frac{P^{(i)}}{Q_i}\right) \tag{78}$$

where $Q = (Q_1, Q_2)$. This implies

$$H_{I,q}(P \times Q) = H_{I,q}(Q) + \sum_j Q_j^2 H_{I,q}(P)$$

and since

$$\left(1 - \sum_k Q_j^2\right) = \frac{q-1}{q}\frac{q}{q-1}\left(1 - \sum_j Q_j^2\right)$$

$$= \frac{q-1}{q} H_{I,q}(Q)$$

$$\sum_j Q_j^2 = 1 - \frac{q-1}{q} H_{I,q}(Q)$$

we get

$$H_{I,q}(P \times Q) = H_{I,q}(Q) + H_{I,q}(P) - \frac{q-1}{q} H_{I,q}(Q) H_{I,q}(P), \tag{79}$$

which is (77) for $\alpha = 2$ and $k = \frac{q}{q-1}$.

We have been told by several experts in physics that the operational significance of the quantities S_α (for $\alpha \neq 1$) in statistical physics seems not to be undisputed.

In contrast, the significance of identification entropy was demonstrated in [4] (see chapter "Identification Entropy"), which is formally close, but essentially different from S_α for two reasons: always $\alpha = 2$ and $k = \frac{q}{q-1}$ is uniquely determined and depends on the alphabet size q!

We also have discussed the coding theoretical meanings of the factors $\frac{q}{q-1}$ and $\left(1 - \sum_{u=1}^{N} P_u^2\right)$.

More recently, we learned from referees that already in 1967 Havrda and Charvát [7] introduced the entropies $\{H_N^\alpha\}$ of *type* α

$$H_N^\alpha(P_1, P_2, \ldots, P_N) = (2^{1-\alpha} - 1)^{-1}\left(\sum_{i=1}^{N} P_i^\alpha - 1\right) \tag{80}$$

$[(P_1, P_2, \ldots, P_N) \in \mathcal{P}([N]), N = 2, 3, \ldots, 0^\alpha = 0]$

$$\lim_{\alpha \to 1} H_N^\alpha(P_1, P_2, \ldots, P_N) = H_N(P_1, P_2, \ldots, P_N),$$

the Boltzmann/Shannon entropy. So, it is reasonable to define

$$H_N^1(P_1, P_2, \ldots, P_N) = H_N(P_1, P_2, \ldots, P_N).$$

This is a generalization of the BGS-entropy different from the Rényi entropies of *order* $\alpha \neq 1$ (which according to [2] were introduced by Schützenberger [8]) given by

$$\alpha H_N(P_1, P_2, \ldots, P_N) = \frac{1}{1-\alpha} \log_2 \sum_{i=1}^{N} P_i^\alpha$$

$[(P_1, P_2, \ldots, P_N) \in \mathcal{P}([N]), N = 2, 3, \ldots]$.

Comparison shows that

$$\alpha H_N(P_1, P_2, \ldots, P_N) = \frac{1}{1-\alpha} \log_2[(2^{1-\alpha} - 1) H_N^\alpha(P_1, P_2, \ldots, P_N) + 1]$$

$[(P_1, P_2, \ldots, P_N) \in \mathcal{P}([N]), N = 2, 3, \ldots]$.

So, while the entropies of order α and the entropies of type α are different for $\alpha \neq 1$, we see that the bijection

$$t \to \frac{1}{1-\alpha} \log_2[(2^{1-\alpha} - 1)t + 1]$$

connects them. Therefore, we may ask what the advantage is in dealing with entropies of type α. We meanwhile also learned that the book [1] gives a comprehensive discussion. Also Daróczy's contribution [6], where "type α" is named "degree α", gives an enlightening analysis.

Note that Rényi entropies ($\alpha \neq 1$) are *additive*, but not subadditive (except for $\alpha = 0$) and not *recursive*, and they have not the *branching* property nor the *sum* property, that is, there exists a measurable function g on $(0, 1)$ such that

$$H_N^\alpha(P_1, P_2, \ldots, P_N) = \sum_{i=1}^{N} g(P_i).$$

Entropies of type α, on the other hand, are *not additive* but do have the subadditivity property and the *sum property* and furthermore are

Additive of degree α:

$$H^{\alpha}_{MN}(P_1Q_1, P_1Q_2, \ldots, P_1Q_N, P_2Q_1, P_2Q_2, \ldots, P_2Q_N, \ldots, P_MQ_1, P_MQ_2, \ldots, P_MQ_N)$$

$$= H^{\alpha}_{M}(P_1, P_2, \ldots, P_M) + H^{\alpha}_{N}(Q_1, Q_2, \ldots, Q_N) + (2^{1-\alpha} - 1)H^{\alpha}_{M}(P_1, P_2, \ldots, P_M) + H^{\alpha}_{N}(Q_1, Q_2, \ldots, Q_N)$$

$[(P_1, P_2, \ldots, P_M) \in \mathcal{P}([M]), (Q_1, Q_2, \ldots, Q_N) \in \mathcal{P}([N]); M = 2, 3, \ldots; N = 2, 3, \ldots]$.

Strong additive of degree α:

$$H^{\alpha}_{MN}(P_1Q_{11}, P_1Q_{12}, \ldots, P_1Q_{1N}, P_2Q_{21}, P_2Q_{22}, \ldots, P_2Q_{2N}, \ldots, P_MQ_{M1}, P_MQ_{M2}, \ldots, P_MQ_{MN})$$

$$= H^{\alpha}_{M}(P_1, P_2, \ldots, P_M) + \sum_{j=1}^{M} P_j^{\alpha} H^{\alpha}_{N}(Q_{j1}, Q_{j2}, \ldots, Q_{jN})$$

$(P_1, P_2, \ldots, P_M) \in \mathcal{P}([M])$, $(Q_{j1}, Q_{j2}, \ldots, Q_{jN}) \in \mathcal{P}([N])$; $j = 1, 2, \ldots, M$; $M = 2, 3, \ldots$; $N = 2, 3, \ldots]$.

Recursive of degree α:

$$H^{\alpha}_{N}(P_1, P_2, \ldots, P_N)$$

$$= H^{\alpha}_{N-1}(P_1 + P_2, P_3, \ldots, P_N) + (P_1 + P_2)^{\alpha} H^{\alpha}_{2}\left(\frac{P_1}{P_1 + P_2}, \frac{P_2}{P_1 + P_2}\right)$$

$[(P_1, P_2, \ldots, P_N) \in \{([N]), N = 3, 4, \ldots \text{ with } P_1 + P_2 > 0]$.

(In consequence, entropies of type α also have the branching property.)

It is clear now that for binary alphabet the ID-entropy is exactly the entropy of type $\alpha = 2$.

However, prior to [11] there are hardly any applications or operational justifications of the entropy of type α.

Moreover, the q-ary case did not exist at all and therefore the name ID-entropy is well justified.

We feel that it must be said that in many papers (with several coauthors) Tsallis at least developed ideas to promote non-standard-equilibrium theory in Statistical Physics using generalized entropies S_α and generalized concepts of inner energy.

Our attention has been drawn also to the papers [5, 9], and [10] with possibilities of connection to our work.

Clear-cut progress was made by Heup in his thesis (see Lecture 21) with a generalization of ID-entropy motivated by L-identification.

References

1. S. Abe, Axioms and uniqueness theorem for Tsallis entropy. Phys. Lett. **A 271**(1–2), 74–79 (2000)
2. J. Aczel, Z. Daroczy, *On Measures of Information and Their Characterizations*. Mathematics in Science and Engineering, vol. 115 (Academic Press, New York, 1975)
3. R. Ahlswede, General theory of information transfer, Preprint 97–118, SFB 343 "Diskrete Strukturen in der Mathematik", Universität Bielefeld, General Theory of Information Transfer and Combinatorics, Report on a Research Project at the ZIF (Center of interdisciplinary studies) in Bielefeld Oct. 1, 2002–August 31, 2003, edit R. Ahlswede with the assistance of L. Bäumer and N. Cai, also Special issue of Discrete Mathematics
4. R. Ahlswede, Identification entropy, in *General Theory of Information Transfer and Combinatorics*. Lecture Notes in Computer Science, vol. 4123 (Springer, Berlin, 2006), pp. 595–613
5. C.C. Campbell, A coding theorem and Rényi's entropy. Inf. Control **8**, 423–429 (1965)
6. Z. Daróczy, Generalized information functions. Inf. Control **16**, 36–51 (1970)
7. J. Havrda, F. Charvát, Quantification method of classical processes, concept of strucural α-entropy. Kybernetika (Prague) **3**, 30–35 (1967)
8. M.P. Schützenberger, Contribution aux applications statistiques de la theorie de l'information, vol. 3, no. 1–2 (Publ. Inst. Statist. Univ. Paris, Paris, 1954), pp. 3–117
9. B.D. Sharma, H.C. Gupta, Entropy as an optimal measure, information theory, in *Proc. Int. CNRS Colloq., Cachan, 1997, Paris*. Colloq. Internat. CNRS, 276 (1978), pp. 151–159
10. F. Topsøe, Game-theoretical equilibrium, maximum entropy and minimum information discrimination, maximum entropy and Bayesian methods. Fund. Theor. Phys. **53**, 15–23 (1992). Published by Kluwer Acad., Paris, France, 1993
11. C. Tsallis, Possible generalization of Boltzmann-Gibbs statistics. J. Stat. Phys. **52**(1–2), 479–487 (1988)
12. C. Tsallis, R.S. Mendes, A.R. Plastino, The role of constraints within generalized nonextensive statistics. Physica A **261**, 534–554 (1998)

L-Identification for Sources

1 Introduction

In [3] (see Part II, chapter "Models with Prior Knowledge of the Receiver") Ahlswede introduced "a general communication model for one sender". Suppose we have a message set $\mathcal{M} = \{1, \ldots, M\}$ whose elements are encoded in such a way that information about them can be transmitted over a channel. If this channel is noiseless, i.e., there occur no errors during the transmission, we speak of (noiseless) source coding. In this case it is common to omit the presence of a channel and speak simply of source coding.

What do we mean by information? In Shannon's classical information transmission problem [15] the decoder is interested in the message which has been encoded by the encoder. However, the decoder may have different goals. In [3] Ahlswede writes:

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

In the above citation every partition $\pi \in \Pi(\mathcal{M})$ is identified with a different decoder. Moreover, the author describes some "seemingly natural families of partitions". We focus on the first three models which highlight the differences between classical information transmission and identification. These are

Model 1: $\Pi_S = \{\pi_{sh}\}, \pi_{sh} = \{\{m\} : m \in \mathcal{M}\}$.
Model 2: $\Pi_I = \{\pi_m : m \in \mathcal{M}\}, \pi_m = \{\{m\}, \mathcal{M} \setminus \{m\}\}$.
Model 3: $\Pi_K = \{\pi_S : |S| = K, S \subset \mathcal{M}\}, \pi_S = \{S, \mathcal{M} \setminus S\}$.

The first model describes Shannon's classical transmission problem. Here the decoder wants to know which message has been encoded by the encoder. Let us assume we are given a probability distribution P on the message set. In source

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_22

coding we consider a *source code* $\mathcal{C} : \mathcal{M} \to \mathcal{Q}^*$. Here $\mathcal{Q}$ is the *q-ary alphabet* $\{0, 1, \ldots, q-1\}$ and $\mathcal{Q}^* = \bigcup_{d=0}^{\infty} \mathcal{Q}^d$. $\mathcal{C}(m)$ is called the *codeword* of the message m. We further assume that $\mathcal{C}$ is a *prefix code*. That is, no codeword is the prefix of another codeword. The goal of source coding is to construct prefix codes which have a small average codeword length. In other words, the mean of the codeword lengths should be as small as possible. It is well-known that this value is lower bounded by Shannon's classical entropy

$$H_q(P) = -\sum_{m=1}^{M} p_m \log_q p_m.$$

There exist codes, e.g. Huffman codes [12] and Shannon-Fano codes [10], which yield an average codeword length of at most $H_q(P) + 1$. The uniform distribution maximizes $H_q(P)$ and it holds that $H_q(1/M, \ldots, 1/M) = \log_q M$.

In the second model the decoder π_m wants to know whether m occurred or not. This is the identification problem introduced for noisy channel coding in [5] (see chapter "Identification via Channels") and analyzed inter alia in [6] (see chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas"), [11, 13]. Identification source coding was introduced in [3], continued in [7] and led to the identification entropy [2]

$$H_{I,q}(P) = \frac{q}{q-1}\left(1 - \sum_{m=1}^{M} p_m^2\right).$$

This entropy function again is maximized by the uniform distribution. Unlike Shannon's entropy it does not grow logarithmically in M but tends to $q/(q-1)$ as M goes to infinity.

A generalization of the identification problem is model 3, which is called *K-identification*. This case arises in several situations. Ahlswede writes: "For instance every person π_S may have a set S of K closest friends and the sender knows that one person $m \in \mathcal{M}$ is sick. All persons π_S want to know whether one of their friends is sick."

Another natural problem is somewhat like the opposite of K-identification. For example, the encoder knows L persons $m_1, \ldots, m_L \in \mathcal{M}$, who have won a lottery. Every participant, a member of $\mathcal{M}$, wants to know whether or not he or she is among the winners. However, the information in which a participant is interested can no longer be represented by a partition of $\mathcal{M}$. We have to partition $\binom{\mathcal{M}}{L}$ and get

$$\Pi_{L,\text{set}} = \{\pi_m \ : \ m \in \mathcal{M}\}, \tag{1}$$

where $\pi_m = \{S_m, \binom{\mathcal{M}}{L} \setminus S_m\}$ and $S_m = \{S \in \binom{\mathcal{M}}{L} \ : \ m \in S\}$. We call this model *$L$-identification for sets*.

One could also think of situations where the L objects, which are known to the encoder, need not be pairwise different. We call this *L-identification for vectors*. The model for this is

$$\Pi_L = \{\pi_m \ : \ m \in \mathcal{M}\}, \tag{2}$$

where $\pi_m = \{A_m, \mathcal{M}^L \setminus A_m\}$ and

$$A_m = \{A \in \mathcal{M}^L \ : \ A \text{ has at least one component equal to } m\}.$$

This can also be applied to K-identification so that we obtain
Model 3': $\Pi_{K,\text{vec}} = \{\pi_A \ : \ A = (a_1, \ldots, a_K) \in \mathcal{M}^K\}$, with

$$\pi_A = \left\{ \bigcup_{i=1}^{K} a_i, \mathcal{M} \setminus \bigcup_{i=1}^{K} \{a_i\} \right\}.$$

This is called *K-identification for vectors* and model 3 *K-identification (for sets)*.

The goal of this thesis is the analysis of L-identification in the case of noiseless coding. We call it *L-identification for sources*. However, the concept of L-identification may also be considered in the case of noisy coding. Moreover, we mainly focus on L-identification for vectors. Thus, if we speak in the remainder of L-identification, we shall always mean L-identification for vectors.

The first section provides basic definitions and notation. In the first subsection of Sect. 2 we give a short introduction into source coding. A *discrete source* is a pair $(\mathcal{U}, P)$, where the *output space* $\mathcal{U}$ is a finite set of cardinality N and P is a probability distribution on $\mathcal{U}$. Further, a *discrete memoryless source* is a pair $(\mathcal{U}^n, P^n)$, where $\mathcal{U}^n$ is the Cartesian product of a finite set $\mathcal{U}$. P^n is a probability distribution on $\mathcal{U}^n$, where the probability of an element $u^n \in \mathcal{U}^n$ is product of the probabilities of its individual components. We further explain what we mean by the *code tree* $T_{\mathcal{C}}$, which corresponds to a given source code $\mathcal{C}$, and provide some notation.

In the second subsection of Sect. 2 we formally define L-identification for sources. Let $L \in \mathbb{N}$ and $(\mathcal{U}^L, P^L)$ be a discrete memoryless source. Due to external constraints (e.g. hardware limitations) all possible *outputs* $u^L = (u_1, \ldots, u_L) \in \mathcal{U}^L$ have to be encoded. This is done by a *q-ary source code* $\mathcal{C}$ on $\mathcal{U}$. That is, every component u_i of u^L is encoded separately.

Following the model in Eq. (2) the goal of L-identification is that every *user* $v \in \mathcal{U}$ shall be able to distinguish whether or not he or she occurs at least once as a component of the output vector u_L. Therefore, we encode all users with the same source code $\mathcal{C}$ and compare sequentially the q-bits of the codeword c_v of the user v with the individual q-bits of the codewords $c_{u_1}, \ldots, c_{u_L}$ of the components of u^L. After every comparison we delete all output components, whose codewords did not coincide during this step with the codeword c_v, from the set of possible candidates. If after some steps all codewords have been eliminated, the L-identification process

terminates with a negative answer. Otherwise we go on until the last q-bit of c_v. The L-identification process terminates with a positive answer if after this last comparison there still are possible candidates left.

The *running time* of q-ary L-identification for given output vector u^L and user v with respect to some code $\mathcal{C}$ is defined as the number of steps until the L-identification process terminates. Since we are given a probability distribution P^L on $\mathcal{U}^L$, we can calculate the mean of the L-identification running time. We call it the *average running time*.

We are interested in several behaviors of the average running time. The first is the *worst-case (average) running time* where we maximize the average running time over all users $v \in \mathcal{U}$. Suppose we have given another probability distribution Q on the set of users $\mathcal{U}$. In this case we calculate the mean of the average running time. This is called the *expected (average) running time*. A special case of this is when $Q = P$. Then we speak of the *symmetric (average) running time*.

We note that the above approach to analyze L-identification can also be used for noiseless K-identification. The only difference between the two models is on which side the L (resp. K) objects are. For L-identification they are on the side of the encoder and for K-identification they are on the side of the decoder. Thus, an immediate conclusion is that the symmetric running time of L- and K-identification is the same if $L = K$. In case of the expected running time we also would have to exchange the probability distributions P and Q. For the worst-case running time such a direct connection has still to be proven.

We begin our analysis of L-identification in Sect. 3 with two new results for the case $L = 1$. This corresponds to identification for sources, which was introduced before. During this thesis we refer to *(1-)identification for sources* if we speak of identification for sources in order to indicate that identification is a special case of L-identification.

The first result in the first subsection of Sect. 3 concerns the case when the q-ary source code $\mathcal{C}$ is a *saturated block code*. This means that all codewords have the same length n and the number of elements equals q^n. We show that for such codes the uniform distribution is optimal for the symmetric running time of (1-)identification. The main part of this subsection is Lemma 220 where we provide a modification for a given probability distribution. If this modification is applied iteratively, it results in the uniform distribution and does not increase the symmetric running time of (1-)identification. This result is used in chapter "An Interpretation of Identification Entropy" in the proof of their Lemma 214.

In Theorem 196 of chapter "Identification for Sources" it is proven that the worst-case running time of binary (1-)identification can be upper bounded by 3 no matter of how big the output space $\mathcal{U}$ is. This was done by an inductive code construction. We show in the second subsection of Sect. 3 how this upper bound can be improved by a slight change of their code construction.

In Sect. 4 we analyze the asymptotic behavior of the symmetric running time of L-identification for the case that P is the uniform distribution. For this we consider the so-called *balanced Huffman codes for the uniform distribution*. These codes

are special cases of the well-known Huffman codes and were introduced in [2] (chapter "Identification Entropy").

In the first subsection of Sect. 4 we point out an interesting connection between balanced Huffman trees and the colexicographic order. This order can be used to construct a balanced Huffman code.

In the second subsection of Sect. 4 we provide Theorem 227, the main result of this section. We prove that if we use balanced Huffman codes for the uniform distribution, the symmetric running time of q-ary L-identification asymptotically equals a rational number $K_{L,q}$, which grows logarithmically in L. In fact, we show that this number is an approximation of the L-th harmonic number.

The main result of this thesis is in Sect. 5 the discovery of the q-ary identification entropy of second degree. We begin this section with the illustration of our approach in finding this entropy function. In order to find a lower bound for 2-identification concerning general distributions we want to apply our asymptotic result of the second subsection of Sect. 4 concerning the uniform distribution. Therefore we first establish a connection between 2-identification inside a given code $\mathcal{C}$ and 2-identification inside the concatenated code $\mathcal{C}^n$. It turns out that not only 2-identification comes into play here but also (1-)identification. In the next step we prove that if n is sufficiently large, 2-identification inside the concatenated code can be lower bounded by 2-identification inside a saturated block code of some given depth. In order to apply Theorem 227 we show that also for 2-identification the uniform distribution is optimal for saturated block codes. With these results we obtain an expression as a lower bound which still depends on (1-)identification. However, the (1-)identification running time appears negatively signed so that we cannot immediately apply its lower bound. This lower bound is the identification entropy $H_{I,q}$ established in [2] (see chapter "Identification Entropy"). During this thesis we refer to $H_{\mathrm{ID}}^{1,q} = H_{I,q}$ since, as we will see, identification entropy is a special case of the q-ary identification entropy of degree L.

In the beginning of the second subsection of Sect. 5 we show that if the underlying probability distribution consists only of q-powers, the previously established lower bound can be attained. This ensures us to define the q-ary identification entropy of second degree by

$$H_{\mathrm{ID}}^{2,q}(P) = 2\frac{q}{q-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^2\right) - \frac{q^2}{q^2-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^3\right).$$

This function obeys some important properties, which appear as desiderata for entropy functions in [1]. It is symmetric, normalized, decisive and expansible. Further, it is lower bounded by the probability distribution where all the probability is concentrated in one point and upper bounded by the uniform distribution. Finally, we establish a grouping behavior, which is a generalization of the grouping behavior of the identification entropy function $H_{\mathrm{ID}}^{1,q}$. With these properties we finally prove that $H_{\mathrm{ID}}^{2,q}$ is indeed a lower bound for the symmetric running time for q-ary 2-identification. Moreover, we show that this bound can be attained if and only if P

consists only of q-powers. As a final result of this subsection we show that balanced Huffman codes are asymptotically optimal for 2-identification.

In the final subsection we provide an upper bound for the worst-case running time by the same code construction which we used in the third subsection of Sect. 5.

In the following Sect. 6 we turn to L-identification for general distributions and define the q-ary identification entropy of degree L by

$$H_{\mathrm{ID}}^{L,q}(P) = -\sum_{l=1}^{L}(-1)^l\binom{L}{l}\frac{q^l}{q^l-1}\left(1-\sum_{u\in\mathcal{U}}p_u^{l+1}\right).$$

We show that also this entropy function is symmetric, normalized, decisive and expansible. It further obeys a grouping behavior, which is a generalized version of the previous grouping behavior for $L = 1, 2$. Unfortunately, we were not able to prove a lower and upper bound. There exist counterexamples for which uniform distribution is not an upper bound. These counterexamples only occur if $N < q$, i.e., the size of the output space is strictly less than the alphabet size. However, in order to show that $H_{\mathrm{ID}}^{L,q}$ is a lower bound for L-identification we only need the bounds for the case $N = q$. We prove this relation under the assumption that in this case uniform distribution is indeed an upper bound. If, additionally, we assume that it is the only distribution which attains this upper bound we can show that there exists a code $\mathcal{C}$ with $H_{\mathrm{ID}}^{L,q}(P) = \mathcal{L}_{\mathcal{C}}^{L,q}(P, P)$ if and only if P consists only of q-powers.

In Sect. 7 we turn to another type of identification namely L-identification for sets. We begin by defining L-identification for sets and point out the differences to L-identification (for vectors). After that we show that if we consider the uniform distribution and balanced Huffman codes, the symmetric running time of L-identification for sets asymptotically equals the symmetric running time of L-identification.

In the final Sect. 8 we state some open problems which arose during the analysis of L-identification.

2 Definitions and Notation

In this section we provide definitions and notations, which are the base for all further calculations. The first subsection is a short overview of source coding. We further introduce code trees, which are useful for visualizing behaviors of a given code. In the second subsection we explain the task of an L-identification code and define the performance behaviors in which we are interested in this thesis.

We begin with some set-theoretical notation. The set of the natural numbers 1 to n is denoted by $[n]$ and the set of all natural numbers from $m + 1$ up to n is denoted by $[m + 1, n]$. However, $[0, 1]$ still denotes the closed real interval from 0 to 1. Let $\mathcal{S}$ be any finite set. Then $2^{\mathcal{S}}$ denotes the power set of $\mathcal{S}$, $\binom{\mathcal{S}}{k}$ denotes the

set of all k-element subsets of $\mathcal{S}$ and $\mathcal{S}^* = \bigcup_{d=0}^{\infty} \mathcal{S}^d$. Further, let P be a probability distribution on $\mathcal{S}$. Then, $supp(P) = \{s \in \mathcal{S} \ : \ P(\{s\}) \neq 0\}$ denotes the *support* of P.

We often have to deal with functions whose arguments are probability distributions on some given finite set. Therefore we formally define a domain for these functions. Following [1] (pp. 26) we define

$$\Delta_n = \{(p_1, \ldots, p_n) \in [0,1]^n \ : \ 0 \leq \sum_{i=1}^{n} p_i \leq 1\}$$

to be the set of all (perhaps incomplete) probability distributions on $[n]$ and

$$\Gamma_n = \{(p_1, \ldots, p_n) \in [0,1]^n \ : \ \sum_{i=1}^{n} p_i = 1\}$$

to be the set of all complete probability distributions on $[n]$. If we want to exclude zero probabilities, we write for $n \geq 2$

$$\mathring{\Delta}_n = \{(p_1, \ldots, p_n) \in (0,1)^n \ : \ 0 < \sum_{i=1}^{n} p_i \leq 1\}$$

and

$$\mathring{\Gamma}_n = \{(p_1, \ldots, p_n) \in (0,1)^n \ : \ \sum_{i=1}^{n} p_i = 1\}.$$

It follows immediately from the above definitions that

$$\Gamma_n = \{(p_1, \ldots, p_n) \in (0,1)^n \ : \ (p_1, \ldots, p_{n-1}) \in \Delta_{n-1} \text{ and } p_n = 1 - \sum_{i=1}^{n-1} p_i\}. \tag{3}$$

This means that Γ_n is a $(n-1)$-dimensional hyperplane in the n-dimensional real space. Hence, if we analyze a function $f : \Gamma_n \to \mathbb{R}$ by differentiation, we only have to consider $n-1$ partial derivatives

$$\frac{\delta}{\delta x_j} \tilde{f}\,(p_1, \ldots, p_{n-1})\,,$$

with $j \in [n-1]$ and where $\tilde{f}(p_1, \ldots, p_{n-1}) = f(p_1, \ldots, p_{n-1}, 1 - \sum_{i=1}^{n-1} p_i)$.

For a mapping $f : \Gamma_n \to \mathbb{R}$ we will write $f(P) = f(p_1, \ldots, p_N)$. Thus, omitting the additional brackets on the right hand side. For a function $g : \Gamma_n^2 \to \mathbb{R}$, however, we retain the brackets and write $g(P, R) = g((p_1, \ldots, p_N), (r_1, \ldots, r_N))$.

2.1 Source Coding and Code Trees

A *discrete source* is a probability space $(\mathcal{U}, 2^{\mathcal{U}}, P)$, where $\mathcal{U}$ is a finite set, called the *output space*. W.l.o.g. we assume that $\mathcal{U} = [N]$ for some $N \in \mathbb{N}$. Further, P is a probability distribution on $\mathcal{U}$ with $p_u = P(\{u\})$. It is called the *output probability distribution*. Often, the indication of $2^{\mathcal{U}}$ is omitted and we will follow this standard and call $(\mathcal{U}, P)$ a discrete source with output space $\mathcal{U} = [N]$ and output probability P. We further introduce the *output RV* $U = id_{\mathcal{U}}$. It follows that $\Pr(U = u) = p_u$.

A discrete *memoryless* source $(\mathcal{U}^n, P^n)$ is characterized by $P_{u^n} = P^n(\{u^n\}) = \prod_{i=1}^n p_{u_i}$ for all $u^n = (u_1, u_2, .., u_n)$. $U^n = id_{\mathcal{U}^n}$ is the output RV for this discrete memoryless source.

For the *alphabet* $\mathcal{Q} = \{0, 1, \ldots, q-1\}$ a mapping $\mathcal{C} : \mathcal{U} \to \mathcal{Q}^*$ is a called *q-ary code* on $\mathcal{U}$ and $\mathcal{C}(u) = c_u = c_{u,1}c_{u,2}\ldots c_{u,\|c_u\|}$ is the q-ary codeword of $u \in \mathcal{U}$. The individual $c_{u,i} \in \mathcal{Q}$ are called q-bits. We also write shortly that $\mathcal{C} = \{c_1, \ldots, c_N\}$. Further, for $u \in \mathcal{U}$ and $k \in [\|c_u\| - 1]$ we define $c_u^k = c_{u,1}\ldots c_{u,k}$ to be the *prefix* of length k of the codeword c_u. In addition we set $c_u^0 = e$, where e is the empty codeword.

A code is called a *prefix code* if no codeword is prefix of another. Formally, for each $c \in \mathcal{C}$ let

$$D(c) = \bigcup_{k=0}^{\|c\|-1} c^k. \tag{4}$$

Then, $\mathcal{C}$ is a prefix code if and only if it holds for all $c, c' \in \mathcal{C}$ that $c \notin D(c')$. For more information on prefix codes we refer to [17]. Hereafter, unless otherwise specified, by a code we shall always understand a prefix code. We also define for some code $\mathcal{C}$ the set of all prefixes of its codewords by

$$D(\mathcal{C}) = \bigcup_{c\in\mathcal{C}} D(c). \tag{5}$$

A *block code* is a code where all codewords have the same length. We further use $\mathcal{C}_{q^n}$ to denote the q-ary block code of size q^n. It is a special block code and called *saturated*.

It is often useful to visualize a code by its *code tree*. Therefore consider a q-ary tree, where all branches with the same branching point are labeled with elements of $\mathcal{Q}$. Such a tree is a code tree $T_{\mathcal{C}}$ of a code $\mathcal{C}$ if there exists a bijective mapping ϕ from the set $\bar{\mathcal{N}}(T_{\mathcal{C}})$ of leaves of $T_{\mathcal{C}}$ onto $\mathcal{C}$ such that $\phi(x)$ equals the labeled path

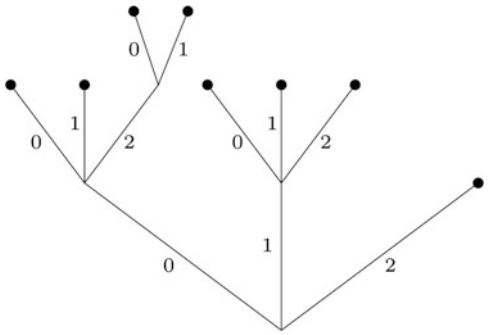


Fig. 1 The ternary code tree of $\mathcal{C} = \{00, 01, 020, 021, 10, 11, 12, 2\}$

from the root of $T_\mathcal{C}$ to leaf x for all $x \in \bar{\mathcal{N}}(T_\mathcal{C})$. Figure 1 shows an example of a code and its corresponding code tree.

We have already used the expression $\bar{\mathcal{N}}(T)$ for the set of leaves or external nodes. In addition, we use $\mathring{\mathcal{N}}(T)$ for the set of branching points or inner nodes of a tree T and $\mathcal{N}(T) = \bar{\mathcal{N}}(T) \cup \mathring{\mathcal{N}}(T)$. The bijective mapping ϕ from before can be extended to $\mathcal{N}(T)$ by mapping every inner node $x \in \mathring{\mathcal{N}}(T)$ to the element in $D(\mathcal{C})$ which corresponds to the labeled path from the root of T to x. Because of this direct connection we do not distinguish between a code and its code tree. We will use $\mathcal{C}$ and $\bar{\mathcal{N}}(T_\mathcal{C})$ equivalently and the same we do for $D(\mathcal{C})$ and $\mathring{\mathcal{N}}(T_\mathcal{C})$.[1] That is, we equivalently use x and $\phi(x)$. For example, $\|x\| = \|\phi(x)\|$. Further, T_x (or $T_{\phi(x)}$) denotes the subtree of T with root in x for some node $x \in \mathcal{N}(T)$. If $\|x\| = 0$, then $T_x = T_e = T$, and if $x \in \bar{\mathcal{N}}(T)$, then $T_x = x$.

Let $\mathcal{C}$ be a source code for the source $(\mathcal{U}, P)$. The *concatenated code* $\mathcal{C}^n$ for the source $(\mathcal{U}^n, P^n)$ is defined as follows. The codeword for each output $u^n = (u_1, \ldots, u_n)$ is the concatenation of the individual codewords of the u_i's. That is

$$c_{u^n} = c_{u_1} \ldots c_{u_n}.$$

If we consider a concatenated code $\mathcal{C}^n$, then $\mathcal{C}$ is called the *basic code*. $\mathcal{C}^n$ can also be obtained by a stepwise construction. Therefore consider the code tree $T_\mathcal{C}$. For each *concatenation step* $1 \leq t \leq n-1$ the new code tree $T_{\mathcal{C}^{t+1}}$ is obtained by replacing each of the leaves of $T_{\mathcal{C}^t}$ with a copy of $T_\mathcal{C}$. Figure 2 shows the first concatenation step of a binary code by means of its code tree. Every node of the concatenated tree where two basic trees are connected is called a *concatenation point*.

[1] This can only be done because we consider prefix codes.

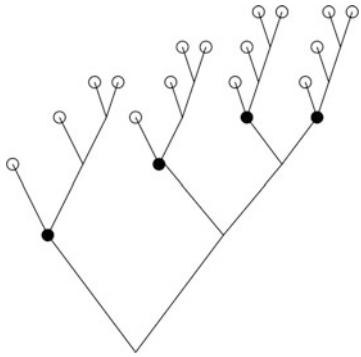


Fig. 2 The concatenated tree $T_{\mathcal{C}^2}$ corresponding to the binary code $\mathcal{C} = \{0, 10, 110, 111\}$

2.2 **L-*Identification***

Consider the discrete memoryless source $(\mathcal{U}^L, P^L)$ together with a source code $\mathcal{C}$ on $\mathcal{U}$. Additionally and in contrast to classical source coding we also introduce the so-called user space $\mathcal{V}$, with $|\mathcal{V}| = |\mathcal{U}|$, together with the *user RV* $V = id_{\mathcal{V}}$. Let $f : \mathcal{V} \to \mathcal{U}$ be a bijective mapping. We encode the *users* v with the same code $\mathcal{C}$ as before. That is, we set $c_v = c_{f(v)}$. W.l.o.g. we assume from now on that $\mathcal{V} = \mathcal{U}$ and $f = id_{\mathcal{U}}$.

The task of L-identification is to decide for every user $v \in \mathcal{U}$ and every output $u^L = (u_1, .., u_L) \in \mathcal{U}^L$ whether or not there exists at least one $l \in [L]$ such that $v = u_l$. To achieve this goal we compare step by step the first, second, third etc. q-bit of c_v with the corresponding q-bits of $c_{u_1}, \ldots, c_{u_L}$. After each step i all u_l with $c_{u_l,i} \neq c_{v,i}$ are eliminated from the set of possible candidates. We continue with step $i + 1$ comparing only those u_l which still are candidates. If at some point during this procedure the last possible candidate is eliminated, the L-identification process stops and returns "No, v is not contained in u^L.". On the other hand, if there are still candidates after the comparison of the last q-bit of c_v, the L-identification process also halts but returns "Yes, v is contained in u^L at position(s) ...". The number of steps until the process halts is called the *L-identification running time* for $(u^L, v) \in \mathcal{U}^L \times \mathcal{U}$.

The algorithm `LID` presented in the appendix in Table 2 accomplishes L-identification. As its input serve the codewords $c_{u_1}, \ldots, c_{u_L}$ and c_v and it returns the triple $(A, s, \mathcal{S})$. Here A is a Boolean variable which is "TRUE" if v is contained in u^L and "FALSE" if not. The second component s equals the number of steps until the algorithm halted and the third component returns the set of positions of the output vector u^L which coincide with the user v. This means that if there exist one or more components of u^L which coincide with v, we also know their exact number and positions. This is not a requirement to L-identification but an extra feature. It follows from the fact that up to the last comparison of q-bits still all possible candidates may not coincide with v.

In Sect. 7 we turn to L-identification for sets and there this feature is not attained since we know that all still possible candidates are pairwise distinct. This means that in some cases L-identification for sets can be faster than L-identification (for vectors). In this case, however, we do not know where the particular user occurred. We explain what we mean by L-identification for sets and point out the differences in greater detail in Sect. 7.

Formally, we define the L-identification running time for given u^L, v and q-ary code $\mathcal{C}$ by

$$\mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) = \mathtt{LID}_2(c_{u_1}, \ldots, c_{u_L}, c_v), \tag{6}$$

where $\mathtt{LID}_2(c_{u_1}, \ldots, c_{u_L}, c_v)$ is the second component of the triple returned by the algorithm LID.

The goal of this thesis is to analyze the expected length of the L-identification running time, also called the *average running time*, for a given user $v \in \mathcal{U}$

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P, v) = \sum_{u^L \in \mathcal{U}^L} P_{u^L} \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v). \tag{7}$$

This can be done in different ways. The first is the worst-case scenario where we are interested in the *worst-case average running time*, which we shortly call the *worst-case running time*,

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P) = \max_{v \in \mathcal{U}} \mathcal{L}_{\mathcal{C}}^{L,q}(P, v). \tag{8}$$

We want to find codes which are as close as possible to the *optimal worst-case running time*

$$\mathcal{L}^{L,q}(P) = \min_{\mathcal{C}} \mathcal{L}_{\mathcal{C}}^{L,q}(P). \tag{9}$$

In the second subsection of Sect. 3 and the third subsection of Sect. 5 we provide upper bounds for $\mathcal{L}^{1,2}(P)$ and $\mathcal{L}^{2,2}(P)$.

Let us assume that also user v is chosen at random according to a probability distribution Q on $\mathcal{U}$. We are now interested in the *expected average running time* or shortly the *expected running time*

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P, Q) = \sum_{v \in \mathcal{U}} Q(\{v\}) \mathcal{L}_{\mathcal{C}}^{L,q}(P, v) \tag{10}$$

and in particular in the *optimal expected running time*

$$\mathcal{L}^{L,q}(P, Q) = \min_{\mathcal{C}} \mathcal{L}_{\mathcal{C}}^{L,q}(P, Q). \tag{11}$$

In this thesis we focus on the special case where $Q = P$ so that Eqs. (10) and (11) become

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P, P) = \sum_{v \in \mathcal{U}} p_v \mathcal{L}_{\mathcal{C}}^{L,q}(P, v) = \sum_{(u^L, v) \in \mathcal{U}^{L+1}} P_{u^L} p_v \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) \tag{12}$$

and

$$\mathcal{L}^{L,q}(P, P) = \min_{\mathcal{C}} \mathcal{L}_{\mathcal{C}}^{L,q}(P, P). \tag{13}$$

We call $\mathcal{L}_{\mathcal{C}}^{L,q}(P, P)$ the *symmetric running time* for a given code $\mathcal{C}$ and $\mathcal{L}^{L,q}(P, P)$ the *optimal symmetric running time*. In Sect. 5 we derive an entropy function for 2-identification. This function provides a lower bound for $\mathcal{L}^{2,q}(P, P)$. In Sect. 6 we discuss an extension of this approach to the case of L-identification for general L. It is clear from the above definitions that

$$\mathcal{L}^{L,q}(P, P) \leq \mathcal{L}^{L,q}(P) \tag{14}$$

so that the bounds we derive in Sect. 6 and the second subsection of Sect. 3, and the second and third subsection of Sect. 5 are lower (resp. upper) bounds for both values.

All the above values also depend on $N = |\mathcal{U}|$. We do not state this fact explicitly since it is contained in both P and $\mathcal{C}$.

3 Two New Results for (1-)Identification

In this section we state two new results for (1-)identification. The first result is about (1-)identification for block codes. In chapter "An Interpretation of Identification Entropy" it is proven that the q-ary identification entropy $H_{\mathrm{ID}}^{1,q}(P)$ is a lower bound for $\mathcal{L}_{\mathcal{C}}^{1,q}(P, P)$. A key step in this proof is to show that if $\mathcal{C}$ is a saturated block code, the running time of identification is minimized by the uniform distribution. This result is provided in the first subsection. Although this may seem obvious the proof is not trivial. Moreover, we will see in Sect. 6 that at least for $L \geq 4$ the uniform distribution is not always optimal for L-identification on block codes.

The second result is about upper bounds for the worst-case running time. In Sect. 5 of chapter "Identification for Sources" it is proved in Theorem 196 that $\mathcal{L}^{1,2}(P) < 3$ by an inductive code construction. We discovered that with a small alteration of their construction this upper bound can be strengthened.

3.1 (1-)Identification for Block Codes

In order to show that the uniform distribution is optimal for (1-)identification on block codes we modify any given probability distribution step by step until we reach the uniform distribution without increasing $\mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(P, P)$. It turns out that not only the uniform distribution is optimal. In fact, all distributions $P = (p_1, \ldots, p_{q^n})$ are optimal for which we are able to partition $\mathcal{U} = [q^n]$ into sets $\mathcal{U}_1, \ldots, \mathcal{U}_{q^{n-1}}$, all of cardinality q, such that $\sum_{u \in \mathcal{U}_i} p_u = 1/q^{n-1}$ for all $i \in [q^{n-1}]$. This is due to the fact that the running time regarding v is the same for all u whose codewords c_u coincide with c_v in all but the last q-bit. The individual steps of modification and their monotone decreasing property are content of

Lemma 220 *Let $n \in \mathbb{N}$, $q \in \mathbb{N}_{\geq 2}$, $k \in \{0, \ldots, n-1\}$ and $t \in \{0, \ldots, q^{n-k-1} - 1\}$. Further, let $P = (p_1, \ldots, p_{q^n})$ and $\tilde{P} = (\tilde{p}_1, \ldots, \tilde{p}_{q^n})$ be probability distributions on $[q^n]$ with*

$$P = (p_1, \ldots, p_{tq^{k+1}}, \underbrace{r_1, \ldots, r_1}_{q^k}, \underbrace{r_2, \ldots, r_2}_{q^k}, \ldots, \underbrace{r_q, \ldots, r_q}_{q^k}, p_{(t+1)q^{k+1}+1}, \ldots, p_{q^n})$$

and

$$\tilde{P} = (p_1, \ldots, p_{tq^{k+1}}, \underbrace{\frac{1}{q}\sum_{i=1}^{q} r_i, \ldots, \frac{1}{q}\sum_{i=1}^{q} r_i}_{q^{k+1}}, p_{(t+1)q^{k+1}+1}, \ldots, p_{q^n}).$$

Then it holds

$$\mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(P, P) - \mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(\tilde{P}, \tilde{P}) = \frac{q^k(q^k - 1)}{2(q-1)} \sum_{i,j=1}^{q} (r_i - r_j)^2 \geq 0.$$

The inequality holds with equality if and only if either $k = 0$ or $r_i = r_j$ for all $i, j \in [q]$.

Proof W.l.o.g. we assume that $t = 0$, such that

$$P = (p_1, \ldots, p_{q^n}) = (r_1, \ldots, r_1, r_2, \ldots, r_2, \ldots, r_q, \ldots, r_q, p_{q^{k+1}+1}, \ldots, p_{q^n})$$

and

$$\tilde{P} = (\tilde{p}_1, \ldots, \tilde{p}_{q^n}) = (\frac{1}{q}\sum_{i=1}^{q} r_i, \ldots, \frac{1}{q}\sum_{i=1}^{q} r_i, p_{q^{k+1}+1}, \ldots, p_{q^n})$$

Also, we use for simplicity the abbreviations

$$L_{u,v} = \mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(u, v) \qquad \alpha_{u,v} = (p_u p_v - \tilde{p}_u \tilde{p}_v) L_{u,v}.$$

It is clear that $L_{u,v} = L_{v,u}$ and hence $\alpha_{u,v} = \alpha_{v,u}$. Also, $\alpha_{u,v} = 0$ for all $u, v \in [q^{k+1} + 1, q^n]$. This yields

$$\mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(P, P) - \mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(\tilde{P}, \tilde{P}) = \sum_{u,v=1}^{q^n} \alpha_{u,v} = \sum_{u,v=1}^{q^{k+1}} \alpha_{u,v} + 2 \sum_{u=1}^{q^{k+1}} \sum_{v=q^{k+1}+1}^{q^n} \alpha_{u,v}.$$

It further holds for $u \in [q^{k+1}]$ and $v \in [q^{k+1} + 1, q^n]$ that

(i) $p_v = \tilde{p}_v$,
(ii) $L_{u,v} = L_{1,v}$, which we denote by L_v,
(iii) $\tilde{p}_u = \frac{1}{q} \sum_{i=1}^{q} r_i$ and
(iv) $\sum_{u=1}^{q^{k+1}} p_u = q^k \sum_{i=1}^{q} r_i$.

From (iii) and (iv) it follows that

$$\sum_{u=1}^{q^{k+1}} \sum_{v=q^{k+1}+1}^{q^n} \alpha_{u,v} = \sum_{v=q^{k+1}+1}^{q^n} p_v L_v \sum_{u=1}^{q^{k+1}} (p_u - \tilde{p}_u) = 0$$

and hence

$$\begin{aligned}
&\mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(P, P) - \mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(\tilde{P}, \tilde{P}) \\
&= \sum_{u,v=1}^{q^{k+1}} \left[p_u p_v - \frac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] L_{u,v} \\
&= \sum_{j,m=1}^{q} \left[r_j r_m - \frac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] \sum_{u=(j-1)q^k+1}^{jq^k} \sum_{v=(m-1)q^k+1}^{mq^k} L_{u,v}.
\end{aligned}$$

Here, the first equality follows from (iii) and the definition of $\tilde{P}$. The second equality is due to the definition of P.

We now take a look at $L_{u,v}$ and see that for $u \in [(j-1)q^k + 1, jq^k]$ and $v \in [(m-1)q^k + 1, mq^k]$ we have

$$L_{u,v} = \begin{cases} n - k & \text{if } j \neq m \\ n - k + \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u, v) & \text{if } j = m. \end{cases}$$

With this observation we get

$$\mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(P,P) - \mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(\tilde{P},\tilde{P}) \tag{15}$$

$$= (n-k)q^{2k} \sum_{j,m=1}^{q} \left[r_j r_m - \frac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] \tag{16}$$

$$+ \sum_{j=1}^{q} \left[r_j^2 - \frac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] \sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) \tag{17}$$

$$= \sum_{j=1}^{q} \left[r_j^2 - \frac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] q^k \left[(q-1)q^k \sum_{l=1}^{k} l q^{-l} + k \right] \tag{18}$$

$$= \frac{q}{q-1} q^k (q^k - 1) \sum_{j=1}^{q} \left[r_j^2 - \frac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right]. \tag{19}$$

The first equality follows from the additional fact that $\sum_{u,v=(j-1)q^k+1}^{jq^k} L_{u,v}$ is invariant in the choice of $j \in [q]$. The partial sum behavior of the geometric series yields the third equality. To understand the second equality we see that

$$\sum_{j,m=1}^{q} \left[r_j r_m - \frac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] = \sum_{j,m=1}^{q} r_j r_m - \left(\sum_{i=1}^{q} r_i \right)^2 = 0.$$

In addition, we have that

$$\sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) = \sum_{v=1}^{q^k} \sum_{l=1}^{k} l |\{u \in [q^k] \,:\, \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) = l\}|.$$

For $l = 1, \ldots, k-1$ the codeword of each element in the above sets has to coincide with c_v in the first $l-1$ q-bits. Those are q^{k-l+1} many. Furthermore, each one of those codewords has to differ from c_v in the lth q-bit. These are $q-1$ out of q. We end up with $(q-1)q^{k-l}$ elements. If $l = k$, also v itself is contained in the corresponding set. As one can see, this is invariant of the choice of $v \in [q^k]$. It follows that

$$\sum_{v=1}^{q^k} \sum_{l=1}^{k} l |\{u \in [q^k] \,:\, \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) = l\}| = q^k \left[\sum_{l=1}^{k-1} l(q-1)q^{k-l} + kq \right]$$

$$= q^k \left[(q-1)q^k \sum_{l=1}^{k} l q^{-l} + k \right].$$

This proves the second equality of Eq. (19). Finally, since

$$\sum_{j=1}^{q}\left[r_j^2-\frac{1}{q^2}\left(\sum_{i=1}^{q}r_i\right)^2\right]=\frac{1}{2q}\sum_{i,j=1}^{q}(r_i-r_j)^2,$$

we obtain the expression to be proven. □

Lemma 220 provides a way to come step by step from any given distribution $P=(p_1,\ldots,p_{q^n})$ to the uniform distribution without increasing the symmetric 2-identification running time on q-ary block codes. In the first step ($t=0$) of the first round ($k=0$) we level out the probabilities $p_1,\ldots,p_q$. In the second step ($t=1$, $k=0$) we level out $p_{q+1},\ldots,p_{2q}$ and so on until in the last step ($t=q^{n-1}-1$) of the first round the remaining probabilities p_{q^n-q+1} up to p_{q^n} are leveled out. We have not changed the symmetric 2-identification running time, and we have constructed a probability distribution which enables us to go on with Lemma 220. This is due to the fact that the first q, the second q up to the last q probabilities are now identical. In round 2 ($k=1$) we begin to level out the first q^2 probabilities, then the second q^2 probabilities up to the last q^2. During these actions Lemma 220 ensures us that the symmetric 2-identification running time does not increase. Again we end up with a distribution which allows us to apply Lemma 220 also in the third round $k=2$ and so on. Finally, in the last round $k=n-1$ we level out the first q^{n-1} identical probabilities and the second and last q^{n-1} identical probabilities and end up with the uniform distribution. We have proven the following

Corollary 221 *Let $n\in\mathbb{N}$ and $q\in\mathbb{N}_{\geq 2}$. Further, let $\mathcal{C}=\mathcal{C}_{q^n}$ and $T=T_{\mathcal{C}}$. Then, for all probability distributions P on $[q^n]$ it holds that*

$$\mathcal{L}_{\mathcal{C}}^{1,q}(P,P)\geq\mathcal{L}_{\mathcal{C}}^{1,q}\left(\left(\frac{1}{q^n},\ldots,\frac{1}{q^n}\right),\left(\frac{1}{q^n},\ldots,\frac{1}{q^n}\right)\right),$$

with equality if and only if $P(T_x)=q^{-\|x\|}$ for all inner nodes $x\in\mathring{\mathcal{N}}(T)$.

3.2 An Improved Upper Bound for Binary Codes

In 4 of chapter "Identification for Sources" we proved in Theorem 196 that $\mathcal{L}^{1,2}(P)<3$ by an inductive code construction. They assumed that w.l.o.g. $p_1\geq p_2\geq\ldots\geq p_N$. In the first step $\mathcal{U}$ is partitioned into $\mathcal{U}_0=[t]$ and $\mathcal{U}_1=[t+1,N]$ such that $\sum_{i=1}^{t}p_i$ is as close as possible to $1/2$. Then, they inductively construct code on $\mathcal{U}_0$ and $\mathcal{U}_1$. Finally, that they prefixed the codewords for all elements in $\mathcal{U}_0$ (resp. $\mathcal{U}_1$) by **0** (resp. **1**).

The proof of this theorem contains some cases differentiation. The worst of these cases is that $\sum_{i=1}^{t} p_i < \frac{1}{2}$ and the user $v_{\max}$ which maximizes $\mathcal{L}_{\mathcal{C}}^{1,2}(P, v)$ is in $\mathcal{U}_1$.[2] In this case we may take up to a certain number additional outputs from $\mathcal{U}_1$ and put them into $\mathcal{U}_0$ in order to speed up the identification process. To do so we define

$$\mathcal{U}_{\max} = \{u \in \mathcal{U} \ : \ c_{u,1} = c_{v_{\max},1}\} \tag{20}$$

and

$$p_{\max} = \sum_{u \in \mathcal{U}_{\max}} p_u. \tag{21}$$

Further, $P_{\max}$ is a probability distribution on $\mathcal{U}_{\max}$ defined by

$$P_{\max,u} = \frac{p_u}{p_{\max}} \tag{22}$$

for all $u \in \mathcal{U}_{\max}$ and $\mathcal{C}_{\max}$ is the code on $\mathcal{U}_{\max}$ which we obtain by deleting the leading bit of all c_u's. With these definitions we get that

$$\begin{aligned}
\mathcal{L}_{\mathcal{C}}^{L,q}(P) &= \sum_{u^L \in \mathcal{U}^L} p_{u_1} \dots p_{u_L} \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v_{\max}) \\
&= 1 + \sum_{l=1}^{L} \binom{L}{l} (1 - p_{\max})^{L-l} \sum_{u^l \in \mathcal{U}_{\max}^l} p_{u_1} \dots p_{u_l} \mathcal{L}_{\mathcal{C}}^{l,q}(u^l, v_{\max}) \\
&= 1 + \sum_{l=1}^{L} \binom{L}{l} (1 - p_{\max})^{L-l} p_{\max}^l \mathcal{L}_{\mathcal{C}}^{l,q}(P_{\max}, v_{\max}) \\
&\leq 1 + \sum_{l=1}^{L} \binom{L}{l} (1 - p_{\max})^{L-l} p_{\max}^l \mathcal{L}_{\mathcal{C}_{\max}}^{l,q}(P_{\max}).
\end{aligned} \tag{23}$$

This simplifies for $L = 1$ and $q = 2$ to

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) \leq 1 + p_{\max} \mathcal{L}_{\mathcal{C}_{\max}}^{1,2}(P_{\max}). \tag{24}$$

This equation provides the induction step for the proof of

Theorem 222 *It holds for all probability distributions P on $\mathcal{U}$ that the worst-case running time for binary (1-)identification can be upper bounded by*

$$\mathcal{L}^{1,2}(P) < \frac{5}{2}.$$

[2] $v_{\max}$ may not be unique, but if there are more than one, it does not matter which of these we choose.

Proof W.l.o.g. we assume that $p_1 \geq p_2 \geq \ldots \geq p_N$. For the induction bases $N = 1, 2$ we have that $\mathcal{L}^{1,2}(P) = 1 < 5/2$ for all P. Now let $N > 2$.

1. *Case* $p_1 \geq \frac{1}{2}$.
   In this case we assign $c_1 = \mathbf{0}$ and $\mathcal{U}_1 = \{2, \ldots, N\}$. Inductively we construct a code $\mathcal{C}' = \{c'_u : u = 2, \ldots, N\}$ on $\mathcal{U}_1$ and we extend this code to a code on $\mathcal{U}$ by setting $c_u = \mathbf{1}c'_u$ for $u \in \mathcal{U}_1$.
   It is clear that $v_{\max} \neq 1$ because in this case $\mathcal{L}^{1,2}(P)$ would equal 1. This is a contradiction since $N > 2$ and thereby we have more than one output whose codeword begins with **1** and each of these outputs results in a running time strictly greater than 1.
   Thus, the maximum is assumed on the "right" side. This yields $p_{\max} \leq 1/2$. Further, by Eq. (24) and the induction hypothesis we have that

$$\mathcal{L}^{1,2}_{\mathcal{C}}(P) < 1 + \frac{1}{2} \cdot \frac{5}{2} = \frac{9}{4} < \frac{5}{2}.$$

2. *Case* $p_1 < \frac{1}{2}$.
   In this case we choose t such that $|1/2 - \sum_{u=1}^{t} p_u|$ is minimized. Now we distinguish again between two subcases.

   (a) *Case* $t = 1$.
   In this case we set $\mathcal{U}_0 = \{1, 2\}$ and $\mathcal{U}_1 = \{3, \ldots, N\}$. Again we inductively construct $\mathcal{C}' = \{c'_u : u = 3, \ldots, N\}$. And we obtain $\mathcal{C}$ by setting $c_1 = \mathbf{00}$, $c_2 = \mathbf{01}$ and $c_u = \mathbf{1}c'_u$ for $u = 3, \ldots, N$.
   If $v_{\max} \in \mathcal{U}_0$, we have that $p_{\max} = p_1 + p_2$ and $\mathcal{C}_{\max} = \{\mathbf{0}, \mathbf{1}\}$. Again by Eq. (24) we obtain

$$\mathcal{L}^{1,2}_{\mathcal{C}}(P) \leq 1 + (p_1 + p_2)\mathcal{L}^{1,2}_{\mathcal{C}_{\max}}(P_{\max}) \leq 2 < \frac{5}{2}.$$

   Otherwise it follows from the definition of t that $p_1 + p_2 > 1/2$. By this we get $p_{\max} < 1/2$ and $\mathcal{C}_{\max} = \mathcal{C}_1$. By induction and Eq. (24) this yields

$$\mathcal{L}^{1,2}_{\mathcal{C}}(P) < 1 + \frac{1}{2} \cdot \frac{5}{2} = \frac{9}{4} < \frac{5}{2}.$$

   (b) *Case* $t \geq 2$.
   We now set $\mathcal{U}_0 = \{1, \ldots, t\}$ and $\mathcal{U}_1 = \{t + 1, \ldots, N\}$ and construct inductively codes $\mathcal{C}' = \{c'_u : u = 1, \ldots, t\}$ and $\mathcal{C}'' = \{c''_u : u = t + 1, \ldots, N\}$. We obtain a code $\mathcal{C}$ on $\mathcal{U}$ by setting

$$c_u = \begin{cases} \mathbf{0}c'_u & \text{for } u = 1, \ldots, t \\ \mathbf{1}c''_u & \text{for } u = t + 1, \ldots, N. \end{cases}$$

(i) *Case* $v_{\max} \in \mathcal{U}_0$.
It follows that $p_{\max} = \sum_{u=1}^{t} p_u$. If $\sum_{u=1}^{t} p_u \leq 1/2$, we get again by induction and Eq. (24) that

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) < 1 + \frac{1}{2} \cdot \frac{5}{2} = \frac{9}{4} < \frac{5}{2}.$$

In the case that $\sum_{u=1}^{t} p_u > 1/2$ we have by the definition of t that

$$\sum_{u=1}^{t} p_u - \frac{1}{2} \leq \frac{1}{2} - \sum_{u=1}^{t-1} p_u.$$

It follows $\sum_{u=1}^{t} p_u \leq (p_t + 1)/2$. Additionally, we have $p_{t-1} < 1/(2(t-1))$ because otherwise $\sum_{u=1}^{t-1} p_u \geq 1/2$. This would be a contradiction to the definition of t. This together implies

$$p_{\max} = \sum_{u=1}^{t} p_u < \frac{1 + 2(t-1)}{4(t-1)}. \tag{25}$$

If $t = 2$, we obtain for the same reasons as in Case 2 that

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) < \frac{5}{2}.$$

If $t = 3$, we get that $\mathcal{C}_{\max} = \mathcal{C}' = \{c_1', c_2', c_3'\}$, with $c_1' = \mathbf{0}$, $c_2' = \mathbf{10}$ and $c_3' = \mathbf{11}$. Further, $p_{\max} = p_1 + p_2 + p_3$ and $P_{\max} = (p_1/p_{\max}, p_2/p_{\max}, p_3/p_{\max})$. Since $p_1 \geq p_2 \geq p_3$ it follows that

$$\frac{p_2 + p_3}{p_{\max}} \leq \frac{2}{3}.$$

This yields

$$\mathcal{L}_{\mathcal{C}_{\max}}^{1,2}(P_{\max}) = 1 + \frac{p_2 + p_3}{p_{\max}} \leq \frac{5}{3}.$$

It now follows from Eqs. (24) and (25) that

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) \leq 1 + \frac{5}{3} p_{\max} < 1 + \frac{5}{3} \cdot \frac{5}{8} = \frac{49}{24} < \frac{5}{2}.$$

For $t \geq 4$ the induction hypothesis and Eq. (25) yield

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) < 1 + \frac{1 + 2(t-1)}{4(t-1)} \cdot \frac{5}{2} \leq 1 + \frac{7}{12} \cdot \frac{5}{2} = \frac{59}{24} < \frac{5}{2}.$$

(ii) *Case* $v_{\max} \in \mathcal{U}_1$.

We get that $p_{\max} = \sum_{u=t+1}^{N} p_u$. If $\sum_{u=t+1}^{N} p_u \leq 1/2$, we get like before

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) < 1 + \frac{1}{2} \cdot \frac{5}{2} = \frac{9}{4} < \frac{5}{2}.$$

If $\sum_{u=t+1}^{N} p_u > 1/2$, it follows that

$$\sum_{u=1}^{t} p_u \geq \frac{1}{2} - \frac{1}{2} p_{t+1}.$$

Since $p_{t+1} \leq \left(\sum_{u=1}^{t} p_u\right)/t$, we further obtain

$$\sum_{u=1}^{t} p_u \geq \frac{t}{2t+1} \geq \frac{2}{5}. \tag{26}$$

Since $p_{\max} = 1 - \sum_{u=1}^{t} p_u$, we finally get by induction and Eq. (26) that

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) < 1 + \frac{3}{5} \cdot \frac{5}{2} = \frac{5}{2}.$$

□

From Theorem 98 in chapter "One Sender Answering Several Questions of Receivers", Part II, and Theorem 222 follows

Corollary 223 *It holds for all probability distributions P on $\mathcal{U}$ that*

$$2\left(1 - \sum_{u \in \mathcal{U}} p_u^2\right) \leq \mathcal{L}^{1,2}(P, P) \leq \mathcal{L}^{1,2}(P) < \frac{5}{2}.$$

4 L-Identification for the Uniform Distribution

In the first subsection we point out an interesting connection between the so-called *balanced Huffman codes for the uniform distribution* and the colexicographic order (see e.g. [14]). This order can be used to construct such codes. In the remaining we refer only to balanced Huffman codes and skip the add on "for the uniform distribution". This is somewhat detached from L-identification but since balanced Huffman codes are crucial for the analysis in the second subsection, we feel that this section is the right place to state this result.

We assume familiarity with the concept of Huffman coding (see [12]) and shall start by recalling the concept of balanced Huffman codes, which was introduced in [2]. Let $N = q^{n-1} + d$, where $0 \leq d \leq (q-1)q^{n-1} - 1$. The q-ary Huffman coding for the uniform distribution of size N yields a code where some codewords have length n and the other codewords have length $n-1$. More precisely, if $0 \leq d < q^{n-1}$, then $q^{n-1} - d$ codewords have length $n-1$ and $2d$ codewords have length n, while in the case $q^{n-1} \leq d \leq (q-1)q^{n-1} - 1$ all codewords have length n. It is well-known that for data compression all Huffman codes are optimal. This is not the case for identification.

In [2] (see chapter "Identification Entropy") it is shown (for $q = 2$) that for identification it is crucial which codewords have length n or, in terms of codetrees, where in the codetree these longer codewords lie. Moreover, those Huffman codes have a shorter expected and worst-case running time for which the longer codewords are distributed along the code tree in such a way that for every inner node the difference between the number of leaves of its left side and the number of leaves of its right side is at most one. In chapter "Identification Entropy" Huffman trees satisfying this property were called *balanced*. By analogy, we shall also say that a q-ary Huffman code is balanced if its corresponding q-ary codetree $\mathcal{H}$ obeys the property that for every inner node $x \in \mathring{\mathcal{N}}(\mathcal{H})$ the difference between the number of leaves of $\mathcal{H}_{xi}$ and $\mathcal{H}_{xj}$ is at most one for all $i, j \in \mathcal{Q}$. We further denote by $\mathcal{H}_{q,N}$ the set of all q-ary balanced Huffman trees with N leaves and the corresponding set of q-ary balanced Huffman codes of size N is denoted by $\mathcal{C}_{q,N}$. If $N = q^n$, there exists only a single balanced Huffman code, namely $\mathcal{C}_{q^n}$. We denote the balanced Huffman tree which corresponds to $\mathcal{C}_{q^n}$ by $\mathcal{H}_{q^n}$.

In identification what is relevant is not the length of a codeword but the length of the maximal common prefix of two or more different codewords. This is why a balanced Huffman code is better for identification than an unbalanced one. It is easy to see by the pigeonhole principle that if we consider Huffman codes with codewords of lengths $n-1$ and n, a balanced Huffman code is optimal for the worst-case running time and we will see in the proof of Theorem 227 that the balancing property is also crucial for the symmetric running time of L-identification.

The q-ary Shannon-Fano coding procedure [10] constructs codes where for every inner node the difference between the sum of the normalized probabilities within its individual branches is as close as possible to $1/q$. It is an easy observation that if we are dealing with uniform distributions, a code is a Shannon-Fano code if and only if it is a balanced Huffman code.

The main result of this section is the examination of the asymptotic behavior of $\mathcal{L}_{\mathcal{C}}^{L,q}(P, P)$ for the case when P is the uniform distribution. We shall prove that this is equal to a rational number $K_{L,q}$ (Theorem 227), which grows logarithmically in L. In fact, we show that $K_{L,2}$ approximates the L-th harmonic number. We note that Theorem 227 also plays a major role in the discovery of the identification entropies, which are discussed in Sects. 5 and 6.

4.1 Colexicographic Balanced Huffman Trees

In this subsection we will show how one can construct a balanced Huffman tree for given q, n and $N = q^{n-1} + d$ for some d by applying the colexicographic order. Therefore, let $k = \lfloor d/q^{n-1} \rfloor \leq q - 2$ and $m = d \mod q^{n-1}$. Since a Huffman code contains only codewords of lengths $n-1$ and n, we begin our construction of a balanced Huffman tree with $\mathcal{H}_{q^{n-1}}$ and extend it into the next level by replacing all its leaves with copies of $\mathcal{H}_k$, which we call extension trees. We call this constructed tree the *base tree* $\mathcal{B}$. Obviously, $\mathcal{B}$ is a balanced Huffman tree. We still have m elements left which have to be inserted into the base tree. It remains to determine which ones of the extension trees will be used for this. Of course, every extension tree can only be used once, because otherwise the balancing property would be violated. Before we explain the construction which provides this, we formalize matters.

Let $A \subseteq \bar{\mathcal{N}}(\mathcal{H}_{q^{n-1}})$ be a set of leaves of $\mathcal{H}_{q^{n-1}}$. Then, we define $\mathcal{B}(A)$ to be the tree which we obtain by replacing all the extension trees of the base tree $\mathcal{B}$ with roots in $\mathcal{A}$ by $\mathcal{H}_{k+1}$. Such a set is called a *valid extension set*, if

$$\Big|\, |\mathcal{B}(A)_{x_1 \ldots x_{\|x\|} i}| - |\mathcal{B}(A)_{x_1 \ldots x_{\|x\|} j}| \,\Big| \leq 1 \tag{27}$$

for all $i, j \in \mathcal{Q}$ and all inner nodes $x \in \mathring{\mathcal{N}}(\mathcal{B})$. See Fig. 3 for examples of a valid and an invalid extension. Equivalently we could have defined that A is a valid extension set if

$$\Big|\, |A_{x,i}| - |A_{x,j}| \,\Big| \leq 1 \tag{28}$$

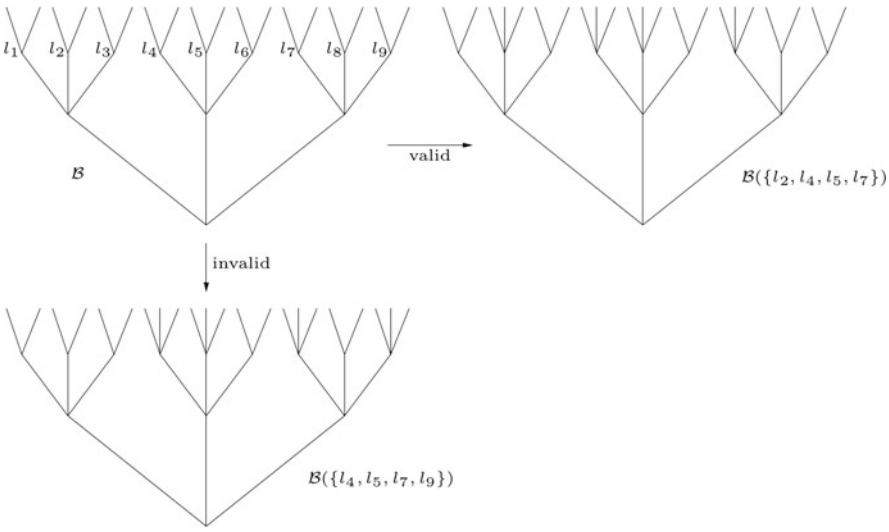


Fig. 3 Examples for a valid and an invalid extension of the ternary base tree $\mathcal{B}$ for $N = 22$

for all $x \in \mathring{\mathcal{N}}(\mathcal{B})$ and all $i, j \in \mathcal{Q}$ and where $A_{x,i} = \{a \in A_x \ : \ a_{\|x\|+1} = i\}$ and $A_x = \{a \in A \ : \ a_1 \dots a_{\|x\|} = x\}$. An immediate conclusion is that if A is a valid extension set, then $\mathcal{B}(A)$ is a balanced Huffman tree.

An easy consequence of the balancing property is the following

Lemma 224 *Let $q^{n-1} < N \leq q^n$, $\mathcal{H} \in \mathcal{H}_{q,N}$ and x be a node of $\mathcal{H}$, then it follows*

$$\left\lfloor \frac{N}{q^{\|x\|}} \right\rfloor \leq |\mathcal{H}_x| \leq \left\lceil \frac{N}{q^{\|x\|}} \right\rceil. \tag{29}$$

The inequality holds with equality for all x if and only if $N = q^n$. Moreover, it simplifies to

$$|\mathcal{H}_x| = q^{n-\|x\|}. \tag{30}$$

For given q and N there may exist many different balanced Huffman trees. We want to point out an interesting case the so-called *colexicographic* balanced Huffman tree. This tree is obtained by taking as the extension set A^{col} the first m codewords of length $n-1$ in colexicographic order.

Let $x, y \in \mathcal{Q}^{n-1}$ and $i_{\max} = \max\{i \in \{1, \dots, n-1\} \ : \ x_i \neq y_i\}$. Then x is said to be less or equal than y in the colexicographic order, denoted by $x \preceq y$, if $x_{i_{\max}} \leq y_{i_{\max}}$. One can easily verify that $(\mathcal{Q}^{n-1}, \preceq)$ is a linearly ordered set since $\mathcal{Q}^{n-1}$ is a product space and the colexicographic order is induced by the trivial linear $\leq$ order on $\mathcal{Q}$. If we denote by c_i the i-th codeword in this order and focus on the k-th q-bits, we observe the following structure.

$$c_{1,k} \dots c_{q^{n-1},k} = \underbrace{Q_k \dots Q_k}_{q^{n-k-1}},$$

where

$$Q_k = \underbrace{0 \dots 0}_{q^{k-1}} \underbrace{1 \dots 1}_{q^{k-1}} \dots \underbrace{(q-1) \dots (q-1)}_{q^{k-1}}.$$

Moreover, the prefixes of length $k-1$ of the codewords within a block Q_k which coincide in the k-th q-bit form the complete $\mathcal{Q}^{k-1}$. And all the codewords in such a block have identical suffixes of length $n-k-1$.

We further define s_k and r_k by $m = s_k q^k + r_k$, where $r_k < q^k$ and $k \in [n-1]$. Finally, r'_k and r''_k are given by $r_k = r'_k q^{k-1} + r''_k$, where $0 \leq r''_k < q^{k-1}$. With this notation we obtain that the k-th q-bits of the first m codewords look like

$$c_{1,k} \dots c_{m,k} = \underbrace{Q_k \dots Q_k}_{s_k} \underbrace{0 \dots 0}_{q^{k-1}} \dots \underbrace{(r'_k - 1) \dots (r'_k - 1)}_{q^{k-1}} \underbrace{r'_k \dots r'_k}_{r''_k}.$$

Let $x \in \mathcal{B}$. With the notation of Eq. (28) we get that A_x^{col} contains exactly q codewords from each of the s_k blocks Q_k each with a different k-th q-bit. In addition, it contains exactly one codeword from each of the small blocks $0 \ldots 0$ to $(r'_k - 1) \ldots (r'_k - 1)$ and at most one codeword from the partial small block $r'_k \ldots r'_k$. This yields

$$|A_{x,i}^{\mathrm{col}}| = \begin{cases} s_k + 1 & \text{if } i = 1, \ldots, r'_k \\ s_k \text{ or } s_k + 1 & \text{if } i = r'_k + 1 \\ s_k & \text{if } i = r'_k + 2, \ldots, q. \end{cases}$$

This together with Eq. (28) shows that A_{col} is a valid extension set. For further information about linear orders see [14].

4.2 An Asymptotic Theorem

The goal of this subsection is to analyze the asymptotic behavior of

$$\mathcal{L}_{\mathcal{C}}^{L,q}\left(\left(\tfrac{1}{N}, \ldots, \tfrac{1}{N}\right), \left(\tfrac{1}{N}, \ldots, \tfrac{1}{N}\right)\right) = \frac{1}{N^{L+1}} \sum_{u_1, \ldots, u_L, v=1}^{N} \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v), \tag{31}$$

with $\mathcal{C} \in \mathcal{C}_{q,N}$. This will be done by applying a different counting method. The above equation suggests to calculate $\mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v)$ for all pairs (u^L, v) individually. Instead we merge all u^L having the same running time regarding some v into sets

$$\mathcal{R}_{\mathcal{C}}^{L,q}(k, v) = \left\{u^L \in \mathcal{U}^L \ : \ \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) = k\right\} \tag{32}$$

for $k \in [\|c_v\|]$. The above defined sets also depend on N. As well as the L-identification functions in the second subsection of Sect. 2 they contain this dependency implicitly via $\mathcal{C}$. Equation (31) now becomes

$$\mathcal{L}_{\mathcal{C}}^{L,q}\left(\left(\tfrac{1}{N}, \ldots, \tfrac{1}{N}\right), \left(\tfrac{1}{N}, \ldots, \tfrac{1}{N}\right)\right) = \frac{1}{N^{L+1}} \sum_{v=1}^{N} \sum_{k=1}^{\|c_v\|} k |\mathcal{R}_{\mathcal{C}}^{L,q}(k, v)|. \tag{33}$$

In order to apply this equation we need to know upper and lower bounds on the cardinalities of these sets. Corollary 226 below provides such bounds and exact values for the case when N is a q-power. The base for this corollary is the following

Lemma 225 *Let $q^{n-1} < N \leq q^n$, $\mathcal{C} \in \mathcal{C}_{q,N}$, $\mathcal{H} = T_{\mathcal{C}}$ and $v \in \mathcal{U}$. Then, for $k \in [\|c_v\| - 1]$ it holds that*

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(k, v)| = \sum_{m=1}^{L} \binom{L}{m} |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}}) \setminus \bar{\mathcal{N}}(\mathcal{H}_{c_v^k})|^m \left(N - |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})|\right)^{L-m}$$

and

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(\|c_v\|, v)| = \sum_{m=1}^{L} \binom{L}{m} |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}})|^m \left(N - |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}})|\right)^{L-m}.$$

Proof In order to simplify notation we shall write $\mathcal{R}(k, v)$ for $\mathcal{R}_{\mathcal{C}}^{L,q}(k, v)$.

1. *Case $k = 1$.*
   The L-identification algorithm terminates after the first step if and only if the codewords of all components of u^L differ already in the first q-bit from $c_{v,1}$. This gives us

   $$\mathcal{R}(1, v) = \left\{u^L \in [q^n]^L \ : \ c_{u_i} \in \bar{\mathcal{N}}(\mathcal{H}) \setminus \bar{\mathcal{N}}(\mathcal{H}_{c_{v,1}}) \ \forall\, i \in [L]\right\}$$

   and therewith

   $$|\mathcal{R}(1, v)| = |\bar{\mathcal{N}}(\mathcal{H}) \setminus \bar{\mathcal{N}}(\mathcal{H}_{c_{v,1}})|^L = (N - |\bar{\mathcal{N}}(\mathcal{H}_{c_{v,1}})|)^L.$$

   This coincides with the first equation of Lemma 225.
2. *Case $k = 2, \ldots, \|c_v\| - 1$.*
   The identification time of u^L and v equals k if and only if it holds for all $i \in [L]$ that $c_{u_i}^k \neq c_v^k$ and that there exists at least one $i \in [L]$ such that $c_{u_i}^{k-1} = c_v^{k-1}$. This consideration yields

   $$\mathcal{R}(k, v) = \Big\{u^L \in [q^n]^L \ : \exists\, i \in [L] \text{ with } c_{u_i} \in \bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}}) \setminus \bar{\mathcal{N}}(\mathcal{H}_{c_v^k}) \text{ and } c_{u_i} \notin \bar{\mathcal{N}}(\mathcal{H}_{c_v^k}) \ \forall\, i \in [L]\Big\}.$$

   In order to count the elements we partition $\mathcal{R}(k, v)$ into L subsets $S_{k,1}, \ldots, S_{k,L}$, where

   $$S_{k,m} = \Big\{u^L \in [q^n]^L \ : \exists\, i_1, \ldots, i_m \in [L] \text{ with } c_{u_{i_1}}, \ldots, c_{u_{i_m}} \in \bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}}) \setminus \bar{\mathcal{N}}(\mathcal{H}_{c_v^k}) \text{ and } c_{u_i} \in \bar{\mathcal{N}}(\mathcal{H}) \setminus \bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}}) \ \forall\, i \in [L] \setminus \{i_1, \ldots, i_m\}\Big\}.$$

If we fix the positions $i_1, \ldots, i_m$, we see that the number of possible vectors is

$$|\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}}) \setminus \bar{\mathcal{N}}(\mathcal{H}_{c_v^k})|^m (N - |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})|)^{L-m}.$$

Since we have no restrictions for these positions, it follows that

$$|S_{k,m}| = \binom{L}{m} |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}}) \setminus \bar{\mathcal{N}}(\mathcal{H}_{c_v^k})|^m \left(N - |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})|\right)^{L-m}.$$

Altogether we obtain

$$\begin{aligned} |\mathcal{R}(k, v)| &= |\bigcup_{m=1}^{L} S_{k,m}| = \sum_{m=1}^{L} |S_{k,m}| \\ &= \sum_{m=1}^{L} \binom{L}{m} |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}}) \setminus \bar{\mathcal{N}}(\mathcal{H}_{c_v^k})|^m \left(N - |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})|\right)^{L-m}. \end{aligned}$$

3. *Case* $k = \|c_v\|$.
In this case also c_v itself may be one of the components of u^L. This yields

$$\mathcal{R}(n, v) = \{u^L \in [q^n]^L \ : \ \exists\, i \in [L] \text{ with } c_{u_i} \in \bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}})\}.$$

According to this we adjust the subsets $S_{n,1}, \ldots, S_{n,L}$, such that

$$\begin{aligned} S_{n,m} = \{u^L \in [q^n]^L \ :& \exists\, i_1, \ldots, i_m \in [L] \text{ with } c_{u_{i_1}}, \ldots, c_{u_{i_m}} \in \bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}}) \\ & \text{and} c_{u_i} \in \bar{\mathcal{N}}(\mathcal{H}) \setminus \bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}}) \ \forall\, i \in [L] \setminus \{i_1, \ldots, i_m\}\}. \end{aligned}$$

Of course, these sets partition $\mathcal{R}(n, 1)$ and since

$$|S_{n,m}| = \binom{L}{m} |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}})|^m (N - |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}})|)^{L-m},$$

for all $m \in [L]$, we obtain the desired result for $|\mathcal{R}(n, v)|$.

□

If we combine Lemma 224 and Lemma 225, we obtain

Corollary 226 *With the same definitions as in Lemma 225 we have the following upper bounds for* $k \in [\|c_v\| - 1]$

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(k, v)| \leq \sum_{m=1}^{L} \binom{L}{m} \left(\left\lceil \frac{N}{q^{k-1}} \right\rceil - \left\lfloor \frac{N}{q^k} \right\rfloor\right)^m \left(N - \left\lfloor \frac{N}{q^{k-1}} \right\rfloor\right)^{L-m}$$

and

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(\|c_v\|, v)| \leq \sum_{m=1}^{L} \binom{L}{m} \left\lceil \frac{N}{q^{\|c_v\|-1}} \right\rceil^m \left(N - \left\lfloor \frac{N}{q^{\|c_v\|-1}} \right\rfloor \right)^{L-m}.$$

Additionally, we get lower bounds for $k \in [\|c_v\| - 1]$

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(k, v)| \geq \sum_{m=1}^{L} \binom{L}{m} \left(\left\lfloor \frac{N}{q^{k-1}} \right\rfloor - \left\lceil \frac{N}{q^k} \right\rceil \right)^m \left(N - \left\lceil \frac{N}{q^{k-1}} \right\rceil \right)^{L-m}$$

and

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(\|c_v\|, v)| \geq \sum_{m=1}^{L} \binom{L}{m} \left\lfloor \frac{N}{q^{\|c_v\|-1}} \right\rfloor^m \left(N - \left\lceil \frac{N}{q^{\|c_v\|-1}} \right\rceil \right)^{L-m}.$$

The above inequalities hold with equality for all $v \in \mathcal{U}$ *if and only if* $N = q^n$. *Moreover, they simplify for all* $k \in [n-1]$ *to*

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(k, v)| = q^{nL} \sum_{m=1}^{L} \binom{L}{m} q^{-km} (q-1)^m (1 - q^{-k+1})^{L-m}$$

and

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(\|c_v\|, v)| = \sum_{m=1}^{L} \binom{L}{m} q^m (q^n - q)^{L-m}.$$

With the above estimates we are now ready to prove the asymptotic theorem for uniform distributions. If we consider the uniform distribution and use a balanced Huffman code for the encoding, the symmetric L-identification running time asymptotically equals a rational number $K_{L,q}$.

Theorem 227 *Let* $L, n \in \mathbb{N}$, $q \in \mathbb{N}_{\geq 2}$, $q^{n-1} < N \leq q^n$, $\mathcal{C} \in \mathcal{C}_{q,N}$ *and* P *be the uniform distribution on* $[N]$. *Then it holds that*

$$\lim_{N \to \infty} \mathcal{L}_{\mathcal{C}}^{L,q}(P, P) = K_{L,q} = -\sum_{l=1}^{L} (-1)^l \binom{L}{l} \frac{q^l}{q^l - 1}.$$

Proof

1. *Case* $N = q^n$.
   It follows from Corollary 226 and Eq. (33) that

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) = \frac{1}{q^{nL}}\left[\sum_{k=1}^{n-1} kq^{nL}\sum_{m=1}^{L}\binom{L}{m}q^{-km}(q-1)^m(1-q^{-k+1})^{L-m} + n\sum_{m=1}^{L}\binom{L}{m}q^m(q^n-q)^{L-m}\right]. \tag{34}$$

   It is easy to check that the second summand together with the leading factor q^{-nL} converges to 0 if n goes to infinity. In fact,

$$\sum_{m=1}^{L}\binom{L}{m}nq^{-m(n-1)}(1-q^{-n+1})^{L-m} \to 0.$$

   This is because $nq^{-m(n-1)} \to 0$ and $(1-q^{-n+1})^{L-m} \to 1$. Thus, we get

$$\begin{aligned}\lim_{n\to\infty}\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) &= \sum_{k=1}^{\infty}k\sum_{m=1}^{L}\binom{L}{m}q^{-km}(q-1)^m(1-q^{-k+1})^{L-m}\\ &= \sum_{m=1}^{L}\sum_{t=0}^{L-m}(-q)^t\binom{L}{m}\binom{L-m}{t}(q-1)^m\sum_{k=1}^{\infty}kq^{-k(m+t)}\\ &= \sum_{m=1}^{L}\sum_{t=0}^{L-m}(-q)^t\binom{L}{m}\binom{L-m}{t}(q-1)^m\frac{q^{m+t}}{(q^{m+t}-1)^2}.\end{aligned} \tag{35}$$

   The second equality follows from $(1-q^{-k+1})^{L-m} = \sum_{t=0}^{L-m}\binom{L-m}{t}(-q)^t q^{-tk}$, while the last equality is a consequence of the geometric series.
   In the following we set $x_{m,t} = (-q)^t\binom{L}{m}\binom{L-m}{t}(q-1)^m$ as well as $z_l = q^l/(q^l-1)^2$ and change the order of summation. This yields

$$\begin{aligned}\lim_{n\to\infty}\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) &= \sum_{m=1}^{L}\sum_{t=0}^{L-m}x_{m,t}z_{m+t} = \sum_{l=1}^{L}z_l\sum_{t=0}^{l-1}x_{l-t,t}\\ &= \sum_{l=1}^{L}\frac{q^l}{(q^l-1)^2}\sum_{t=0}^{l-1}(-q)^t\binom{L}{l-t}\binom{L-l+t}{t}(q-1)^{l-t}\end{aligned}$$

$$= \sum_{l=1}^{L} \binom{L}{l} \frac{q^l}{(q^l-1)^2} \sum_{t=0}^{l-1} \binom{l}{t} (-q)^t (q-1)^{l-t}$$

$$= \sum_{l=1}^{L} \binom{L}{l} \frac{q^l}{(q^l-1)^2} \left((-1)^l - (-q)^l \right)$$

$$= -\sum_{l=1}^{L} (-1)^l \binom{L}{l} \frac{q^l}{q^l-1}.$$

2. *Case* $q^{n-1} < N < q^n$.
For this case we obtain

$$\begin{aligned}
&\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) \\
&\leq \frac{1}{N^{L+1}} \sum_{v=1}^{N} \left[\sum_{k=1}^{\|c_v\|-1} k \sum_{m=1}^{L} \binom{L}{m} \left(\lceil \frac{N}{q^{k-1}} \rceil - \lfloor \frac{N}{q^k} \rfloor \right)^m \left(N - \lfloor \frac{N}{q^{k-1}} \rfloor \right)^{L-m} \right] \\
&+ \frac{1}{N^{L+1}} \sum_{v=1}^{N} \left[\|c_v\| \sum_{m=1}^{L} \binom{L}{m} \lceil \frac{N}{q^{\|c_v\|-1}} \rceil^m (N - \lfloor \frac{N}{q^{\|c_v\|-1}} \rfloor)^{L-m} \right] \\
&\leq \frac{1}{N} \sum_{v=1}^{N} \left[\sum_{k=1}^{\|c_v\|-1} k \sum_{m=1}^{L} \binom{L}{m} \left(\frac{q-1}{q^k} + \frac{2}{N} \right)^m \left(1 - \frac{q}{q^k} + \frac{1}{N} \right)^{L-m} \right] \\
&+ \frac{1}{N} \sum_{v=1}^{N} \left[\|c_v\| \sum_{m=1}^{L} \binom{L}{m} (q^{-\|c_v\|+1} + \frac{1}{N})^m (1 - q^{-\|c_v\|+1} + \frac{1}{N})^{L-m} \right].
\end{aligned} \tag{36}$$

The first inequality is obtained by the insertion of the upper bound in Corollary 226 into Eq. (33). $\lceil N/q^k \rceil \leq N/q^k + 1$ and $\lfloor N/q^k \rfloor \geq N/q^k - 1$ yield the second inequality. We now divide this case into two subcases.

(a) *Case* $2q^{n-1} \leq N < q^n$.
In this case all codewords have length n. Hence Eq. 36 reduces to

$$\begin{aligned}
\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) \leq &\sum_{k=1}^{n-1} k \sum_{m=1}^{L} \binom{L}{m} \left(\frac{q-1}{q^k} + \frac{2}{N} \right)^m \left(1 - \frac{q}{q^k} + \frac{1}{N} \right)^{L-m} \\
&+ n \sum_{m=1}^{L} \binom{L}{m} (q^{-n+1} + \frac{1}{N})^m \left(1 - \frac{q}{q^n} + \frac{1}{N} \right)^{L-m}.
\end{aligned} \tag{37}$$

As in the case $N = q^n$ the second summand goes to zero as N goes to infinity. Thus, we only have to consider the first summand. In fact, we can reduce this case to the previous one by applying the binomial theorem. We obtain

$$\left(\frac{q-1}{q^k}+\frac{2}{N}\right)^m = \left(\frac{q-1}{q^k}\right)^m + \sum_{t=0}^{m-1}\binom{m}{t}\left(\frac{q-1}{q^k}\right)^t\left(\frac{2}{N}\right)^{m-t}$$

and

$$\left(1-\frac{q}{q^k}+\frac{1}{N}\right)^{L-m} = \left(1-\frac{q}{q^k}\right)^{L-m} + \sum_{s=0}^{L-m-1}\binom{L-m}{s}\frac{(1-\frac{q}{q^k})^s}{N^{L-m-s}}.$$

In the following we use

$$A = \sum_{t=0}^{m-1}\binom{m}{t}\left(\frac{q-1}{q^k}\right)^t\left(\frac{2}{N}\right)^{m-t}$$

and

$$B = \sum_{s=0}^{L-m-1}\binom{L-m}{s}(1-q^{-k+1})^s\frac{1}{N^{L-m-s}}.$$

With this notation the right hand side of Eq. (37) asymptotically equals

$$\sum_{k=1}^{n-1} k \sum_{m=1}^{L}\binom{L}{m}\left[\left(\frac{q-1}{q^k}\right)^m\left(1-q^{-k+1}\right)^{L-m} + \left(\frac{q-1}{q^k}\right)^m B + \left(1-q^{-k+1}\right)^{L-m} A + AB\right]. \tag{38}$$

If we focus on the second summand in the square brackets, we see that

$$\begin{aligned}
&\sum_{k=1}^{n-1} k \sum_{m=1}^{L}\binom{L}{m}\left(\frac{q-1}{q^k}\right)^m B \\
&= \sum_{m=1}^{L}\sum_{s=0}^{L-m-1}\binom{L-m}{s}\binom{L}{m}\frac{(q-1)^m}{N^{L-m-s}}\sum_{k=1}^{n-1} kq^{-km}\left(1-\frac{q}{q^k}\right)^{L-m} \\
&= \sum_{m=1}^{L}\sum_{s=0}^{L-m-1}\sum_{r=0}^{L-m}(-q)^r\binom{L-m}{r}\binom{L-m}{s}\binom{L}{m}\frac{(q-1)^m}{N^{L-m-s}}\sum_{k=1}^{n-1}\frac{k}{q^{k(m+r)}} \\
&= \sum_{m=1}^{L}\sum_{s=0}^{L-m-1}\sum_{r=0}^{L-m}\frac{N^{m+s}}{N^L}\frac{\alpha(m,s,r)}{(q^{m+r}-1)^2}\left(q^{m+r}-\frac{(q^{m+r}-1)n+q^{m+r}}{q^{n(m+r)}}\right),
\end{aligned}$$

where $\alpha(m,s,r) = (-q)^r \binom{L-m}{r}\binom{L-m}{s}\binom{L}{m}(q-1)^m$. The last equality follows from the partial sum behavior of the geometric series. This expression tends to zero as N (resp. $n \approx \log_q N$) goes to infinity because $L - m - s \geq 1$.
In the same way it can be shown that the third and the fourth summand of Eq. (38) also tend to zero. Thus, we end up with exactly the same expression like Eq. (35). This proves the upper bound for this case. By using the same arguments and the lower estimates in Corollary 226 one can easily show the matching lower bound.

(b) *Case* $q^{n-1} < N < 2q^{n-1}$.
In this case $N = q^{n-1} + d$, with $0 < d < q^{n-1}$, and there exist exactly $q^{n-1} - d$ codewords of length $n-1$ and $2d$ codewords of length n. Then, Eq. (36) becomes

$$
\begin{aligned}
&\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) \\
&\leq \frac{q^{n-1}-d}{N}\left[\sum_{k=1}^{n-2} k \sum_{m=1}^{L}\binom{L}{m}\left(\frac{q-1}{q^k}+\frac{2}{N}\right)^m\left(1-\frac{q}{q^k}+\frac{1}{N}\right)^{L-m}\right. \\
&\qquad\left. + (n-1)\sum_{m=1}^{L}\binom{L}{m}\left(\frac{q^2}{q^n}+\frac{1}{N}\right)^m\left(1-\frac{q^2}{q^n}+\frac{1}{N}\right)^{L-m}\right] \\
&\quad + \frac{2d}{N}\left[\sum_{k=1}^{n-1} k \sum_{m=1}^{L}\binom{L}{m}\left(\frac{q-1}{q^k}+\frac{2}{N}\right)^m\left(1-\frac{q}{q^k}+\frac{1}{N}\right)^{L-m}\right. \\
&\qquad\left. + n\sum_{m=1}^{L}\binom{L}{m}\left(\frac{q}{q^n}+\frac{1}{N}\right)^m\left(1-\frac{q}{q^n}+\frac{1}{N}\right)^{L-m}\right] \\
&= \sum_{k=1}^{n-2} k \sum_{m=1}^{L}\binom{L}{m}\left(\frac{q-1}{q^k}+\frac{2}{N}\right)^m\left(1-\frac{q}{q^k}+\frac{1}{N}\right)^{L-m} \\
&\quad + \left(\frac{q^n}{q}-d\right)\frac{n-1}{N}\sum_{m=1}^{L}\binom{L}{m}\left(\frac{q^2}{q^n}+\frac{1}{N}\right)^m\left(1-\frac{q^2}{q^n}+\frac{1}{N}\right)^{L-m} \\
&\quad + 2d\frac{n-1}{N}\sum_{m=1}^{L}\binom{L}{m}\left(\frac{q(q-1)}{q^n}+\frac{2}{N}\right)^m\left(1-\frac{q^2}{q^n}+\frac{1}{N}\right)^{L-m} \\
&\quad + 2d\frac{n}{N}\sum_{m=1}^{L}\binom{L}{m}\left(\frac{q}{q^n}+\frac{1}{N}\right)^m\left(1-\frac{q}{q^n}+\frac{1}{N}\right)^{L-m}.
\end{aligned}
$$

For the same reason as in the preceding cases the last three summands tend to zero as $N \to \infty$ and since the first summand asymptotically equals the first summand of Eq. (37), the upper bound also in this last case is settled.

Omitting the details we limit ourselves to remark that also in this case the matching lower bound can be easily obtained by a perfectly analogous argument.

Thus, the proof of the theorem is complete. □

A natural question regards the asymptotic growth of $K_{L,q}$ with respect to L. Table 1 shows some values of $K_{L,2}$. This motivates the assumption that $K_{L,2}$ grows logarithmically in L. In fact, this assumption proves true by the following considerations. First, we see that

$$K_{L,2} = -\sum_{l=1}^{L}(-1)^l\binom{L}{l}\frac{2^l}{2^l-1} = 1 - \sum_{l=1}^{L}\frac{(-1)^l\binom{L}{l}}{2^l-1}.$$

By using the geometric series we get

$$K_{L,2} - 1 = -\sum_{l=1}^{L}\frac{(-1)^l\binom{L}{l}}{2^l}\sum_{k=0}^{\infty}2^{-kl} = -\sum_{k=0}^{\infty}\sum_{l=1}^{L}\binom{L}{l}(-1)^l 2^{-(k+1)l}.$$

The binomial theorem now yields

$$K_{L,2} - 1 = -\sum_{k=0}^{\infty}((1-2^{-(k+1)})^L - 1) = \sum_{k=1}^{\infty}(1-(1-2^{-k})^L).$$

If we now set $\xi_k = (1-2^{-k})$, we obtain

$$\begin{aligned} K_{L,2} - 1 = \sum_{k=1}^{\infty}(1-\xi_k^L) &= \sum_{k=1}^{\infty}(1-\xi_k)(1+\xi_k+\xi_k^2+\ldots+\xi_k^{L-1}) \\ &= \sum_{k=1}^{\infty}\frac{1}{2^k}(1+\xi_k+\xi_k^2+\ldots+\xi_k^{L-1}). \end{aligned}$$

Figure 4 shows that this expression is an approximation by the upper sum of the integral

$$\int_0^1(1+x+x^2+\ldots+x^{L-1})dx = 1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{L} = H_L,$$

Table 1 The growth of $K_{L,2}$ in L

L	1	2	2^2	2^3	2^5	2^{10}	2^{13}
$K_{L,2} \approx$	2	2, 6667	3, 5048	4, 4211	6, 3552	11, 3335	14, 3328
$\frac{K_{L,2}-1}{\log L} \approx$	*	1, 6667	1, 2524	1, 1404	1, 0710	1, 0333	1, 0256

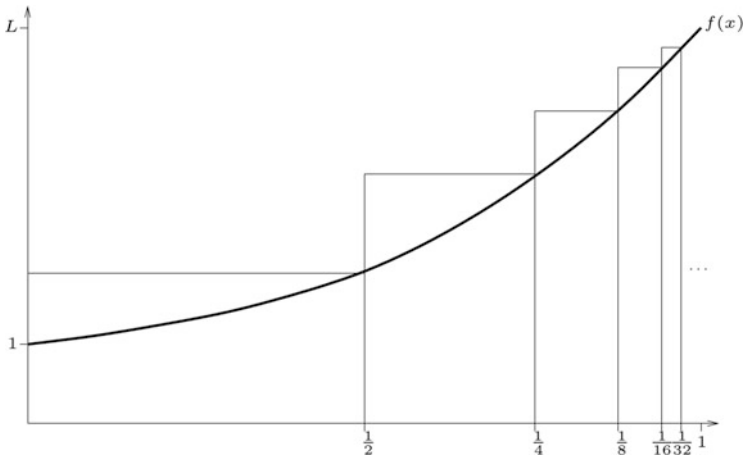


Fig. 4 $K_{L,2} - 1$ approximates the integral of $f(x) = 1 + x + x^2 + \ldots + x^{L-1}$

where H_L denotes the L-th harmonic number. Since H_L grows logarithmically with respect to L, so does $K_{L,2}$.

Strehl [16] generalized this result for the case $q > 2$. His result is the content of the following

Proposition 228 (Strehl 2006, [16]) *It holds that*

$$\lim_{L\to\infty} \frac{H_L}{K_{L,q}} = \ln q,$$

where H_L denotes the L-th harmonic number and ln *is the natural logarithm.*

5 Two-Identification for General Distributions

In the previous section we have seen how L-identification behaves for the uniform distribution. In this section we turn to general distributions and establish a lower bound for 2-identification.

Let us focus on the case $L = 2$, $N = q^n$, $P = (1/q^n, \ldots, 1/q^n)$ and $\mathcal{C} = \mathcal{C}_{q^n}$. Every q-ary comparison, which is done during 2-identification for u^2 and v is itself an l-identification ($l \in [2]$) between the t-th q-bit of the codewords of the l still possible candidates and $c_{v,t}$. The running time of each of those "small" identifications is 1 no matter of the value of l. In fact, we have applied up to n "small" identifications within the code $\mathcal{C}_q$ in order to perform the original 2-identification within $\mathcal{C}_{q^n}$.

It is clear that $\mathcal{C}_{q^n} = \mathcal{C}_q^n$. Further, let $r_{t+1,l}$ be the probability that after the tth comparison there are still l possible candidates left. We can now calculate 2-identification running time within $\mathcal{C}_q^n$ by

$$\begin{aligned}
&\mathcal{L}_{\mathcal{C}_q^n}^{2,q}\left(\left(\frac{1}{q^n},\ldots,\frac{1}{q^n}\right),\left(\frac{1}{q^n},\ldots,\frac{1}{q^n}\right)\right)\\
&\quad= 1+\sum_{t=1}^{n-1}\sum_{l=1}^{2}\binom{2}{l}r_{t+1,l}\mathcal{L}_{\mathcal{C}_q}^{l,q}\left(\left(\frac{1}{q},\ldots,\frac{1}{q}\right),\left(\frac{1}{q},\ldots,\frac{1}{q}\right)\right)\\
&\quad= 1+2\sum_{t=1}^{n-1}r_{t+1,1}+\sum_{t=1}^{n-1}r_{t+1,2}.
\end{aligned}$$

Here, the binomial coefficient in the first equality occurs since in the case $l = 1$ either u_1 or u_2 is the leftover candidate. We have to take into account both possibilities. As stated before l-identification running time within $\mathcal{C}_q$ always equals 1. This proves the second equality. This approach yields an alternative proof of Theorem 227 for $L = 2$ and $|\mathcal{U}| = q^n$. However, we stop this analysis here and will come back to it later.

The above observations lead us to the attempt of doing the same for any given source code $\mathcal{C}$. Namely, to consider the discrete memoryless source $((\mathcal{U}^n)^2, (P^n)^2)$ together with the concatenated code $\mathcal{C}^n$ and try to establish a connection between the 2-identification running time within $\mathcal{C}^n$ and the l-identification running times within $\mathcal{C}$. This relation is the content of Lemma 229. It turns out that we also have to consider (1-)identification within the basic code. This fact makes further analysis more sophisticated, especially for the general case of Sect. 6.

In order to apply Theorem 227 we firstly let n go to infinity. The result of this procedure is stated in Corollary 230. It is a consequence of Lemma 229. Furthermore, we show that from a particular concatenation step on we can lower bound all further concatenated codes to a saturated code $\mathcal{C}_{q^{K_n}}$ of some given length K_n. This is done in the proof of Lemma 233. Finally, Corollary 232 states that the uniform distribution is optimal for 2-identification within a block code. Altogether at the end of the first subsection we obtain

$$\mathcal{L}_{\mathcal{C}}^{2,q}(P,P) \geq (1-\sum_{u\in\mathcal{U}}p_u^3)\left(2\frac{q}{q-1}-\frac{q^2}{q^2-1}\right)-2\left(\frac{1-\sum_{u\in\mathcal{U}}p_u^3}{1-\sum_{u\in\mathcal{U}}p_u^2}-1\right)\mathcal{L}_{\mathcal{C}}^{1,q}(P,P).$$

as a lower bound for 2-identification.

Unfortunately, (1-)identification appears negatively signed so that we cannot immediately apply the lower bound $\mathcal{L}_{\mathcal{C}}^{1,q}(P,P) \geq H_{\mathrm{ID}}^{1,q}(P)$, which has been proven in chapter "An Interpretation of Identification Entropy". In the same work it has been shown that this lower bound is attainable if P consists only of q-powers. Proposition 235 at the beginning of the second subsection proves this equality also for 2-identification. This is the base for the definition of the *q-ary identification*

entropy of second degree

$$H_{\mathrm{ID}}^{2,q}(P) = 2\frac{q}{q-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^2\right) - \frac{q^2}{q^2-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^3\right).$$

In the remaining part of the second subsection we prove some fundamental properties of this function. There are symmetry, expansibility, normalization, decisiveness, bounding between 0 and the uniform distribution and a special grouping behavior. Using these properties we prove Theorem 236 where we show that $H_{\mathrm{ID}}^{2,q}(P)$ is a lower bound for 2-identification. We end this part with a corollary which states that if we consider the uniform distribution on $\mathcal{U}$, balanced Huffman codes are asymptotically optimal for 2-identification.

Finally, we establish an upper bound for the binary case in the third Subsection. The code construction in the proof coincides with the one used for (1)-identification in the second subsection of Sect. 3

5.1 An Asymptotic Approach

Lemma 229 *Let $\mathcal{U}$ be a finite set, $q \in \mathbb{N}_{\geq 2}$, P be a probability distribution on $\mathcal{U}$ and $\mathcal{C}$ be a prefix code. It holds that*

$$\mathcal{L}_{\mathcal{C}^n}^{2,q}(P^n, P^n) = \mathcal{L}_{\mathcal{C}}^{2,q}(P, P)\left(1 + \sum_{t=1}^{n-1}\left(\sum_{u\in\mathcal{U}} p_u^3\right)^t\right)$$
$$+ 2\mathcal{L}_{\mathcal{C}}^{1,q}(P, P)\left(\sum_{t=1}^{n-1}\left(\sum_{u\in\mathcal{U}} p_u^2\right)^t - \sum_{t=1}^{n-1}\left(\sum_{u\in\mathcal{U}} p_u^3\right)^t\right).$$

Proof It is clear that while we are in the first basic tree we have to apply 2-identification and there are three possibilities of what might happen.

1. Both elements u_1^n and u_2^n do not coincide with v^n.
   The reason would be that their first components $u_{1,1}, u_{2,1}$ do not coincide with v_1. This stops the identification process.
2. Only one element, e.g. u_1^n, coincides with v^n.
   This would be because $u_{1,1} = v_1$ and $u_{2,1} \neq v_1$. Then, we continue with applying (1-)identification in the next tree (resp. code).
3. Both elements coincide with v^n.
   In this case also in the next tree 2-identification would have to be applied.

The main idea now is to exploit the fact that the symmetric 2-identification running time is an expectation. Therefore we introduce X_{t+1} as the RV which

indicates how many components of (U_1^n, U_2^n) are still candidates at step t. For all $t \in \{1, \ldots, n-1\}$ we define

$$X_{t+1} = \begin{cases} 0 & \text{if } U_1^t \neq V^t \neq U_2^t \ / \\ 1 & \text{otherwise} \\ 2 & \text{if } U_1^t = U_2^t = V^t \end{cases}$$

and we set $X_1 = 2$. In order to calculate the corresponding probabilities we use the facts that U_1, U_2 and V are i.i.d. With this we get

$$\begin{aligned} \Pr(X_{t+1} = 2) &= \Pr(U_1^t = U_2^t = V^t) \\ &= \sum_{u^t \in \mathcal{U}^t} p_{u^t}^3 = \sum_{u_1, \ldots, u_t \in \mathcal{U}} (p_{u_1} \ldots p_{u_t})^3 = \left(\sum_{u \in \mathcal{U}} p_u^3 \right)^t \end{aligned}$$

and

$$\begin{aligned} \Pr(X_{t+1} = 1) &= 2 \Pr(U_1^t = V^t \text{ and } U_2^t \neq V^t) \\ &= 2 \sum_{u^t \in \mathcal{U}^t} p_{u^t}^2 (1 - p_{u^t}) \\ &= 2 \left[\sum_{u_1, \ldots, u_t \in \mathcal{U}} (p_{u_1} \ldots p_{u_t})^2 - \sum_{u_1, \ldots, u_t \in \mathcal{U}} (p_{u_1} \ldots p_{u_t})^3 \right] \\ &= 2 \left[\left(\sum_{u \in \mathcal{U}} p_u^2 \right)^t - \left(\sum_{u \in \mathcal{U}} p_u^3 \right)^t \right]. \end{aligned}$$

As stated before the symmetric 2-identification running time is an expectation. Since for the first time-step $X_1 = 2$ and for all other time-steps the case $X_t = 0$ leads to the termination of the identification process before time-step t, we obtain

$$\begin{aligned} \mathcal{L}_{\mathcal{C}^n}^{2,q}(P^n, P^n) &= \sum_{t=1}^{n} \mathbb{E}(\mathcal{L}_{\mathcal{C}}^{X_t, q}(P, P)) = \sum_{t=0}^{n-1} \mathbb{E}(\mathcal{L}_{\mathcal{C}}^{X_{t+1}, q}(P, P)) \\ &= \mathcal{L}_{\mathcal{C}}^{2,q}(P, P) + \sum_{t=1}^{n-1} \Pr(X_{t+1} = 1) \mathcal{L}_{\mathcal{C}}^{1,q}(P, P) \\ &\quad + \sum_{t=1}^{n-1} \Pr(X_{t+1} = 2) \mathcal{L}_{\mathcal{C}}^{2,q}(P, P) \end{aligned}$$

$$= \mathcal{L}_{\mathcal{C}}^{2,q}(P,P)\left(1+\sum_{t=1}^{n-1}\left(\sum_{u\in\mathcal{U}}p_u^3\right)^t\right)$$
$$+2\mathcal{L}_{\mathcal{C}}^{1,q}(P,P)\left(\sum_{t=1}^{n-1}\left(\sum_{u\in\mathcal{U}}p_u^2\right)^t-\sum_{t=1}^{n-1}\left(\sum_{u\in\mathcal{U}}p_u^3\right)^t\right).$$

□

If we now establish the limit for n going to infinity and apply the geometric series for $k = 2, 3$ we obtain

$$\sum_{t=1}^{\infty}\left(\sum_{u\in\mathcal{U}}p_u^k\right)^t = \frac{1}{1-\sum_{u\in\mathcal{U}}p_u^k}-1$$

and thus,

$$\lim_{n\to\infty}\mathcal{L}_{\mathcal{C}^n}^{2,q}(P^n,P^n) = \frac{\mathcal{L}_{\mathcal{C}}^{2,q}(P,P)}{1-\sum_{u\in\mathcal{U}}p_u^3}+2\left(\frac{1}{1-\sum_{u\in\mathcal{U}}p_u^2}-\frac{1}{1-\sum_{u\in\mathcal{U}}p_u^3}\right)\mathcal{L}_{\mathcal{C}}^{1,q}(P,P).$$

This proves

Corollary 230 *Let $\mathcal{U}$ be a finite set, $q \in \mathbb{N}_{\geq 2}$, P be a probability distribution on $\mathcal{U}$ and $\mathcal{C}$ be prefix code. It then holds that*

$$\mathcal{L}_{\mathcal{C}}^{2,q}(P,P) = (1-\sum_{u\in\mathcal{U}}p_u^3)\lim_{n\to\infty}\mathcal{L}_{\mathcal{C}^n}^{2,q}(P^n,P^n)-2\left(\frac{1-\sum_{u\in\mathcal{U}}p_u^3}{1-\sum_{u\in\mathcal{U}}p_u^2}-1\right)\mathcal{L}_{\mathcal{C}}^{1,q}(P,P).$$

Let us go back to the case where $\mathcal{U} = [q]$, $P = (1/q,\ldots,1/q)$ and $\mathcal{C} = \mathcal{C}_q$. In this case $\mathcal{L}_{\mathcal{C}}^{l,q}(P,P) = 1$ for $l \in [2]$. It follows immediately from Corollary 230 that

$$\lim_{n\to\infty}\mathcal{L}_{\mathcal{C}^n}^{2,q}(P^n,P^n) = 2\frac{q}{q-1}-\frac{q^2}{q^2-1}. \tag{39}$$

This is the promised alternative proof of Theorem 227 for $L = 2$ and $|\mathcal{U}| = q^n$.

What we do now is to lower bound the expression $\lim_{n\to\infty}\mathcal{L}_{\mathcal{C}^n}^{2,q}(P^n,P^n)$. In Lemma 233 we show that we can limit ourselves to typical sequences (see [8]). Then we cut the codetree at some given depth and fill up the shorter branches to this depth with zero probability elements in order to obtain a saturated tree, resp. a block code. This does not increase the symmetric identification running time.

In Theorem 227 of Sect. 4, we have shown how L-identification and in particular 2-identification behaves asymptotically on block codes if we consider the uniform distribution. To use this result we have to show that for 2-identification uniform distribution is optimal for block codes. The following lemma provides a way

for coming from any probability distribution to the uniform distribution without increasing the symmetric identification running time.

Lemma 231 *Let $n \in \mathbb{N}$, $q \in \mathbb{N}_{\geq 2}$, $k \in \{0, \ldots, n-1\}$ and $t \in \{0, \ldots, q^{n-k-1} - 1\}$. Further, let $P = (p_1, \ldots, p_{q^n})$ and $\tilde{P} = (\tilde{p}_1, \ldots, \tilde{p}_{q^n})$ be probability distributions on $[q^n]$ with*

$$P = (p_1, \ldots, p_{tq^{k+1}}, \underbrace{r_1, \ldots, r_1}_{q^k}, \underbrace{r_2, \ldots, r_2}_{q^k}, \ldots, \underbrace{r_q, \ldots, r_q}_{q^k}, p_{(t+1)q^{k+1}+1}, \ldots, p_{q^n})$$

and

$$\tilde{P} = (p_1, \ldots, p_{tq^{k+1}}, \underbrace{\frac{1}{q}\sum_{i=1}^{q} r_i, \ldots, \frac{1}{q}\sum_{i=1}^{q} r_i}_{q^{k+1}}, p_{(t+1)q^{k+1}+1}, \ldots, p_{q^n}).$$

Then it holds that

$$\mathcal{L}^{2,q}_{\mathcal{C}_{q^n}}(P, P) - \mathcal{L}^{2,q}_{\mathcal{C}_{q^n}}(\tilde{P}, \tilde{P}) \geq 0.$$

The inequality holds with equality if and only if either $k = 0$ or $r_i = r_j$ for all $i, j \in [q]$.

Proof W.l.o.g. we further assume that $t = 0$ such that for $i \in [q]$

$$p_{(i-1)q^k+1} = p_{(i-1)q^k+2} = \ldots = p_{iq^k} = r_i.$$

Also, we use for simplicity the abbreviations $L_{u_1u_2,v} = \mathcal{L}^{2,q}_{\mathcal{C}_{q^n}}((u_1, u_2), v)$ and $\alpha_{u_1u_2,v} = (p_{u_1}p_{u_2}p_v - \tilde{p}_{u_1}\tilde{p}_{u_2}\tilde{p}_v)L_{u_1u_2,v}$. With this notation we obtain

$$\begin{aligned}
&\mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}(P, P) - \mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}(\tilde{P}, \tilde{P}) \\
&= \sum_{u_1,u_2,v=1}^{q^n} \alpha_{u_1u_2,v} \\
&= \sum_{v=1}^{q^n} \left[\sum_{u_1,u_2=1}^{q^{k+1}} \alpha_{u_1u_2,v} + 2 \sum_{u_1=1}^{q^{k+1}} \sum_{u_2=q^{k+1}+1}^{q^n} \alpha_{u_1u_2,v} + \sum_{u_1,u_2=q^{k+1}+1}^{q^n} \alpha_{u_1u_2,v} \right] \\
&= \sum_{i=1}^{6} R_i, \qquad (40)
\end{aligned}$$

where the second equality comes from the fact that $L_{u_1u_2,v} = L_{u_2u_1,v}$ and where

$$R_1 = \sum_{u_1,u_2,v=1}^{q^{k+1}} \alpha_{u_1u_2,v} \qquad R_2 = \sum_{u_1,u_2=1}^{q^{k+1}} \sum_{v=q^{k+1}+1}^{q^n} \alpha_{u_1u_2,v}$$

$$R_3 = 2\sum_{u_1,v=1}^{q^{k+1}} \sum_{u_2=q^{k+1}+1}^{q^n} \alpha_{u_1u_2,v} \qquad R_4 = 2\sum_{u_1=1}^{q^{k+1}} \sum_{u_2,v=q^{k+1}+1}^{q^n} \alpha_{u_1u_2,v}$$

$$R_5 = \sum_{u_1,u_2=q^{k+1}+1}^{q^n} \sum_{v=1}^{q^{k+1}} \alpha_{u_1u_2,v} \qquad R_6 = \sum_{u_1,u_2,v=q^{k+1}+1}^{q^n} \alpha_{u_1u_2,v}.$$

As one might expect the above summands disappear, except for R_1 and R_3. This is obvious for R_6 since $p_u = \tilde{p}_u$ for all $u \in [q^{k+1}+1, q^n]$.

If $u_1, u_2 \in [q^{k+1}+1, q^n]$, we have on the one hand that $L_{u_1u_2,v} = L_{u_1u_2,1}$ for all $v \in [q^{k+1}]$. We denote this by $L_{u_1u_2}$. On the other hand $p_{u_i} = \tilde{p}_{u_i}$ for $i = 1, 2$.

This yields

$$\begin{aligned} R_5 &= \sum_{u_1,u_2=q^{k+1}+1}^{q^n} \sum_{v=1}^{q^{k+1}} L_{u_1u_2} p_{u_1} p_{u_2} \left[p_v - \frac{1}{q} \sum_{i=1}^{q} r_i \right] \\ &= \sum_{u_1,u_2=q^{k+1}+1}^{q^n} L_{u_1u_2} p_{u_1} p_{u_2} \left[\sum_{v=1}^{q^{k+1}} p_v - q^k \sum_{i=1}^{q} r_i \right] = 0. \end{aligned}$$

Here, the final equality follows from $\sum_{v=1}^{q^{k+1}} p_v = \sum_{i=1}^{q} q^k r_i$.

If $u_2, v \in [q^{k+1}+1, q^n]$ and $u_1 \in [q^{k+1}]$, we see that $L_{u_1u_2,v} = L_{1u_2,v}$ and $p_{u_2} = \tilde{p}_{u_2}$ as well as $p_v = \tilde{p}_v$. Thus, proceeding as before we have that $R_4 = 0$.

If $u_1, u_2 \in [q^{k+1}]$ and $v \in [q^{k+1}+1, q^n]$, it follows that $L_{u_1u_2,v} = L_{11,v}$, which is denoted by L_v, and $p_v = \tilde{p}_v$. With this we get

$$\begin{aligned} R_2 &= \sum_{u_1,u_2=1}^{q^{k+1}} \sum_{v=q^{k+1}+1}^{q^n} L_v p_v \left[p_{u_1} p_{u_2} - \frac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] \\ &= \sum_{v=q^{k+1}+1}^{q^n} L_v p_v \left[\sum_{u_1,u_2=1}^{q^{k+1}} p_{u_1} p_{u_2} - q^{2k} \left(\sum_{i=1}^{q} r_i \right)^2 \right] = 0. \end{aligned}$$

Here, $\sum_{u_1,u_2=1}^{q^{k+1}} p_{u_1} p_{u_2} = \left(\sum_{i=1}^{q} q^k r_i \right)^2$ yields the final equality. Altogether we end up with

$$\mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}(P, P) - \mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}(\tilde{P}, \tilde{P}) = R_1 + R_3.$$

We begin our remaining examinations with R_3. Similar as before we get $L_{u_1u_2,v} = L_{u_11,v}$, which we denote by $L_{u_1,v}$, and $p_{u_2} = \tilde{p}_{u_2}$ if $u_1, v \in [q^{k+1}]$ and $u_2 \in [q^{k+1}+1, q^n]$. We obtain

$$\begin{aligned}\frac{1}{2}R_3 &= \sum_{u_1,v=1}^{q^{k+1}} \sum_{u_2=q^{k+1}+1}^{q^n} L_{u_1,v}p_{u_2}\left[p_{u_1}p_v - \frac{1}{q^2}\left(\sum_{i=1}^{q} r_i\right)^2\right] \\ &= \sum_{u_1,v=1}^{q^{k+1}} L_{u_1,v}\left[p_{u_1}p_v - \frac{1}{q^2}\left(\sum_{i=1}^{q} r_i\right)^2\right] \sum_{u_2=q^{k+1}+1}^{q^n} p_{u_2} \\ &= (1-q^k\sum_{i=1}^{q} r_i) \sum_{u,v=1}^{q^{k+1}} L_{u,v}\left[p_u p_v - \frac{1}{q^2}\left(\sum_{i=1}^{q} r_i\right)^2\right].\end{aligned}$$

We set $A = \sum_{u,v=1}^{q^{k+1}} L_{u,v}\left[p_u p_v - \frac{1}{q^2}\left(\sum_{i=1}^{q} r_i\right)^2\right]$ and separate the different subtrees with roots in level $n-k-1$ in which u and v can occur. We get

$$A = \sum_{s,t=1}^{q} \sum_{u=(s-1)q^k+1}^{sq^k} \sum_{v=(t-1)q^k+1}^{tq^k} L_{u,v}\left[r_s r_t - \frac{1}{q^2}\left(\sum_{i=1}^{q} r_i\right)^2\right].$$

Since it holds for $s, t \in [q]$, $u \in [(s-1)q^k+1, sq^k]$ and $v \in [(t-1)q^k+1, tq^k]$ that

$$L_{u,v} = \begin{cases} n-k & \text{if } s \neq t \\ n-k+\mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) & \text{if } s = t, \end{cases}$$

the above equation becomes

$$\begin{aligned}A &= (n-k)\left[\sum_{s,t=1}^{q} q^{2k} r_s r_t - q^{2k}\left(\sum_{i=1}^{q} r_i\right)^2\right] \\ &\quad + \sum_{s=1}^{q} \sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v)\left[r_s^2 - \frac{1}{q^2}\left(\sum_{i=1}^{q} r_i\right)^2\right] \\ &= \frac{1}{2q}\sum_{i,j=1}^{q} (r_i - r_j)^2 \sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v).\end{aligned}$$

The second equation follows on the one hand from $\sum_{s,t=1}^{q} r_s r_t = \left(\sum_{i=1}^{q} r_i\right)^2$. From this follows that the first summand is 0. On the other hand

$$\sum_{s=1}^{q} r_s^2 - \frac{1}{q}\left(\sum_{i=1}^{q} r_i\right)^2 = \frac{1}{2q}\sum_{i,j=1}^{q}(r_i - r_j)^2.$$

By applying Corollary 226 we obtain

$$\begin{aligned}
\sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) &= q^k \sum_{l=1}^{k} l\,|\mathcal{R}^{1,q}_{\mathcal{C}_{q^k}}(q^k, l, 1)| \\
&= q^k \left[\sum_{l=1}^{k-1} l q^{k-l}(q-1) + kq\right] \\
&= q^k \left[q^k(q-1)\sum_{l=1}^{k} l q^{-l} + k\right] \\
&= q^k \left[q^k(q-1)\frac{q(q^k-1) - k(q-1)}{q^k(q-1)^2} + k\right] \\
&= \frac{q}{q-1} q^k (q^k - 1).
\end{aligned}$$

Putting all this together we get

$$R_3 = \frac{1}{q-1} q^k (q^k - 1)(1 - q^k \sum_{i=1}^{q} r_i) \sum_{i,j=1}^{q} (r_i - r_j)^2 \geq 0.$$

This equals 0 if and only if either $k = 0$ or $r_i = r_j$ for all $i, j \in [q]$ or $\sum_{i=1}^{q} r_i = q^{-k}$. The last condition is equivalent to $p_i = 0$ for all $i \in [q^{k+1} + 1, q^n]$.

We now turn to R_1. With the same notation as before we have

$$\begin{aligned}
R_1 &= \sum_{u_1,u_2,v=1}^{q^{k+1}} L_{u_1u_2,v}\left[p_{u_1}p_{u_2}p_v - \frac{1}{q^3}\left(\sum_{i=1}^{q} r_i\right)^3\right] \\
&= \sum_{s_1,s_2,t=1}^{q} \sum_{r=1}^{2} \sum_{u_r=(s_r-1)q^k+1}^{s_r q^k} \sum_{v=(t-1)q^k+1}^{tq^k} L_{u_1u_2,v}\left[r_{s_1}r_{s_2}r_t - \frac{1}{q^3}\left(\sum_{i=1}^{q} r_i\right)^3\right] \\
&= \sum_{s_1,s_2,t=1}^{q}\left[r_{s_1}r_{s_2}r_t - \frac{1}{q^3}\left(\sum_{i=1}^{q} r_i\right)^3\right] \sum_{r=1}^{2} \sum_{u_r=(s_r-1)q^k+1}^{s_r q^k} \sum_{v=(t-1)q^k+1}^{tq^k} L_{u_1u_2,v}.
\end{aligned}$$

For $u_r \in [(s_r-1)q^k+1, s_rq^k]$ and $v \in [(t-1)q^k+1, tq^k]$ it holds that

$$L_{u_1u_2,v} = \begin{cases} n-k & \text{if } s_1 \neq t \text{ and } s_2 \neq t \\ n-k+\mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u_1,v) & \text{if } s_1 = t \text{ and } s_2 \neq t \\ n-k+\mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u_2,v) & \text{if } s_1 \neq t \text{ and } s_2 = t \\ n-k+\mathcal{L}^{2,q}_{\mathcal{C}_{q^k}}((u_1,u_2),v) & \text{if } s_1 = s_2 = t. \end{cases}$$

If we insert the above equations into R_1, we get

$$\begin{aligned} R_1 &= (n-k)\left[\sum_{s_1,s_2,t=1}^{q} q^{3k} r_{s_1}r_{s_2}r_t - q^{3k}\left(\sum_{i=1}^{q} r_i\right)^3\right] \\ &\quad + \sum_{s_1=1}^{q}\sum_{s_2=1,s_2\neq s_1}^{q} q^k \sum_{u_1,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u_1,v)\left[r_{s_1}^2 r_{s_2} - \frac{1}{q^3}\left(\sum_{i=1}^{q} r_i\right)^3\right] \\ &\quad + \sum_{s_2=1}^{q}\sum_{s_1=1,s_1\neq s_2}^{q} q^k \sum_{u_2,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u_2,v)\left[r_{s_1} r_{s_2}^2 - \frac{1}{q^3}\left(\sum_{i=1}^{q} r_i\right)^3\right] \\ &\quad + \sum_{s=1}^{q} \sum_{u_1,u_2,v=1}^{q^k} \mathcal{L}^{2,q}_{\mathcal{C}_{q^k}}((u_1,u_2),v)\left[r_s^3 - \frac{1}{q^3}\left(\sum_{i=1}^{q} r_i\right)^3\right] \\ &= 2q^k\left[\sum_{s=1}^{q}\sum_{t=1,t\neq s}^{q} r_s^2 r_t - \frac{q-1}{q^2}\left(\sum_{i=1}^{q} r_i\right)^3\right] \sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) \\ &\quad + \left[\sum_{s=1}^{q} r_s^3 - \frac{1}{q^2}\left(\sum_{i=1}^{q} r_i\right)^3\right] \sum_{u_1,u_2,v=1}^{q^k} \mathcal{L}^{2,q}_{\mathcal{C}_{q^k}}((u_1,u_2),v). \end{aligned}$$

If all r_i's are zero, we obtain $R_1 = 0$. We exclude this case and normalize the probabilities $r_1, \ldots, r_q$ by setting $\bar{r}_i = r_i / \sum_{j=1}^{q} r_j$ for $i \in [q]$. This yields

$$\begin{aligned} R_1 = \left(\sum_{i=1}^{q} r_i\right)^3 \Bigg[& 2q^k\left(\sum_{s}\sum_{t\neq s} \bar{r}_s^2 \bar{r}_t - \frac{q-1}{q^2}\right) \sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) \\ & + \left(\sum_{s} \bar{r}_s^3 - \frac{1}{q^2}\right) \sum_{u_1,u_2,v=1}^{q^k} \mathcal{L}^{2,q}_{\mathcal{C}_{q^k}}((u_1,u_2),v)\Bigg]. \end{aligned}$$

We have already seen during the calculations of R_3 that

$$\sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) = \frac{q}{q-1} q^k (q^k - 1).$$

By applying Corollary 226 we further get that

$$\begin{aligned}
&\sum_{u_1,u_2,v=1}^{q^k} \mathcal{L}^{2,q}_{\mathcal{C}_{q^k}}((u_1,u_2),v) \\
&\quad = q^k \sum_{l=1}^{k} l |\mathcal{R}^{2,q}_{\mathcal{C}_{q^k}}(q^k, l, 1)| \\
&\quad = q^k \left[\sum_{l=1}^{k-1} l q^{2k} \left(2q^{-l}(q-1)(1-q^{-l+1}) + q^{-2l}(q-1)^2 \right) \right] \\
&\qquad + q^k k \left(2q(q^k - q) + q^2 \right) \\
&\quad = q^k \left[\sum_{l=1}^{k} l q^{2k} \left(2q^{-l}(q-1)(1-q^{-l+1}) + q^{-2l}(q-1)^2 \right) \right] \\
&\qquad + q^k k (2q^k - 1) \\
&\quad = (q-1) q^{3k} \left[2 \sum_{l=1}^{k} l q^{-l} - (q+1) \sum_{l=1}^{k} l q^{-2l} \right] \\
&\qquad + k q^k (2q^k - 1) \\
&\quad = (q-1) q^{3k} \left[2 \frac{q(q^k-1) - k(q-1)}{q^k (q-1)^2} - (q+1) \frac{q^2(q^{2k}-1) - k(q^2-1)}{q^{2k}(q^2-1)^2} \right] \\
&\qquad + k q^k (2q^k - 1) \\
&\quad = 2 \frac{q}{q-1} q^{2k} (q^k - 1) - \frac{q^2}{q^2-1} q^k (q^{2k} - 1) \\
&\quad = \frac{q}{q-1} q^k (q^k - 1) \frac{(q+2)q^k - q}{q+1}.
\end{aligned}$$

Applying this result we obtain

$$R_1 = \left(\sum_{i=1}^{q} r_i\right)^3 \frac{q}{q-1} q^k (q^k-1)\left[2q^k\left(\sum_s \bar{r}_s^2 - \sum_s \bar{r}_s^3 - \frac{q-1}{q^2}\right) + \frac{(q+2)q^k - q}{q+1}\left(\sum_s \bar{r}_s^3 - \frac{1}{q^2}\right)\right]$$

$$= -\left(\sum_{i=1}^{q} r_i\right)^3 \frac{q}{q-1} q^k (q^k-1)\left[\frac{q}{q+1}(q^k+1)\sum_s \bar{r}_s^3 - 2q^k \sum_s \bar{r}_s^2 + \frac{(2q+1)q^k - 1}{q(q+1)}\right].$$

It remains to show that

$$\frac{q}{q+1}(q^k+1)\sum_s \bar{r}_s^3 - 2q^k \sum_s \bar{r}_s^2 + \frac{(2q+1)q^k - 1}{q(q+1)} \leq 0.$$

The left hand side obviously equals 0 if $\bar{r}_1 = \ldots = \bar{r}_q = 1/q$, i.e. $r_1 = \ldots = r_q$. Let us define $f : \Delta_{q-1} \to \mathbb{R}$ by

$$f(x_1, \ldots, x_{q-1}) = a_1\left[\sum_{s=1}^{q-1} x_s^3 + \left(1 - \sum_{s=1}^{q-1} x_s\right)^3\right] - a_2\left[\sum_{s=1}^{q-1} x_s^2 + \left(1 - \sum_{s=1}^{q-1} x_s\right)^2\right],$$

where $a_1 = q(q^k+1)/(q+1)$ and $a_2 = 2q^k$. We will show that $(1/q, \ldots, 1/q)$ is the only extremal point of f in Γ_q and that it is a local maximum. The first partial derivative for $j \in [q-1]$ is

$$\frac{\delta}{\delta x_j} f(x_1, \ldots, x_{q-1}) = 3a_1\left(x_j^2 - (1 - \sum_{i=1}^{q-1} x_i)^2\right) - 2a_2\left(x_j - (1 - \sum_{i=1}^{q-1} x_i)\right)$$

$$= \left(x_j - (1 - \sum_{i=1}^{q-1} x_i)\right)\left[3a_1\left(x_j + 1 - \sum_{i=1}^{q-1} x_i\right) - 2a_2\right].$$

It follows that the gradient $\nabla f = \mathbf{0}$ if and only if either $x_j = 1 - \sum_{i=1}^{q-1} x_i$ for all $j \in [q-1]$, which yields $x_1 = \ldots = x_{q-1} = 1/q$, or $3a_1(x_j + 1 - \sum_{i=1}^{q-1} x_i) - 2a_2 = 0$

for all $j \in [q-1]$. Since

$$3a_1(1 - \sum_{i=1, i\neq j}^{q-1} x_i) - 2a_2 \leq 3\frac{q}{q+1}(q^k+1) - 4q^k < -q^k + 3 \leq 0,$$

the latter is impossible. We conclude that the only extremal point of f is $(1/q, .., 1/q)$. Further, the second partial derivatives are

$$\frac{\delta^2}{\delta x_k \delta x_j} f(x_1, \ldots, x_{q-1}) = \begin{cases} 6a_1(1 - \sum_{i=1}^{q-1} x_i) - 2a_2 & \text{if } k \neq j \\ 6a_1(1 - \sum_{i=1, i\neq j}^{q-1} x_i) - 4a_2 & \text{if } k = j \end{cases}$$

such that

$$\frac{\delta^2}{\delta x_k \delta x_j} f\left(\frac{1}{q}, .., \frac{1}{q}\right) = \begin{cases} \frac{6a_1}{q} - 2a_2 & \text{if } k \neq j \\ \frac{12a_1}{q} - 4a_2 & \text{if } k = j. \end{cases}$$

Since $(6a_1/q) - 2a_2 = [6(q^k+1)/(q+1)] - 4q^k \leq -2(q^k - 1) < 0$, we see that $(1/q, .., 1/q)$ is a global maximum. With this we obtain that $R_1 \geq 0$, with equality if and only if either $k = 0$ or $r_i = r_j$ for all $i, j \in [q]$. Remember that $R_3 \geq 0$. It equals zero if and only if either $k = 0$ or $r_i = r_j$ for all $i, j \in [q]$ or $p_i = 0$ for $i \in [q^{k+1} + 1, q^n]$. Further, $\mathcal{L}^{2,q}_{\mathcal{C}_{q^n}}(P, P) - \mathcal{L}^{2,q}_{\mathcal{C}_{q^n}}(\tilde{P}, \tilde{P}) = R_1 + R_3$. It follows that this difference is not negative. Moreover, it equals 0 if and only if either $k = 0$ or $r_i = r_j$ for all $i, j \in [q]$. This concludes the proof. □

By applying Lemma 231 in the same way as Lemma 220 in the first subsection of Sect. 3 we obtain

Corollary 232 *Let $n \in \mathbb{N}$ and $q \in \mathbb{N}_{\geq 2}$. Further, let $\mathcal{C} = \mathcal{C}_{q^n}$ and $T = T_{\mathcal{C}}$. Then it holds for all probability distributions P on $[q^n]$ that*

$$\mathcal{L}^{2,q}_{\mathcal{C}}(P, P) \geq \mathcal{L}^{2,q}_{\mathcal{C}}\left(\left(\frac{1}{q^n}, \ldots, \frac{1}{q^n}\right), \left(\frac{1}{q^n}, \ldots, \frac{1}{q^n}\right)\right).$$

The inequality holds with equality if and only if $P(T_x) = q^{-\|x\|}$ for all inner nodes $x \in \mathring{\mathcal{N}}(T)$.

Before we come to Lemma 233, we provide a short excurs on *δ-typical sequences*. These are defined e.g. in [9] Definition 2.8 (p. 33). We will change some of the notation of this definition in order to harmonize it with the notation used in this thesis and related papers.

"For any distribution P on $\mathcal{U}$, a sequence $u^n \in \mathcal{U}^n$ is called *P-typical* with constant δ if

$$\left| \frac{1}{n} < u^n | a > - p_a \right| \leq \delta \tag{41}$$

for every $a \in \mathcal{U}$ and, in addition, no $a \in \mathcal{U}$ with $p_a = 0$ occurs in u^n. The set of such sequences will be denoted by $\mathcal{T}^n_{P,\delta}$."

Here, the value of $< u^n | a >$ is the number of appearances of a as a component of u^n. In words, a sequence $u^n \in \mathcal{U}^n$ is called *P-typical* with constant δ if for all $a \in \mathcal{U}$ the difference between the relative frequency of a in u^n and the actual probability of a with respect to P is at most δ.

Lemma 2.12 in [9] and its subsequent remark state that

$$P^n(\mathcal{T}^n_{P,\delta}) \geq 1 - \frac{|\mathcal{U}|}{4n\delta^2} \tag{42}$$

Further, it follows from Eq. (41) for all $u^n \in \mathcal{T}^n_{P,\delta}$ that

$$P^n_{u^n} = \prod_{a \in \mathcal{U}} p_a^{<u^n|a>} \leq \prod_{a \in supp(P)} p_a^{n(p_a - \delta)} = 2^{-n\left(H(P) + \delta \sum_{a \in supp(P)} \log p_a\right)}. \tag{43}$$

Here, $H(P) = -\sum_{a \in supp(P)} p_a \log p_a$ is Shannon's classical entropy. In the following we use $M_P = -\sum_{a \in supp(P)} \log p_a$. It holds that $0 \leq M_P < \infty$ with equality on the left hand side if and only if $supp(P) = 1$. We exclude this case in our further analysis. It follows that for all $\epsilon > 0$ exists $\delta > 0$ such that on the one hand it holds that

$$P^n((\mathcal{T}^n_{P,\delta})^c) \leq \frac{|\mathcal{U}| M_P}{4n\epsilon^2}. \tag{44}$$

On the other hand it holds for all $u^n \in \mathcal{T}^n_{P,\delta}$ that

$$P^n_{u^n} \leq 2^{-n(H(P) - \epsilon)}. \tag{45}$$

To see this choose $\delta = \epsilon / M_P$ and apply Eqs. (42) and (43). Things are now settled to prove

Lemma 233 *Let P be probability distribution on $\mathcal{U}$ with $|supp(P)| > 1$. For all $\epsilon > 0$ and all q-ary prefix codes $\mathcal{C}$ over $\mathcal{U}$ there exist sequences $\alpha_n(\epsilon) = \alpha_n \to 0$ and $K_n(\epsilon) = K_n \to \infty$ such that*

$$\mathcal{L}^{2,q}_{\mathcal{C}^n}(P^n, P^n) \geq (1 - \alpha_n)^3 \mathcal{L}^{2,q}_{\mathcal{C}_{qK_n}} \left(\left(\frac{1}{q^{K_n}}, \ldots, \frac{1}{q^{K_n}} \right), \left(\frac{1}{q^{K_n}}, \ldots, \frac{1}{q^{K_n}} \right) \right)$$

holds for all sufficiently large n.

Proof The proof of this theorem follows the same guidelines as the proof of Lemma 214 in chapter "An Interpretation of Identification Entropy". However, we changed some of its steps in order to obtain a more explanatory proof.

We begin the proof without explicitly specifying K_n and α_n. This will be done later. We partition $\mathcal{U}^n$ according to the given code $\mathcal{C}^n$ into $\mathcal{U}_1^n = \{u^n \in \mathcal{U}^n : \|c_{u^n}\| \leq K_n\}$ and $\mathcal{U}_2^n = \mathcal{U}^n \setminus \mathcal{U}_1^n$. Since $\mathcal{C}^n$ is a q-ary prefix code, we have that

$$|\mathcal{U}_1^n| \leq q^{K_n}. \tag{46}$$

For $\epsilon > 0$ we choose $\delta = \epsilon / M_P$ and obtain

$$\begin{aligned} P^n(\mathcal{U}_1^n) &= P^n(\mathcal{U}_1^n \cap \mathcal{T}_{P,\delta}^n) + P^n(\mathcal{U}_1^n \cap (\mathcal{T}_{P,\delta}^n)^c) \\ &\leq |\mathcal{U}_1^n \cap \mathcal{T}_{P,\delta}^n| 2^{-n(H(P)-\epsilon)} + P^n((\mathcal{T}_{P,\delta}^n)^c) \\ &\leq q^{K_n} 2^{-n(H(P)-\epsilon)} + \frac{|\mathcal{U}| M_P}{4n\epsilon^2}. \end{aligned}$$

The first inequality follows by Eq. (45). Equations (44) and (46) yield the second inequality.

We now set $K_n = \left\lfloor \frac{n(H(P)-2\epsilon)}{\log q} \right\rfloor$ as well as $\alpha_n = 2^{-n\epsilon} + \frac{|\mathcal{U}| M_P}{4n\epsilon^2}$ and obtain

$$P^n(\mathcal{U}_1^n) \leq \alpha_n$$

and thus

$$P^n(\mathcal{U}_2^n) \geq 1 - \alpha_n. \tag{47}$$

We will now construct a new source code by cutting all codewords in $\mathcal{U}_2^n$ back to length K_n. Formally, we define the new source $\tilde{\mathcal{U}} = \tilde{\mathcal{U}}_1 \cup \tilde{\mathcal{U}}_2$, where $\tilde{\mathcal{U}}_1 = \mathcal{U}_1^n$ and $\tilde{\mathcal{U}}_2$ is defined as follows. Let $\cong$ be an equivalence relation on $\mathcal{U}_2^n$ with $u^n \cong v^n :\Leftrightarrow c_{u^n}^{K_n} = c_{v^n}^{K_n}$ and let $\mathcal{E}_1, \ldots, \mathcal{E}_m$ be the equivalence classes. Further, we associate with every equivalence class $\mathcal{E}_i$ the object e_i and define $\tilde{\mathcal{U}}_2 = \{e_1, \ldots, e_m\}$. Moreover, we define a probability distribution $\tilde{P}$ on $\tilde{\mathcal{U}}$ by $\tilde{P}(u^n) = P(u^n)$ for all $u^n \in \tilde{\mathcal{U}}_1$ and $\tilde{P}(e_k) = \sum_{u^n \in \mathcal{E}_k} P(u^n)$ for $k \in [m]$. Finally, we obtain a new code $\tilde{\mathcal{C}} : \tilde{\mathcal{U}} \to \mathcal{Q}^*$ by $\tilde{c}_{u^n} = c_{u^n}$ if $u^n \in \tilde{\mathcal{U}}_1$ and $\tilde{c}_{e_k}$ will be the common prefix of length K_n of the objects in $\mathcal{E}_k$. This construction step is visualized in Fig. 5. It follows that

$$\mathcal{L}_{\mathcal{C}^n}^{2,q}(P^n, P^n) \geq \mathcal{L}_{\tilde{\mathcal{C}}}^{2,q}(\tilde{P}, \tilde{P}). \tag{48}$$

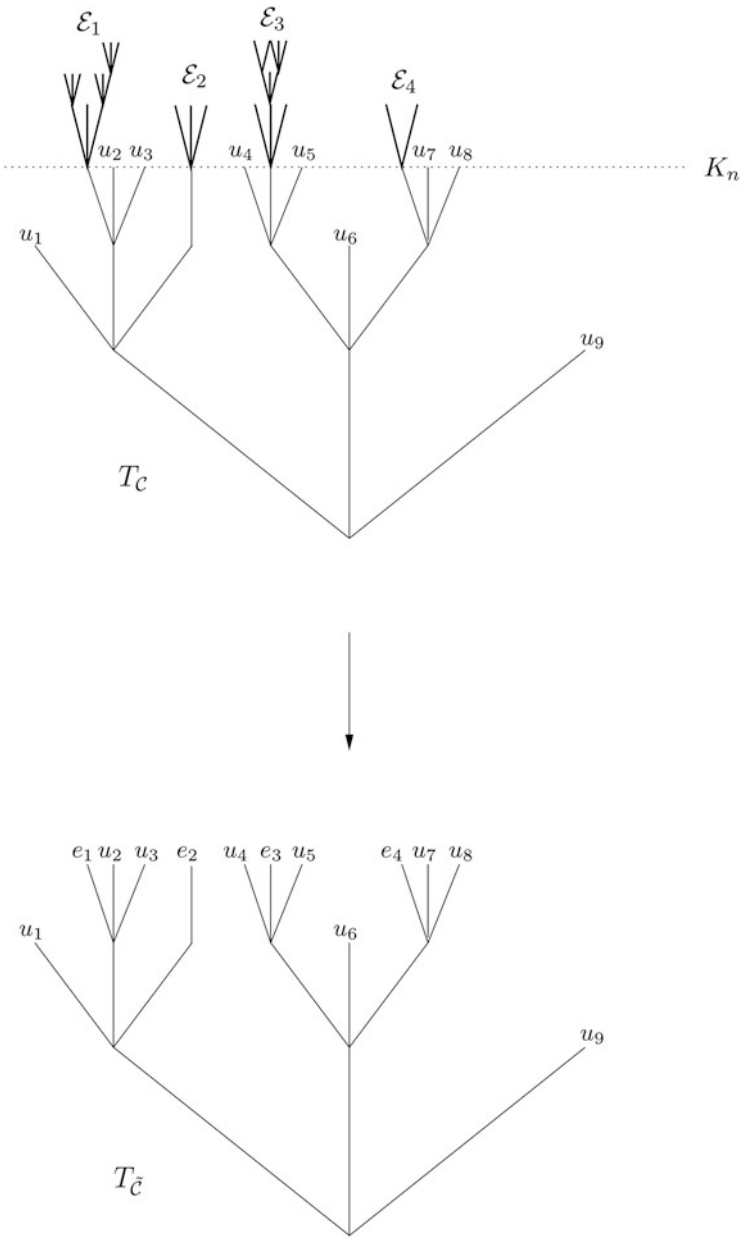


Fig. 5 The cutting of $T_{\mathcal{C}}$ at depth K_n yields $T_{\tilde{\mathcal{C}}}$ with $\tilde{\mathcal{U}}_1 = \{u_1, u_2, \ldots, u_9\}$ and $\tilde{\mathcal{U}}_2 = \{e_1, e_2, e_3, e_4\}$

The next step is to focus only on the $\tilde{\mathcal{U}}_2$-part of $\tilde{\mathcal{U}}$. Again we operate without increasing the symmetric 2-identification running time since

$$\begin{aligned}
\mathcal{L}^{2,q}_{\tilde{\mathcal{C}}}(\tilde{P},\tilde{P}) &= \sum_{\tilde{u}_1,\tilde{u}_2,\tilde{v}\in\tilde{\mathcal{U}}} \tilde{P}(\tilde{u}_1)\tilde{P}(\tilde{u}_2)\tilde{P}(\tilde{v})\mathcal{L}^{2,q}_{\tilde{\mathcal{C}}}((\tilde{u}_1,\tilde{u}_2),\tilde{v}) \\
&\geq \sum_{\tilde{u}_1,\tilde{u}_2,\tilde{v}\in\tilde{\mathcal{U}}_2} \tilde{P}(\tilde{u}_1)\tilde{P}(\tilde{u}_2)\tilde{P}(\tilde{v})\mathcal{L}^{2,q}_{\tilde{\mathcal{C}}}((\tilde{u}_1,\tilde{u}_2),\tilde{v}) \\
&= \sum_{i_1,i_2,j=1}^{m} \tilde{P}(e_{i_1})\tilde{P}(e_{i_2})\tilde{P}(e_j)\mathcal{L}^{2,q}_{\tilde{\mathcal{C}}}((e_{i_1},e_{i_2}),e_j) \\
&= \left(\sum_{k=1}^{m}\tilde{P}(e_k)\right)^3 \sum_{i_1,i_2,j=1}^{m} \tilde{P}_2(e_{i_1})\tilde{P}_2(e_{i_2})\tilde{P}_2(e_j)\mathcal{L}^{2,q}_{\tilde{\mathcal{C}}_2}((e_{i_1},e_{i_2}),e_j).
\end{aligned}$$

Here, $\tilde{P}_2$ is a probability distribution on $\tilde{\mathcal{U}}_2$ defined by

$$\tilde{P}_2(e_j) = \frac{\tilde{P}(e_j)}{\sum_{k=1}^{m}\tilde{P}(e_k)}$$

for $j \in [m]$. Further, $\tilde{\mathcal{C}}_2$ is the restriction of $\tilde{\mathcal{C}}$ to $\tilde{\mathcal{U}}_2$.

Since

$$\sum_{k=1}^{m}\tilde{P}(e_k) = \sum_{k=1}^{m}\sum_{u^n\in\mathcal{E}_k} P^n(u^n) = P^n(\mathcal{U}_2^n),$$

we obtain by Eq. (47) that

$$\mathcal{L}^{2,q}_{\tilde{\mathcal{C}}}(\tilde{P},\tilde{P}) \geq (1-\alpha_n)^3\mathcal{L}^{2,q}_{\tilde{\mathcal{C}}_2}(\tilde{P}_2,\tilde{P}_2). \tag{49}$$

Although $\tilde{\mathcal{C}}_2$ is a block code with codewords of length K_n it may be—and maybe by far—not saturated. To achieve this property we extend $\tilde{\mathcal{U}}_2$ to a set of cardinality q^{K_n}, assign zero probabilities to the additional elements and use for them codewords from $\mathcal{Q}^{K_n} \setminus \tilde{\mathcal{C}}_2$. We now obey the conditions of Corollary 232 by which we obtain

$$\mathcal{L}^{2,q}_{\tilde{\mathcal{C}}_2}(\tilde{P}_2,\tilde{P}_2) \geq \mathcal{L}^{2,q}_{\mathcal{C}_{qK_n}}((\frac{1}{q^{K_n}},\ldots,\frac{1}{q^{K_n}}),(\frac{1}{q^{K_n}},\ldots,\frac{1}{q^{K_n}})). \tag{50}$$

The inequalities (48), (49) and (50) finally yield the statement of the lemma. □

By applying Theorem 227 and Lemma 233 to Corollary 230 we obtain

Corollary 234 *Let $\mathcal{U}$ be a finite set, $q \in \mathbb{N}_{\geq 2}$, P be a probability distribution on $\mathcal{U}$ with $|supp(P)| > 1$ and $\mathcal{C}$ be a q-ary prefix code. It then holds that*

$$\mathcal{L}_{\mathcal{C}}^{2,q}(P, P) \geq (1-\sum_{u\in\mathcal{U}} p_u^3)\left(2\frac{q}{q-1} - \frac{q^2}{q^2-1}\right) - 2\left(\frac{1-\sum_{u\in\mathcal{U}} p_u^3}{1-\sum_{u\in\mathcal{U}} p_u^2} - 1\right)\mathcal{L}_{\mathcal{C}}^{1,q}(P, P).$$

5.2 *The* q*-ary Identification Entropy of Second Degree*

Since (1-)identification appears negatively signed, we can not immediately apply its lower bound $\mathcal{L}_{\mathcal{C}}^{1,q}(P, P) \geq H_{\mathrm{ID}}^{1,q}(P)$ (see [4], chapter "An Interpretation of Identification Entropy"). But we can show that the bound of Corollary 234 is attained if P consists only of q-powers and $\mathcal{C}$ is a code with $\|c_u\| = -\log_q p_u$.

Proposition 235 *Let P be a probability distribution on $\mathcal{U}$ which only consists of q-powers and $\mathcal{C}$ be a q-ary prefix code, where $\|c_u\| = -\log_q p_u$ for all $u \in \mathcal{U}$. It then holds that*

$$\mathcal{L}_{\mathcal{C}}^{2,q}(P, P) = 2\frac{q}{q-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^2\right) - \frac{q^2}{q^2-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^3\right).$$

Proof It is an immediate consequence from the condition $\|c_u\| = -\log_q p_u$ for all $u \in \mathcal{U}$ that

$$P(T_x) = q^{-\|x\|} \tag{51}$$

holds for all $x \in \mathcal{N}(T)$, where $T = T_{\mathcal{C}}$. We now introduce for all $v \in \mathcal{U}$ and $k = 1, \ldots, \|c_v\|$ the set

$$\tilde{\mathcal{R}}_{\mathcal{C}}^{1,q}(k, v) = \mathcal{R}_{\mathcal{C}}^{1,q}(1, v) \,\dot\cup\, \ldots \,\dot\cup\, \mathcal{R}_{\mathcal{C}}^{1,q}(k-1, v). \tag{52}$$

Proceeding as in the proof of Theorem 227 we obtain

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P, P) = \sum_{v\in\mathcal{U}} p_v \sum_{k=1}^{\|c_v\|} k \sum_{(u_1,u_2)\in\mathcal{R}_{\mathcal{C}}^{2,q}(k,v)} p_{u_1} p_{u_2}.$$

In the following we use $S_{k,v} = \sum_{(u_1,u_2)\in\mathcal{R}_{\mathcal{C}}^{2,q}(k,v)} p_{u_1} p_{u_2}$. With the notation of Eq. (52) it holds that

$$S_{k,v} = 2 \sum_{u_1\in\mathcal{R}_{\mathcal{C}}^{1,q}(k,v)} \sum_{u_2\in\bar{\mathcal{R}}_{\mathcal{C}}^{1,q}(k,v)} p_{u_1} p_{u_2} + \sum_{u_1,u_2\in\mathcal{R}_{\mathcal{C}}^{1,q}(k,v)} p_{u_1} p_{u_2}.$$

Here, the equality holds because there exists either one component for which (1-)identification against v takes exactly k time-steps and the other yields a (1-)identification time regarding v of at most $k-1$ or both components have a (1-)identification time regarding v of k.

1. *Case* $k = 1, \ldots, \|c_v\| - 1$.
   In this case we have that $\mathcal{R}_{\mathcal{C}}^{1,q}(k, v) = \bar{T}_{c_v^{k-1}} \setminus \bar{T}_{c_v^k}$ and $\bar{\mathcal{R}}_{\mathcal{C}}^{1,q}(k, v) = \mathcal{U} \setminus \bar{T}_{c_v^{k-1}}$. This together with Eq. (51) yields

$$\sum_{u\in\mathcal{R}_{\mathcal{C}}^{1,q}(k,v)} p_u = P(T_{c_v^{k-1}}) - P(T_{c_v^k}) = q^{-k+1} - q^{-k} = q^{-k}(q-1)$$

   and

$$\sum_{u\in\bar{\mathcal{R}}_{\mathcal{C}}^{1,q}(k,v)} p_u = 1 - P(T_{c_v^{k-1}}) = 1 - q^{-k+1}.$$

   Thus,

$$\begin{aligned} S_{k,v} &= 2q^{-k}(q-1)(1-q^{-k+1}) + q^{-2k}(q-1)^2 \\ &= (1-q^{-k})^2 - (1-q^{-k+1})^2. \end{aligned}$$

2. *Case* $k = \|c_v\|$.
   In this case we have that $\mathcal{R}_{\mathcal{C}}^{1,q}(\|c_v\|, v) = \bar{T}_{c_v^{\|c_v\|-1}}$ and $\bar{\mathcal{R}}_{\mathcal{C}}^{1,q}(\|c_v\|, v) = \mathcal{U} \setminus \bar{T}_{c_v^{\|c_v\|-1}}$. Equation (51) yields

$$\sum_{u\in\mathcal{R}_{\mathcal{C}}^{1,q}(\|c_v\|,v)} p_u = P(T_{c_v^{\|c_v\|-1}}) = q^{-\|c_v\|+1}$$

   and

$$\sum_{u\in\bar{\mathcal{R}}_{\mathcal{C}}^{1,q}(\|c_v\|,v)} p_u = 1 - P(T_{c_v^{\|c_v\|-1}}) = 1 - q^{-\|c_v\|+1}.$$

Thus, we obtain

$$S_{\|c_v\|,v} = 2q^{-\|c_v\|+1}(1-q^{-\|c_v\|+1}) + q^{-2(\|c_v\|-1)} = 1-(1-q^{-\|c_v\|+1})^2.$$

Together, the above two cases yield

$$\begin{aligned}\sum_{k=1}^{\|c_v\|} kS_{k,v} &= \sum_{k=1}^{\|c_v\|-1} k\left[(1-q^{-k})^2-(1-q^{-k+1})^2\right] + \|c_v\|\left[1-(1-q^{-\|c_v\|+1})^2\right] \\ &= \sum_{k=1}^{\|c_v\|-1} k(1-q^{-k})^2 + \|c_v\| - \sum_{k=1}^{\|c_v\|} k(1-q^{-k+1})^2.\end{aligned}$$

If we take a look at the first sum plus $\|c_v\|$, we see that

$$\begin{aligned}\sum_{k=1}^{\|c_v\|-1} k(1-q^{-k})^2 + \|c_v\| &= \sum_{k=1}^{\|c_v\|-1} k(1-2q^{-k}+q^{-2k}) + \|c_v\| \\ &= \sum_{k=1}^{\|c_v\|} k - 2\sum_{k=1}^{\|c_v\|-1} kq^{-k} + \sum_{k=1}^{\|c_v\|-1} kq^{-2k}.\end{aligned}$$

Further, we obtain

$$\begin{aligned}\sum_{k=1}^{\|c_v\|} k(1-q^{-k+1})^2 &= \sum_{k=1}^{\|c_v\|} k(1-2q^{-k+1}+q^{-2k+2}) \\ &= \sum_{k=1}^{\|c_v\|} k - 2\sum_{k=1}^{\|c_v\|} kq^{-k+1} + \sum_{k=1}^{\|c_v\|} kq^{-2k+2}.\end{aligned}$$

Subtracting the second from the first result we get

$$\begin{aligned}\sum_{k=1}^{\|c_v\|} kS_{k,v} &= 2(q-1)\sum_{k=1}^{\|c_v\|} kq^{-k} - (q^2-1)\sum_{k=1}^{\|c_v\|} kq^{-2k} \\ &\quad + \|c_v\|q^{-\|c_v\|}(2-q^{-\|c_v\|}) \\ &= 2\frac{q}{q-1}(1-p_v) - 2\|c_v\|p_v - \frac{q^2}{q^2-1}(1-p_v^2) + \|c_v\|p_v^2 \\ &\quad + \|c_v\|p_v(2-p_v) \\ &= 2\frac{q}{q-1}(1-p_v) - \frac{q^2}{q^2-1}(1-p_v^2).\end{aligned}$$

Here, the first equality follows from the previously calculated sums. The second equality holds since by assumption $q^{-\|c_v\|} = p_v$ for all $v \in \mathcal{U}$ and since we have for $j = 1, 2$ that

$$\sum_{k=1}^{\|c_v\|} kq^{-jk} = \frac{1}{(q^j-1)^2}[q^j - (q^j(\|c_v\|+1) - \|c_v\|)q^{-j\|c_v\|}]$$
$$= \frac{q^l}{(q^l-1)^2}(1-p_v^l) - \frac{\|c_v\|}{q^l-1}p_v^l.$$

Finally the above calculations yield

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) = \sum_{v\in\mathcal{U}} p_v \sum_{k=1}^{\|c_v\|} kS_{k,v}$$
$$= 2\frac{q}{q-1}\left(1 - \sum_{v\in\mathcal{U}} p_v^2\right) - \frac{q^2}{q^2-1}\left(1 - \sum_{v\in\mathcal{U}} p_v^3\right).$$

□

This result encourages us in the believe that the right side of the equation in Proposition 235 is in general a lower bound for 2-identification. As we will see soon it obeys some fundamental properties for entropy functions. Therefore, we define $H_{\mathrm{ID}}^{2,q} : \Gamma_N \to \mathbb{R}$ by

$$H_{\mathrm{ID}}^{2,q}(P) = 2\frac{q}{q-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^2\right) - \frac{q^2}{q^2-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^3\right). \tag{53}$$

We call it the *q-ary identification-entropy of second degree*. Its role as a lower bound for 2-identification is expressed in

Theorem 236 *Let $\mathcal{U}$ be a finite set and $q \in \mathbb{N}_{\geq 2}$. It holds for all probability distributions P on $\mathcal{U}$ and all q-ary prefix codes $\mathcal{C}$ that*

$$\mathcal{L}_{\mathcal{C}}^{2,q}(P,P) \geq H_{\mathrm{ID}}^{2,q}(P),$$

where equality is attained if and only if P consists only of q-powers, and $\mathcal{C}$ is a prefix code, with $\|c_u\| = -\log_q p_u$ for all $u \in \mathcal{U}$.

Before we prove Theorem 236, we will first analyze the functional properties of $H_{\mathrm{ID}}^{2,q}$. A list of desiderata for entropy functions can be found in [1], pp. 50. We now show that entropy function obeys important ones of them.

Theorem 237 *The following properties hold for $H_{\mathrm{ID}}^{2,q}(P)$.*

1. Symmetry:

$$H_{\mathrm{ID}}^{2,q}(p_1,\ldots,p_N) = H_{\mathrm{ID}}^{2,q}(p_{\pi(1)},\ldots,p_{\pi(N)}), \tag{54}$$

where π is a permutation on $[N]$.

2. Expansibility:

$$H_{\mathrm{ID}}^{2,q}(p_1,\ldots,p_N) = H_{\mathrm{ID}}^{2,q}(p_1,\ldots,p_N,0). \tag{55}$$

3. Decisiveness:

$$H_{\mathrm{ID}}^{2,q}(1,0,\ldots,0) = 0.$$

4. Normalization:

$$H_{\mathrm{ID}}^{2,q}\left(\frac{1}{q},\ldots,\frac{1}{q}\right) = 1. \tag{56}$$

5. Bounds:

$$H_{\mathrm{ID}}^{2,q}(1,0,\ldots,0) \le H_{\mathrm{ID}}^{2,q}(P) \le H_{\mathrm{ID}}^{2,q}\left(\frac{1}{N},\ldots,\frac{1}{N}\right). \tag{57}$$

6. Grouping Behavior: *For $m \le N$ let*

(a) $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_m$ be a partition of $\mathcal{U}$ of non-empty sets
(b) $Q = (Q_1,\ldots,Q_m)$ be the probability distribution on $[m]$ defined by $Q_i = \sum_{u\in\mathcal{U}_i} p_u$
(c) P_i is the probability distribution on $\mathcal{U}_i$ defined by $p_{i,u} = p_u/Q_i$ for all $i \in [m]$ and $u \in \mathcal{U}_i$.

It then holds that

$$H_{\mathrm{ID}}^{2,q}(P) = H_{\mathrm{ID}}^{2,q}(Q) + \sum_{i=1}^{m}\left[2Q_i^2(1-Q_i)H_{\mathrm{ID}}^{1,q}(P_i) + Q_i^3 H_{\mathrm{ID}}^{2,q}(P_i)\right]. \tag{58}$$

Proof Symmetry, expansibility and decisiveness follow directly from the definition of $H_{\mathrm{ID}}^{2,q}$. Further, the normalization property follows from

$$H_{\mathrm{ID}}^{2,q}\left(\frac{1}{q},\ldots,\frac{1}{q}\right) = 2\frac{q}{q-1}\left(1-\frac{1}{q}\right) - \frac{q^2}{q^2-1}\left(1-\frac{1}{q^2}\right) = 1.$$

Bounds: Let $f(p_1,\ldots,p_{N-1}) = H_{\mathrm{ID}}^{2,q}(p_1,\ldots,p_{N-1},1-\sum_{i=1}^{N-1} p_i)$. We will show that the gradient $\nabla f(p_1,\ldots,p_{N-1}) = \mathbf{0}$ if and only if $(p_1,\ldots,p_{N-1}) =$

$(1/N, \ldots, 1/N)$. For that we set $p_N = 1 - \sum_{i=1}^{N-1} p_i$ and obtain that it holds for all $j \in [N-1]$ that

$$\frac{\delta}{\delta p_j} f(p_1, \ldots, p_{N-1}) = -4\frac{q}{q-1}(p_j - p_N) + 3\frac{q^2}{q^2-1}(p_j^2 - p_N^2).$$

It follows immediately that $\nabla f(1/N, \ldots, 1/N) = \mathbf{0}$.

Assume now that for any $P' \neq (1/N, \ldots, 1/N)$ it holds that $\nabla f(P') = \mathbf{0}$. It follows that there exists $j \in [N-1]$ such that $p_j \neq p_N$. If we now take a look at $\frac{\delta}{\delta p_j} f(P')$, we see that

$$\frac{\delta}{\delta p_j} f(P') = 0 \qquad \Longleftrightarrow \qquad 3\frac{q}{q+1}(p_j + p_N) = 4.$$

This is a contradiction because $\frac{q}{q+1}(p_j + p_N)$ is clearly smaller than 1.

In order to ensure that $(1/N, \ldots, 1/N)$ is indeed a maximum we show that the Hessian is negative definite. In fact, we will obtain a stronger result namely that all second derivatives $\frac{\delta^2}{\delta p_k \delta p_j} f(1/N, \ldots, 1/N)$ are strictly negative.

$$\frac{\delta^2}{\delta p_k \delta p_j} f\left(\frac{1}{N}, \ldots, \frac{1}{N}\right) = \begin{cases} 4\frac{q}{q-1}\left(\frac{3q}{N(q+1)} - 2\right) & \text{if } k = j \\ \\ 2\frac{q}{q-1}\left(\frac{3q}{N(q+1)} - 2\right) & \text{if } k \neq j. \end{cases}$$

From $q \geq 2$ now follows that $\frac{3q}{N(q+1)} - 2 < 0$ if $N \geq 2$. And for $N = 1$ we are in the trivial case, where $H_{\mathrm{ID}}^{2,q}(1) = 0$.

Grouping Behavior:

We use

$$S_i = 2Q_i^2(1 - Q_i)H_{\mathrm{ID}}^{1,q}(P_i) + Q_i^3 H_{\mathrm{ID}}^{2,q}(P_i),$$

for all $i \in [m]$ and observe that

$$\begin{aligned} S_i &= 2Q_i^2(1 - Q_i)\frac{q}{q-1}\left(1 - \frac{1}{Q_i^2}\sum_{u \in \mathcal{U}_i} p_u^2\right) \\ &\quad + Q_i^3\left[2\frac{q}{q-1}\left(1 - \frac{1}{Q_i^2}\sum_{u \in \mathcal{U}_i} p_u^2\right) - \frac{q^2}{q^2-1}\left(1 - \frac{1}{Q_i^3}\sum_{u \in \mathcal{U}_i} p_u^3\right)\right] \\ &= 2\frac{q}{q-1}\left(Q_i^2 - \sum_{u \in \mathcal{U}_i} p_u^2\right) - \frac{q^2}{q^2-1}\left(Q_i^3 - \sum_{u \in \mathcal{U}_i} p_u^3\right). \end{aligned}$$

By summing the S_i's up we obtain

$$\sum_{i=1}^{m} S_i = 2\frac{q}{q-1}\left(\sum_{i=1}^{m} Q_i^2 - \sum_{u\in\mathcal{U}} p_u^2\right) - \frac{q^2}{q^2-1}\left(\sum_{i=1}^{m} Q_i^3 - \sum_{u\in\mathcal{U}} p_u^3\right)$$

and thus

$$\begin{aligned} H_{\mathrm{ID}}^{2,q}(Q) + \sum_{i=1}^{m} S_i &= 2\frac{q}{q-1}\left(1 - \sum_{i=1}^{m} Q_i^2\right) - \frac{q^2}{q^2-1}\left(1 - \sum_{i=1}^{m} Q_i^3\right) \\ &\quad + 2\frac{q}{q-1}\left(\sum_{i=1}^{m} Q_i^2 - \sum_{u\in\mathcal{U}} p_u^2\right) - \frac{q^2}{q^2-1}\left(\sum_{i=1}^{m} Q_i^3 - \sum_{u\in\mathcal{U}} p_u^3\right) \\ &= 2\frac{q}{q-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^2\right) - \frac{q^2}{q^2-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^3\right) \\ &= H_{\mathrm{ID}}^{2,q}(P). \end{aligned}$$

□

In order to prove Theorem 236 we need a decomposition formula for the 2-identification running time. It turns out that the decomposition of the 2-identification running time behaves mainly in the same way as the grouping behavior of the q-ary identification entropy of second degree. We prove this formula in its general form since we will also need this lemma in the next section.

Lemma 238 *For all $i \in \mathcal{Q}$ let*

1. $\mathcal{U}_i = \{u \in \mathcal{U} \,:\, c_{u,1} = i\}$
2. $Q_i = \sum_{u\in\mathcal{U}_i} p_u$
3. *P_i be a probability distribution on $\mathcal{U}_i$ defined by $p_{i,u} = \frac{p_u}{Q_i}$ for all $u \in \mathcal{U}_i$*
4. *$\mathcal{C}^{(i)} : \mathcal{U}_i \to \mathcal{Q}^*$ be the code on $\mathcal{U}_i$ defined by $c_u^{(i)} = c_{u,2}c_{u,3}\ldots c_{u,\|c_u\|}$ for all $u \in \mathcal{U}_i$.*

Then it holds that

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) = 1 + \sum_{i\in\mathcal{Q}}\sum_{l=1}^{L}\binom{L}{l} Q_i^{l+1}(1-Q_i)^{L-l}\mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}(P_i,P_i).$$

For $L = 2$ this becomes

$$\mathcal{L}_{\mathcal{C}}^{2,q}(P,P) = 1 + \sum_{i\in\mathcal{Q}}\left[2Q_i^2(1-Q_i)\mathcal{L}_{\mathcal{C}^{(i)}}^{1,q}(P_i,P_i) + Q_i^3\mathcal{L}_{\mathcal{C}^{(i)}}^{2,q}(P_i,P_i)\right].$$

Proof We observe that

$$\begin{aligned}\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) &= \sum_{u^L\in\mathcal{U}^L}\sum_{v\in\mathcal{U}} P_{u^L}^L p_v \mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v)\\ &= \sum_{i\in\mathcal{Q}}\sum_{v\in\mathcal{U}_i}\sum_{u^L\in\mathcal{U}^L} P_{u^L}^L p_v \mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v).\end{aligned}$$

Since $\mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v) = \mathcal{L}_{\mathcal{C}}^{L,q}((u_1,\ldots,u_L),v) = \mathcal{L}_{\mathcal{C}}^{L,q}((u_{\pi(1)},\ldots,u_{\pi(L)}),v)$ for all permutations π on $[L]$, we get for all $i\in\mathcal{Q}$

$$\begin{aligned}&\sum_{v\in\mathcal{U}_i}\sum_{u^L\in\mathcal{U}^L} P_{u^L}^L p_v \mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v)\\ &= \sum_{l=0}^{L}\binom{L}{l}\sum_{v\in\mathcal{U}_i}\sum_{u_1,\ldots,u_l\in\mathcal{U}_i}\sum_{u_{l+1},\ldots,u_L\in\mathcal{U}\setminus\mathcal{U}_i} P_{u^L}^L p_v \mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v)\\ &= \sum_{l=0}^{L}\binom{L}{l}(1-Q_i)^{L-l}\sum_{u_1,\ldots,u_l,v\in\mathcal{U}_i} p_{u_1}\ldots p_{u_l}p_v(1+\mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}((u_1,\ldots,u_l),v))\\ &= Q_i\sum_{l=0}^{L}\binom{L}{l}Q_i^l(1-Q_i)^{L-l} + \sum_{l=1}^{L}\binom{L}{l}(1-Q_i)^{L-l}Q_i^{l+1}\mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}(P_i,P_i)\\ &= Q_i + \sum_{l=1}^{L}\binom{L}{l}(1-Q_i)^{L-l}Q_i^{l+1}\mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}(P_i,P_i).\end{aligned}$$

The second equality follows since $\mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v) = 1+\mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}((u_1,\ldots,u_l),v)$ holds if $u_1,\ldots,u_l,v\in\mathcal{U}_i$ and $u_{l+1},\ldots,u_L\in\mathcal{U}\setminus\mathcal{U}_i$. Adding this up for $i\in\mathcal{Q}$ we obtain the desired result. □

As one can see there is a strong relation between the above decomposition formula for 2-identification and the grouping behavior of the identification entropy of second degree. In the following inductive proof of Theorem 236 we exploit this relation in order to apply the induction step.

Proof of Theorem 236. For $L=1$ the statement follows for all $N\in\mathbb{N}$ from Theorem 202 in chapter "Identification Entropy". As the induction base for N we have to consider all the cases $N=1,..,q$ and since here $\mathcal{L}_{\mathcal{C}}^{2,q}(P,P)=1$, we have to show that $H_{\mathrm{ID}}^{2,q}(P)\le 1$. It follows by the expansibility property (55) of the second degree identification entropy function that we only have to consider the case $N=q$. Further, the maximality of the uniform distribution (57) and the

normalization property (56) yield

$$H_{\mathrm{ID}}^{2,q}(p_1,\dots,p_q) \le H_{\mathrm{ID}}^{2,q}\left(\frac{1}{q},\dots,\frac{1}{q}\right) = 1.$$

We set $Q = (Q_0,\dots,Q_{q-1})$ and use the same notation as in Lemma 238. The inequality of Theorem 236 now follows from

$$\begin{aligned}\mathcal{L}_{\mathcal{C}}^{2,q}(P,P) &= 1 + \sum_{i\in\mathcal{Q}}\left[2Q_i^2(1-Q_i)\mathcal{L}_{\mathcal{C}^{(i)}}^{1,q}(P_i,P_i) + Q_i^3\mathcal{L}_{\mathcal{C}^{(i)}}^{2,q}(P_i,P_i)\right]\\ &\ge H_{\mathrm{ID}}^{2,q}(Q) + \sum_{i\in\mathcal{Q}}\left[2Q_i^2(1-Q_i)H_{\mathrm{ID}}^{1,q}(P_i) + Q_i^3 H_{\mathrm{ID}}^{2,q}(P_i)\right]\\ &= H_{\mathrm{ID}}^{2,q}(P). \end{aligned} \tag{59}$$

Here, the equality of the first line follows from Lemma 238. The inequality is a consequence of the induction step together with the normalization property (56) and the established bounds (57) of $H_{\mathrm{ID}}^{2,q}$. Finally, the grouping behavior (58) of $H_{\mathrm{ID}}^{2,q}$ yields the second equality.

The fact that this lower bound is attained for every q-ary prefix code $\mathcal{C}$ for which equality (51) holds has already been proven by Proposition 235. If we instead have that the inequality of Theorem 236 holds with equality, then also the inequality of Eq. (59) is in fact an equality and thus

(i) $H_{\mathrm{ID}}^{2,q}(Q) = 1$
(ii) $H_{\mathrm{ID}}^{1,q}(P_i) = \mathcal{L}_{\mathcal{C}_i}^{1,q}(P_i,P_i)$
(iii) $H_{\mathrm{ID}}^{2,q}(P_i) = \mathcal{L}_{\mathcal{C}_i}^{2,q}(P_i,P_i)$.

We have seen in the proof of the bounds of the entropy function that the uniform distribution is the only point where the first derivative of the identification entropy function equals zero and thus $(1/q,\dots,1/q)$ is the only point for which $H_{\mathrm{ID}}^{2,q}(Q) = 1$. Together with (i) this means that we get for all $i \in \mathcal{Q}$ that

$$Q_i = \frac{1}{q} \tag{60}$$

The crucial part is now (ii). For all $i \in \mathcal{Q}$ we obtain from Eq. (60) and the definitions of P_i and $\mathcal{C}^{(i)}$ (see Lemma 238) that for $u \in \mathcal{U}_i$ we have

$$p_u = Q_i\, p_{i,u} = \frac{p_{i,u}}{q} \tag{61}$$

and

$$\|c_u\| = \|c_u^{(i)}\| + 1. \tag{62}$$

Moreover, Theorem 213 in chapter "An Interpretation of Identification Entropy" stated that for (1-)identification an equality between the running time and identification entropy is only attained if and only if the probability distribution consists only of q-powers and the lengths of the codewords equal the negative logarithm of the probability of their corresponding elements. Thus it follows from (ii) that all the $p_{i,u}$'s are q-powers and that $\|c_u^{(i)}\| = -\log_q p_{i,u}$. Together with Eqs. (61) and (62) we finally obtain that P consists only of q-powers and that

$$\|c_u\| = -\log_q p_{i,u} + 1 = -\log_q \frac{p_{i,u}}{q} = -\log_q p_u.$$

□

In Theorem 227 we have shown for the uniform distribution that if $\mathcal{C}$ is a balanced Huffman code, its symmetric 2-identification running time asymptotically equals

$$K_{2,q} = 2\frac{q}{q-1} - \frac{q^2}{q^2-1}.$$

Since

$$\begin{aligned} H_{\mathrm{ID}}^{2,q}\left(\frac{1}{N},\ldots,\frac{1}{N}\right) &= 2\frac{q}{q-1}\frac{N-1}{N} - \frac{q^2}{q^2-1}\frac{N^2-1}{N^2} \\ &= 2\frac{q}{q-1} - \frac{q^2}{q^2-1} - 2\frac{q}{q-1}\frac{1}{N} + \frac{q^2}{q^2-1}\frac{1}{N^2} \end{aligned}$$

and thus

$$\lim_{N\to\infty} H_{\mathrm{ID}}^{2,q}\left(\frac{1}{N},\ldots,\frac{1}{N}\right) = K_{2,q},$$

we get

Corollary 239 *Considering the uniform distribution, balanced Huffman codes are asymptotically optimal for 2-identification.*

5.3 An Upper Bound for Binary Codes

In this subsection we establish an upper bound for $q = 2$. As said in the introduction of this section this is done mainly by the same code construction as in the second subsection of Sect. 3. We define $\mathcal{U}_{\max}$, $p_{\max}$ and $P_{\max}$ according to Eqs. (20), (21) and (22). Further, Eq. (23) becomes

$$\mathcal{L}_{\mathcal{C}}^{2,2}(P) \leq 1 + 2(1-p_{\max})p_{\max}\mathcal{L}_{\mathcal{C}_{\max}}^{1,2}(P_{\max}) + p_{\max}^2\mathcal{L}_{\mathcal{C}_{\max}}^{2,2}(P_{\max}). \tag{63}$$

We prove now by induction over N the following

Theorem 240 *It holds for all probability distributions P on $\mathcal{U}$ that the worst-case running time for binary 2-identification can be upper bounded by*

$$\mathcal{L}^{2,2}(P) < \frac{55}{16}.$$

Proof W.l.o.g. we assume that $p_1 \geq p_2 \geq \ldots \geq p_N$. As induction base serve the cases $N = 1, 2$ for which the running time always equals 1.

In order to apply the upper bound for (1-)identification, we use the same code construction as in Theorem 222. We partition $\mathcal{U}$ into sets $\mathcal{U}_0$ and $\mathcal{U}_1$, which differ from case to case. We choose t such that $|\frac{1}{2} - \sum_{u=1}^{t} p_u|$ is minimal and set

$$\mathcal{U}_0 = \begin{cases} \{1\} & \text{if } p_1 \geq \frac{1}{2} \\ \{1,2\} & \text{if } p_1 < \frac{1}{2} \text{ and } t = 1 \\ \{1,\ldots,t\} & \text{if } p_1 < \frac{1}{2} \text{ and } t \geq 2. \end{cases}$$

Once we have chosen $\mathcal{U}_0$ and $\mathcal{U}_1 = \mathcal{U} \setminus \mathcal{U}_0$ we inductively construct codes $\mathcal{C}_i$ on $\mathcal{U}_i$. Note that $\mathcal{C}_0 = \emptyset$ if $p_1 \geq 1/2$. From these codes we derive a code $\mathcal{C}$ on $\mathcal{U}$ by prefixing all codewords in $\mathcal{C}_i$ with i.

1. *Case* $p_1 \geq \frac{1}{2}$.

   For the same reason as in the proof of Theorem 222 we have that the element $v_{\max}$, which maximizes $\mathcal{L}_{\mathcal{C}}^{2,2}(P, v)$, is in $\mathcal{U}_1$. It follows by induction, Eq. (63) and Theorem 222 that

   $$\mathcal{L}_{\mathcal{C}}^{2,2}(P) < 1 + 5(1 - p_{\max})p_{\max} + \frac{55}{16}p_{\max}^2.$$

   Since the right hand side is monotone increasing in $p_{\max}$ and $p_{\max} \leq 1/2$ we obtain

   $$\mathcal{L}_{\mathcal{C}}^{2,2}(P) < 1 + \frac{5}{4} + \frac{55}{16} \cdot \frac{1}{4} = \frac{199}{64} < \frac{55}{16}.$$

   In the following, whenever there occurs the case that $p_{\max} \leq 1/2$ we obtain for the same reasons as above that $\mathcal{L}_{\mathcal{C}}^{2,2}(P) < 199/64 < 55/16$.

2. *Case* $p_1 < \frac{1}{2}$.

   (a) *Case* $t = 1$.
   We obtain by the definition of t that $\sum_{u=1}^{4} p_u > 1/2$. If $v_{\max} \in \mathcal{U}_0$ it follows that $\mathcal{L}_{\mathcal{C}}^{2,2}(P) \leq 2$. Further, we get for $v_{\max} \in \mathcal{U}_1$ that $p_{\max} < 1/2$.

   (b) *Case* $t \geq 2$.

(i) *Case* $v_{\max} \in \mathcal{U}_0$.
We have $p_{\max} = \sum_{u=1}^{t} p_u$. If $t = 2$, we again get that $p_{\max} \leq 1/2$ and if $t = 3$ we get as with Case 2 of the proof of Theorem 222 that

$$\mathcal{L}^{1,2}_{\mathcal{C}_{\max}}(P_{\max}) = 1 + \frac{p_2 + p_3}{p_{\max}} \leq \frac{5}{3}.$$

Further, for the same reasons we obtain

$$\mathcal{L}^{2,2}_{\mathcal{C}_{\max}}(P_{\max}) = 1 + \frac{2(p_2 + p_3)}{p_{\max}} \leq \frac{7}{3}.$$

Applying the above two equations together with Eqs. (25) and (63) yields

$$\mathcal{L}^{2,2}_{\mathcal{C}}(P) \leq 1 + 2 \cdot \frac{3}{8} \cdot \frac{5}{8} \cdot \frac{5}{3} + \frac{25}{64} \cdot \frac{7}{3} = \frac{517}{192} < \frac{55}{16}.$$

For $t \geq 4$ we get by Eq. (25) that

$$p_{\max} < \frac{7}{12}$$

if $p_{\max} \geq 1/2$. This together with the induction hypothesis and Theorem 222 yields

$$\mathcal{L}^{2,2}_{\mathcal{C}}(P) < 1 + 5 \cdot \frac{5}{12} \cdot \frac{7}{12} + \frac{49}{144} \cdot \frac{55}{16} = \frac{7799}{2304} < \frac{55}{16}.$$

(ii) *Case* $v_{\max} \in \mathcal{U}_1$.
We get $p_{\max} = \sum_{u=t+1}^{N} p_u$. Now, Eq. (26) yields

$$p_{\max} = 1 - \sum_{u=1}^{t} p_u \leq \frac{3}{5}.$$

From this it follows together with the induction hypothesis and Theorem 222 that

$$\mathcal{L}^{2,2}_{\mathcal{C}}(P) < 1 + 5 \cdot \frac{2}{5} \cdot \frac{3}{5} + \frac{9}{25} \cdot \frac{55}{16} = \frac{55}{16}.$$

□

We have established a lower and an upper bound for binary 2-identification so that we close this section with

Corollary 241 *It holds for all probability distributions P on $\mathcal{U}$ that*

$$4\left(1 - \sum_{u \in \mathcal{U}} p_u^2\right) - \frac{4}{3}\left(1 - \sum_{u \in \mathcal{U}} p_u^3\right) \leq \mathcal{L}^{2,2}(P, P) \leq \mathcal{L}^{2,2}(P) < \frac{55}{16}.$$

6 L-Identification for General Distributions

We now try to generalize the results of the preceding section. We begin with the definition of the q-ary identification entropy of degree L. Again, this function obeys some important desiderata for entropy functions. However, we did not succeed in proving the analogous lower and upper bounds for these entropies. In fact, there exist counterexamples to the natural conjecture that the uniform distribution is an upper bound. In order to show that $H_{\mathrm{ID}}^{L,q}$ is a lower bound for L-identification we only need the bounds for the case where the size of the output space equals the size of the alphabet. We show that we can prove $H_{\mathrm{ID}}^{L,q} \leq \mathcal{L}_{\mathcal{C}}^{L,q}(P, P)$ if we assume that in this case the uniform distribution is indeed an upper bound. Moreover, if we assume that for $N = q$ the uniform distribution is the only distribution for which the upper bound of $H_{\mathrm{ID}}^{L,q}$ is attained, we can show that again if and only if P consists only of q-powers we get that there exists a code $\mathcal{C}$ such that $H_{\mathrm{ID}}^{L,q}(P) = \mathcal{L}_{\mathcal{C}}^{L,q}(P, P)$.

Definition 242 Let $\mathcal{U}$ be a finite set with $|\mathcal{U}| = N$, $L \in \mathbb{N}$, $q \geq 2$ and $P = (p_1, \ldots, p_N) \in \Gamma_N$. Then the **q**-*ary identification entropy of degree L* $H_{\mathrm{ID}}^{L,q} : \Gamma_N \to \mathbb{R}$ is defined by

$$H_{\mathrm{ID}}^{L,q}(P) = -\sum_{l=1}^{L}(-1)^l \binom{L}{l} \frac{q^l}{q^l - 1}\left(1 - \sum_{u\in\mathcal{U}} p_u^{l+1}\right).$$

It is an easy observation that for $L = 1$ the above function equals the identification entropy established in chapter "Identification Entropy". Also for $L = 2$ it coincides with the identification entropy of second degree from the second subsection of Sect. 5.

This function again obeys important desiderata for entropies from [1]. It is clearly symmetric, expansible and decisive. It is also normalized. This follows from

$$H_{\mathrm{ID}}^{L,q}\left(\frac{1}{q}, \ldots, \frac{1}{q}\right) = -\sum_{l=1}^{L}(-1)^l \binom{L}{l} = 1. \tag{64}$$

Another interesting property is that $H_{\mathrm{ID}}^{L,q}$ obeys a grouping behavior which is a generalized version of the grouping behavior of the q-ary identification entropy of the second degree. With the same definitions as in 6. of Theorem 237 we obtain

$$H_{\mathrm{ID}}^{L,q}(P) = H_{\mathrm{ID}}^{L,q}(Q) + \sum_{i=1}^{m}\sum_{l=1}^{L}\binom{L}{l} Q_i^{l+1}(1 - Q_i)^{L-l} H_{\mathrm{ID}}^{l,q}(P_i). \tag{65}$$

To see this we set

$$S_i = \sum_{l=1}^{L} \binom{L}{l} Q_i^{l+1} (1 - Q_i)^{L-l} H_{\mathrm{ID}}^{l,q}(P_i),$$

for all $i \in [m]$ and observe that S_i equals

$$\begin{aligned}
&-\sum_{l=1}^{L} \binom{L}{l} Q_i^{l+1} (1 - Q_i)^{L-l} \sum_{k=1}^{l} (-1)^k \binom{l}{k} \frac{q^k}{q^k - 1} (1 - Q_i^{-(k+1)} \sum_{u \in \mathcal{U}_i} p_u^{k+1}) \\
&= -\sum_{k=1}^{L} (-1)^k \frac{q^k}{q^k - 1} (1 - Q_i^{-(k+1)} \sum_{u \in \mathcal{U}_i} p_u^{k+1}) \sum_{l=k}^{L} \binom{L}{l} \binom{l}{k} Q_i^{l+1} (1 - Q_i)^{L-l} \\
&= -\sum_{k=1}^{L} (-1)^k \binom{L}{k} \frac{q^k}{q^k - 1} (Q_i^{k+1} - \sum_{u \in \mathcal{U}_i} p_u^{k+1}) \sum_{l=k}^{L} \binom{L-k}{l-k} Q_i^{l-k} (1 - Q_i)^{L-l} \\
&= -\sum_{k=1}^{L} (-1)^k \binom{L}{k} \frac{q^k}{q^k - 1} (Q_i^{k+1} - \sum_{u \in \mathcal{U}_i} p_u^{k+1}).
\end{aligned}$$

Here, the last equality follows from

$$\sum_{l=k}^{L} \binom{L-k}{l-k} Q_i^{l-k} (1 - Q_i)^{L-l} = \sum_{l=0}^{L-k} \binom{L-k}{l} Q_i^{l} (1 - Q_i)^{L-l-k} = 1.$$

If we now replace k by l, we obtain

$$\sum_{i=1}^{m} S_i = -\sum_{l=1}^{L} (-1)^l \binom{L}{l} \frac{q^l}{q^l - 1} (\sum_{i=1}^{m} Q_i^{l+1} - \sum_{u \in \mathcal{U}} p_u^{l+1}).$$

This yields

$$\begin{aligned}
&H_{\mathrm{ID}}^{L,q}(Q) + \sum_{i=1}^{m} S_i \\
&= -\sum_{l=1}^{L} (-1)^l \binom{L}{l} \frac{q^l}{q^l - 1} \left(1 - \sum_{i=1}^{m} Q_i^{l+1} + \sum_{i=1}^{m} Q_i^{l+1} - \sum_{u \in \mathcal{U}} p_u^{l+1}\right) \\
&= H_{\mathrm{ID}}^{L,q}(P).
\end{aligned}$$

The crucial part are the lower and upper bound. It is natural for an entropy function that it is minimized if the probability is 1 for a single object and upper bounded by the uniform distribution. However, we encountered counterexamples such as $L \geq 4$, $q \geq 15$ and $N = 2$ or $L \geq 5$, $q \geq 100$ and $N = 3$. We conjecture that it holds at least for $N \geq q$ and all L and q that

$$H_{\mathrm{ID}}^{L,q}(1, 0, \ldots, 0) \leq H_{\mathrm{ID}}^{L,q}(P) \leq H_{\mathrm{ID}}^{L,q}\left(\frac{1}{N}, \ldots, \frac{1}{N}\right). \tag{66}$$

This claim, in fact just in the case $N = q$, would suffice to prove that $H_{\mathrm{ID}}^{L,q}$ is a lower bound for L-identification. We did not succeed in proving this claim in general for all L and q and will discuss this problem in greater detail in Sect. 8. Before we turn to the cases for which we were able to prove the desired bounds, we state

Proposition 243 *If Eq. (66) holds for $N = q$, we get*

$$H_{\mathrm{ID}}^{L,q}(P) \leq \mathcal{L}^{L,q}(P, P).$$

Proof We will use induction over L and N. As the induction base for L serves the case $L = 1$ for which it has been proven in chapter "Identification Entropy" that identification entropy (of first degree) is a lower bound for (1-)identification. Also the case $L = 2$ has been settled in the preceding Sect. 5.

The induction base for N is the case $N = q$. By the expansibility property this case settles all necessary induction bases $1, \ldots, q$. Trivially, if $\mathcal{C} = \mathcal{Q}$, we get that $\mathcal{L}_{\mathcal{C}}^{L,q}(P) = 1$. Since we have assumed that Eq. (66) holds, Eq. (64) proves this induction base.

To prove the proposition we partition $\mathcal{U}$ according to some given code $\mathcal{C}$ into $\mathcal{U}_0, \ldots, \mathcal{U}_{q-1}$, where $\mathcal{U}_i = \{u \in \mathcal{U} : c_{u,1} = i\}$. Further, let Q be a probability distribution on $\mathcal{Q}$ defined by $Q_i = \sum_{u \in \mathcal{U}} p_u$ and P_i be probability distributions on $\mathcal{U}$ defined by $P_{i,u} = p_u / Q_i$ for all $u \in \mathcal{U}$. With these definitions we obtain

$$\begin{aligned}
\mathcal{L}_{\mathcal{C}}^{L,q}(P, P) &= 1 + \sum_{i \in \mathcal{Q}} \sum_{l=1}^{L} \binom{L}{l} Q_i^{l+1} (1 - Q_i)^{L-l} \mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}(P_i^l, P_i) \\
&\geq H_{\mathrm{ID}}^{L,q}(Q) + \sum_{i=1}^{m} \sum_{l=1}^{L} \binom{L}{l} Q_i^{l+1} (1 - Q_i)^{L-l} H_{\mathrm{ID}}^{l,q}(P_i) \\
&= H_{\mathrm{ID}}^{L,q}(P).
\end{aligned} \tag{67}$$

Here the first equality follows from Lemma 238, the inequality from the normalization property (64), the assumed bounds (66) and the induction base. The final equality is a consequence of the grouping behavior (65). □

As stated before there are some cases for which we can prove Eq. (66). In fact, we prove more, namely

Proposition 244 *$H_{\mathrm{ID}}^{L,2}(P)$ is strictly concave for $L \leq 20$.*

Proof Let

$$f(p) = H_{\mathrm{ID}}^{L,2}(p, 1-p) = -\sum_{l=1}^{L}(-1)^l \binom{L}{l} \frac{2^l}{2^l - 1}\left(1 - p^{l+1} - (1-p)^{l+1}\right).$$

If we now look at all derivatives, we see that for $k = 1$

$$\frac{\delta^k}{\delta^k p} f(p) = \sum_{l=1}^{L}(-1)^l \binom{L}{l} \frac{2^l}{2^l - 1} \frac{(l+1)!}{(l-k+1)!}\left(p^{l-k+1} - (1-p)^{l-k+1}\right)$$

and for all $k \in \{2, \ldots, L+1\}$

$$\frac{\delta^k}{\delta^k p} f(p) = \sum_{l=k-1}^{L}(-1)^l \binom{L}{l} \frac{2^l}{2^l - 1} \frac{(l+1)!}{(l-k+1)!}\left(p^{l-k+1} + (-1)^k (1-p)^{l-k+1}\right).$$

A first observation is that if k is odd, we get

$$\frac{\delta^k}{\delta^k p} f\left(\frac{1}{2}\right) = 0.$$

If we sort $f(p)$ with respect to the power of p, we get

$$\begin{aligned} f(p) = {} & ((-1)^L - 1)\frac{2^L}{2^L - 1} p^{L+1} \\ & + \left((L+1)\frac{2^L}{2^L - 1} - (1 + (-1)^L) L \frac{2^{L-1}}{2^{L-1} - 1}\right) p^L + \sum_{l=1}^{L-1} \alpha_l p^l, \end{aligned}$$

for some α_l. This yields that for even L we have a polynomial of degree L with

$$\frac{\delta^L}{\delta^L p} f(p) = \left((L+1)\frac{2^L}{2^L - 1} - 2L \frac{2^{L-1}}{2^{L-1} - 1}\right) L! < 0$$

and for odd L we have a polynomial of degree $L+1$ with

$$\frac{\delta^{L+1}}{\delta^{L+1} p} f(p) = -2\frac{2^L}{2^L - 1}(L+1)! < 0.$$

Since for even (resp. odd) L the L-th (resp. $(L+1)$th) derivative is a strictly negative constant, we know that the $(L-2)$-th (resp. $(L-1)$th) derivative is a concave function. To show that it is also strictly negative it suffices to show that it is negative for $p = 1/2$ since the $(L-1)$-th (L-th) derivative is zero only at this point. This step can then be iterated and if we can show that all even derivatives are strictly negative at $p = 1/2$, we finally obtain that $H_{\mathrm{ID}}^{L,2}$ is a concave function. For $L = 2, \ldots, 20$ the values of all even derivatives at $p = 1/2$ have been computed and turn out to be strictly negative. □

For $L \geq 21$ there occur positive values within the even derivatives so that we cannot prove concavity via this argument. Nevertheless, also for these cases the graphs of the identification entropy functions let us assume that they are still concave. Since the binary identification entropy of degrees up to 20 are concave and symmetric, we obtain

Corollary 245 *Let $L \leq 20$ it then holds that*

$$H_{\mathrm{ID}}^{L,2}(1,0,\ldots,0) \leq H_{\mathrm{ID}}^{L,2}(P) \leq H_{\mathrm{ID}}^{L,2}\left(\frac{1}{N},\ldots,\frac{1}{N}\right),$$

with equality on the right hand side if and only if $P = (1/N, \ldots, 1/N)$.

The cases proved above and especially the strong connection between the grouping behavior (65) and Lemma 238 provide us with strong believe that the q-ary identification entropy of degree L is indeed a lower bound for the symmetric L-identification running time. But there are two other encouraging facts about the connection between those two concepts. The first is that we get for the uniform distribution the same result like for 2-identification. In fact, we have

$$H_{\mathrm{ID}}^{L,q}\left(\frac{1}{N},\ldots,\frac{1}{N}\right) = -\sum_{l=1}^{L}(-1)^l\binom{L}{l}\frac{q^l}{q^l-1}(1-\frac{1}{N^l})$$

yielding

$$\lim_{N\to\infty} H_{\mathrm{ID}}^{L,q}\left(\frac{1}{N},\ldots,\frac{1}{N}\right) = -\sum_{l=1}^{L}(-1)^l\binom{L}{l}\frac{q^l}{q^l-1}$$

and thus, if $\mathcal{C} \in \mathcal{C}_{q,N}$,

$$\lim_{N\to\infty} H_{\mathrm{ID}}^{L,q}\left(\frac{1}{N},\ldots,\frac{1}{N}\right) = \lim_{N\to\infty} \mathcal{L}_{\mathcal{C}}^{L,q}\left(\frac{1}{N},\ldots,\frac{1}{N}\right).$$

Therefore, a proof of Eq. (66) would also imply that for the case of the uniform distribution balanced Huffman codes are asymptotically optimal for L-identification.

The second encouraging fact is stated in the following

Proposition 246 *Let P be a probability distribution on $\mathcal{U}$ which consists only of q-powers and $\mathcal{C}$ be a code for $(\mathcal{U}, P)$ with $\|c_u\| = -\log_q p_u$ for all $u \in \mathcal{U}$. Then for all L and q it holds that*

$$H_{\mathrm{ID}}^{L,q}(P) = \mathcal{L}_{\mathcal{C}}^{L,q}(P, P).$$

Proof We first introduce for all $v \in \mathcal{U}$ and $k = 1, \ldots, \|c_v\|$ the following sets

- $\mathcal{U}_{v,k}^{L} = \{u^L \in \mathcal{U}^L \,:\, \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) = k\}$
- $\mathcal{U}_{v,k} = \{u \in \mathcal{U} \,:\, \mathcal{L}_{\mathcal{C}}^{1,q}(u, v) = k\}$
- $\bar{\mathcal{U}}_{v,k} = \mathcal{U}_{v,1} \dot{\cup} \ldots \dot{\cup} \mathcal{U}_{v,k-1}$

With this notation we obtain

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P, P) = \sum_{v \in \mathcal{U}} p_v \sum_{k=1}^{\|c_v\|} k \sum_{u^L \in \mathcal{U}_{v,k}^{L}} P_{u^L}^{L}.$$

We use $S_k = \sum_{u^L \in \mathcal{U}_{v,k}^{L}} p_{u_1} \ldots p_{u_L}$ and obtain

$$S_k = \sum_{l=1}^{L} \binom{L}{l} \sum_{u_1, \ldots, u_l \in \mathcal{U}_{v,k}} \sum_{u_{l+1}, \ldots, u_L \in \bar{\mathcal{U}}_{v,k}} p_{u_1} \ldots p_{u_L}.$$

Here, the second equality holds because there has to be at least one output for which identification against v takes exactly k time-steps while all others (or none if $l = L$) have an identification time regarding v of at most $k - 1$.

1. *Case $k = 1, \ldots, \|c_v\| - 1$.*
   In this case we have that $\mathcal{U}_{v,k} = \bar{T}_{c_v^{k-1}} \setminus \bar{T}_{c_v^k}$ and $\bar{\mathcal{U}}_{v,k} = \bar{T}_{\mathcal{C}} \setminus \bar{T}_{c_v^{k-1}}$. This yields

   $$\sum_{u \in \mathcal{U}_{v,k}} p_u = P(T_{c_v^{k-1}}) - P(T_{c_v^k}) = q^{-k+1} - q^{-k} = q^{-k}(q-1)$$

   and

   $$\sum_{u \in \bar{\mathcal{U}}_{v,k}} p_u = 1 - P(T_{c_v^{k-1}}) = 1 - q^{-k+1}$$

   and therewith

   $$S_k = \sum_{l=1}^{L} \binom{L}{l} q^{-kl}(q-1)^l (1 - q^{-k+1})^{L-l} = (1 - q^{-k})^L - (1 - q^{-k+1})^L.$$

2. *Case* $k = \|c_v\|$.
In this case we have that $\mathcal{U}_{v,\|c_v\|} = \bar{T}_{c_v^{\|c_v\|-1}}$ and $\bar{\mathcal{U}}_{v,\|c_v\|} = \bar{T}_{\mathcal{C}} \setminus \bar{T}_{c_v^{\|c_v\|-1}}$. We obtain

$$\sum_{u \in \mathcal{U}_{v,\|c_v\|}} p_u = P(T_{c_v^{\|c_v\|-1}}) = q^{-\|c_v\|+1}$$

and

$$\sum_{u \in \bar{\mathcal{U}}_{v,\|c_v\|}} p_u = 1 - P(T_{c_v^{\|c_v\|-1}}) = 1 - q^{-\|c_v\|+1}$$

and therewith

$$S_{\|c_v\|} = \sum_{l=1}^{L} \binom{L}{l} q^{-(\|c_v\|-1)l}(1-q^{-\|c_v\|+1})^{L-l} = 1 - (1-q^{-\|c_v\|+1})^L.$$

Combining the above two cases yields

$$\begin{aligned}\sum_{k=1}^{\|c_v\|} k S_k &= \sum_{k=1}^{\|c_v\|-1} k\left[(1-q^{-k})^L - (1-q^{-k+1})^L\right] + \|c_v\|\left[1-(1-q^{-\|c_v\|+1})^L\right] \\ &= \sum_{k=1}^{\|c_v\|-1} k(1-q^{-k})^L + \|c_v\| - \sum_{k=1}^{\|c_v\|} k(1-q^{-k+1})^L.\end{aligned}$$

We set $A = \sum_{k=1}^{\|c_v\|-1} k(1-q^{-k})^L + \|c_v\|$ and $B = \sum_{k=1}^{\|c_v\|} k(1-q^{-k+1})^L$. We then get for A

$$\begin{aligned}A &= \sum_{k=1}^{\|c_v\|-1} k \sum_{l=0}^{L} \binom{L}{l}(-1)^{L-l} q^{-(L-l)k} + \|c_v\| \\ &= \sum_{l=0}^{L} \binom{L}{l}(-1)^{L-l} \sum_{k=1}^{\|c_v\|-1} k q^{-(L-l)k} + \|c_v\| \\ &= \sum_{l=0}^{L-1} \binom{L}{l}(-1)^{L-l} \sum_{k=1}^{\|c_v\|-1} k q^{-(L-l)k} + \sum_{k=1}^{\|c_v\|} k \\ &= \sum_{l=0}^{L-1} \binom{L}{l}(-1)^{L-l} \sum_{k=1}^{\|c_v\|} k q^{-(L-l)k} - \sum_{l=0}^{L-1} \binom{L}{l}(-1)^{L-l}\|c_v\| q^{-(L-l)\|c_v\|} + \sum_{k=1}^{\|c_v\|} k\end{aligned}$$

and for B respectively

$$\begin{aligned} B &= \sum_{k=1}^{\|c_v\|} k \sum_{l=0}^{L} \binom{L}{l} (-1)^{L-l} q^{-(L-l)(k-1)} \\ &= \sum_{l=0}^{L} \binom{L}{l} (-1)^{L-l} q^{L-l} \sum_{k=1}^{\|c_v\|} k q^{-(L-l)k} \\ &= \sum_{l=0}^{L-1} \binom{L}{l} (-1)^{L-l} q^{L-l} \sum_{k=1}^{\|c_v\|} k q^{-(L-l)k} + \sum_{k=1}^{\|c_v\|} k. \end{aligned}$$

Subtracting B from A yields

$$\begin{aligned} \sum_{k=1}^{\|c_v\|} k S_k = &\sum_{l=0}^{L-1} \binom{L}{l} (-1)^{L-l} (1 - q^{L-l}) \sum_{k=1}^{\|c_v\|} k q^{-(L-l)k} \\ &- \sum_{l=0}^{L-1} \binom{L}{l} (-1)^{L-l} \|c_v\| q^{-(L-l)\|c_v\|}. \end{aligned}$$

Since

$$\sum_{k=1}^{\|c_v\|} k q^{-(L-l)k} = \frac{q^{L-l}}{(q^{L-l}-1)^2} (1 - q^{-(L-l)\|c_v\|}) - \frac{\|c_v\| q^{-(L-l)\|c_v\|}}{q^{L-l}-1}$$

and by assumption of the theorem $q^{-\|c_v\|} = p_v$ for all $v \in \mathcal{U}$, we finally obtain

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P, P) = \sum_{v \in \mathcal{U}} p_v \sum_{k=1}^{\|c_v\|} k S_k = -\sum_{l=1}^{L} \binom{L}{l} (-1)^l \frac{q^l}{q^l - 1} (1 - \sum_{v \in \mathcal{U}} p_v^{l+1}) = H_{\mathrm{ID}}^{L,q}(P).$$

□

According to the previous results for (1-) and 2-identifications it seems natural that the equality of Proposition 246 is only assumed for the mentioned cases and that we have a strict inequality between the q-ary identification entropy of degree L and the symmetric L-identification running time if P does not consists only of q-powers. The following proposition formalizes this if we assume that for $N = q$ the uniform distribution maximizes $H_{\mathrm{ID}}^{L,q}$ and that

$$H_{\mathrm{ID}}^{L,q}(P') < H_{\mathrm{ID}}^{L,q}\left(\frac{1}{q}, \ldots, \frac{1}{q}\right)$$

for all other distributions $P' \neq (1/q, \ldots, 1/q)$.

Proposition 247 *Let P be a probability distribution on $\mathcal{U}$ for which it holds that*

$$H_{\mathrm{ID}}^{L,q}(P) = \mathcal{L}_{\mathcal{C}}^{L,q}(P, P).$$

We further assume that $H_{\mathrm{ID}}^{L,q}(P') < H_{\mathrm{ID}}^{L,q}((1/q, \ldots, 1/q))$ for all $P' \neq (1/q, \ldots, 1/q)$. It then follows that P consists only of q-powers and $\mathcal{C}$ is a code for $(\mathcal{U}, P)$ with $\|c_u\| = -\log_q p_u$ for all $u \in \mathcal{U}$.

Proof As induction base serves the case $L = 1$, which has been proven in Theorem 213 in chapter "An Interpretation of Identification Entropy". For the induction steps it now follows from the assumptions of that the inequality in Eq. 67 becomes an equality so that we have

$$\begin{aligned} &1 + \sum_{i \in \mathcal{Q}} \sum_{l=1}^{L} \binom{L}{l} Q_i^{l+1} (1 - Q_i)^{L-l} \mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}(P_i^l, P_i) \\ &= H_{\mathrm{ID}}^{L,q}(Q) + \sum_{i=1}^{m} \sum_{l=1}^{L} \binom{L}{l} Q_i^{l+1} (1 - Q_i)^{L-l} H_{\mathrm{ID}}^{l,q}(P_i). \end{aligned} \tag{68}$$

For the definitions of Q_i, P_i and $\mathcal{C}^{(i)}$ see again Lemma 238. From this equation follows

(i) $H_{\mathrm{ID}}^{L,q}(Q) = 1$
(ii) $H_{\mathrm{ID}}^{l,q}(P_i) = \mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}(P_i^l, P_i)$ for $l \in [L]$

On the one hand it follows from the assumptions and i) that $Q = (1/q, \ldots, 1/q)$ and on the other hand it follows from the induction hypothesis and ii) that P_i consists only of q-powers and that $\|c_u^{(i)}\| = -\log_q p_{i,u}$. Since $p_u = Q_i p_{i,u} = p_{i,u}/q$ for all $u \in \mathcal{U}_i$, we obtain that also P consists only of q-powers and finally $\|c_u\| = -\log_q p_{i,u} + 1 = -\log_q \frac{p_{i,u}}{q} = -\log_q p_u$ for all $u \in \mathcal{U}_i$. □

7 L-Identification for Sets

Like before the discrete source $(\mathcal{U}, P)$ together with a source code $\mathcal{C}$ forms the basis for our analysis of L-identification for sets. Unlike in the second subsection of Sect. 2, however, we do not consider as the output space the discrete memoryless source $(\mathcal{U}^L, P^L)$ but the discrete source $(\tilde{\mathcal{U}}, \tilde{P})$, where $\tilde{\mathcal{U}} = \binom{\mathcal{U}}{L}$. We write $\tilde{P}_S$ for $\tilde{P}(\{S\})$. The task of L-identification for sets is in principle the same as before. It has to be able to distinguish for all users $v \in \mathcal{U}$ and all outputs $S \in \tilde{\mathcal{U}}$ whether there exists an element u in S with $u = v$ or not.

In this section we will analyze the asymptotic behavior of the symmetric running time of L-identification for sets for the case when $\tilde{P}$ is the uniform distribution on

$\tilde{\mathcal{U}}$ and also the users are chosen uniformly. We will see that it asymptotically equals the symmetric running time of L-identification (for vectors) and thus $K_{L,q}$, which was examined in the second subsection of Sect. 4.

It is clear that L-identification for sets can be seen as a special case of our preliminary L-identification (for vectors) as we exclude all vectors with two or more identical components. This fact changes the running time of L-identification in the following way. Again, we compare q-bit by q-bit the codewords of the elements of S to the corresponding q-bit of c_v and after every step we cancel out all elements which do not coincide. Suppose after some step k during the identification process we are left over with the same amount of possible candidates as there are codewords in $\bar{\mathcal{N}}(T_{c_v^k})$. Since we are considering sets and not vectors, we know that each of the elements which belong to the codewords in $\bar{\mathcal{N}}(T_{c_v^k})$ are elements of S and so does v itself. At such a point we terminate the identification process and answer: "Yes, v is in S!". Figure 6 shows an example of such an event for $N = 17$ and $L = 9$. In this example v equals u_1. This is indicated by the thick path from the root to u_1. After the first q-ary comparison u_5 and u_7 are deleted from the set of possible candidates but there are more than seven codewords which begin with $\mathbf{0}$ so that v still might be not contained in S. After the second comparison u_2 and u_9 are canceled and we still have more codewords in $\bar{\mathcal{N}}(T_{\mathbf{00}})$ than possible candidates. After the third step, however, u_6 is not longer a candidate. This leaves us with four possible candidates. Since $|\bar{\mathcal{N}}(T_{\mathbf{000}})| = 4$, we know that v has to be an element of S and terminate the L-identification process.

The L-identification algorithm `LID` now becomes the L-identification algorithm for sets. It is called `LIDforSets` and stated in Table 3 in the appendix. Now let $S = \{u_1, u_2, \ldots, u_L\} \in \tilde{\mathcal{U}}$ we then define the L-identification time for S, an user v and a q-ary code $\mathcal{C}$ by

$$\tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(S, v) = \texttt{LIDforSets}_2(c_{u_1}, \ldots, c_{u_L}, c_v), \tag{69}$$

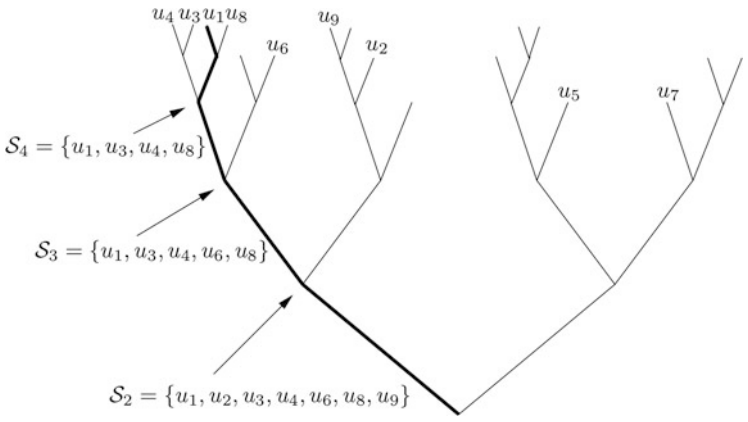


Fig. 6 An example when the 9-identification process terminates because $|\mathcal{S}_4| = |\{u \in \mathcal{U} : c_u^3 = c_v^3\}| = 4$. For the definition of $\mathcal{S}_i$ see Table 3 in the appendix

where $\texttt{LIDforSets}_2(c_{u_1}, \ldots, c_{u_L}, c_v)$ is the second component of the return pair of the algorithm `LIDforSets`.

In the same way as in the second subsection of Sect. 2 we now define the *average running time* for a given user $v \in \mathcal{U}$ by

$$\tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}, v) = \sum_{S \in \tilde{\mathcal{U}}} \tilde{P}_S \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(S, v), \tag{70}$$

the *worst-case running-time* by

$$\tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}) = \max_{v \in \mathcal{U}} \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}, v) \tag{71}$$

and if we have a probability distribution Q on $\mathcal{U}$, we define the *expected running time* by[3]

$$\tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}, Q) = \sum_{v \in \mathcal{U}} Q(\{v\}) \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}, v). \tag{72}$$

In both scenarios we are again interested in the optimal running time. That is

$$\tilde{\mathcal{L}}^{L,q}(\tilde{P}) = \min_{\mathcal{C}} \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}) \tag{73}$$

and

$$\tilde{\mathcal{L}}^{L,q}(\tilde{P}, Q) = \min_{\mathcal{C}} \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}, Q). \tag{74}$$

We will now take a look at the asymptotic behavior of $\tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}, Q)$ for the case when both $\tilde{P}$ and Q are uniform distributions on $\tilde{\mathcal{U}}$, resp. $\mathcal{U}$, and that $\mathcal{C} \in \mathcal{C}_{q,N}$ is a balanced Huffman code. In this case we call $\tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}, Q)$ as before the *symmetric running time* for L-identification for sets. In order to simplify notation we shall write $\bar{P} = \left(\binom{N}{L}^{-1}, \ldots, \binom{N}{L}^{-1}\right)$. Equation (72) then becomes

$$\tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}\left(\bar{P}, (\frac{1}{N}, \ldots, \frac{1}{N})\right) = \frac{1}{N\binom{N}{L}} \sum_{S \in \tilde{\mathcal{U}}} \sum_{v \in \mathcal{U}} \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(S, v). \tag{75}$$

It turns out that

$$\lim_{N \to \infty} \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}\left(\bar{P}, (\frac{1}{N}, \ldots, \frac{1}{N})\right) = \lim_{N \to \infty} \mathcal{L}_{\mathcal{C}}^{L,q}\left((\frac{1}{N}, \ldots, \frac{1}{N}), (\frac{1}{N}, \ldots, \frac{1}{N})\right) \tag{76}$$

[3]Remember that all those functions implicitly depend also on $N = |\mathcal{U}|$ via $\mathcal{C}$, $\tilde{P}$ and Q.

and thus equals the same rational number $K_{L,q}$ which has been examined in the second subsection of Sect. 4. This may be somewhat surprising at first glance since the output spaces $\mathcal{U}^L$ and $\tilde{\mathcal{U}}$ as well as the underlying algorithms differ from each other. Yet, it becomes clear if we take into account that these differences "disappear" if N goes to infinity. By this we mean that the cardinality of the family of sets, which cause the algorithm `LIDforSets` to terminate with a positive answer before it reaches the last step, is so small that its probability goes to zero as N tends to infinity. The same is true for the set of all vectors which have more than one identical component. We will now formalize the above explanations in order to prove Eq. (76).

Let $f : \mathcal{U}^L \to \bigcup_{l=1}^{L} \binom{\mathcal{U}}{l}$ be defined by $f(u^L) = \bigcup_{i=1}^{L} \{u_i\}$. Further, let $\mathcal{U}' \subset \mathcal{U}$ be the set of all vectors whose components are pairwise distinct. It follows that the restriction $f|_{\mathcal{U}'}$ is a surjective mapping from $\mathcal{U}'$ onto $\tilde{\mathcal{U}}$ and that $|f^{-1}(S)| = L!$ for all $S \in \tilde{\mathcal{U}}$. This yields $|\mathcal{U}'| = L!\binom{N}{L}$ and

$$\begin{aligned}&\mathcal{L}_{\mathcal{C}}^{L,q}\left(\left(\frac{1}{N},\ldots,\frac{1}{N}\right),\left(\frac{1}{N},\ldots,\frac{1}{N}\right)\right)\\&= \frac{1}{N^{L+1}}\sum_{v\in\mathcal{U}}\left[\sum_{u^L\in\mathcal{U}'}\mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v)+\sum_{u^L\in\mathcal{U}\setminus\mathcal{U}'}\mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v)\right].\end{aligned} \tag{77}$$

Since $\sum_{u^L\in\mathcal{U}\setminus\mathcal{U}'}\mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v) \leq (1+\log_q N)L!\binom{N}{L}$, it follows that the second summand multiplied by $1/N^L$ tends to zero for $N \to \infty$.

We now turn to $\tilde{\mathcal{U}}$ and assume that $N = q^n$ such that $\mathcal{C} = \mathcal{C}_{q^n}$.[4] We define $\tilde{\mathcal{U}}' \subset \tilde{\mathcal{U}}$ to be the family of sets S for which there exists at least one leaf in each subtree with root in level $n-1$ which is not contained in S. We use $T = T_{\mathcal{C}_{q^n}}$ and obtain

$$\tilde{\mathcal{U}}' = \{S \in \tilde{\mathcal{U}} \,:\, \bar{\mathcal{N}}(T_x) \setminus S \neq \emptyset \;\forall x \in \mathcal{Q}^{n-1}\}.$$

It follows that from the nature of the algorithms `LID` and `LIDforSets` that for all $v \in \mathcal{U}$, $S \in \tilde{\mathcal{U}}'$ and $u^L \in f^{-1}(S)$ we have that

$$\tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(S,v) = \mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v). \tag{78}$$

It is clear that if $L < q$, we get that $\tilde{\mathcal{U}}' = \tilde{\mathcal{U}}$ and if $L \geq q$, we obtain that

$$\tilde{\mathcal{U}} \setminus \tilde{\mathcal{U}}' = \bigcup_{x\in\mathcal{Q}^{n-1}}\left(\bar{\mathcal{N}}(T_x) \cup \binom{\tilde{\mathcal{U}}\setminus\bar{\mathcal{N}}(T_x)}{L-q}\right).$$

[4]The analysis for $N \neq q^n$, which we omit, involves the same calculations but includes some more case distinctions.

From this follows that

$$|\tilde{\mathcal{U}} \setminus \tilde{\mathcal{U}}'| \leq \sum_{x \in \mathcal{Q}^{n-1}} \left| \left(\bar{\mathcal{N}}(T_x) \cup \binom{\tilde{\mathcal{U}} \setminus \bar{\mathcal{N}}(T_x)}{L-q} \right) \right|$$
$$= q^{n-1} \left(q + \binom{N-q}{L-q} \right) = N + \frac{N}{q} \binom{N-q}{L-q}.$$

This yields

$$\frac{1}{N\binom{N}{L}} \sum_{v \in \mathcal{U}} \sum_{S \in \tilde{\mathcal{U}} \setminus \tilde{\mathcal{U}}'} \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(S, v) \leq \frac{1}{\binom{N}{L}} \log_q N |\tilde{\mathcal{U}} \setminus \tilde{\mathcal{U}}'|$$
$$\leq \frac{1}{\binom{N}{L}} \log_q N \left(N + \frac{N}{q} \binom{N-q}{L-q} \right).$$

The right hand side of the third line tends to zero as N goes to infinity. We return to L-identification for vectors and similar to the definition of $\tilde{\mathcal{U}}'$ we define

$$\mathcal{U}'' = \{u^L \in \mathcal{U}' \, : \, \forall x \in \mathcal{Q}^{n-1} \, \exists \, w \in \bar{\mathcal{N}}(T_x) \text{ and } l \in [L] \text{ s.t. } w \neq u_l\}$$

and for similar reasons as above we obtain that for $N \to \infty$

$$\frac{1}{N^{L+1}} \sum_{v \in \mathcal{U}} \sum_{u^L \in \mathcal{U}' \setminus \mathcal{U}''} \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) \to 0.$$

Finally, we can partition $\mathcal{U}'' = \bigcup_{S \in \tilde{\mathcal{U}}'} f^{-1}(S)$ and get

$$\frac{1}{N^{L+1}} \sum_{v \in \mathcal{U}} \sum_{u^L \in \mathcal{U}''} \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) = \frac{1}{N^{L+1}} \sum_{v \in \mathcal{U}} \sum_{S \in \tilde{\mathcal{U}}'} \sum_{u^L \in f^{-1}(S)} \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v)$$
$$= \frac{L!}{N^{L+1}} \sum_{v \in \mathcal{U}} \sum_{S \in \tilde{\mathcal{U}}'} \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(S, v),$$

where the last equality follows from Eq. (78). Since $L!/N^L$ asymptotically equals $1/\binom{N}{L}$, we finally proved

Theorem 248 *Let $L, n \in \mathbb{N}$, $q \in \mathbb{N}_{\geq 2}$, $q^{n-1} < N \leq q^n$, $\mathcal{C} \in \mathcal{C}_{q,N}$ and $\bar{P}$ be the uniform distribution on $\tilde{\mathcal{U}}$. Then it holds that*

$$\lim_{N \to \infty} \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q} \left(\bar{P}, (\frac{1}{N}, \ldots, \frac{1}{N}) \right) = \lim_{N \to \infty} \mathcal{L}_{\mathcal{C}}^{L,q} \left((\frac{1}{N}, \ldots, \frac{1}{N}), (\frac{1}{N}, \ldots, \frac{1}{N}) \right) = K_{L,q},$$

where $K_{L,q} \in \mathbb{R}$ is defined in Theorem 227.

8 Open Problems

In this final section we will give an overview of some open problems which arose during the study of L-identification. We begin with three types of problems concerning L-identification (for vectors). The first is to settle the induction base in the proof of Proposition 243. It is the only fragment left in order to completely prove that q-ary identification entropy $H_{\mathrm{ID}}^{L,q}$ of degree L is a lower bound for L-identification.

The second problem is a generalization of Lemmas 220 and 231 where we proved that concerning block codes the uniform distribution is optimal for (1-) and 2-identification. At least for $L \geq 4$ this is not longer true in general as there exist simple counterexamples. However, we claim that if the size of the block is sufficiently large, again uniform distribution becomes optimal.

The second subsection covers L-identification for sets. We have seen in Sect. 7 that for the uniform distribution L-identification for sets behaves in the same way as L-identification (for vectors) if the cardinality of the output space tends to infinity. Unfortunately we have not made any major discoveries if we turn to general distributions.

8.1 Induction Base for the Proof of Proposition 243

The most important problem is to settle for all L and q the induction base $N = q$ of the proof of Proposition 243. With the solution of this problem we would obtain that the q-ary identification entropy $H_{\mathrm{ID}}^{L,q}$ of degree L is a lower bound for L-identification. In the following we establish a chain of problems which are partly subproblems. Figure 7 visualizes this chain.

Problem 1

Show that it holds for all L, q and probability distributions P on $[q]$ that

$$H_{\mathrm{ID}}^{L,q}(P) \leq 1. \tag{79}$$

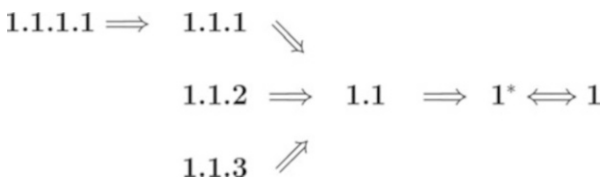


Fig. 7 The logical chain of the problems leading to a proof of Proposition 243

Since $H_{\mathrm{ID}}^{L,q}$ is normalized (see Eq. (64)), the above problem is equivalent to

Problem 1*

Show that it holds for all L, q and probability distributions P on $[q]$ that

$$H_{\mathrm{ID}}^{L,q}(P) \leq H_{\mathrm{ID}}^{L,q}\left(\frac{1}{q}, \ldots, \frac{1}{q}\right). \tag{80}$$

We have claimed in Sect. 6 that Eq. (66) holds which solves problem 1* in the more general form where $N \geq q$. This yields

Problem 1.1

Show that it holds for all L, q and probability distributions P on $[N]$, where $N \geq q$, that

$$H_{\mathrm{ID}}^{L,q}(1, 0, \ldots, 0) \leq H_{\mathrm{ID}}^{L,q}(P) \leq H_{\mathrm{ID}}^{L,q}\left(\frac{1}{N}, \ldots, \frac{1}{N}\right).$$

We provide three approaches which possibly are suitable for solving Problem 1.1. The first is somewhat in the spirit of Lemmas 220 and 231 where we step by step adjust an arbitrary probability distribution so that it becomes the uniform distribution without increasing the symmetric L-identification running time. For this let $P \neq (1/N, \ldots, 1/N)$ be a probability distribution on $[N]$. Remember that we assumed $N \geq q$. Clearly, there exists an element, say 1, for which $p_1 > 1/N$ and an element, say 2, for which $p_2 < 1/N$. We now construct a new probability distribution $\bar{P}$ by setting $\bar{p}_1 = \bar{p}_2 = (p_1 + p_2)/2$ and $\bar{p}_i = p_i$, for all $i \in \{3, \ldots, N\}$. If we can show that $H_{\mathrm{ID}}^{L,q}(\bar{P}) - H_{\mathrm{ID}}^{L,q}(P) \geq 0$, we would have solved problem 1.1 since we can come arbitrarily close to $(1/N, \ldots, 1/N)$ by applying the above construction iteratively and sufficiently many times. Thus we state

Problem 1.1.1

Show that it holds for all L, q and probability distributions P on $[N]$, where $N \geq q$, that

$$H_{\mathrm{ID}}^{L,q}(\bar{P}) - H_{\mathrm{ID}}^{L,q}(P) \geq 0,$$

where $\bar{P}$ is defined by $\bar{p}_1 = \bar{p}_2 = (p_1 + p_2)/2$ and $\bar{p}_i = p_i$ for all $i \in \{3, \ldots, N\}$.

We begin the calculation of this difference and obtain

$$\begin{aligned} &H_{\mathrm{ID}}^{L,q}(\bar{P}) - H_{\mathrm{ID}}^{L,q}(P) \\ &= \sum_{l=1}^{L} (-1)^l \binom{L}{l} \frac{q^l}{q^l - 1} \left(\frac{1}{2^l}(p_1 + p_2)^{l+1} - p_1^{l+1} - p_2^{l+1}\right) \end{aligned}$$

$$= \sum_{l=1}^{L} (-1)^l \binom{L}{l} \sum_{t \geq 0} q^{-tl} \left(\frac{1}{2^l} (p_1 + p_2)^{l+1} - p_1^{l+1} - p_2^{l+1} \right)$$

$$= \sum_{i=1}^{2} p_i \sum_{t \geq 0} \left[\left(1 - \frac{p_1 + p_2}{2q^t} \right)^L - (1 - \frac{p_i}{q^t})^L \right].$$

Note that while the first summand is positive the second one is negative. Yet the positive summand is weighted by p_1 which is greater than p_2 by which the negative summand is weighted. We therefore feel that the following problem may be a good candidate for solving the main problem 1. One has to keep in mind that $N \geq q$ is crucial so this fact has to come in play.

Problem 1.1.1.1

Show that if $N \geq q$, $p_1 + p_2 \leq 1$ and $p_1 > 1/N > p_2$, we get that

$$\sum_{i=1}^{2} p_i \sum_{t \geq 0} \left[\left(1 - \frac{p_1 + p_2}{2q^t} \right)^L - (1 - \frac{p_i}{q^t})^L \right] \geq 0.$$

We also could try to prove problem 1.1 via the direct way. For this consider an probability distribution P on $[N]$ (still $N \geq q$) and assume w.l.o.g. that

$$p_1 \geq p_2 \geq \ldots \geq p_{n_1} > \frac{1}{N} > p_{n_1+1} \geq \ldots \geq p_{n_2} \tag{81}$$

and $p_{n_2+1} = \ldots = p_N = 1/N$. With the same calculations as above we obtain

$$H_{\mathrm{ID}}^{L,q} \left(\frac{1}{N}, \ldots, \frac{1}{N} \right) - H_{\mathrm{ID}}^{L,q}(P) = \sum_{i=1}^{n_2} p_i \sum_{t \geq 0} \left[\left(1 - \frac{1}{Nq^t} \right)^L - (1 - \frac{p_i}{q^t})^L \right].$$

Again the first n_1 summands are positive and weighted by the greater weights $p_1, \ldots, p_{n_1}$. We obtain

Problem 1.1.2

Show that if $N \geq q$ and if $(p_1, \ldots, p_N)$ obeys equation (81), we get that

$$\sum_{i=1}^{n_2} p_i \sum_{t \geq 0} \left[\left(1 - \frac{1}{Nq^t} \right)^L - (1 - \frac{p_i}{q^t})^L \right] \geq 0.$$

Another approach would be to follow the proof of the bounds for the q-ary identification entropy of second degree (see Theorem 237). In this proof we analyze the first derivative of the entropy function and showed that there exists only one extremal point namely a maximum at $(1/N, \ldots, 1/N)$. As we have mentioned in

the definition section we only have to consider $N-1$ partial derivatives and obtain for $v \in [N-1]$

$$\frac{\delta}{\delta p_v} H_{\mathrm{ID}}^{L,q} = \sum_{l=1}^{L} (-1)^l \binom{L}{l} \frac{q^l}{q^l - 1} (l+1) \left(p_v^l - (1 - \sum_{u=1}^{N-1} p_u)^l \right).$$

This obviously is zero if $p_1 = \ldots = p_{N-1} = 1/N$. We are left with

Problem 1.1.3

Show that $p_1 = \ldots = p_{N-1} = 1/N$ is the only point in Δ_{N-1} which is for all $v \in [N-1]$ the root of

$$\sum_{l=1}^{L} (-1)^l \binom{L}{l} \frac{q^l}{q^l - 1} (l+1) \left(p_v^l - (1 - \sum_{u=1}^{N-1} p_u)^l \right).$$

8.2 **L**-*Identification for Block Codes*

In the first subsection of Sect. 3 and Corollary 232 we proved that concerning block codes the uniform distribution is optimal for the symmetric running time of (1-) and 2-identification. This, however, is not longer true at least for $L \geq 4$. We can show this by an easy example. Therefore consider $q = 2$, $N = 4$, $L = 4$ and $\mathcal{C} = \mathcal{C}_{2^2}$. It follows with the notation of the second subsection of Sect. 4 that

$$\begin{aligned}
&\mathcal{L}_{\mathcal{C}}^{4,2}\left((\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}), (\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}) \right) \\
&= \frac{1}{4^4} \left(|\mathcal{R}_{\mathcal{C}}^{4,2}(1,1)| + 2|\mathcal{R}_{\mathcal{C}}^{4,2}(2,1)| \right) \\
&= \frac{1}{4^4} \left(2^4 + 22^4(2^4 - 1) \right) \\
&= \frac{31}{16}.
\end{aligned}$$

We now take the probability distribution $P = (1/8, 1/8, 3/8, 3/8)$. The assignment of the individual probabilities to the codewords (resp. the corresponding outputs) is depicted in Fig. 8. We obtain

$$\begin{aligned}
&\mathcal{L}_{\mathcal{C}}^{4,2}\left((\tfrac{1}{8}, \tfrac{1}{8}, \tfrac{3}{8}, \tfrac{3}{8}), (\tfrac{1}{8}, \tfrac{1}{8}, \tfrac{3}{8}, \tfrac{3}{8}) \right) \\
&= \sum_{i=1}^{2} \sum_{v=2(i-1)+1}^{2i} p_v \sum_{l=0}^{4} \binom{4}{l} \sum_{u_1,\ldots,u_l=2(i-1)+1}^{2i} \sum_{u_{l+1},\ldots,u_4 \in [4] \setminus \{2(i-1)+1, 2i\}} P_{u^4} \mathcal{L}(u^4, v)
\end{aligned}$$

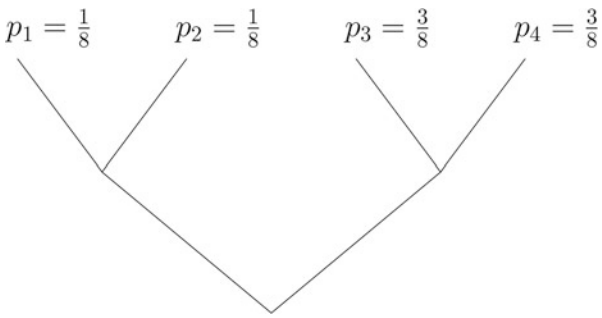


Fig. 8 An example for 4-identification on block codes which has a faster symmetric running time than the uniform distribution

$$= \frac{1}{4}\left(\frac{3^4}{2^8} + \frac{1}{2^7}\sum_{l=1}^{4}\binom{4}{l}3^{4-l}\right) + \frac{3}{4}\left(\frac{1}{2^8} + \frac{1}{2^7}\sum_{l=1}^{4}\binom{4}{l}3^{l}\right)$$

$$= \frac{491}{256} < \frac{496}{256} = \frac{31}{16} = \mathcal{L}_{\mathcal{C}}^{4,2}(\frac{1}{4}, \ldots, \frac{1}{4}).$$

This inconsistency disappears for 4-identification already for the next level, where $N = 8$. In general we claim for all L that if the block code is large enough, the uniform distribution becomes optimal again. This is the content of

Problem 2

Show that for all L exists $n_L \in \mathbb{N}$ such that it holds for all $n \geq n_L$ and all probability distributions P on $[q^n]$ that

$$\mathcal{L}_{\mathcal{C}_{q^n}}^{L,q}(P, P) \geq \mathcal{L}_{\mathcal{C}_{q^n}}^{L,q}\left((\frac{1}{q^n}, \ldots, \frac{1}{q^n}), (\frac{1}{q^n}, \ldots, \frac{1}{q^n})\right).$$

Of course, we cannot solve this problem by applying generalized versions of Lemmas 220 and 231. Since these lemmas are applied to small subtrees in the beginning, we would get that during the first modifications of some given probability the symmetric running time would increase if we level out the corresponding probabilities. But we think that these small increases are absorbed by later steps where we level out bigger and bigger subtrees. A big help in order to solve problem 2 would be if we could establish an exact expression for the differences like we have done in Lemma 220. With this we would hopefully be able to solve problem 2. However, already for $L = 2$ we do not have such an expression. Like before in the corresponding lemmas for (1-) and 2-identification let $n \in \mathbb{N}$, $q \in \mathbb{N}_{\geq 2}$, $k \in \{0, \ldots, n-1\}$ and $t \in \{0, \ldots, q^{n-k-1} - 1\}$. Further, let $P = (p_1, \ldots, p_{q^n})$

and $\tilde{P} = (\tilde{p}_1, \ldots, \tilde{p}_{q^n})$ be probability distributions on $[q^n]$ with

$$P = (p_1, \ldots, p_{tq^{k+1}}, \underbrace{r_1, \ldots, r_1}_{q^k}, \underbrace{r_2, \ldots, r_2}_{q^k}, \ldots, \underbrace{r_q, \ldots, r_q}_{q^k}, p_{(t+1)q^{k+1}+1}, \ldots, p_{q^n})$$

and

$$\tilde{P} = (p_1, \ldots, p_{tq^{k+1}}, \underbrace{\frac{1}{q}\sum_{i=1}^{q} r_i, \ldots, \frac{1}{q}\sum_{i=1}^{q} r_i}_{q^{k+1}}, p_{(t+1)q^{k+1}+1}, \ldots, p_{q^n}).$$

Problem 2.1

Establish for $L \geq 2$ an exact expression for the difference

$$\mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}(P, P) - \mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}(\tilde{P}, \tilde{P}).$$

8.3 **L**-*Identification for Sets for General Distributions*

The basic problem if we turn to general distributions is that the connection between a probability distribution P on $\mathcal{U}$ and a distribution $\tilde{P}$ on $\tilde{\mathcal{U}} = \binom{\mathcal{U}}{L}$ is not as straight forward as it is if we consider the discrete memoryless source $(\mathcal{U}^L, P^L)$, where the probability of a vector is the product of the probabilities of its components. In order to establish such a connection we provide

Definition 249 Let P be a probability distribution on $\mathcal{U}$. Then we define its *correlated distribution* $P^{(L)}$ on $\tilde{\mathcal{U}}$ by setting

$$P_S^{(L)} = \sum_{\pi \in \Pi_L} \prod_{l=1}^{L} \frac{p_{s_{\pi(l)}}}{1 - \sum_{m=1}^{l-1} p_{s_{\pi(m)}}}$$

for all $S = \{s_1, \ldots, s_L\} \in \tilde{\mathcal{U}}$ and where Π_L is the set of all permutations on $[L]$.

This probability equals the probability of a set S which is filled step by step with elements from $\mathcal{U}$ according to P. The first element, say $u_1 \in \mathcal{U}$, is chosen with probability p_{u_1}. Now we normalize the probabilities of the remaining elements by dividing with $1 - p_{u_1}$ and chose the next element, say u_2, with probability $p_{u_2}/(1 - p_{u_1})$ and so on until S contains L elements. The fact that different choosing sequences result in the same set S is taken into account by the sum over all permutations of [L].

Problem 5

Establish an identification entropy for L-identification for sets which provides a lower bound for $\tilde{\mathcal{L}}^{L,q}(P^{(L)}, P)$?

We have seen that a crucial part in the discovery of the q-ary identification entropy of degree L and its role as a lower bound for L-identification is the Decomposition Lemma 238. We have

Problem 5.1

Establish a decomposition formula for $\tilde{\mathcal{L}}^{L,q}(P^{(L)}, P)$ which is suitable to finding a solution for problem 5?

Appendix

Table 2 The L-identification algorithm

```
LID{
  S_1 := [L];
  for i from 1 to ||c_v|| − 1 do {
    if (∀ l ∈ S_i : c_{u_l,i} ≠ c_{v,i}) then {
      return ("FALSE",i,∅);
    }
    else {
      set S_{i+1} := {l ∈ S_i : c_{u_l,i} = c_{v,i}};
    }
  }

  if (∀ l ∈ S_{||c_v||} : c_{u_l,||c_v||} ≠ c_{v,||c_v||}) then {
    return ("FALSE",||c_v||,∅);
  }
  else {
    set S := {l ∈ S_{||c_v||} : c_{u_l,||c_v||} = c_{v,||c_v||}};
    return ("TRUE",||c_v||,S);
  }
}
```

Table 3 The L-identification algorithm for sets

```
LIDforSets {
  S_1 := S
  for i from 1 to ||c_v|| do {
    if (∀ u ∈ S_i : c_{u,i} ≠ c_{v,i}) then {
      return ("FALSE", i)
    }
    else {
      set S_{i+1} := {u ∈ S_i : c_{u,i} = c_{v,i}}
      if |S_{i+1}| = |N̄(T_{c_v^i})| then {
        return ("TRUE", i)
      }
    }
  }
}
```

References

1. J. Aczel, Z. Daroczy, On measures of information and their characterizations. Math. Sci. Eng. **115**, ii–xii, 1–234 (1975)
2. R. Ahlswede, *Identification Entropy, General Theory of Information Transfer and Combinatorics*. Lecture Notes in Computer Science, vol. 4123 (Springer, Berlin, 2006), pp. 595–613
3. R. Ahlswede, General theory of information transfer: updated. Discret. Appl. Math. **156**(9), 1348–1388 (2008)
4. R. Ahlswede, N. Cai, An interpretation of identification entropy. IEEE Trans. Inf. Theory **52**(9), 4198–4207 (2006)
5. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**(1), 15–29 (1989)
6. R. Ahlswede, G. Dueck, Identification in the presence of feedback – a discovery of new capacity formulas. IEEE Trans. Inf. Theory **35**, 30–39 (1989)
7. R. Ahlswede, B. Balkenhol, C. Kleinewächter, Identification for sources, in *General Theory of Information Transfer and Combinatorics*, ed. by R. Ahlswede et al. Lecture Notes in Computer Science, vol. 4123 (Springer, Berlin, 2006)
8. T.M. Cover, J.A. Thomas, *Elements of Information Theory* (Wiley-Interscience, Hoboken, 1991)
9. I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems* (Academic Press, Cambridge, 1981)
10. R.M. Fano, *Transmission of Information* (MIT Press, Cambridge; Wiley, Hoboken, 1961)
11. T.S. Han, S. Verdú, New results in the theory and application of identification via channels. IEEE Trans. Inform. Theory **IT-38**, 14–25 (1992)
12. D.A. Huffman, A method for the construction of minimum redundancy codes. Proc. IRE **40**, 1098–1101 (1952)
13. C. Kleinewächter, On identification, in *General Theory of Information Transfer and Combinatorics*. Lecture Notes of Computer Science, vol. 4123 (2006)
14. J.G. Rosenstein, *Linear Orderings*. Pure and Applied Mathematics, vol. 98 (Academic Press, Cambridge, 1982)

15. C.E. Shannon, A mathematical theory of communication. Bell Syst. Techn. J. **27**, 379–423, 623–656 (1948)
16. V. Strehl, Private communication, Computer Science Department (Informatik 8), University Erlangen-Nürnberg (2006)
17. R. Veldhuis, M. Breeuwer, *An Introduction to Source Coding* (Prentice Hall, Upper Saddle River, 1993)

Testing of Hypotheses and Identification

Marat Burnashev

Identification became a quite popular topic in Information Theory. It is a very good example of interaction between Mathematical Statistics (Testing of Hypotheses) and Information Theory. There are still many interesting open problems in it. I am going to devote a good part of these lectures to that topic. In order to come smoothly to identification problems we will need first some knowledge on testing of hypotheses and related notions.

1 Preliminaries: Testing of Hypotheses and L_1-Distance

Since later in identification we will deal only with finite alphabets all necessary preliminaries will be given here also only for finite alphabets.

L_1-Distance
Let $Y = (y_1, y_2, \ldots, y_K)$ be some finite set (alphabet). A *probability measure* P on Y is defined by any collection of *nonnegative values* $\{P(y_1), P(y_2), \ldots, P(y_K)\}$ satisfying the condition:

$$\sum_{i=1}^{K} P(y_i) = 1.$$

This lecture was given by Marat Burnashev in Bielefeld in 2001. Rudolf Ahlswede used his notes for his own lecture.

The original version of this chapter was revised. A correction to this chapter can be found at https://doi.org/10.1007/978-3-030-65072-8_28

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_23

The L_1-*distance* between a pair of probability measures P and Q on Y is defined as

$$\|P - Q\| = \sum_{i=1}^{K} |P(y_i) - Q(y_i)|$$

(sometimes it is called a *variational distance* and sometimes the factor 1/2 in front of the sum is used).

The *support of a measure P* is defined as

$$\text{supp } P = \{y \in Y : P(y) > 0\}$$

(i.e. it consists of all points y of the set Y where the measure P is strongly positive).

Example If $P = Q$ (i.e., $P(y) = Q(y)$ for any $y \in Y$) then obviously $\|P - Q\| = 0$. ▲

Example If the supports of measures P and Q do not intersect (i.e. measures P and Q are *orthogonal* to each other) then $\|P - Q\| = 2$. ▲

The L_1-distance between a probability measure P and some collection of probability measures $\{Q_\alpha, \alpha \in \mathcal{A}\}$ (where $\mathcal{A}$ is an arbitrary index set) is defined as

$$\|P - \{Q_\alpha, \alpha \in \mathcal{A}\}\| = \inf_{\alpha \in \mathcal{A}} \|P - Q_\alpha\|$$

We will need one more operation with probability measures.

Let $\{Q_\alpha, \alpha \in \mathcal{A}\}$ (where $\mathcal{A}$ is an arbitrary index set) be some collection of probability measures. *Convex hull* of the family $\{Q_\alpha, \alpha \in \mathcal{A}\}$ (denoted by conv $\{Q_\alpha, \alpha \in \mathcal{A}\}$ or simply conv $\{Q_\alpha\}$) is the set of all possible finite convex linear combinations of measures from $\{Q_\alpha\}$. In other words, conv $\{Q_\alpha\}$ consists of all measures F that can be represented in the form

$$F = \sum_{i \in \mathcal{A}} c_i Q_i$$

with

$$c_i \geq 0 \quad \text{for all } i\,, \quad \text{and} \quad \sum_{i \in \mathcal{A}} c_i = 1\,,$$

and only finite number of c_i's are not equal to zero.

We denote by $\mathcal{P}(Y)$ the set of *all* probability measures on Y. The set $\mathcal{P}(Y)$ is a *complete* set with respect to L_1-distance (metric). In other words, if $\{Q_n,\ n = 1, 2, \dots, \}$ is a *fundamental* sequence (i.e., $\|Q_k - Q_m\| \leq \epsilon$ if $k, m \geq n(\epsilon)$ for any $\epsilon > 0$) then it has a *limit* $Q_0 \in \mathcal{P}(Y)$ (i.e., it is also a *probability* measure).

Example Let on the set $Y = (y_1, y_2, \ldots, y_K)$ be given the collection of probability measures $\{Q_i\ ,\ i = 1, \ldots, K\}$ such that $Q_i(y_i) = 1$ for any $i = 1, \ldots, K$ (it is clear that then $Q_i(y_j) = 0$ for any $i \neq j$). Then $\mathrm{conv}\{Q_i\} = \mathcal{P}(Y)$. ▲

Testing of Simple Hypotheses

Let on a finite space Y there are given two probability measures: P and Q. We get *one* observation $y \in Y$ according to some distribution F. We know that an unknown distribution F satisfies one of the following two possibilities (hypotheses):

$$H_0 :\ F = P\ ,$$

or

$$H_1 :\ F = Q$$

In other words, due to hypothesis H_0 the distribution F coincides with P and due to hypothesis H_1 the distribution F is equal to Q.

Remark Any hypothesis H is called *simple* if the observation distribution F is completely determined provided that H is a true hypothesis (in contrast to *composite* hypothesis when provided H is true we know only that F belongs to some *class* of distributions). Therefore we consider here the testing of *two simple hypotheses* problem.

Now we want to decide in some optimal way which of those hypotheses has really taken place (i.e., to *test* those hypotheses). Any *non-randomized* way of decision making is determined by some set $\mathcal{D} \subseteq Y$ such that

$$y \in \mathcal{D} \Rightarrow H_0\ ,$$

$$y \in \mathcal{D}^c = Y \setminus \mathcal{D} \Rightarrow H_1$$

(so if our observation belongs to $\mathcal{D}$ we make a decision in favor of hypothesis H_0 and if it belongs to $\mathcal{D}^c$ we make a decision in favor of hypothesis H_1).

Since we test here *two* hypotheses there are possible *two* kinds of errors:

the *1-st kind error* when we make a decision in favor of hypothesis H_1 while hypothesis H_0 is true, with probability

$$\alpha(\mathcal{D}) = P(\mathcal{D}^c)\ ,$$

and the *2-nd kind error* when we make a decision in favor of hypothesis H_0 while hypothesis H_1 is true, with probability

$$\beta(\mathcal{D}) = Q(\mathcal{D})$$

Of course, we would like to minimize both error probabilities, but usually it is impossible (except for orthogonal measures P and Q) and therefore we should find some compromise between them. For that reason, in order to find the *optimal way* of testing those hypotheses we need some *quality criteria* (i.e. some characteristics allowing us to compare different ways of decision making).

There are widely known several such quality criteria, but we will be interested in one particular quality criterion: we want to have the test that minimizes the *sum of both error probabilities*. Probably, this criterion is not very popular in classic mathematical statistics, but it fits the best our needs in Identification. In particular, we will be interested in the value:

$$\delta(P, Q) = \min_{\mathcal{D}}\{\alpha(\mathcal{D}) + \beta(\mathcal{D})\}$$

Next result gives a simple answer to that question.

Lemma 250 *When testing two simple hypotheses P and Q for minimal possible sum $\alpha + \beta$ the following equality holds true:*

$$\delta(P, Q) = \min_{\mathcal{D}}\{\alpha(\mathcal{D}) + \beta(\mathcal{D})\} = 1 - \frac{1}{2}\|P - Q\| \ , \tag{1}$$

where $\|P - Q\|$ is the L_1-distance between measures P and Q.

Proof For any set $\mathcal{D} \subseteq Y$ we have

$$\alpha(\mathcal{D}) + \beta(\mathcal{D}) = P(Y \setminus \mathcal{D}) + Q(\mathcal{D}) = 1 - [P(\mathcal{D}) - Q(\mathcal{D})] \geq$$

$$\geq 1 - \sup_{\mathcal{D}} \left\{ \sum_{y \in \mathcal{D}} (P(y) - Q(y)) \right\} = 1 - \sum_{\{y: P(y) > Q(y)\}} (P(y) - Q(y)) =$$

$$= 1 - \frac{1}{2}\|P - Q\| \ ,$$

where we used easy-to-check equalities

$$\|P - Q\| = \sum_{\{y: P(y) > Q(y)\}} (P(y) - Q(y)) + \sum_{\{y: Q(y) > P(y)\}} (Q(y) - P(y))$$

and

$$\sum_{\{y: P(y) > Q(y)\}} (P(y) - Q(y)) = \sum_{\{y: Q(y) > P(y)\}} (Q(y) - P(y)) \ .$$

>From that proof it is easy to see also that the optimal set $\mathcal{D}_{\text{opt}}$ to make a decision in favor of hypothesis H_0 has the form

$$\mathcal{D}_{\text{opt}} = \{y : P(y) > Q(y)\}$$

Moreover, we can add to this set any points y with $P(y) = Q(y)$. □

Testing of Simple Hypothesis Against a Composite One

Now we come to the main part of our introduction. Let be, on a finite space Y, some probability measure P and some *collection* Ω of probability measures Q. The collection Ω may be finite or infinite, but contains more then one element. We assume that an unknown distribution F satisfies one of the following two hypotheses:

$$H_0 : \ F = P \ ,$$

or

$$H_1 : \ F \in \Omega = \{Q\} \ .$$

In other words, due to hypothesis H_0 the distribution F coincides with P and due to hypothesis H_1 the distribution F belongs to the class Ω (i.e., coincides with some distribution Q from Ω).

When dealing with discrete sets (like our finite set Y) and composite hypothesis we should enlarge our ways of decision making (i.e., allowing to make a *randomized decisions*).

When testing two hypotheses on the set Y any *randomized decision (test)* is determined by choosing some *decision function* $0 \le \varphi(y) \le 1 \ , \ y \in Y$. Each value $\varphi(y)$ of that function is simply the probability to make a decision in favor of hypothesis H_0 when y is observed.

For non-randomized tests the function $\varphi(y)$ takes only values 0 and 1. Clear that in our case of testing two simple hypotheses we had $\varphi(y) = 1 \, , \ y \in \mathcal{D}$, and $\varphi(y) = 0 \, , \ y \in \mathcal{D}^c$.

Coming back to our problem of testing a simple hypothesis H_0 against a composite one H_1, suppose that we chose some decision function $\varphi(y) \, , \ y \in Y$. Then we can also make two kinds of errors: *1-st kind* and *2-nd kind*. We define their probabilities as:

probability of the *1-st kind error* when we make a decision in favor of hypothesis H_1 while hypothesis H_0 is true

$$\alpha(\varphi) = E_P \, (1 - \varphi(Y)) = \sum_{y \in Y} (1 - \varphi(y)) \, P(y) \ ;$$

and probability of the *2-nd kind error* when we make a decision in favor of hypothesis H_0 while hypothesis H_1 is true

$$\beta(\varphi) = \sup_{Q \in \Omega} E_Q \varphi(Y) = \sup_{Q \in \Omega} \sum_{y \in Y} \varphi(y) Q(y)$$

As can be seen in defining the 2-nd kind of error probability we consider the *worst* possible case in error probability.

Now we want to have the test (i.e., decision function $\varphi(y)$) that minimizes the sum of both error probabilities (it is some kind of the *minimax* problem of hypotheses testing). In particular, we will be interested in the value:

$$\delta(P, \Omega) = \min_{\varphi} \delta(P, \Omega, \varphi) = \min_{\varphi} \{\alpha(\varphi) + \beta(\varphi)\}$$

Next result gives the answer to that question.

Theorem 251 *When testing a simple hypothesis P against a composite hypothesis Ω the minimal possible sum $\alpha + \beta$ satisfies the following relation:*

$$\delta(P, \Omega) = \min_{\varphi} \{\alpha(\varphi) + \beta(\varphi)\} = 1 - \frac{1}{2} \| P - \operatorname{conv} \Omega \| , \tag{2}$$

where $\operatorname{conv} \Omega$ *denotes the convex hull of all probability measures Q from the collection Ω.*

Let us give first some statistical meaning to the relation (2). Let us take some finite convex linear combination $\sum \pi_i Q_i$ of some elements Q_i from Ω (of course, we get then some element of conv Ω). We may think that we put some *prior distribution* on the set Ω and replace the composite hypothesis Ω by the simple hypothesis $\sum \pi_i Q_i$. Then clearly $\delta(P, \Omega) \geq \delta(P, \sum \pi_i Q_i)$ and moreover due to Lemma 250 we have

$$\delta(P, \Omega) \geq \delta(P, \sum \pi_i Q_i) = 1 - \frac{1}{2} \| P - \sum \pi_i Q_i \|$$

Since the last relation holds true for any finite linear combination $\sum \pi_i Q_i$ it then follows that

$$\delta(P, \Omega) \geq 1 - \frac{1}{2} \| P - \operatorname{conv} \Omega \|$$

Therefore, in order to establish the relation (2) it remains to show the sign $\leq$ in it.

Remark When we put some *prior distribution* $\pi(Q)$, $Q \in \Omega$, and define the 2-nd kind of error probability as

$$\beta(\varphi) = E_\pi \, E_Q \varphi(Y) = \sum_{Q \in \Omega} \pi(Q) \sum_{y \in Y} \varphi(y) Q(y) \ ,$$

we get a *bayesian* statement of the testing problem. From that point of view Theorem 251 is some variation of one of the most crucial principles in mathematical statistics that can be stated here as follows: the *minimax* statement of the problem is equivalent to the *bayesian* statement with the *worst (least-favorable) prior distribution* on $\overline{\text{conv } \Omega}$, where $\overline{\text{conv } \Omega}$ denotes the closure of the set conv Ω in metric L_1 (remind that conv $\Omega \subseteq \mathcal{P}(Y)$ and the set of all probability measures $\mathcal{P}(Y)$ on finite Y is complete in metric L_1).

Unfortunately, Theorem 251 (like any other that kind of result in mathematical statistics) is an *existence* result since it shows only the existence of the least-favorable distribution and says nothing about how to find it (it is usually a difficult problem).

Sketch of the Proof of Sign $\leq$ ***in Theorem 251*** We have

$$\delta(P, \Omega) = \inf_\varphi \sup_{Q \in \Omega} \{\alpha(\varphi) + \beta(\varphi, Q) \ \leq \inf_\varphi \sup_\pi \{\alpha(\varphi) + \beta(\varphi, \pi)\} \ .$$

Assume now that we can change the order of *inf* and *sup* in the last relation (without changing the value of the resulting function). Then we can continue it as

$$\delta(P, \Omega) \leq \sup_\pi \inf_\varphi \{\alpha(\varphi) + \beta(\varphi, \pi)\} = \sup_\pi \left\{1 - \frac{1}{2} \| P - \sum \pi_i Q_i \| \right\} =$$

$$= 1 - \frac{1}{2} \| P - \text{conv } \Omega \| \ ,$$

that completes the proof of Theorem 251. □

It remains to understand when we can change that order. But it is the well-known *minimax* theorem in dynamic programming, convex analysis, etc. It is easy to check that in our case the spaces of φ and π are separable and complete, and moreover the function has necessary properties.

Historical Notes Wald's book [15] is a classic book on decision functions (including relations between minimax and bayesian approaches). Book [14] is another good source for testing of hypotheses. All essential for us results can be found also in [13]. Relations (1)–(2) appeared first in the paper [6] and both easily follow from [13, 15].

2 Measures Separated in L_1-Metrics

Let Y still be a finite alphabet. Probability measures P and Q on Y are *orthogonal* if $\|P - Q\| = 2$. We consider first some properties of families of *"almost orthogonal"* probability measures on Y.

Definition 252 A collection $(P_i\,, i = 1, \ldots, M)$ of probability measures P_i on Y is called (M, δ)-**pairwise separated** collection (family) if for any $i \neq j$ the following condition is satisfied:

$$\|P_i - P_j\| = \sum_Y |P_i(y) - P_j(y)| \geq 2(1 - \delta)\ . \tag{3}$$

Definition 253 A collection $(P_i\,, i = 1, \ldots, M)$ of probability measures P_i on Y is called (M, δ)—**completely separated** collection (family) if for any i the following condition is satisfied:

$$\|P_i - \mathrm{conv}\{P_j\,,\, j \neq i\}\| = \min_{\mathbf{c}} |P_i - \sum_{j \neq i} c_j P_j| \geq 2(1 - \delta)\ , \tag{4}$$

where $\mathrm{conv}\{\mathcal{A}\}$ means the convex hull of measures from the family $\mathcal{A}$ and minimum is taken over all probability vectors $\mathbf{c} = (c_1, \ldots, c_M)$.

It is clear that any δ-completely separated collection of measures is a δ-pairwise separated collection as well.

It will turn out later that a completely separated collection of measures is essentially equivalent to some identification-code that is of our main interest in this topic. It is natural then to investigate first the maximum possible cardinality of a completely separated family of measures.

Denote the cardinality of the set Y by N and let $M_c(N, \delta)$ and $M_p(N, \delta)$ be the maximal possible cardinalities of completely separated and pairwise separated collections on the set Y, respectively.

It is obvious that

$$M_c(N, \delta) \leq M_p(N, \delta), \quad 0 \leq \delta \leq 1\ . \tag{5}$$

Since as usual in that kind of problems we are not able to find the values $M_c(N, \delta)$ and $M_p(N, \delta)$, we will get some estimates for them through lowerbounding the value $M_c(N, \delta)$ and upperbounding the value $M_p(N, \delta)$. All bounds below are oriented to the case when N is large, δ is small and moreover $N\delta^2$ is also large.

Proposition 254 *For any $N \geq 2$ and $0 < \delta < 1$ the following bounds are valid:*

$$M_c(N, \delta) \geq \exp\{N\delta^2/(2e^2)\}\ , \tag{6}$$

$$M_p(N, \delta) \leq (2/(1 - \delta))^{N-1}\ . \tag{7}$$

Remark Upperbound (7) will be essentially improved further but in that proof we will use upperbound (7)).

Proof An analog of the inequality (6) (with a smaller constant instead of $1/(2e^2)$ was proved in [1, Statement 1] by an *"exhaustive"* method. Similar inequality in a more complicated situation was obtained in [3, Theorem 1] also by the "exhaustive" method. In order to prove the inequality (6) here for the sake of variety we use some *"random choice"* arguments.

We put as every measure $P_i\,, i = 1, \ldots, M$, an equiprobable distribution on some subset $\mathcal{D}_i \subset Y$ of cardinality ϵN (parameter $\epsilon < \delta$ will be chosen later). Assume that the cardinality of the intersection of any pair of regions $\mathcal{D}_i$ and $\mathcal{D}_j\,,\ i \neq j$, does not exceed $\epsilon\delta N$. Then we get for any i

$$\|P_i - \mathrm{conv}\{P_j\,,\ j \neq i\}\| \geq 2\left[P_i\,(\mathcal{D}_i) - \mathrm{conv}\{P_j\,,\ j \neq i\}\,(\mathcal{D}_i)\right] =$$

$$= 2 - 2\sup_{j \neq i} P_j\,(\mathcal{D}_i) \geq 2(1-\delta)\ .$$

Therefore such collection $(P_i\,, i = 1, \ldots, M)$ would be a (M, δ)—completely separated collection. Now, we choose every region $\mathcal{D}_i$ randomly from equiprobable elements of the set Y. Moreover all regions are chosen independently from each other. Then for the probability P that there will be a pair $\mathcal{D}_i$ and $\mathcal{D}_j\,,\ i \neq j$, with $|\mathcal{D}_i \cap \mathcal{D}_j| > \epsilon\delta N$ the following estimate is valid

$$P \leq M(M-1)\sum_{i > \epsilon\delta N} \binom{\epsilon N}{i}\binom{N-\epsilon N}{\epsilon N - i}\left[2\binom{N}{\epsilon N}\right]^{-1}$$

In order to simplify the last sum we use a standard way. Notice that for $i > \epsilon\delta N$ we have

$$\binom{\epsilon N}{i+1}\binom{N-\epsilon N}{\epsilon N - i - 1}\left[\binom{\epsilon N}{i}\binom{N-\epsilon N}{\epsilon N - i}\right]^{-1} \leq \frac{\epsilon(1-\delta)^2}{\delta(1-2\epsilon+\delta\epsilon)} \leq \frac{\epsilon}{\delta}\ .$$

Therefore replacing that sum by the geometrical progression we get

$$P \leq M(M-1)\delta\binom{\epsilon N}{\epsilon\delta N}\binom{N-\epsilon N}{\epsilon N - \epsilon\delta N}\left[2(\delta-\epsilon)\binom{N}{\epsilon N}\right]^{-1}\ . \tag{8}$$

We can loosen this bound with the easy verifiable inequality

$$\binom{K-a}{a-i}\Big/\binom{K}{a} \leq \left(\frac{a}{K-a}\right)^i\left(\frac{K-a}{K}\right)^{a-i}\ .$$

Indeed

$$\binom{K-a}{a-i}/\binom{K}{a} =$$

$$= \frac{(a-i+1)\dots a}{(K-a+1)\dots(K-a+i)} \frac{(K-2a+i+1)\dots(K-a)}{(K-a+i+1)\dots K} \leq$$

$$\leq \left(\frac{a}{K-a}\right)^i \left(\max_{1\leq j\leq a-i} \frac{K-2a+i+j}{K-a+i+j}\right)^{a-i} \leq \left(\frac{a}{K-a}\right)^i \left(\frac{K-a}{K}\right)^{a-i}$$

Then, (8) becomes

$$P \leq \frac{\delta M(M-1)}{2(\delta-\epsilon)} \exp\left\{N\delta\epsilon \ln \frac{e\epsilon}{\delta}\right\} . \tag{9}$$

We put now $\epsilon = [\epsilon_0 N]/N$, $\epsilon_0 = \delta e^{-2}$. Since $0 \leq (\epsilon_0 - \epsilon)N \leq 1$, we get from (9) after some simple calculations, assuming $N\delta \geq 15$

$$P \leq M(M-1)\exp\{-N\delta^2 e^{-2}\} . \tag{10}$$

Now if the right side of (10) is less than 1, then there exists a collection of M regions having the required properties, from where the lower bound (6) follows provided $N\delta \geq 15$. Since there are always N orthogonal measures on Y (and using that fact when $N\delta < 15$), we get that lower bound (6) is valid for any $N \geq 2$ and $0 \leq \delta \leq 1$.

We now prove the upper bound (7). Let $P = (x_1, \dots, x_N)$ be any probability measure on Y and let $\mathcal{P}_N$ be the set of all probability measures on Y:

$$\mathcal{P}_N = \left\{P : x_i \geq 0, \ \sum_{i=1}^N x_i = 1\right\} .$$

The set $\mathcal{P}_N$ has a "volume" v_N :

$$v_N = \int \dots \int_{A(N-1,1)} dx_1 \dots dx_{N-1} = \frac{1}{(N-1)!} , \tag{11}$$

where the following notation was used

$$A(K,a) = \left\{x \in R^K : x_i \geq 0 \ , \ i = 1, \dots, K \ ; \ \sum_{i=1}^K x_i \leq a\right\} .$$

Indeed

$$v_N = \int_0^a \int \dots \int_{A(N-2,a-x_{N-1})} dx_1 \dots dx_{N-2} dx_{N-1}$$
$$= \int_0^a v_{N-1}(a - x_{N-1})^{N-2} dx_{N-1} = v_{N-1} a^{N-1}/(N-1) = v_N a^{N-1} .$$

Therefore $v_N = v_{N-1}/(N-1)$, $v_2 = 1$, from where formula (11) follows.

We fix now any measure $P_0 = (z_1, \dots, z_N)$ and consider the set $D(P_0)$ of all measures P such that $\|P - P_0\| \le 1 - \delta$. It is possible to show (see Appendix) that $D(P_0)$ has the minimal "volume" $w_N(\delta)$ when $z_1 = 1$, $z_2 = \dots = z_N = 0$ (i.e., when P_0 is an extreme point of the set $\mathcal{P}_N$). Then, $\|P - P_0\| = 2 - 2x_1$ and with the help of (11) we get ($b = A(N-1, 1) \cap \{x : x_1 \ge (1+\delta)/2\}$)

$$w_N(\delta) = \int \dots \int_B dx_1 \dots dx_{N-1}$$
$$= v_{N-1} \int_{(1+\delta)/2}^1 (1 - x_1)^{N-2} dx_1 = v_N (2/(1-\delta))^{N-1} . \tag{12}$$

Since $M_p(N, \delta) \le v_N / w_N(\delta)$, we now get the upper bound (7) from (11)–(12). □

The following result improves the upper bound (7).

Proposition 255 *For any $0 < \delta < 1$ the following inequality holds true:*

$$M_p(N, \delta) \le N + \frac{1}{\delta^2} + \frac{1}{2\delta^2} \exp\left\{ \frac{\delta^2 N}{(1 - \sqrt{\delta})^3} \ln \frac{2e}{\delta^2} \right\} . \tag{13}$$

To prove Proposition 255 we will need the following lemma.

Lemma 256 *Let μ_i, $i = 1, \dots, M$, be the collection of δ-pairwise separated probability measures on the set Y of cardinality N.*

(a) If $\max_{y,i} \mu_i(y) \le \mu$ then

$$M \le \frac{(1-\delta)\mu N}{(1 - \delta\mu N)_+} . \tag{14}$$

(b) If $\mu_i(y) \ge \mu \ge \delta/N$ for all $y \in Y_i = \{y : \mu_i(y) > 0\}$, $i = 1, \dots, M$, then

$$M \le \frac{(1-\delta)\mu N}{2\delta} \exp\left\{ \frac{\delta}{\mu} \ln \frac{2e\mu N}{\delta(1-\delta)} \right\} . \tag{15}$$

Proof We start with inequality (14). It is clear that

$$\sum_{i=1}^{M}\sum_{j=1}^{M}||\mu_i-\mu_j|| \geq 2(1-\delta)M(M-1). \tag{16}$$

On the other hand, using the representation

$$||\mu-\nu|| = 2\left(1-\sum_{y\in Y}\min\{\mu(y),\nu(y)\}\right)$$

(which follow from simple formula $|a-b| = a+b-2\min\{a,b\}$) and inequality $\min\{\mu_i(y),\mu_j(y)\} \geq \mu_i(y)\mu_j(y)/\mu$, we get applying Cauchy–Buniakovsky inequality:

$$\begin{aligned}\sum_{i=1}^{M}\sum_{j=1}^{M}\|\mu_i-\mu_j\| &\leq 2\left[M^2-\frac{1}{\mu}\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{Y}\mu_i(y)\mu_j(y)\right]\\ &= 2\left[M^2-\frac{1}{\mu}\sum_{Y}\left(\sum_{i=1}^{M}\mu_i(y)\right)^2\right]\\ &\leq 2\left[M^2-\frac{1}{\mu N}\left(\sum_{Y}\sum_{i=1}^{M}\mu_i(y)\right)^2\right] 2M^2\left(1-\frac{1}{\mu N}\right). \end{aligned}\tag{17}$$

Now from (16)–(17) the inequality (14) follows.

We will prove now the inequality (15). Assume first that $\mu \geq \delta$. Then obviously $Y_i \bigcap Y_j = \emptyset$, $i \neq j$ and therefore $M \leq N$. In order to show that the right side of (15) is greater than N it is sufficient to show that

$$0 \leq \frac{\delta}{\mu}\ln\frac{2e\mu N}{\delta(1-\delta)} + \ln\frac{(1-\delta)\mu}{2\delta} =$$

$$= \left(1-\frac{\delta}{\mu}\right)\ln(1-\delta) + \left(1+\frac{\delta}{\mu}\right)\ln(\frac{\mu}{\delta}) + \frac{\delta}{\mu}\ln(2eN) - \ln 2 \ .$$

Using simple inequality $\ln(1/x) \geq 1-x$ we get that the last expression is greater or equal than

$$1-\ln 2-\frac{\delta^2}{\mu^2}+\frac{\delta}{\mu}\ln(2N) \geq 0\ ,\ \ N \geq 2\ ,\ \delta \leq \mu \leq 1\ ,$$

from where validity of (15) for $\delta \leq \mu$ follows.

Therefore we assume now that $\mu < \delta$. Since

$$\mu|Y_i \cap Y_j| \le \sum_Y \min\{\mu_i(y), \mu_j(y)\} \le \delta \,,$$

then $|Y_i \cap Y_j| \le [\delta/\mu] = T$, $i \ne j$; $T \ge 1$. Therefore the number of measures μ_i with $|Y_i| > T$ does not exceed $\binom{N}{T+1}$ and the number of measures μ_i with $|Y_i| \le T$ obviously does not exceed $\binom{N}{T} M_p(T, \delta)$. Therefore

$$M \le \binom{N}{T+1} + \binom{N}{T} M_p(T, \delta) \le \frac{N}{T} \binom{N}{T} M_p(T, \delta) \ . \tag{18}$$

Using inequality (7) to upperbound the value $M_p(T, \delta)$ and simple inequality $\binom{n}{k} \le (en/k)^k$ we get from (18)

$$M \le \frac{N(1-\delta)}{2} \exp\left\{ T \ln \frac{2eN}{T(1-\delta)} - \ln T \right\} \ .$$

The exponent function in the right side of last inequality is $\cap$-convex on T and moreover its derivative at $T = \delta/\mu$ is positive for $N \ge 2$. Therefore we can upperbound it replacing T by δ/μ , that gives the inequality (15). □

Proof of Proposition 255 Let $(\mu_i \,;\, i = 1, \ldots, M)$ be a set of δ-pairwise separated measures on the set Y of cardinality N. It is clear that it contains not more than N measures μ_i with $\max_y \mu_i(y) > \delta$. Therefore below we consider only measures μ_i with $\max_y \mu_i(y) \le \delta$.

Fix now some μ and q such that $0 < \mu \le \delta < q < 1$ and put $Y_i(\mu) = \{y \in Y : \mu_i(y) \ge \mu\}$. Consider first all measures μ_i such that $\mu_i\left(Y_i^c(\mu)\right) \ge 1 - q$. Denoting $M_1(\mu, q)$ the total number of such measures we introduce on their basis new probability measures ν_i with supports $Y_i^c(\mu)$:

$$\nu_i(y) = \frac{\mu_i(y)}{\mu_i(Y_i^c(\mu))} \ , \ y \in Y_i^c(\mu) \ ; \ i = 1, \ldots, M_1 \ .$$

For these measures for any i , $j = 1, \ldots, M_1$; $i \ne j$, the following relations hold true:

$$\|\nu_i - \nu_j\| \ge 2\left\{1 - \frac{\delta}{(1-q)}\right\} \ , \ \max_{i,y} \nu_i(y) < \frac{\mu}{\mu_i(Y_i^c(\mu))} \le \frac{\mu}{(1-q)} \ .$$

Therefore from Lemma 256a) we get the upperbound for the value $M_1(\mu, q)$:

$$M_1(\mu, q) \le \frac{\mu N}{\left((1-q)^2 - \delta\mu N\right)_+} \ . \tag{19}$$

For remaining $M - M_1(\mu, q) = M_2(\mu, q)$ measures μ_i , $i = 1, \ldots, M_2$, we have $\mu_i(Y_i(\mu)) \geq q$. Since all measures values do not exceed δ then in every set $Y_i(\mu)$ there exists a subset $Y_i'(\mu)$ such that $q \leq \mu_i(Y_i'(\mu)) \leq \delta + q$. Introduce the new probability measures σ_i with supports $Y_i'(\mu)$:

$$\sigma_i(y) = \frac{\mu_i(y)}{\mu_i(Y_i'(\mu))} \ , \ a \in Y_i'(\mu) \ ; \ i = 1, \ldots, M_2 \ .$$

For these measures for any i , $j = 1, \ldots, M_2$; $i \neq j$, and $y \in Y_i'(\mu)$ we have

$$\|\sigma_i - \sigma_j\| \geq 2\left(1 - \frac{\delta}{q}\right) \ , \ \sigma_i(y) \geq \frac{\mu}{\delta + q} \ .$$

Therefore from Lemma 256(b) we get the upperbound for the value $M_2(\mu, q)$:

$$M_2(\mu, q) \leq \frac{\mu N}{2\delta} \exp\left\{\frac{\delta(\delta + q)}{\mu q} \ln \frac{2e\mu N q}{\delta(q - \delta)}\right\} \ , \tag{20}$$

provided

$$\mu/(\delta + q) \geq \delta/Nq \ . \tag{21}$$

It follows from (19)–(20) and the remark made at the Proposition's proof beginning that

$$M_p(N, \delta) \leq N + \frac{\mu N}{\left((1 - q)^2 - \delta\mu N\right)_+} + \tag{22}$$

$$+ \frac{\mu N}{2\delta} \exp\left\{\frac{\delta(\delta + q)}{\mu q} \ln \frac{2e\mu N q}{\delta(q - \delta)}\right\} \ .$$

Choosing now $\mu\delta N = (1 - \sqrt{\delta})^3(1 + \sqrt{\delta})$, $q = \sqrt{\delta}$, we get from (21)–(22) (and from (14) for $\delta^2 N \leq 1 - \delta$) that

$$M_p(N, \delta) \leq N + \frac{1}{\delta^2} + \frac{1}{2\delta^2} \exp\left\{\frac{\delta^2 N}{(1 - \sqrt{\delta})^3} \ln \frac{2e}{\delta^2}\right\} \ ,$$

for $\delta \leq 1/4$. If $1/4 < \delta < 1$ then this upperbound is weaker than upperbound (7) that concludes the proof of Proposition 255. □

Statements of Propositions 254–255 are combined in the following theorem.

Theorem 257 *For any $N \geq 2$ and $0 < \delta < 1$ the following bounds are valid :*

$$\exp\{N\delta^2/(2e^2)\} \ \leq M_c(N, \delta) \ \leq$$

$$\leq M_p(N, \delta) \ \leq \ N + \frac{1}{\delta^2} + \frac{1}{2\delta^2} \exp\left\{\frac{\delta^2 N}{(1 - \sqrt{\delta})^3} \ln \frac{2e}{\delta^2}\right\} \ . \tag{23}$$

Remark Lower and upper bounds (23) are essentially oriented to the case when $\delta \to 0$ and $N\delta^2 \to \infty$. In that case, as can be seen from (23), the upperbound contains an additional factor of order $\ln(1/\delta)$. From that factor point of view it is not known even which of bounds (23) (lower or upper?) can be improved. And the reason for that is quite serious. In order to understand it, assume that each probability measure μ_i is the equiprobable distribution on set Y_i and moreover, $|Y_i| = \delta N$, $i = 1, \ldots, M$. Then $\|\mu_i - \mu_j\| = 2(1 - |Y_i \cap Y_j|/(\delta N) \geq 2(1 - \delta)$ if $|Y_i \cap Y_j| \leq \delta^2 N$ for any $i \neq j$. Now what is the largest possible number M of such measures?

This question is equivalent to the following question from coding theorem. We consider a *constant weight* $w = \delta N$ binary code of length N with minimal distance $\geq (1 - \delta)w$. What is the maximal cardinality $M(N, \delta)$ of such code? The best known (in coding theory) upperbound for the number $M(N, \delta)$ (it is the so called *Johnson's bound* or some its small improvements) contains already that factor $\ln(1/\delta)$! (see also [3]).

Historical Notes Notions of pairwise and completely separated families were introduced in [8]. They reflect some geometrical approach to identification problem considered in [1, 2]. Proposition 254 was proved in [8]. Proposition 255 was proved in [4].

3 Identification Codes or "How Large is the Set of all Output Measures for Noisy Channel?"

Identification Codes

We remind the reader of the ID-code definition (see Definition 11)

Remark As can be seen it is allowed here for regions $\mathcal{D}_i$ to intersect between each other (for $\delta > 0$). We can generalize also ID-codes allowing to use randomized decisions (such possibility will be used once below). We didn't include such randomization in the definition of ID-codes in order to avoid some over-complication (nevertheless, all upperbounds for the maximal cardinalities below remain valid for such generalized ID-codes as well).

The notion of ID-codes was first introduced in [1], where it was shown that the maximal cardinality $M(n, \delta, W)$ of ID-code of length n satisfies the following lowerbound.

Proposition 258

$$\lim_{n\to\infty} \frac{\ln \ln M(n, \delta, W)}{n} \geq C(W) , \qquad 0 < \delta \leq 1 , \tag{24}$$

where $C(W)$ is the capacity of the channel W in natural units.

The converse of the inequality (24)

$$\lim_{n\to\infty} \frac{\ln\ln M(n,\delta,W)}{n} \le C(W), \quad 0 \le \delta \le \delta_0 \ , \tag{25}$$

where δ_0 is some unspecified positive constant, was obtained in [11]. It should be mentioned that earlier in [1] some weaker form of the inequality (25) was proved for $M(n,\delta_n,W)$, where $\delta_n \le e^{-\epsilon n}$ and ϵ is any positive constant.

It follows from (24)–(25) that

$$\lim_{n\to\infty} \frac{\ln\ln M(n,\delta,W)}{n} = C(W), \quad 0 < \delta \le \delta_0 \ . \tag{26}$$

Such a double exponential growth rate of the cardinality of ID-code (in contrast to an ordinary exponential one for usual codes) created a certain theoretical and practical interest to these codes [1–4, 8, 11, 12]. ID-codes are also close to some cryptography problems [3, 4].

Our more general geometrical approach to ID-codes was introduced in [8]. That approach is based on a certain equivalence between ID-codes and some families of "almost orthogonal" measures. Such an approach (and its natural connection with testing of composite hypotheses in mathematical statistics) not only enlarges the research analytical apparatus but also enables us to strengthen some results from [1, 11]. In particular, we will show that the inequality (25) remains valid for any $0 \le \delta < 1/2$.

As a result we will be able to prove the following

Theorem 259

$$\lim_{n\to\infty} \frac{\ln\ln M(n,\delta,W)}{n} = \begin{cases} 0 & , \quad \delta = 0 \quad ; \\ mathcalC(W) & , \ 0 < \delta < 1/2 \ ; \\ \infty & , \ 1/2 \le \delta \le 1 \ . \end{cases} \tag{27}$$

Before going to the most interesting case of $0 < \delta < 1/2$ let us consider the remaining cases.

If $\delta = 0$ then obviously there are not more than $|A|^n$ orthogonal measures on the output B^n of the channel and therefore $M(n,\delta=0) \le |A|^n$ from where the first line of (27) follow.

If $\delta \ge 1/2$ then we can use any number M of input measures $\{P_i\,,\ i = 1,\dots,M\}$ (e.g., all of them may even coincide) and make the following stupid decision at the output of channel: independently of the output we make a decision in favor of measure P_i with probability $1/2$ and against it with the same probability $1/2$. Such a strange ID-code with randomized decision rules will provide all error probabilities equal to $1/2$ from where the last line of (27) follows (this is the example from [1, p. 16].

Let us consider now the most interesting case $0 < \delta < 1/2$.

How Large is the Set of all Output Measures for Noisy Channel?
For a noisy channel we are able to choose only any *input* distribution P on A^n and so we are able to get *output* distributions only of the form $Q = W^{(n)}P$. Therefore the set $\mathcal{P}_{\text{out}}(B^n)$ of all possible output distributions on B^n is $\mathcal{P}_{\text{out}}(B^n) = W^{(n)}\mathcal{P}(A^n)$.

In order to stay in our geometrical framework we need to introduce a natural definition of pairwise and completely separated families of probability measures for noisy channel W.

Definition 260 (Definition 252') A collection $(P_i, i = 1, \ldots, M)$ of probability measures P_i on A^n (or collection $(Q_i, i = 1, \ldots, M)$ of probability measures $Q_i = W^{(n)}P_i$ on B^n) is called (M, n, δ, W)—**pairwise separated** collection if for any $i \neq j$ the following condition is satisfied:

$$\left\| W^{(n)}P_i - W^{(n)}P_j \right\| = \|Q_i - Q_j\| \geq 2(1-\delta) . \tag{28}$$

Definition 261 (Definition 253') A collection $(P_i, i = 1, \ldots, M)$ of probability measures P_i on A^n (or collection $(Q_i, i = 1, \ldots, M)$ of probability measures $Q_i = W^{(n)}P_i$ on B^n) is called (M, n, δ, W)—**completely separated** collection if for any i the following condition is satisfied:

$$\left\| W^{(n)}P_i - \text{conv}\left\{ W^{(n)}P_j ,\ j \neq i \right\} \right\| =$$

$$= \left\| Q_i - \text{conv}\left\{ Q_j ,\ j \neq i \right\} \right\| \geq 2(1-\delta) . \tag{29}$$

Proposition 262 *For any (M, n, δ, W)—completely separated collection $\{P_i\}$ (or $\{Q_i\}$) it is possible to define regions $\{\mathcal{D}_i \subseteq B^n\}$ such that $\{P_i, \mathcal{D}_i\}$ will be (M, n, δ, W)—ID-code. Conversely, any (M, n, δ, W)—ID-code is $(M, n, 2\delta, W)$—completely separated collection.*

Proof Let $\{P_i\}$ (and therefore $\{Q_i\}$ as well) be (M, n, δ, W)—completely separated collection. Then as the region $\mathcal{D}_i$ we choose the set D giving the minimum to the sum of error probabilities when testing Q_i against all remaining measures $\{Q_j ,\ j \neq i\}$. Due to (2) that sum will not exceed δ and therefore each error probability also will not exceed δ. It means that such collection $\{Q_i , \mathcal{D}_i\}$ will be (M, n, δ, W)—ID-code. Converse part of proposition is obvious. □

Remark The factor 2 that comes out when from (M, n, δ, W)—ID-code we get $(M, n, 2\delta, W)$—completely separated collection is the reason why the value $\delta = 1/2$ becomes the critical value for the maximal cardinality $M(n, \delta, W)$ of ID-code (cf. Theorem 259).

We denote also by $M_p(n, \delta, W)$ and $M_c(n, \delta, W)$ maximal possible cardinalities of pairwise and completely δ-separated families of probability measures for

noisy channel W, respectively. Then from Proposition 262 similar to (5) we have for $0 < \delta < 1/2$:

$$M_c(n, \delta, W) \le M(n, \delta, W) \le M_c(n, 2\delta, W) \le M_p(n, 2\delta, W) \quad . \tag{30}$$

Due to (30) instead of investigation the maximal cardinality $M(n, \delta, W)$ of ID-codes (and so dealing with error probabilities and regions $\mathcal{D}_i$) it is sufficient to investigate the maximal cardinality $M_c(n, 2\delta, W)$ of completely separated family of measures on the output set $\mathcal{P}_{\text{out}}(B^n) = W^{(n)}\mathcal{P}(A^n)$.

We start that investigation first with getting the lower bound (24). Its validity follows from (30) and the following

Proposition 263 *For any $\delta > 0$ and $R < C$ there exists $n_0(R, \delta)$ such that for any $n > n_0(R, \delta)$ the following inequality holds true :*

$$M_c(n, \delta, W) \ge \exp\left\{\frac{\delta^2 e^{nR}}{20}\right\} \ ; \ R < C \ , \ \delta > 0 \ , \ n > n_0(R, \delta) \ .$$

Proof Due to the direct part of the well-known *coding theorem* [9, 10] for any $R < C$, $\lambda > 0$, there exists some $n_0(R, \lambda)$ such that for any $n > n_0(R, \lambda)$ there exist $M > e^{nR}$ input blocks $i \in A^n$ generating measures $\{Q_i\}$ from $\mathcal{P}_{\text{out}}(B^n) = W^{(n)}\mathcal{P}(A^n)$ and regions $\mathcal{D}_i \subseteq B^n$, $i = 1, \dots, M$, such that for any $i, j = 1, \dots, M$; $i \ne j$, the following conditions are satisfied

$$\mathcal{D}_i \cap \mathcal{D}_j = \emptyset \ , \quad Q_i(\mathcal{D}_i) \ge 1 - \lambda \ .$$

Now we are almost in the situation of Sect. 2 where we constructed δ-completely separated family of measures on the alphabet Y (i.e. for the noiseless channel). Instead of the alphabet Y we have now the family of measures $\{Q_i\}$ and their "almost" supports $\{\mathcal{D}_i\}$. We can make that analogy complete if we replace (for a while) every measure Q_i by another probability measure Q_i' in the following way:

$$Q_i'(y) = \frac{Q_i(y)}{Q_i(\mathcal{D}_i)} \quad \text{for} \quad y \in \mathcal{D}_i \ ; \quad Q_i'(y) = 0 \quad \text{for} \quad y \notin \mathcal{D}_i \ ,$$

(another words, we bound the measure Q_i to the set $\mathcal{D}_i$ and normalize it to have the probability measure). Now we are exactly in the situation of Sect. 2 with the collection of sets $\{\mathcal{D}_i\}$ (or the collection of measures $\{Q_i'\}$) instead of the alphabet Y. It is easy to understand now that any δ-completely separated collection build on the "alphabet" $\{Q_i'\}$ generates a similar $(1 - (1 - \delta)(1 - \lambda) = \delta + \lambda - \delta\lambda)$—completely separated collection build on the "alphabet" $\{Q_i\}$. Therefore denoting $\epsilon = \delta + \lambda - \delta\lambda$, we get from Proposition 254

$$M_c(n, \epsilon, W) \ge \exp\left\{\frac{(\epsilon - \lambda)^2 e^{nR}}{2(1 - \lambda)^2 e^2}\right\} \ ; \ R < C \ , \ \epsilon > 0 \ , \ n > n_0(R, \lambda) \ .$$

Putting finally $\lambda = \epsilon/10$, we get the assertion of Proposition 263. □

Remark It is not difficult to upper bound the function $n_0(R, \lambda)$. For example, from [10, Problem 5.23] follows that

$$n_0(R, \lambda) \leq \frac{(1 + 4 \ln^2 B)}{(C - R)_+^2} \ln \frac{4}{\lambda} \quad .$$

We switch now to the proof of converse of (24) for any $0 < \delta < 1/2$.

It is a good example of a problem that, *if formulated in a right form*, will be reduced to an essentially technical (in mathematical sense) task. In order to ask ourselves such a *right question*, let $f_i(y)$ be the generated probability measure on the channel output $Y = B^n$ when input block $i \in A^n$ is used. Assume that we put some prior distribution $\{\pi_i\}$ on the input set A^n and therefore get the output distribution $Q_\pi(y) = \sum \pi_i f_i(y)$ on B^n. Generally, dimension of the input distribution $\{\pi_i\}$ is equal to A^n. But do we really need to use *all blocks* from A^n in order to approximate (in L_1-metrics) the generated distribution $Q_\pi(y) = \sum \pi_i f_i(y)$ on B^n ? And the answer is *no*. It turns out that for that purpose it is sufficient to use only the order of e^{nC} input blocks, where $C = C(W)$ is the channel capacity.

Essentially, this kind of result was first obtained in [11, Lemma 1] where authors showed that any output distribution can be rather accurately approximated by using some input distribution $\{\pi_i\}$ on A^n whose masses take values on a lattice with a span of the order of e^{-nC}.

We shall give here another proof of an equivalent (but a more geometrical) result from [8]. It can be formulated in the following way.

Proposition 264 *For any $\delta > 0$ there exists $n_0(\delta)$ such that for any output measure Q on B^n, $n \geq n_0(\delta)$, there exist $B^{n\delta} e^{nC}$ input blocks $i \in A^n$ such that for their generated measures $\{Q_i\}$ the following inequality holds true :*

$$\left\| Q - \mathrm{conv} \left\{ Q_i, i = 1, \ldots, B^{n\delta} e^{nC} \right\} \right\| \leq \delta \ . \tag{31}$$

In other words, any output distribution Q can be arbitrary closely approximated by using only the order of e^{nC} of input blocks. We should emphasize here that blocks used for that purpose, generally speaking, *depend on measure* Q, but the number $n_0(\delta)$ *does not depend on* Q,

Putting aside for a while the proof of that Proposition, we shall show now how to get necessary corollaries from it.

We notice first that any channel W (or $W^{(n)}$) acts like a "compressing" operator in the following sense.

Lemma 265 *For any channel W, any pair of input distributions P_1, P_2 and corresponding pair of output distributions Q_1, Q_2 the following inequality holds true:*

$$\|Q_1 - Q_2\| \leq \|P_1 - P_2\| \ . \tag{32}$$

Proof Indeed

$$\|Q_1 - Q_2\| = \sum_Y \left| \sum_X W(y|x)\,(p_1(x) - p_2(x)) \right|$$

$$\leq \sum_X |p_1(x) - p_2(x)| \sum_Y W(y|x)$$

$$= \sum_X |p_1(x) - p_2(x)| = \|P_1 - P_2\|. \qquad \square$$

Now we can prove the converse result. It will follow from (30) and the following

Proposition 266 *For any* $0 < \delta < 1/2$ *the following inequality holds true:*

$$\lim_{n\to\infty} \frac{\ln \ln M_p(n, 2\delta, W)}{n} \leq C(W)\ , \quad 0 < \delta < 1/2\ . \tag{33}$$

Proof Let $\{Q_i,\ i = 1, \ldots, M\}$ be some 2δ—pairwise separated collection of output measures. We fix some small $\epsilon > 0$ such that $2(\delta + \epsilon) < 1$. By virtue of Proposition 264 any measure Q_i can be ϵ-approximated by another output measure Q_i' where the measure Q_i' is generated by some $B^{n\epsilon}e^{nC}$ input blocks from A^n. Clear that the collection $\{Q_i',\ i = 1, \ldots, M\}$ is $2(\delta + \epsilon)$—pairwise separated. Since the channel $W^{(n)}$ is a "compressing" operator (Lemma 265), the maximal number of $2(\delta + \epsilon)$—pairwise separated measures, generated by every collection of $N = B^{n\epsilon}e^{nC}$ input blocks, is upperbounded by formula (7). The total number of collections with the cardinality N on the alphabet A^n does not exceed A^{nN}. Therefore, for the maximal possible cardinality $M_p(n, 2\delta, W)$ of pairwise separated collection we get ($N = B^{n\epsilon}e^{nC}$, $2(\delta + \epsilon) < 1$)

$$M_p(n, 2\delta, W) < (2/(1 - 2\delta - 2\epsilon))^N A^{nN} \leq$$

$$\leq \exp\left\{ \left(\ln \frac{2}{(1 - 2\delta - 2\epsilon)} + n \ln A \right) e^{n(C + \epsilon \ln B)} \right\} \quad , \quad n > n_0(\epsilon)\ ,$$

from where relation (33) and (25) follow. $\square$

Remark It was sufficient for us to use here the upperbound (7) instead of a much tighter upperbound (13). The reason is that *log log* kind of asymptotic in *Identification* (e.g., (33)) is very insensitive to their difference. Absolutely different situation is in a closely related *Authentication* [4].

Now for the purpose of completeness we shall show also how [11, Lemma 1] follows from Proposition 264.

Corollary 267 *Any output distribution* Q *on* B^n *can be arbitrary closely approximated by using some input distribution* $\{p_i, i = 1, \ldots, N\}, N \sim e^{nC}$, *on* A^n, *taking values only of the form* $p_i = j/N$, j*—integer.*

Proof For any $\epsilon > 0$ and any distribution $\{p_i, i = 1, \ldots, N\}$ there exists some distribution $\{\hat{p}_i\}$, taking values only of the form $\hat{p}_i = j\epsilon/N$ with $|p_i - \hat{p}_i| \leq \epsilon/N$, and moreover,

$$\|p - \hat{p}\| = \sum_{i=1}^{N} |p_i - \hat{p}_i| \leq \epsilon \ . \tag{34}$$

Indeed, we choose $\hat{p}_1 = j\epsilon/N$ with the minimal possible $|p_1 - \hat{p}_1|$. Then $|p_1 - \hat{p}_1| \leq \epsilon/(2N)$. Now we choose $\hat{p}_2$ of the same form with the minimal possible $|(p_1 + p_2) - (\hat{p}_1 + \hat{p}_2)|$. Then $|(p_1 + p_2) - (\hat{p}_1 + \hat{p}_2)| \leq \epsilon/(2N)$ and $|p_2 - \hat{p}_2| \leq \epsilon/(2N)$. Repeating this process, we get as a result the collection $\{\hat{p}_i\}$, having desired properties. Since the channel is a compressing operator, for measures Q and $\hat{Q}$, generated by distributions $\{p_i\}$ and $\{\hat{p}_i\}$, respectively, due to (34), we have $\|Q - \hat{Q}\| \leq \epsilon$. From this result and Proposition 264 the assertion of Corollary 267 follows. □

It remained us only to establish the validity of Proposition 264. The proof below is essentially technical and is based on "random choice" of approximating input distribution.

Proof of Proposition 264 We shall need a few simple pure statistical lemmas.

Let Y be a finite alphabet and there are given K probability distributions $f_i(y)$, $i = 1, \ldots, K$, on it with prior probabilities $\{p_i\}$. Consider also the "averaged" distribution

$$p(y) = \sum_{i=1}^{K} p_i f_i(y) \ .$$

We choose now randomly and independently s distributions from $\{f_i\}$ with respective probabilities $\{p_i\}$ (with returns) and put

$$\hat{p}(y) = \frac{1}{s} \sum_{i=1}^{K} \nu_i f_i(y) \ ,$$

where ν_i = number of measures f_i among s chosen distributions. □

Lemma 268 *The following estimate on the variance of* $\hat{p}(y)$ *holds :*

$$E|p(y) - \hat{p}(y)|^2 \leq \frac{1}{s} \sum_{j=1}^{K} p_j(1 - p_j) f_j^2(y) \ , \ y \in Y \ . \tag{35}$$

Proof Let μ_l be the index of measure chosen from $\{f_i\}$ on the lth step. Then $\mu_1, \dots, \mu_s$ are i.i.d.r.v.'s, and moreover

$$Ef_{\mu_l} = p(y) \ , \ \sum_{j=1}^{K} v_j f_j(y) = \sum_{l=1}^{s} f_{\mu_l}(y) \ .$$

Therefore,

$$\begin{aligned} E|\hat{p}(y) - p(y)|^2 &= E\left|\frac{1}{s}\sum_{l=1}^{s}\left(f_{\mu_l}(y) - p(y)\right)\right|^2 \\ &= \frac{1}{s}E\left(f_{\mu_1} - p(y)\right)^2 \\ &= \frac{1}{s}\left(Ef_{\mu_1}^2(y) - p^2(y)\right) \\ &= \frac{1}{s}\left\{\sum_{j=1}^{K} p_j f_j^2(y) - \left[\sum_{j=1}^{K} p_j f_j(y)\right]^2\right\} \\ &\le \frac{1}{s}\sum_{j=1}^{K} p_j(1-p_j)f_j^2(y). \end{aligned}$$

□

Now for $\epsilon > 0$ and $i = 1, \dots, K$, we choose some sets $Y_i(\epsilon)$ such that $P_i\{Y_i(\epsilon)\} \ge 1 - \epsilon \ , \ i = 1, \dots, K$, and put

$$K(y) = \{i : y \in Y_i(\epsilon)\} \ , \quad Y(\epsilon) = \bigcup_{i=1}^{K} Y_i(\epsilon) \ .$$

Lemma 269 *For any $\epsilon > 0$ the following estimate is valid :*

$$\sum_{Y} E|\hat{p}(y) - p(y)| \le 2\epsilon + \left(\frac{|Y(\epsilon)|}{s}\max_{\substack{j=1,\dots,K \\ y\in Y_j(\epsilon)}} f_j(y)\right)^{1/2} \tag{36}$$

($|A|$ means the cardinality of the set A).

Proof We have

$$\sum_Y E|\hat{p}(y) - p(y)| \le \sum_{Y(\epsilon)} E\left|\sum_{j\in K(y)}\left(\frac{\nu_j}{s} - p_j\right) f_j(y)\right| + \sum_Y E\left|\sum_{j\notin K(y)}\left(\frac{\nu_j}{s} - p_j\right) f_j(y)\right| = \Sigma_1 + \Sigma_2 \ .$$

Using (35) and Cauchy-Buniakovsky inequality we get for Σ_1

$$\begin{aligned}
\Sigma_1 &\le \sum_{Y(\epsilon)}\left[E\left|\sum_{j\in K(y)}\left(\frac{\nu_j}{s} - p_j\right) f_j(y)\right|^2\right]^{1/2} \\
&\le \sum_{Y(\epsilon)}\left(\frac{1}{s}\sum_{j\in K(y)} p_j f_j^2(y)\right)^{1/2} \\
&\le \left(\frac{|Y(\epsilon)|}{s}\sum_{Y(\epsilon)}\sum_{j\in K(y)} p_j f_j^2(y)\right)^{1/2} \\
&\le \left(\frac{|Y(\epsilon)|}{s}\sum_{j=1}^{K} p_j \sum_{Y_j} f_j^2(y)\right)^{1/2} \\
&\le \left(\frac{|Y(\epsilon)|}{s}\max_{\substack{j=1,\dots,K \\ y\in Y_j(\epsilon)}} f_j(y)\right)^{1/2} \ .
\end{aligned}$$

For Σ_2 we have obviously

$$\begin{aligned}
\Sigma_2 &\le \sum_Y \sum_{j\notin K(y)} E\left|\frac{\nu_j}{s} - p_j\right| f_j(y) \\
&\le 2\sum_Y \sum_{j\notin K(y)} p_j f_j(y) \\
&= 2\sum_{j=1}^{K} p_j \sum_{Y\setminus Y_j(\epsilon)} f_j(y) \\
&\le 2\epsilon \ .
\end{aligned}$$

Combining these two inequalities we get (25). □

In the setting of a noisy channel measure $f_i(y)$ represents the conditional distribution of the channel output $y \in B^n$ given the "input" $i \in A^n$. As sets $Y_i(\epsilon) \subseteq B^n$ some "supports" of measures $f_i(y)$ will be chosen. In that meaning Lemma 269 is supposed to be used for the same type codeblocks i. The next result generalizes it for codeblocks of different types.

Let L classes of measures $\{f_{li}(y)\,,\ i = 1, \dots, K(l)\}$, $l = 1, \dots, L$, with prior probabilities $\{p_{li}\}$ are given and

$$p(y) = \sum_{l=1}^{L} \sum_{i=1}^{K(l)} p_{li}\, f_{li}(y)\ .$$

In order to approximate $p(y)$ we choose in every class l randomly and independently s_l measures f_{li} proportionally to the distribution p_{li}. Moreover, for every pair l, i we choose also some set $Y_{li}(\epsilon)$ such that $P_{li}\{Y_{li}\} \geq 1 - \epsilon$ and put (ν_{li} = number of measures f_{li} among s_l chosen distributions)

$$\hat{p}(y) = \sum_{l=1}^{L} \frac{p_l}{s_l} \sum_{i=1}^{K(l)} \nu_{li}\, f_{li}(y)\ ,\quad p_l = \sum_{i=1}^{K(l)} p_{li}\ ,\quad Y_l(\epsilon) = \bigcup_{i=1}^{K(l)} Y_{li}(\epsilon)\ .$$

Lemma 270 *For any $\epsilon > 0$ the following estimate is valid :*

$$\sum_{Y} E|\hat{p}(y) - p(y)| \leq 2\epsilon + \left(\sum_{l=1}^{L} \frac{p_l |Y_l(\epsilon)|}{s_l} \max_{\substack{i=1,\dots,K(l) \\ y \in Y_{li}(\epsilon)}} f_{li}(y) \right)^{1/2}\ . \tag{37}$$

Proof We have from Lemma 269:

$$\sum_{Y} E|\hat{p}(y) - p(y)| \leq 2\epsilon + \sum_{l=1}^{L} p_l \left(\frac{|Y_l(\epsilon)|}{s_l} \max_{\substack{i=1,\dots,K(l) \\ y \in Y_{li}(\epsilon)}} f_{li}(y) \right)^{1/2}$$

$$\leq 2\epsilon + \left(\sum_{l=1}^{L} \frac{p_l |Y_l(\epsilon)|}{s_l} \max_{\substack{i=1,\dots,K(l) \\ y \in Y_{li}(\epsilon)}} f_{li}(y) \right)^{1/2}\ .$$

Now, we return to the channel $W^{(n)}$ with input A^n and output B^n alphabets. Application of Lemma 270 to that case is rather standard and it is based on considering codeblocks from A^n of the same type (composition) (see [9, Ch. 1.2 and 2.1] or [1]). Necessary formulas (38) below can be also found in [9] or [1].

As usual, we partition all codeblocks from A^n on classes l consisting of codeblocks of the same composition (type). In every pair l, i the parameter i will denote the block's index inside the class l. An input block (l, i) generates the distribution f_{li} on the channel output B^n. As a set $Y_{li} \subseteq B^n$ we choose a

set of "almost 1" probability and consisting of approximately equiprobable points. More precisely, we fix some arbitrary small $\epsilon > 0$ and let P_l denotes the type of codeblock (l, i) [10] and Q_{li} denotes the output distribution generated by that codeblock. Then for any $n \geq n_0(\epsilon)$ it is possible to choose some sets $Y_{li}(\epsilon) \subseteq B^n$ with $Q_{li}\{Y_{li}(\epsilon)\} \geq 1 - \epsilon$ such that simultaneously for all l, i the following conditions will be satisfied:

$$|Y_{li}(\epsilon)| < \exp\{n(1+\epsilon)H(W|P_l)\}\,,$$

$$\max_{\substack{i=1,\ldots,K(l)\\ y\in Y_{li}(\epsilon)}} f_{li}(y) < \exp\{-n(1-\epsilon)H(W|P_l)\}\,, \tag{38}$$

$$|Y_l(\epsilon)| = \left|\bigcup_i Y_{li}(\epsilon)\right| < \exp\{n(1+\epsilon)H(W P_l)\}\ .$$

Now we put

$$s_l = \epsilon^{-2}\exp\{n\,[I(P_l, W) + \epsilon H(W P_l) + \epsilon H(W|P_l)]\}\ . \tag{39}$$

Then taking into account that $H(W P_l) - H(W|P_l) = I(P_l, W) \leq C$, $H(W P_l) + H(W|P_l) \leq 2\ln B$ and that the total number of classes $L \leq (n+1)^{|A|}$, we have from (38)–(39) for $n > n_0(\epsilon)$

$$s = \sum_{l=1}^{L} s_l \leq \epsilon^{-2} L \exp\{n[C + 2\epsilon \ln B]\} \leq B^{3\epsilon n} e^{nC}\ . \tag{40}$$

Similarly we have from (37)–(39)

$$\sum_{B^n} E|\hat{p}(y) - p(y)| \leq 3\epsilon\ . \tag{41}$$

Now from (40)–(41) Proposition 264 follows. □

On a "Basis" for the Set of all Output Measures

In connection with Proposition 264 some natural question arises: is it possible to choose some *universal collection* of $N \sim e^{nC}$ input blocks (like a basis) such that using them it is possible arbitrary closely to approximate *any* output distribution? The answer, generally speaking, is *no*. To see this consider the example of the binary symmetrical channel with crossover probability $p < 1/2$.

Let x_0 be the all-zero input block of the length n and $x_1, \ldots, x_N$ be all possible input blocks of the Hamming weight $w(x_i) \geq d$ (d will be chosen later). Let also $\{Q_i, i = 0, 1, \ldots, N\}$ be the generated measures on the channel output B^n, respectively. We shall evaluate now how large should be d such that it will

be impossible to approximate sufficiently accurately the measure Q_0 with the help of all remaining measures $\{Q_i, i = 1, \dots, N\}$. It is convenient to use here some statistical interpretation and the formula (1). Let us consider the problem of testing hypothesis Q_0 against $\{Q_i, i = 1, \dots, N\}$. Notice first that if the vector x_i, $i = 0, 1, \dots, N$, is transmitted then for the Hamming weight $w(y)$ of output vector y we have

$$E\,(w(y)\,|x_i) = w(x_i) + p\,(n - 2w(x_i))\ ,$$

$$E\Big\{\big[w(y) - E\big(w(y)|x_i\big)\big]^2 \big| x_i\Big\} = p(1-p)n\ .$$

Now we choose as the decision set $\mathcal{D}_0$ of acceptance hypothesis Q_0 the following set:

$$\mathcal{D}_0 = \Big\{y : w(y) \le pn + [2znp(1-p)]^{1/2}\Big\}\ ,$$

where the parameter $z = z(\delta) > 0$ will be chosen later. Then, due to the central limit theorem, we have

$$Q_0(Y^n \setminus \mathcal{D}_0) = \Phi\left(-\sqrt{2z}\right) + o(1)\ ,\quad n \to \infty\ , \tag{42}$$

$$Q_i(\mathcal{D}_0) = \Phi\left[-\frac{w(x_i)(1-2p)}{\sqrt{np(1-p)}} + \sqrt{2z}\right] + o(1)\ ,\quad i = 1, \dots, N\ .$$

Now, if $w^2(x_i)(1-2p)^2 \ge 8znp(1-p)$, $i = 1, \dots, N$, then due to (42) we have from formula (1)

$$\begin{aligned}
&\|Q_0 - \mathrm{conv}\{Q_i,\ i = 1, \dots, N\}\| \\
&\ge 2\left[1 - Q_0(Y^n \setminus \mathcal{D}_0) - \mathrm{conv}\{Q_i,\ i = 1, \dots, N\}(\mathcal{D}_0)\right] \\
&\ge 2\left[1 - 2\Phi\left(-\sqrt{2z}\right)\right] + o(1)\ ,\qquad n \to \infty.
\end{aligned} \tag{43}$$

We use now simple upperbound $2\Phi(-x) \le e^{-x^2/2}$, $x > 0$, and put $z = \ln(1/\delta)$. Then we get from (43)

$$\|Q_0 - \mathrm{conv}\{Q_i,\ i = 1, \dots, N\}\| \ge 2(1-\delta)\ . \tag{44}$$

Therefore, if the minimum distance d of arbitrary code $\{x_1, \dots, x_N\}$ satisfies the condition $d^2(1-2p)^2 \ge 8np(1-p)\ln(1/\delta)$, then the collection $\{Q_i, i = 1, \dots, N\}$ is (N, δ)—completely separated. It is easy to understand that the maximal cardinality N of such a collection has the order of $\exp\{n\ln 2 - a(\delta, p)n^{1/2}\ln n\}$, i.e., it differs negligibly from the cardinality 2^n of the whole space of input blocks.

Completely and Pairwise Separated Collections of Measures

As we know already, ID-codes and completely separated collections of measures are essentially equivalent. But it is rather difficult to check the complete separability of measures (e.g., condition (4)). It is much easier to check the pairwise separability of measures. Then the following question naturally arises: if we are given some collection of δ—pairwise separated measures, is it possible then to select some of its sub-collection such that it will be ϵ—completely separated (with ϵ close to δ) and will have "practically" the same cardinality?

Or, equivalently, under which conditions a collection of δ—pairwise separated measures is simultaneously a collection of ϵ—completely separated measures (with ϵ slightly worse than δ)?

>From statistics point of view similar question can be formulated in the following way. Let some set $\{Q_i, i = 0, 1, \ldots, N\}$ of probability measures be given. Assume that when we test hypothesis Q_0 against any simple hypothesis Q_i we can achieve sum of error probabilities $\delta(Q_0, Q_i) \leq \delta$. Then due to formula (1) we have $\|Q_0 - Q_i\| \geq 2(1-\delta)$ for any $i = 1, \ldots, N$, . What can we say about the minimal sum of error probabilities when we test the hypothesis Q_0 against *all* hypothesis Q_i, $i = 1, \ldots, N$, ? Or, due to formula (2), what can we say about $\|Q_0 - \mathrm{conv}\{Q_i, i = 1, \ldots, N\}\|$? Clear that this new sum $\delta(Q_0, \mathrm{conv}\{Q_i\})$ will be $\geq \delta$, but when we can expect that it will be rather close to δ?

Before giving a partial answer to this question we bring two examples, showing what kind of results can be expected here.

Example Let there be given the equiprobable distribution P_0 and measures P_i, $i = 1, \ldots, |X|$, on the alphabet X, such that $P_i\{x_i\} = 1$. Then $\|P_0 - P_i\| = 2 - 2|X|^{-1} \geq 2 - 2\delta$, if $|X| \geq 1/\delta$. On the other hand, it is clear that $\|P_0 - \mathrm{conv}\ \{P_i,\ i = 1, \ldots, |X|\}\| = 0$, i.e., this collection of measures is not completely separable for any $\delta < 1$. From ID-codes point of view for any choice of a nonempty region $\mathcal{D}_0$ we shall have $P_i(\mathcal{D}_0) = 1$ for some $i > 0$. But if we remove the measure P_0 from that collection, the remaining orthogonal measures will represent completely separated collection. ▲

It becomes clear from the next (more meaningful) example that it is hardly possible to get some easy-to check necessary and sufficient conditions in this problem. The infinite alphabet used in that example is not crucial. It is always possible to approximate it arbitrary precisely by a sufficiently large finite alphabet.

Example Let P_i, $i = 1, \ldots, N$ be Gaussian measures in Euclidean space R^n with the identity covariance matrix I_n and the mean vector $A^{1/2}\mathbf{e}_i$, where $\mathbf{e}_i$ is the n—dimensional vector with the ith coordinate 1 and all the remaining coordinates are zeros. That model corresponds to the observation of one of N orthogonal signals with energy A in white Gaussian noise. Then with the help of the formula (7) it is

simple to get

$$\|P_i - P_j\| = 2\left[1 - 2P\left\{\xi > (A/2)^{1/2}\right\}\right] = \tag{45}$$

$$= 2\left[2\Phi\left((A/2)^{1/2}\right) - 1\right],\ i \neq j\ ,$$

where ξ is the Gaussian RV with parameters $(0, 1)$ and $\Phi(x)$ is the standard distribution function of this RV.

It is hardly possible to calculate explicitly the distance between the measure P_0 and the convex combination of all remaining measures. But it is possible to upperbound it. Putting on the set $P_j,\ j \neq i$, the equiprobable prior distribution and using the simple inequality

$$\|P - Q\|^2 \leq E_P\left(\frac{dQ}{dP}\right)^2 - 1\ ,$$

we get after some simple calculations (see details in [6])

$$\|P_i - \mathrm{conv}\{P_j,\ j \neq i\}\|^2 \leq (N-1)^{-1}e^{2A}\ . \tag{46}$$

We can see from (45) that when A is sufficiently large, measures $\{P_i,\ i = 1, \ldots, N\}$ are pairwise almost orthogonal and hence are well pairwise separated. But from (46) it follows that if the number of measures N is much larger than $\exp\{2A\}$, then this collection of measures is not completely separated for small δ. Particularly, for such N it is impossible to discriminate hypotheses P_i and $\{P_j,\ j \neq i\}$ with small error probabilities. In fact, it is possible to show [7] that the true critical value here is $\exp\{A\}$. ▲

We present now a sufficient condition for a complete separability of measures which is rather easy to check in some cases.

Proposition 271 *Let there be given probability measures $P_i,\ i = 1, \ldots, N,$ on a finite alphabet X. For any subset $A \subseteq X$ we denote*

$$\epsilon_{ij}(A) = \sum_A P_i(x)P_j(x)\ .$$

Then for any i the following estimate holds true:

$$\|P_i - \mathrm{conv}\{P_j,\ j \neq i\}\| \geq 2\max_A\left[P_i(A) - \left(|A|\max_{j\neq i}\epsilon_{ij}(A)\right)^{1/2}\right] \tag{47}$$

and $\{P_i,\ i = 1, \ldots, N\}$ *is an* (N, δ)*—completely separated collection with*

$$\delta = \max_i \min_A \left[P_i(X \setminus A) + \left(|A| \max_{j \neq i} \epsilon_{ij}(A) \right)^{1/2} \right] . \tag{48}$$

Proof Let $Q = \sum c_i P_i$ be any convex combination of measures $P_i,\ i = 1, \ldots, N$, and $A \subseteq X$. Using representation $|p - q| = p + q - 2\min\{p, q, \}$ and the Cauchy-Bunyakowski inequality we carry out the following chain of calculations:

$$\begin{aligned}
\|P_1 - Q\| &= 2 \left[1 - \sum_X \min\{P_1(x), Q(x)\} \right] \\
&\geq 2 \left[P_1(A) - \sum_A \min\{P_1(x), Q(x)\} \right] \\
&\geq 2 \left[P_1(A) - \sum_A (P_1(x) Q(x))^{1/2} \right] \\
&\geq 2 \left[P_1(A) - \left(|A| \sum_A P_1(x) Q(x) \right)^{1/2} \right] \\
&\geq 2 \left[P_1(A) - \left(|A| \max_{i \geq 2} \epsilon_{1i}(A) \right)^{1/2} \right] ,
\end{aligned}$$

from where formulas (47)–(48) follow. □

Appendix

We show that the set $D(P_0)$ in the proof of the inequality (8) has the minimum volume when P_0 is an extreme point of the set $\mathcal{P}_N$.

Proof Let $P_i\ ,\ i = 1, \ldots, N$, denotes the N—dimensional vector whose ith coordinate is 1 and all the remaining coordinates are 0. We fix arbitrary $a > 0$ and consider sets

$$V_i = \{x \in \mathcal{P}_N : \|P_i - x\| \leq a\}\ ,\ i = 1, \ldots, N\ .$$

Let $c = (c_1, \ldots, c_N)$ be an arbitrary probability vector. Introduce sets

$$V(c) = \sum_{i=1}^{N} c_i V_i = \left\{ y \in \mathcal{P}_N : y = \sum_{i=1}^{N} c_i y_i \ ; y \in V_i \ , i = 1, \ldots, N \right\} \ ,$$

$$D(c) = \{ y \in \mathcal{P}_N : \|c - y\| \le a \} \ .$$

Obviously, sets V_i , $V(c)$ and $D(c)$ are convex. It is simple also to check that $V(c) \subseteq D(c)$. Designating $v(A)$ the set's A volume and taking into account that volumes $v(V_i)$, $i = 1, \ldots, N$ are equal, we get by virtue of Brunn-Minkowski theorem [5]

$$v(D(c)) \ge v(V(c)) \ge \left[\sum_{i=1}^{N} c_i v^{1/N}(V_i) \right]^N = v(V_1) \ . \qquad \square$$

References

1. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**(1), 15–29 (1989)
2. R. Ahlswede, G. Dueck, Identification in the presence of feedback—a discovery of new capacity formulas. IEEE Trans. Inf. Theory **35**, 30–39 (1989)
3. L.A. Bassalygo, M.V. Burnashev, Estimate for the maximal number of messages for a given probability of succesful deception. Probl. Inf. Trans. **30**(2), 42–48 (1994).
4. L.A. Bassalygo, M.V. Burnashev, Authentication, identification and pairwise separated measures. Probl. Inf. Trans. **32**(1), 41–47 (1996)
5. Ju.D. Burago, V.A. Zalgaller. *Geometrical Inequalities* [in Russian] (Nauka, Leningrad, 1980)
6. M.V. Burnashev, Minimax Detection of inaccurately known signal in the background of white gaussian noise. Theory Prob. Its Appl. **24**(1), 106–118 (1979)
7. M.V. Burnashev, I.A. Begmatov, On a signal detection problem leading to stable distributions. Theory Prob. Its Appl. **35**(3), 1169–1172 (1990)
8. M.V. Burnashev, S. Verdu, Measures Separated in L_1 metrics and ID–codes. Probl. Inf. Trans. **30** (3), 3–14 (1994)
9. I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems* (Academic Press, New York, 1981)
10. R.G. Gallager. *Information Theory and Reliable Communication* (Wiley, New York, 1968)
11. T.S. Han, S. Verdú, New results in the theory and application of identification via channels. IEEE Trans. Inf. Theory **IT-38**, 14–25 (1992)
12. T.S. Han, S. Verdu, Approximation theory of output statistics. IEEE Trans. Inf. Theory, **39**(3), 752–772 (1993)
13. P.J. Huber, V. Strassen, Minimax tests and the Neyman–Pearson Lemma for capacities. Ann. Stat. **1**(2), 251–263 (1973)
14. E.L. Leman. *Testing Statistical Hypotheses* (Chapman & Hall, New York, 1959)
15. A. Wald. *Statistical Decision Functions* (Wiley, New York, 1950)

On Logarithmically Asymptotically Optimal Testing of Hypotheses and Identification

We introduce a new aspect of the influence of the information-theoretical methods on the statistical theory. The procedures of the probability distributions identification for $K(\geq 1)$ random objects each having one from the known set of $M(\geq 2)$ distributions are studied. N-sequences of discrete independent RV's represent results of N observations for each of K objects. On the base of such samples decisions must be made concerning probability distributions of the objects. For $N \to \infty$ the exponential decrease of the test's error probabilities is considered. The reliability matrices of logarithmically asymptotically optimal procedures are investigated for some models and formulations of the identification problems. The optimal subsets of reliabilities which values may be given beforehand and conditions guaranteeing positiveness of all the reliabilities are investigated.

"In statistical literature such a problem is referred to as one of classification or discrimination, but identification seems to be more appropriate"

Radhakrishna Rao [27].

1 Problem Statement

Let $\mathbf{X}_k = (X_{k,n},\ n \in [N]),\ k \in [K]$, be $K(\geq 1)$ sequences of N discrete i.i.d. RV's representing possible results of N observations, respectively, for each of K randomly functioning objects.

For $k \in [K], n \in [N]$, $X_{k,n}$ assumes values $x_{k,n}$ in the finite set $\mathcal{X}$ of cardinality $|\mathcal{X}|$. Let $\mathcal{P}(\mathcal{X})$ be the space of all possible distributions on $\mathcal{X}$. There are $M(\geq 2)$ probability distributions $G_1, \ldots, G_M$ from $\mathcal{P}(\mathcal{X})$ in inspection, some of which are assigned to the vectors $\mathbf{X}_1, \ldots, \mathbf{X}_K$. This assignment is unknown and must be determined on the base of N-samples (results of N independent observations) $\mathbf{x}_k = (x_{k,1}, \ldots, x_{k,N})$, where $x_{k,n}$ is a result of the nth observation of the kth object.

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_24

When $M = K$ and all objects are different (any two objects cannot have the same distribution), there are $K!$ possible decisions. When objects are independent, there are M^K possible combinations.

Bechhofer, Kiefer, and Sobel presented investigations on sequential multiple-decision procedures in [7]. This book is concerned principally with a particular class of problems referred to as ranking problems.

Chapter "Models with Prior Knowledge at the Sender" of the book by Ahlswede and Wegener [5] is devoted to statistical identification and ranking problems.

We study models considered in [7] and [5] and variations of these models inspired by the pioneering papers by Ahlswede and Dueck [4] (see chapter "Identification via Channels" in Part I) and by Ahlswede [1], applying the concept of optimality developed in [9, 16, 22–24, 28] for the models with $K = 1$.

Consider the following family of error probabilities of a test

$$\alpha^{(N)}_{m_1,m_2,\ldots,m_K|l_1,l_2,\ldots,l_K},\ (m_1,m_2,\ldots,m_K) \neq (l_1,l_2,\ldots,l_K),\ m_k,l_k \in [M],\ k \in [K],$$

which are the probabilities of decisions $l_1, l_2, \ldots, l_K$ when actual indices of the distributions of the objects were, respectively, $m_1, m_2, \ldots, m_K$.

The probabilities to reject all K hypotheses when they are true are the following

$$\alpha^{(N)}_{m_1,m_2,\ldots,m_K|m_1,m_2,\ldots,m_K} = \sum_{(l_1,l_2,\ldots,l_K)\neq(m_1,m_2,\ldots,m_K)} \alpha^{(N)}_{m_1,m_2,\ldots,m_K|l_1,l_2,\ldots,l_K}.$$

We study exponential decrease of the error probabilities when $N \to \infty$ and define (using logarithms and exponents to the base e)

$$\limsup_{N\to\infty} -\frac{1}{N}\log \alpha^{(N)}_{m_1,m_2,\ldots,m_K|l_1,l_2,\ldots,l_K} = E_{m_1,m_2,\ldots,m_K|l_1,l_2,\ldots,l_K} \geq 0. \tag{1}$$

These are exponents of error probabilities which we call reliabilities (in association with Shannon's reliability function [15]). We shall examine the matrix $\mathbf{E} = \{E_{m_1,m_2,\ldots,m_K|l_1,l_2,\ldots,l_K}\}$ and call it the reliability matrix.

Our criterion of optimality is: given M, K and values of a part of reliabilities to obtain the best (the largest) values for others. In addition it is necessary to describe the conditions under which all these reliabilities are positive. The procedure that realizes such testing is identification, which following Birgé [9], we call "logarithmically asymptotically optimal" (LAO).

Let $N(x|\mathbf{x})$ be the number of repetitions of the element $x \in \mathcal{X}$ in the vector $\mathbf{x} \in \mathcal{X}^N$, and let

$$Q = \{Q(x) = N(x|\mathbf{x})/N,\ x \in \mathcal{X}\}$$

is the distribution, called "the empirical distribution" (ED) of the sample $\mathbf{x}$ in statistics, in information theory called "the type" [14, 15] and in algebraic literature "the composition".

Denote the space of all empirical distributions for given N by $\mathcal{P}^{(N)}(\mathcal{X})$ and by $\mathcal{T}_Q^{(N)}$ the set of all vectors of the ED $Q \in \mathcal{P}^{(N)}(\mathcal{X})$.

Consider for $k \in [K]$, $m \in [M]$, relative entropies

$$D(Q_k||G_m) = \sum_{x\in\mathcal{X}} Q_k(x) \log \frac{Q_k(x)}{G_m(x)},$$

and entropies

$$H(Q_k) = -\sum_{x\in\mathcal{X}} Q_k(x) \log Q_k(x).$$

We shall use the following relations for the probability of the vector $\mathbf{x}$ when G_m is the distribution of the object:

$$G_m^{(N)}(\mathbf{x}) = \prod_{n=1}^{N} G_m(x_n) = \exp\{-N[D(Q||G_m) + H(Q)]\}.$$

For $m_k \in [M]$, $k \in [K]$, when the objects are independent and G_{m_k} is the distribution of the kth object:

$$P^{(N)}_{m_1,m_2,\ldots,m_K}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K) = \exp\{-N[\sum_{k=1}^{K} D(Q_k||G_{m_k}) + H(Q_k)]\}. \tag{2}$$

The equalities follow from the independence of N observations of K objects and from the definitions of relative entropies and entropies. It should be noted that the equality (2) is valid even when its left part is equal to 0, in that case for one of $\mathbf{x}_k$ the distribution Q_k is not absolutely continuous relative to G_{m_k} and $D(Q_k||G_{m_k}) = \infty$.

Our arguments will be based on the following fact: the "maximal likelihood" test accepts as the solution values $m_1, m_2, \ldots, m_k$, which maximize the probability $P^{(N)}_{m_1,m_2,\ldots,m_K}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K)$, but from (2) we see that the same solution can be obtained by minimization of the sum $\sum_{k=1}^{K}[D(Q_k||G_{m_k}) + H(Q_k)]$, that is the comparison with the help of relative entropies of the ED's of observed vectors with their hypothetical distributions may be helpful.

In this lecture we consider the following models.

1. K objects are different, they have different distributions among $M \geq K$ possibilities. For simplicity we restrict ourselves to the case $K = 2, M = 2$. It is the identification problem in formulations of the books [7] and [5].
2. K objects are independent, that is some of them may have the same distributions. We consider an example for $K, M = 2$. It is surprising, but this model has not been considered earlier in the literature.

3. We investigate one object, $K = 1$, and M possible probability distributions. The question is whether the mth distribution occurred or not. This is the problem of identification of distributions in the spirit of chapter "Identification via Channels".
4. Ranking, or ordering problem [1]. We have one vector of observations $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ and M hypothetical distributions. The receiver wants to know whether the index of the true distribution of the object is in $\{1, 2, \ldots, r\}$ or in $\{r + 1, \ldots, M\}$.
5. r-identification of distribution [1]. Again $K = 1$. One wants to identify the observed object as a member either of the subset $\mathcal{S}$ of $[M]$, or of its complement, with r being the number of elements in $\mathcal{S}$.

Section 2 of this lecture presents necessary notions and results on hypothesis testing. The models of identification for independent objects are considered in 3 and for different objects in 4. Section 5 is devoted to the problem of identification of an object distribution and 6 to the problems of r-identification and ranking. Some results are illustrated by numerical examples and graphs. Many directions of further research are indicated in the course of the text and in the 7.

2 Background

The study of interdependence of exponential rates of decrease, as the sample size N goes to the infinity, of the error probabilities $\alpha_{1|2}^{(N)}$ of the "first kind" and $\alpha_{2|1}^{(N)}$ of the "second kind" was started by the works of Hoeffding [23], Csiszár and Longo [16], Tusnády [28], Longo and Sgarro [24], Birgé [9], and for multiple hypotheses by Haroutunian [22]. Similar problems for Markov dependence of experiments were investigated by Natarajan [26], Haroutunian [21], Gutman [18] and others. As it was remarked by Blahut in his book [11], it is unfortunately confusing that the errors are denoted type I and type II, while the hypotheses are subscripted 0 and 1. The word "type" is also used in another sense to refer to the type of a measurement or the type of a vector. For this reason we do not use the names "0" and "1" for hypotheses and the name "type" for errors. Note that in [10, 11, 17] an application of the methods of hypothesis testing to the proper problems of information theory is developed.

It will be very interesting to combine investigation of described models with the approach initiated by the paper of Ahlswede and Csiszár [3] and developed by many authors, particularly, for the exponentially decreasing error probabilities by Han and Kobayashi [20].

In [8] Berger formulated the problem of remote statistical inference. Zhang and Berger [29] studied a model of an estimation system with compressed information. Similar problems were examined by Ahlswede and Burnashev [2] and by Han and AmariAmari, S. [19]. In the paper of Ahlswede, Yang and Zhang [6] identification in channels via compressed data was considered. Fu and Shen [17] studied hypothesis testing for an arbitrarily varying source.

Our further considerations will be based on the results from [22] on multiple hypotheses testing, so now we expose briefly corresponding formulations and proofs. In our terms it is the case of one object ($K = 1$) and M possible distributions (hypotheses) $G_1, \ldots, G_M$. A test $\varphi(\mathbf{x})$ on the base of N-sample $\mathbf{x} = (x_1, \ldots, x_N)$ determines the distribution. Since experiments are independent the probability of the sample $\mathbf{x}$ if the distribution is G_m will be

$$G_m^{(N)}(\mathbf{x}) = \prod_{n=1}^{N} G_m(x_n), \ m \in [M].$$

We study error probabilities $\alpha_{m|l}^{(N)}$ for $m, l \in [M]$. Here $\alpha_{m|l}^{(N)}$ is the probability that the distribution G_l was accepted instead of true distribution G_m. For $m = l$ the probability to reject G_m when it is true, is denoted by $\alpha_{m|m}^{(N)}$ thus:

$$\alpha_{m|m}^{(N)} = \sum_{l:l\neq m} \alpha_{m|l}^{(N)}.$$

This probability is called [12] the test's "error probability of the kind m". The matrix $\{\alpha_{m|l}^{(N)})\}$ is sometimes called the "power of the test" [12].

In this lecture we suppose that the list of possible hypotheses is complete. Remark that, as it was noted by Rao [27], the case, when the objects may have also some distributions different from $G_1, \ldots, G_M$, is interesting too.

Let us analyze the reliability matrix

$$\mathbf{E} = \begin{pmatrix} E_{1|1} \ \ldots \ E_{1|l} \ \ldots \ E_{1|M} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ E_{m|1} \ldots E_{m|l} \ldots E_{m|M} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ E_{M|1} \ldots E_{M|l} \ldots E_{M|M} \end{pmatrix}$$

with components

$$E_{m|l} = \limsup_{N\to\infty} -\frac{1}{N} \log \alpha_{m|l}^{(N)}, \ m, l \in [M].$$

According to this definition and the definition of $\alpha_{m|l}^{(N)}$ we can derive that

$$E_{m|m} = \min_{l:m\neq l} E_{m|l}. \tag{3}$$

Really,

$$E_{m|m} = \limsup_{N\to\infty} -\frac{1}{N}\log\sum_{l:m\neq l}\alpha_{m|l}^{(N)}$$

$$= \limsup_{N\to\infty} -\frac{1}{N}\log\max_{l:m\neq l}\alpha_{m|l}^{(N)} + \limsup_{N\to\infty} -\frac{1}{N}\log\left[\left(\sum_{l:m\neq l}\alpha_{m|l}^{(N)}\right)/\max_{l:m\neq l}\alpha_{m|l}^{(N)}\right]$$

$$= \min_{l:m\neq l} E_{m|l}.$$

The last equality is a consequence of the fact that for all m and N

$$1 \le (\sum_{l:m\neq l}\alpha_{m|l}^{(N)})/\max_{l:m\neq l}\alpha_{m|l}^{(N)} \le M-1.$$

In the case $M = 2$, the reliability matrix is

$$\mathbf{E} = \begin{pmatrix} E_{1|1} & E_{1|2} \\ E_{2|1} & E_{2|2} \end{pmatrix} \tag{4}$$

and it follows from (3) that there are only two different values of elements, namely

$$E_{1|1} = E_{1|2} \text{ and } E_{2|1} = E_{2|2}, \tag{5}$$

so in this case the problem is to find the maximal possible value of one of them, given the value of the other.

In the case of M hypotheses for given positive and finite $E_{1|1}, E_{2|2}, \ldots, E_{M-1,M-1}$ let us consider the regions of distributions

$$\mathcal{R}_l = \{Q : D(Q||G_l) \le E_{l|l}\}, \qquad l \in [M-1], \tag{6}$$

$$\mathcal{R}_M = \{Q : D(Q||G_l) > E_{l|l},\ l \in [M-1]\} = \mathcal{P}(\mathcal{X}) - \bigcup_{l=1}^{M-1}\mathcal{R}_l, \tag{7}$$

$$\mathcal{R}_l^{(N)} = \mathcal{R}_l \bigcap \mathcal{P}^{(N)}, \qquad l \in [M]. \tag{8}$$

Let

$$E_{l|l}^* = E_{l|l}^*(E_{l|l}) = E_{l|l},\ l \in [M-1], \tag{9}$$

$$E_{m|l}^* = E_{m|l}^*(E_{l|l}) = \inf_{Q\in\mathcal{R}_l} D(Q||G_m),\ m \in [M],\ m \neq l,\ l \in [M-1], \tag{10}$$

$$E^*_{m|M} = E^*_{m|M}(E_{1|1}, \ldots, E_{M-1,M-1}) = \inf_{Q \in \mathcal{R}_M} D(Q||G_m),\ m \in [M-1], \tag{11}$$

$$E^*_{M|M} = E^*_{M|M}(E_{1|1}, \ldots, E_{M-1,M-1}) = \min_{l \in [M-1]} E^*_{M|l}. \tag{12}$$

If some distribution G_m is not absolutely continuous relative to G_l the reliability $E^*_{m|l}$ will be equal to the infinity, this means that corresponding $\alpha^{(N)}_{m|l} = 0$ for some large N.

The principal result of [22] is:

Theorem 272 *If all the distributions G_m are different and all elements of the matrix $\{D(G_l||G_m)\}$, $l, m \in [M]$, are positive, but finite, two statements hold:*

(i) when the positive numbers $E_{1|1}, E_{2|2}, \ldots, E_{M-1,M-1}$ satisfy conditions

$$E_{1|1} < \min_{l \in [2,M]} D(G_l||G_1),$$

$$\vdots \tag{13}$$

$$E_{m|m} < \min[\min_{l \in [m-1]} E^*_{m|l}(E_{l|l}), \min_{l \in [m+1,M]} D(G_l||G_m)],\ m \in [2, M-1],$$

then there exists a LAO sequence of tests, the reliability matrix of which $E^* = \{E^*_{m|l}\}$ is defined in (9)–(12) and all elements of it are positive;

(ii) even if one of conditions (13) is violated, then the reliability matrix of any such test has at least one element equal to zero (that is the corresponding error probability does not tend to zero exponentially).

The essence of the proof of Theorem 272 consists in construction of the following optimal tests sequence. Let the decision l will be taken when $\mathbf{x}$ gets into the set

$$\mathcal{B}^{(N)}_l = \bigcup_{Q \in \mathcal{R}^{(N)}_l} \mathcal{T}^{(N)}_Q, \qquad l \in [M],\ N = 1, 2, \ldots. \tag{14}$$

The non-coincidence of the distributions G_m and the conditions (13) guarantee that the sets from (14) are not empty, they meet conditions

$$\mathcal{B}^{(N)}_l \bigcap \mathcal{B}^{(N)}_m = \emptyset, \qquad l \neq m,$$

and

$$\bigcup_{l=1}^{M} \mathcal{B}^{(N)}_l = \mathcal{X}^N,$$

and so they define a sequence of tests, which proves to be LAO.

For the simplest particular case $M = 2$ elements of the reliability matrix (4) satisfy equalities (5) and for given $E_{1|1}$ from (5) and (7) we obtain the value of $E^*_{2|1} = E^*_{2|2}$:

$$E^*_{2|1}(E_{1|1}) = \inf_{Q:D(Q||G_1)\le E_{1|1}} D(Q||G_2). \tag{15}$$

Here, according to (13), we can take $E_{1|1}$ from $(0, D(G_2\|G_1))$ and $E^*_{2|1}(E_{1|1})$ will range between $D(G_1||G_2)$ and 0.

3 Identification Problem for Model with Independent Objects

We begin with study of the second model. To illustrate possibly arising developments and essential features we consider a particular case $K = 2, M = 2$. It is clear that the case with $M = 1$ is trivial. The reliability matrix is (see (1))

$$\mathbf{E} = \begin{pmatrix} E_{1,1|1,1} & E_{1,1|1,2} & E_{1,1|2,1} & E_{1,1|2,2} \\ E_{1,2|1,1} & E_{1,2|1,2} & E_{1,2|2,1} & E_{1,2|2,2} \\ E_{2,1|1,1} & E_{2,1|1,2} & E_{2,1|2,1} & E_{2,1|2,2} \\ E_{2,2|1,1} & E_{2,2|1,2} & E_{2,2|2,1} & E_{2,2|2,2} \end{pmatrix}.$$

Let us denote by $\alpha^{(1)}_{m_1|l_1}, \alpha^{(2)}_{m_2|l_2}$ and $E^{(1)}_{m_1|l_1}, E^{(2)}_{m_2|l_2}$ the error probabilities and the reliabilities as in (4) for, respectively, the first and the second objects.

Lemma 273 *If* $0 < E^{(i)}_{1|1} < D(G_2||G_1)$, $i = 1, 2$, *then the following equalities hold true:*

$$E_{m_1,m_2|l_1,l_2} = E^{(1)}_{m_1|l_1} + E^{(2)}_{m_2|l_2}, \quad \text{if } m_1 \ne l_1,\ m_2 \ne l_2, \tag{16}$$

$$E_{m_1,m_2|l_1,l_2} = E^{(i)}_{m_i|l_i}, \quad \text{if } m_{3-i} = l_{3-i},\ m_i \ne l_i,\ i = 1, 2, \tag{17}$$

Proof From the independence of the objects it follows that

$$\alpha^{(N)}_{m_1,m_2|l_1,l_2} = \alpha^{(N,1)}_{m_1|l_1}\alpha^{(N,2)}_{m_2|l_2}, \quad \text{if } m_1 \ne l_1,\ m_2 \ne l_2, \tag{18}$$

$$\alpha^{(N)}_{m_1,m_2|l_1,l_2} = \alpha^{(N,i)}_{m_i|l_i}(1 - \alpha^{(N,3-i)}_{m_{3-i}|l_{3-i}}), \quad \text{if } m_{3-i} = l_{3-i},\ m_i \ne l_i,\ i = 1, 2, \tag{19}$$

According to (1), from (18) we obtain (16), from (19) and the conditions of positiveness of $E^{(i)}_{1|1}$ and $E^{(i)}_{2|2}$, $i = 1, 2$, (17) follows. □

Theorem 274 *If the distributions G_1 and G_2 are different, the strictly positive elements $E_{1,1|1,2}$, $E_{1,1|2,1}$ of the reliability matrix* $\mathbf{E}$ *are given and bounded above:*

$$E_{1,1|1,2} < D(G_2||G_1), \text{ and } E_{1,1|2,1} < D(G_2||G_1), \tag{20}$$

then the other elements of the matrix $\mathbf{E}$ *are defined as follows:*

$$\begin{aligned} E_{2,1|2,2} &= E_{1,1|1,2}, & E_{1,2|2,2} &= E_{1,1|2,1}, \\ E_{2,2|1,1} &= E_{1,2|1,1} + E_{2,1|1,1}, & E_{2,1|1,2} &= E_{2,1|1,1} + E_{1,2|2,2}, \\ E_{1,2|2,1} &= E_{1,2|1,1} + E_{1,2|2,2}, & E_{1,1|2,2} &= E_{1,1|1,2} + E_{1,1|2,1}, \end{aligned}$$

$$\begin{aligned} E_{1,2|1,1} = E_{2,2|2,1} &= \inf_{Q:D(Q||G_1)\le E_{1,1|1,2}} D(Q||G_2), \\ E_{2,1|1,1} = E_{2,2|1,2} &= \inf_{Q:D(Q||G_1)\le E_{1,1|2,1}} D(Q||G_2), \\ E_{m_1,m_2|m_1,m_2} &= \min_{(l_1,l_2)\neq(m_1,m_2)} E_{m_1,m_2|l_1,l_2}, \ m_1, m_2 = 1, 2. \end{aligned} \tag{21}$$

If one of the inequalities (20) is violated, then at least one element of the matrix $\mathbf{E}$ *is equal to 0.*

Proof The last equalities in (21) follow (as (3)) from the definition of

$$\alpha^{(N)}_{m_1,m_2|m_1,m_2} = \sum_{(l_1,l_2)\neq(m_1,m_2)} \alpha^{(N)}_{m_1,m_2|l_1,l_2}, \qquad m_1, m_2 = 1, 2.$$

Let us consider the reliability matrices of each of the objects X_1 and X_2

$$\mathbf{E^{(1)}} = \begin{pmatrix} E^{(1)}_{1|1} & E^{(1)}_{1|2} \\ E^{(1)}_{2|1} & E^{(1)}_{2|2} \end{pmatrix} \qquad \text{and} \qquad \mathbf{E}^{(2)} = \begin{pmatrix} E^{(2)}_{1|1} & E^{(2)}_{1|2} \\ E^{(2)}_{2|1} & E^{(2)}_{2|2} \end{pmatrix}.$$

From (5) we know that $E^{(i)}_{1|1} = E^{(i)}_{1|2}$ and $E^{(i)}_{2|1} = E^{(i)}_{2|2}$, $i = 1, 2$. From (20) it follows that $0 < E^{(1)}_{1|1} < D(G_2||G_1)$, $0 < E^{(2)}_{1|1} < D(G_2||G_1)$. Really, if $0 < E_{1,1|1,2} < D(G_2||G_1)$, but $E^{(2)}_{1|1} \ge D(G_2||G_1)$, then from (19) and (1) we arrive to

$$\limsup_{N\to\infty} -\frac{1}{N}\log(1 - \alpha^{(N,2)}_{1|2}) < 0,$$

therefore index N_0 exists, such that for sub-sequence of $N > N_0$ we will have $1 - \alpha^{(N,2)}_{1|2} > 1$. But this is impossible because $\alpha^{(N,2)}_{1|2}$ is the probability and must be positive.

Using Lemma 273 we can deduce that the reliability matrix $\mathbf{E}$ can be obtained from matrices $\mathbf{E}^{(1)}$ and $\mathbf{E}^{(2)}$ as follows:

$$\mathbf{E} = \begin{pmatrix} \min(E^{(1)}_{1|2}, E^{(2)}_{1|2}) & E^{(2)}_{1|2} & E^{(1)}_{1|2} & E^{(1)}_{1|2} + E^{(2)}_{1|2} \\ E^{(2)}_{2|1} & \min(E^{(1)}_{1|2}, E^{(2)}_{2|1}) & E^{(1)}_{1|2} + E^{(2)}_{2|1} & E^{(1)}_{1|2} \\ E^{(1)}_{2|1} & E^{(1)}_{2|1} + E^{(2)}_{1|2} & \min(E^{(1)}_{2|1}, E^{(2)}_{1|2}) & E^{(2)}_{1|2} \\ E^{(1)}_{2|1} + E^{(2)}_{2|1} & E^{(1)}_{2|1} & E^{(2)}_{2|1} & \min(E^{(1)}_{2|1}, E^{(2)}_{2|1}) \end{pmatrix},$$

in other words, providing, that conditions (20) are fulfilled, we find that

$$\begin{aligned} &E_{1,1|1,2} = E^{(2)}_{1|2} = E^{(2)}_{1|1}, && E_{1,1|2,1} = E^{(1)}_{1|2} = E^{(1)}_{1|1}, \\ &E_{2,1|2,2} = E_{1,1|1,2} = E^{(2)}_{1|2}, && E_{1,2|2,2} = E_{1,1|2,1} = E^{(1)}_{1|2}, \\ &E_{1,2|1,1} = E_{2,2|2,1} = E^{(2)}_{2|1}, && E_{2,1|1,1} = E_{2,2|1,2} = E^{(1)}_{2|1}, \\ &E_{2,2|1,1} = E^{(1)}_{2|1} + E^{(2)}_{2|1}, && E_{2,1|1,2} = E^{(1)}_{2|1} + E^{(2)}_{1|2}, \\ &E_{1,2|2,1} = E^{(1)}_{1|2} + E^{(2)}_{2|1}, && E_{1,1|2,2} = E^{(1)}_{1|2} + E^{(2)}_{1|2}, \end{aligned} \tag{22}$$

and

$$E_{m_1,m_2|m_1,m_2} = \min\{E^{(1)}_{m_1|m_1}, E^{(2)}_{m_2|m_2}\}, \qquad m_1, m_2 = 1, 2,$$

From Theorem 272 we know that if $E^{(i)}_{1|1} \in (0, D(G_2||G_1))$, $i = 1, 2$, then the tests of both objects are LAO and the elements $E^{(i)}_{2|1}$, $i = 1, 2$, can be calculated (see (15)) by

$$E^{(i)}_{2|1} = \inf_{Q: D(Q||G_1) \le E^{(i)}_{1|1}} D(Q||G_2), \quad i = 1, 2, \tag{23}$$

and if $E^{(i)}_{1|1} \ge D(G_2||G_1)$, then $E^{(i)}_{2|1} = 0$.

According to (22) and (23), we obtain, that when (20) takes place, the elements of the matrix $\mathbf{E}$ are determined by relations (21). When one of the inequalities (20) is violated, then from (23) and the first and the third lines of (22) we see, that some elements in the matrix $\mathbf{E}$ must be equal to 0 (namely, either $E_{1,2|1,1}$, or $E_{2,1|1,1}$ and others).

Now let us show that the compound test for two objects is LAO, that is it is optimal. Suppose that for given $E_{1,1|1,2}$ and $E_{1,1|2,1}$ there exists a test with matrix $\mathbf{E}'$, such that it has at least one element exceeding the respective element of the matrix $\mathbf{E}$. Comparing elements of matrices $\mathbf{E}$ and $\mathbf{E}'$ different from $E_{1,1|1,2}$ and $E_{1,1|2,1}$, from (22) we obtain that either $E_{1,2|1,1} < E'_{1,2|1,1}$, or $E_{2,1|1,1} < E'_{2,1|1,1}$,

i.e. either $E^{(2)}_{2|1} < E^{(2)'}_{2|1}$, or $E^{(1)}_{2|1} < E^{(1)'}_{2|1}$. It is contradiction to the fact, that LAO tests have been used for the objects X_1 and X_2.

When it is demanded to take the same values for the reliabilities of the first and the second objects $E^{(1)}_{1|2} = E^{(2)}_{1|2} = a_1$ and, consequently, $E^{(1)}_{2|1} = E^{(2)}_{2|1} = a_2$, then the matrix $\mathbf{E}$ will take the following form

$$\mathbf{E} = \begin{pmatrix} a_1 & a_1 & a_1 & 2a_1 \\ a_2 & \min(a_1, a_2) & a_1 + a_2 & a_1 \\ a_2 & a_1 + a_2 & \min(a_1, a_2) & a_1 \\ 2a_2 & a_2 & a_2 & a_2 \end{pmatrix}. \qquad \square$$

4 Identification Problem for Models with Different Objects

The K objects are not independent, they have different distributions, and so the number M of the distributions is not less than K. This is the model studied in [7]. For brevity we consider the case $K = 2$, $M = 2$. The matrix of reliabilities will be the following:

$$\mathbf{E} = \begin{pmatrix} E_{1,2|1,2} & E_{1,2|2,1} \\ E_{2,1|1,2} & E_{2,1|2,1} \end{pmatrix}. \tag{24}$$

Since the objects are strictly dependent this matrix coincides with the reliability matrix of the first object (see (4))

$$\mathbf{E^{(1)}} = \begin{pmatrix} E^{(1)}_{1|1} & E^{(1)}_{1|2} \\ E^{(1)}_{2|1} & E^{(1)}_{2|2} \end{pmatrix},$$

because the distribution of the second object is uniquely defined by the distribution of the first one.

We can conclude that among 4 elements of the reliability matrix of two dependent objects only 2 elements are distinct, the second of which is defined by given $E^{(1)}_{1|1} = E_{1,2|1,2}$.

From symmetry it follows that the reliability matrix of the second object also may determine the matrix (24).

5 Identification of the Probability Distribution of an Object

Let we have one object, $K = 1$, and there are known $M \geq 2$ possible distributions. The question is whether rth distribution occurred, or not. There are two error probabilities for each $r \in [M]$ the probability $\alpha^{(N)}_{m=r|l\neq r}$ to accept l different from r,

when r is in reality, and the probability $\alpha^{(N)}_{m\neq r|l=r}$ that r is accepted, when it is not correct.

The probability $\alpha^{(N)}_{m=r|l\neq r}$ is already known, it coincides with the probability $\alpha^{(N)}_{r|r}$ which is equal to $\sum\limits_{l:l\neq r}\alpha^{(N)}_{r|l}$. The corresponding reliability $E_{m=r|l\neq r}$ is equal to $E_{r|r}$ which satisfies the equality (3).

We have to determine the dependence of $E_{m\neq r|l=r}$ upon given $E_{m=r|l\neq r}=E_{r|r}$, which can be assigned values satisfying conditions (13), this time we will have the conditions:

$$0<E_{r|r}<\min_{l:l\neq r}D(G_l\|G_r),\quad r\in[M].$$

We need the probabilities of different hypotheses. Let us suppose that the hypotheses $G_1,\dots,G_M$ have, say, probabilities $\Pr(r),\ r\in[M]$. The only supposition we shall use is that $\Pr(r)>0,\ r\in[M]$. We will see, that the result formulated in the following theorem does not depend on values of $\Pr(r),\ r\in[M]$, if they all are strictly positive.

Now we can make the following reasoning for each $r\in[M]$:

$$\alpha^{(N)}_{m\neq r|l=r}=\frac{\Pr^{(N)}(m\neq r,l=r)}{\Pr(m\neq r)}=\frac{1}{\sum\limits_{m:m\neq r}\Pr(m)}\sum_{m:m\neq r}\Pr^{(N)}(m,r).$$

From here we see that for $r\in[M]$

$$\begin{aligned}E_{m\neq r|l=r}&=\limsup_{N\to\infty}\left(-\frac{1}{N}\log\alpha^{(N)}_{m\neq r|l=r}\right)\\&=\limsup_{N\to\infty}\frac{1}{N}\left(\log\sum_{m:m\neq r}\Pr(m)-\log\sum_{m:m\neq r}\alpha^{(N)}_{m|r}\Pr(m)\right)\\&=\min_{m:m\neq r}E^*_{m|r}.\end{aligned}\tag{25}$$

Using (25) by analogy with the formula (15) we conclude (with $\mathcal{R}_r$ defined as in (6) for each r including $r=M$ by the values of $E_{r|r}$ from $(0,\min\limits_{l:l\neq r}D(G_l||G_r))$) that

$$\begin{aligned}E_{m\neq r|l=r}(E_{r|r})&=\min_{m:m\neq r}\inf_{Q\in\mathcal{R}_r}D(Q\|G_m)\\&=\min_{m:m\neq r}\inf_{Q:D(Q\|G_r)\le E_{r|r}}D(Q\|G_m),\ r\in[M].\end{aligned}\tag{26}$$

We can summarize this result in

Theorem 275 *For the model with different distributions, for the given sample* **x** *we define its ED* Q*, and when* $Q \in \mathcal{R}_r^{(N)}$ *we accept the hypothesis* r*. Under condition that the probabilities of all* M *hypotheses are positive the reliability of such test* $E_{m\neq r|l=r}$ *for given* $E_{m=r|l\neq r} = E_{r|r}$ *is defined by (26).*

For presentation of examples let us consider the set $\mathcal{X} = \{0, 1\}$ with only 2 elements. Let 5 probability distributions are given on $\mathcal{X}$:

$$G_1 = \{0.1,\ 0.9\}$$
$$G_2 = \{0.65,\ 0.35\}$$
$$G_3 = \{0.45,\ 0.55\}$$
$$G_4 = \{0.85,\ 0.15\}$$
$$G_5 = \{0.23,\ 0.77\}$$

On Fig. 1, the results of calculations of $E_{m\neq r|l=r}$ as function of $E_{m=r|l\neq r}$ are presented.

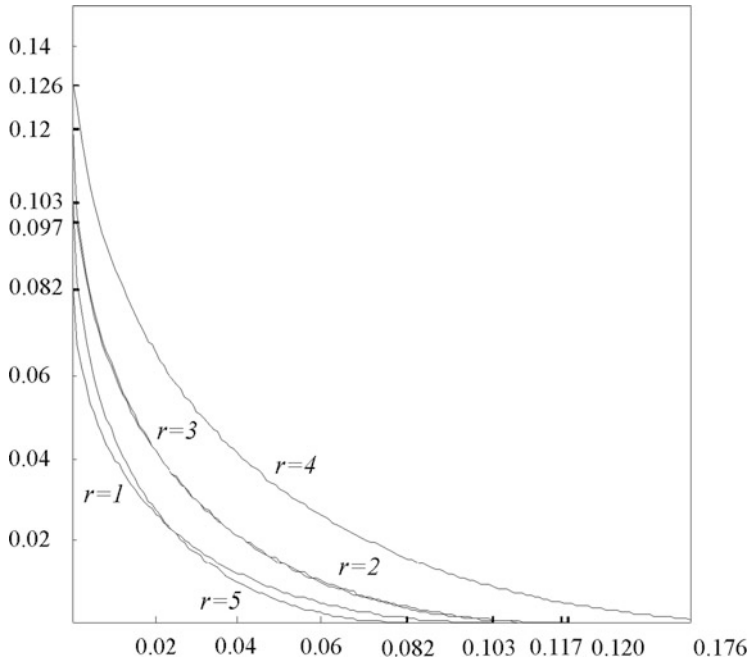


Fig. 1 $E_{m\neq r|l=r}$ as function of $E_{m=r|l\neq r}$ for $r = 1, 2, 3, 4, 5$

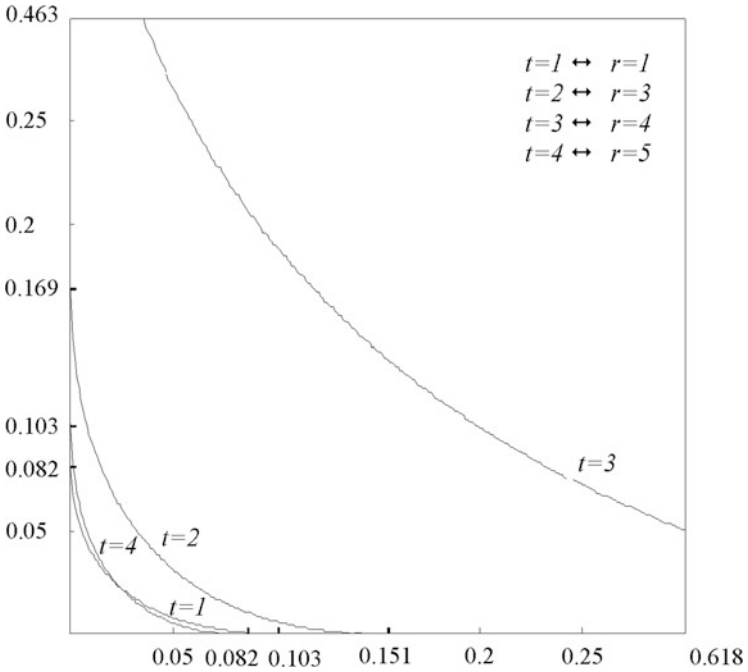


Fig. 2 $E_{m\neq t|l=t}$ as function of $\left[E_{t|t}\right]$ for $t = 1, 2, 3, 4$

The elements of the matrix of relative entropies of all pairs of distributions are used for calculation of conditions (13) for this example.

$$\{D(G_m\|G_l)\}_{m\in[5]}^{l\in[5]} = \begin{pmatrix} 0 & 0.956 & 0.422 & 2.018 & 0.082 \\ 1.278 & 0 & 0.117 & 0.176 & 0.576 \\ 0.586 & 0.120 & 0 & 0.618 & 0.169 \\ 2.237 & 0.146 & 0.499 & 0 & 1.249 \\ 0.103 & 0.531 & 0.151 & 1.383 & 0 \end{pmatrix}.$$

In Figs. 2 and 3 the results of calculations of the same dependence are presented for 4 distributions taken from previous 5.

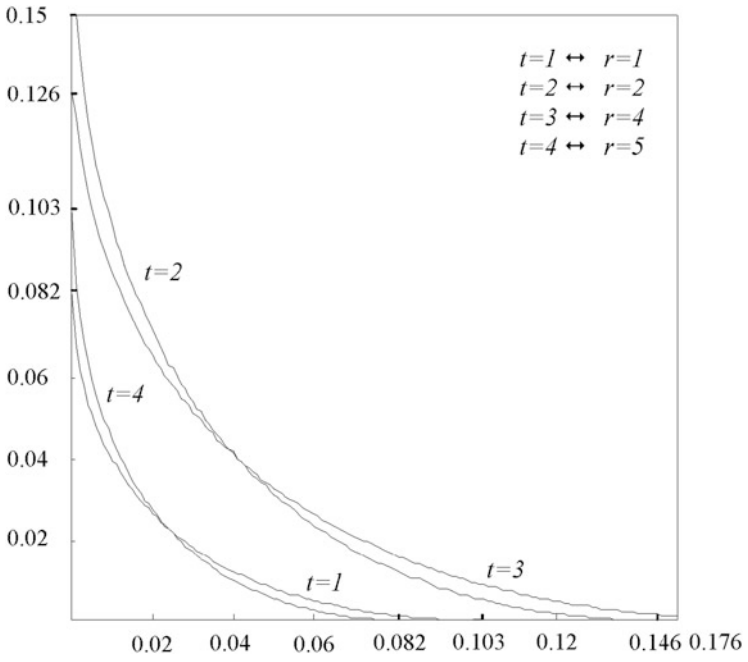


Fig. 3 $E_{m\neq t|l=t}$ as function of $\left[E_{t|t}\right]$ for $t = 1, 2, 3, 4$

6 r-Identification and Ranking Problems

The model was introduced in [1] and named K-identification. Since in this lecture the letter K is already used we speak of r-identification. Given N-sample $\mathbf{x}$ of measurements of the object the problem is to answer to the question: is the distribution of the object in the part $\mathcal{S}$ of M possible distributions or in its complement, here r is the number of elements of the set $\mathcal{S}$.

Again we can make decision on the base of the ED Q of the sample $\mathbf{x}$ and suppose that before experiments all hypotheses have some positive probabilities

$$\Pr(1), \ldots, \Pr(M).$$

Using (6)–(8) with some $E_{1,1}, \ldots, E_{M-1,M-1}$ meeting the conditions (13) when $Q \in \bigcup\limits_{l\in\mathcal{S}} \mathcal{R}_l^{(N)}$ decision "l is in $\mathcal{S}$" follows.

The model of ranking is the particular case of the model of r-identification with $\mathcal{S} = \{1, 2, \dots, r\}$. But conversely the r-identification problem without loss of generality may be considered as the ranking problem, to this end we can renumber the hypotheses placing the hypotheses of $\mathcal{S}$ in the r first places. Because these two models are mathematically equivalent we shall speak below only of the ranking model.

It is enough to consider the cases $r \leq \lceil M/2 \rceil$, because in the cases of larger r we can replace $\mathcal{S}$ with its complement. Remark that the case $r = 1$ was considered in 5.

We study two error probabilities of a test: the probability $\alpha^{(N)}_{m\leq r|l>r}$ to make incorrect decision when m is not greater than r and the probability $\alpha^{(N)}_{m>r|l\leq r}$ to make error when m is greater than r. The corresponding reliabilities are

$$E_1(r) = E_{m\leq r|l>r} \text{ and } E_2(r) = E_{m>r|l\leq r}, \ 1 \leq r \leq \lceil M/2 \rceil. \tag{27}$$

With supposition (6) we have

$$\begin{aligned} \alpha^{(N)}_{m\leq r|l>r} &= \frac{\Pr^{(N)}(m \leq r, l > r)}{\Pr(m \leq r)} \\ &= \frac{1}{\sum\limits_{m\leq r} \Pr(m)} \sum_{m\leq r} \sum_{l>r} \Pr^{(N)}(m, l) \\ &= \frac{1}{\sum\limits_{m\leq r} \Pr(m)} \sum_{m\leq r} \sum_{l>r} \alpha^{(N)}_{m|l} \Pr(m). \end{aligned} \tag{28}$$

The definition (27) of $E_1(r)$ and the equality (28) give

$$\begin{aligned} E_1(r) &= \limsup_{N\to\infty} -\frac{1}{N} \log \alpha^{(N)}_{m\leq r|l>r} \\ &= \limsup_{N\to\infty} -\frac{1}{N} \left[\log \sum_{m\leq r} \sum_{l>r} \Pr(m) \alpha^{(N)}_{m|l} - \log \sum_{m\leq r} \Pr(m) \right] \\ &= \min_{m\leq r, l>r} E_{m|l}. \end{aligned} \tag{29}$$

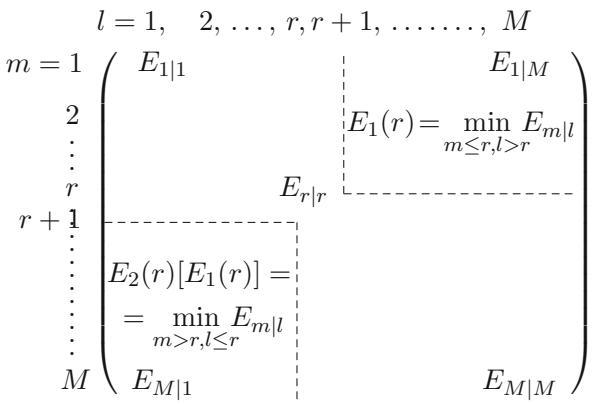


Fig. 4 Calculation of $E_2(r)\,[E_1(r)]$

Analogously, at the same time

$$\begin{aligned} E_2(r) &= \limsup_{N\to\infty} -\frac{1}{N} \log \alpha^{(N)}_{m>r|l\le r} \\ &= \limsup_{N\to\infty} -\frac{1}{N} \left[\log \sum_{m>r} \sum_{l\le r} \alpha^{(N)}_{m|l} - \log \sum_{m>r} \Pr(m) \right] \\ &= \min_{m>r, l\le r} E_{m|l}. \end{aligned} \tag{30}$$

For any test the value of $E_1(r)$ must satisfy the condition (compare (3) and (29))

$$E_1(r) \ge \min_{m:m\le r} E_{m|m}. \tag{31}$$

Thus for any test meeting all inequalities from (13) for $m \le r$ and inequality (31) the reliability $E_2(r)$ may be calculated with the equality (30). For given value of $E_1(r)$ the best $E_2(r)$ will be obtained if we use liberty in selection of the biggest values for reliabilities $E_{m|m}$, $r < m \le M-1$, satisfying for those m-s conditions (13). These reasonings may be illuminated by Fig. 4 and resumed as follows:

Theorem 276 *When the probabilities of the hypotheses are positive, for given $E_1(r)$ for $m \le r$ not exceeding the expressions on the right in (13), $E_2(r)$ may*

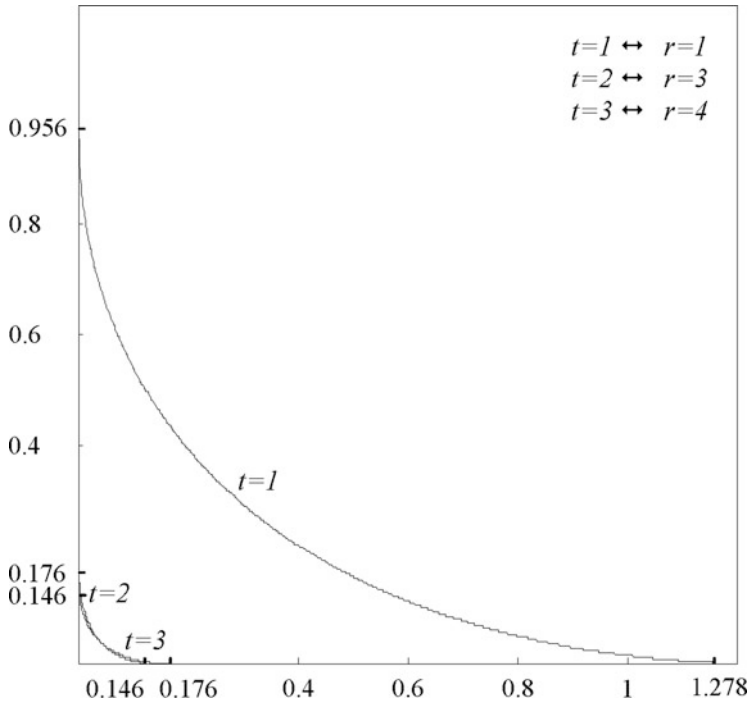


Fig. 5 $E_{m\neq t|l=t}$ as function of $\left[E_{t|t}\right]$ for $t = 1, 2, 3$

be calculated in the following way:

$$E_2(r)\,[E_1(r)] = \max_{\{E_{m|l},\ m,l\in[M]\}:\ \min\limits_{m\le r,\ l>r} E^*_{m|l}=E_1(r)} \left[\min_{m>r,\ l\le r} E^*_{m|l}\right] \tag{32}$$

with $E^*_{m|l}$ *defined in (9)–(12).*

Remark One can see from (32) that for $r = 1$ we arrive to (26) for $r = 1$.

In Figs. 5 and 6 for 2 subsets by 3 distributions taken from 5 defined for Fig. 1 the results of calculation of the dependence (26) and in Figs. 7 and 8 the corresponding results of the formula (33) are presented.

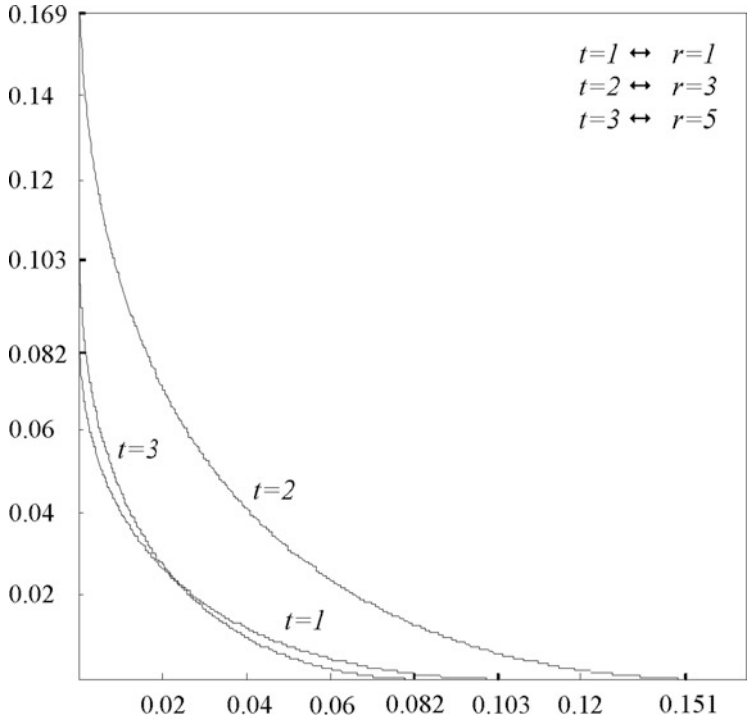


Fig. 6 $E_{m\neq t|l=t}$ as function of $\left[E_{t|t}\right]$ for $t = 1, 2, 3$

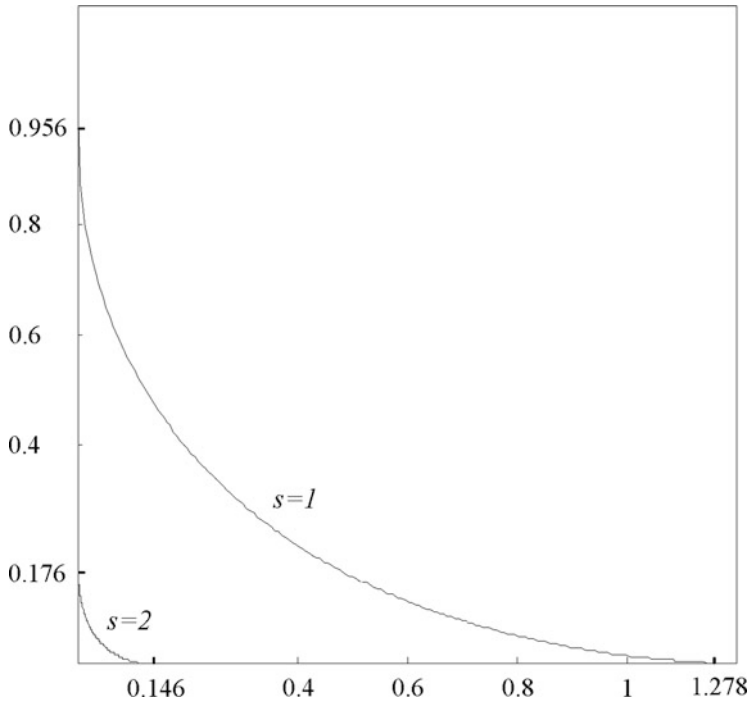


Fig. 7 $E_2(r)$, $E_1(r)$

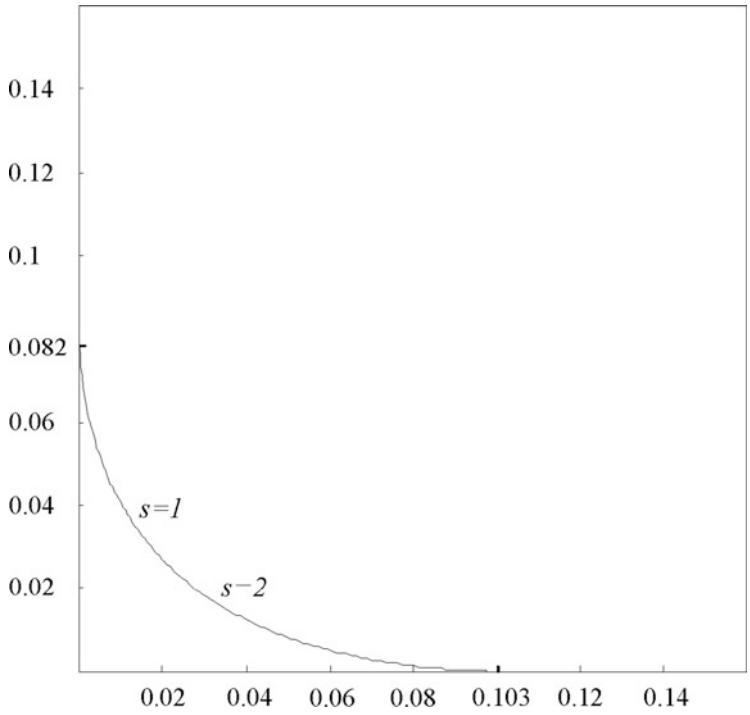


Fig. 8 $E_2(r)$, $E_1(r)$

7 Conclusion and Extensions of Problems

The lecture is a contribution to influence of the information theory methods on statistical theory. We have shown by simple examples what questions arise in different models of statistical identification.

Problems and results of the lecture may be extended in several directions some of which have been already noted above.

It is necessary to examine models in which measurements are described by more general classes of RV's and processes [18, 19, 21, 26].

One of the directions is connected with the use of compressed data of measurements [2, 6, 8, 19, 29].

One may see perspectives in application of identification approach and methods to the authentication theory [25] and steganography [13].

References

1. R. Ahlswede, General theory of information transfer, Preprint 97–118, SFB 343 "Diskrete Strukturen in der Mathematik", Universität Bielefeld, General Theory of Information Transfer and Combinatorics, Report on a Research Project at the ZIF (Center of interdisciplinary studies) in Bielefeld Oct. 1, 2002–August 31, 2003, edit R. Ahlswede with the assistance of L. Bäumer and N. Cai, also Special issue of Discrete Mathematics
2. R. Ahlswede, M. Burnashev, On minimax estimation in the presence of side information about remote data. Ann. Stat. **18**(1), 141–171 (1990)
3. R. Ahlswede, I. Csiszár, Hypotheses testing with communication constraints. IEEE Trans. Inf. Theory **32**(4), 533–542 (1986)
4. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**(1), 15–29 (1989)
5. R. Ahlswede, I. Wegener, *Search Problems* (Wiley, New York, 1987)
6. R. Ahlswede, E. Yang, Z. Zhang, Identification via compressed data. IEEE Trans. Inf. Theory **43**(1), 48–70 (1997)
7. R.E. Bechhofer, J. Kiefer, M. Sobel, *Sequential Identification and Ranking Procedures*. University of Chicago Press, Chicago (1968)
8. T. Berger, Decentralized estimation and decision theory, in *Presented at IEEE Seven Springs Workshop on Information Theory, Mt. Kisco, NY* (1979)
9. L. Birgé, Vitesse maximales de décroissance des erreurs et tests optimaux associés. Z. Wahrsch. verw Gebiete **55**, 261–273 (1981)
10. R.E. Blahut, Hypothesis testing and information theory. IEEE Trans. Inform. Theory, **IT-20**, 405–417 (1974)
11. R.E. Blahut, *Principles and Practice of Information Theory* (Addison-Wesley, Reading, 1991)
12. A.A. Borovkov, *Mathematical Statistics (in Russian)* (Nauka, Novosibirsk, 1997)
13. C. Cachin, An information-theoretic model for steganography, in *Proceedings of the 2nd Workshop on Information Hiding (David Ausmith, ed.)*. Lecture Notes in computer Science (Springer, Berlin, 1998)
14. I. Csiszár, Method of types. IEEE Trans. Inf. Theory **44**(6), 2505–2523 (1998)
15. I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems* (Academic Press, New York, 1981)
16. I. Csiszár, G. Longo, On the error exponent for source coding and for testing simple statistical hypotheses. Studia Sc. Math. Hungarica **6**, 181–191 (1971)
17. F.W. Fu, S.Y. Shen, Hypothesis testing for arbitrarily varying source with exponents type constraint. IEEE Trans. Inf. Theory **44**(2), 892–895 (1998)
18. M. Gutman, Asymptotically optimal classification for multiple test with empirically observed statistics. IEEE Trans. Inf. Theory **35**(2), 401–408 (1989)
19. T.S. Han, S. Amari, Statistical inference under multiterminal data compression. IEEE Trans Inf. Theory **44**(6), 2300–2324 (1998)
20. T.S. Han, K. Kobayashi, Exponential-type error probabilities for multiterminal hypothesis testing. IEEE Trans. Inf. Theory **35**(1), 2–13 (1989)
21. E.A. Haroutunian, On asymptotically optimal testing of hypotheses concerning Markov chain (in Russian). Izvestia Acad. Nauk Armenian SSR. Seria Mathem. **22**(1), 76–80 (1988)
22. E.A. Haroutunian, Logarithmically asymptotically optimal testing of multiple statistical hypotheses. Probl. Control Inf. Theory **19**(5–6), 413–421 (1990)
23. W. Hoeffding, Asymptotically optimal tests for multinomial distributions. Ann. Math. Stat. **36**, 369–401 (1965)
24. G. Longo, A. Sgarro, The error exponent for the testing of simple statistical hypotheses, a combinatorial approach. J. Comb. Inf. Syst. Sci. **5**(1), 58–67 (1980)
25. U.M. Maurer, Authentication theory and hypothesis testing. IEEE Trans. Inf. Theory **46**(4), 1350–1356 (2000)

26. S. Natarajan, Large deviations, hypotheses testing, and source coding for finite Markov chains. IEEE Trans. Inf. Theory **31**(3), 360–365 (1985)
27. R.C. Rao, *Linear Statistical Inference and its Applications* (Wiley, New York, 1965)
28. G. Tusnady, On asymptotically optimal tests. Ann. Stat. **5**(2), 385–393 (1977)
29. Z. Zhang, T. Berger, Estimation via compressed information. IEEE Trans. Inf. Theory **34**(2), 198–211 (1988)

On Error Exponents in Quantum Hypothesis Testing

In the simple quantum hypothesis testing problem, upper bounds on the error probabilities are shown based on a key operator inequality between a density operator and its pinching. Concerning the error exponents, the upper bounds lead to a non-commutative analogue of the Hoeffding bound, which is identical with the classical counterpart if the hypotheses, composed of two density operators, are mutually commutative. The upper bounds also provide a simple proof of the direct part of the quantum Stein's lemma.

1 Introduction

Quantum hypothesis testing is a fundamental problem in quantum information theory, because it is one of the most simple problems where the difficulty derived from non-commutativity of operators appears. It is also closely related to other topics in quantum information theory, as in classical information theory. Actually, its relation with quantum channel coding is discussed in [7, 15].

Let us outline briefly significant results in classical hypothesis testing for probability distributions $p^n(\cdot)$ versus $q^n(\cdot)$, where $p^n(\cdot)$ and $q^n(\cdot)$ are i.i.d. extensions of some probability distributions $p(\cdot)$ and $q(\cdot)$ on a finite set $\mathcal{X}$. In the classical case, the asymptotic behaviors of the first kind error probability α_n and the second kind error probability β_n for the optimal test were studied thoroughly as follows.

First, when α_n satisfies the constant constraint $\alpha_n \leq \varepsilon$ ($\varepsilon > 0$), the error exponent of βn for the optimal test, say $\beta_n^*(\varepsilon)$, is written asymptotically as

$$\limsup_{n\to\infty} \frac{1}{n} \log \beta_n^* = -D(p||q) \tag{1}$$

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_25

for any ε, where $D(p||q)$ is the relative entropy. The equality (1) is called Stein's lemma (see e.g. [4, p.115]), and the quantum analogue of (1) was established recently [8, 14].

Next, when α_n satisfies the exponential constraint $\alpha_n \leq e^{-nr}$ $(r > 0)$, the error exponent of β_n for the optimal test is asymptotically determined by

$$\limsup_{n\to\infty} \frac{1}{n} \log \beta_n^{\dagger}(r) = - \min_{p': D(p'||p) \leq r} D(p'||q) \tag{2}$$

$$= - \max_{0<s\leq 1} \frac{\Psi(s) - (1-s)r}{s} \tag{3}$$

where the function $\Psi(s)$ is defined as

$$\Psi(s) \triangleq - \log \sum_{x\in\mathcal{X}} p(x)^{1-s} q(x)^s. \tag{4}$$

Historically speaking, (2) and the test achieving it were shown in [9], followed by another expression (3) (see [3]), which we call the Hoeffding bound here. In quantum hypothesis testing, the error exponent of $1 - \beta_n$ was studied in [14] to obtain a similar result to (3), which led to the strong converse property in quantum hypothesis testing. Concerning quantum fixed-length pure state source coding, the error exponent of erroneously decoded probability was determined in [5], where the optimality of the error exponent similar to (3) was discussed.

In this lecture (see [13]), a quantum analogue of the Hoeffding bound (3), (4) is introduced to derive a bound on the error exponent in quantum hypothesis testing. As a by-product of the process to derive the exponent, a simple proof of the quantum Stein's lemma is also given.

2 Definition and Main Results

Let $\mathcal{H}$ be a Hilbert space which represents a physical system in interest. We assume $\dim\mathcal{H} < \infty$ for mathematical simplicity. Let us denote the set of linear operators on $\mathcal{H}$ as $\mathcal{L}(\mathcal{H})$ and define the set of density operators on $\mathcal{H}$ by

$$\mathcal{S}(\mathcal{H}) \triangleq \{\rho \in \mathcal{L}(\mathcal{H}) : \rho = \rho^* \geq 0, \mathrm{Tr}[\rho] = 1\}. \tag{5}$$

We study the hypothesis testing problem for the null hypothesis

$$H_0 : \rho_n \triangleq \rho^{\otimes n} \in \mathcal{S}(\mathcal{H}^{\otimes n})$$

versus the alternative hypothesis

$$H_1 : \sigma_n \triangleq \sigma^{\otimes n} \in \mathcal{S}(\mathcal{H}^{\otimes n})$$

where $\rho^{\otimes n}$ and $\sigma^{\otimes n}$ are the nth tensor powers of arbitrarily given density operators ρ and σ in $\mathcal{S}(\mathcal{H})$.

The problem is to decide which hypothesis is true based on the data drawn from a quantum measurement, which is described by a positive operator valued measure (POVM) on $\mathcal{H}^{\otimes n}$, i.e., a resolution of identity $\sum_i M_{n,i} = I_n$ by non-negative operators $M_n = \{M_{n,i}\}$ on $\mathcal{H}^{\otimes n}$. If a POVM consists of projections on $\mathcal{H}^{\otimes n}$, it is called a projection valued measure (PVM). In the hypothesis testing problem, however, it is sufficient to treat a two-valued POVM $\{M_0, M_1\}$, where the subscripts 0 and 1 indicate the acceptance of H_0 and H_1, respectively. Thus, an operator $A_n \in \mathcal{L}(\mathcal{H}^{\otimes n})$ satisfying inequalities $0 \leq A_n \leq I_n$ is called a test in the sequel, since A_n is identified with the POVM $\{A_n, I_n - A_n\}$. For a test A_n, the error probabilities of the first kind and the second kind are, respectively, defined by

$$\alpha_n(A_n) \triangleq \mathrm{Tr}[\rho_n(I_n - A_n)]$$

$$\beta_n(A_n) \triangleq \mathrm{Tr}[\sigma_n A_n].$$

Let us define the optimal value for $\beta_n(A_n)$ under the constant constraint on $\alpha_n(A_n)$

$$\beta_n^*(\varepsilon) \triangleq \min\{\beta_n(A_n) : \ A_n : \text{ test }, \alpha_n(A_n) \leq \varepsilon\} \tag{6}$$

and let

$$D(\rho||\sigma) \triangleq \mathrm{Tr}[\rho(\log\rho - \log\sigma)] \tag{7}$$

which is called the quantum relative entropy. Then we have the following theorem, which is one of the most essential theorems in quantum information theory.

Proposition 277 (The Quantum Stein's Lemma) *For all* $0 < \varepsilon < 1$*, it holds that*

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n^*(\varepsilon) = -D(\rho||\sigma). \tag{8}$$

The first proof of (8) was composed of two inequalities, the direct part and the converse part. The direct part, concerned with existence of good tests, claims that

$$\forall\, 0 < \varepsilon \leq 1, \qquad \limsup_{n\to\infty} \frac{1}{n} \log \beta_n^*(\varepsilon) \leq -D(\rho||\sigma) \tag{9}$$

and it was given by Hiai and Petz [8]. In this lecture, the main focus is on the direct part. Note that the direct part (9) is equivalent to the existence of a sequence of tests

$\{A_n\}$ such that

$$\lim_{n\to\infty} \alpha_n(A_n) = 0 \quad \text{and} \quad \limsup_{n\to\infty} \frac{1}{n} \log \beta_n(A_n) \le -D(\rho||\sigma) \tag{10}$$

(see [14]). On the other hand, the converse part, concerned with nonexistence of too good tests, asserts that

$$\forall\, 0 < \varepsilon < 1, \qquad \liminf_{n\to\infty} \frac{1}{n} \log \beta_n^*(\varepsilon) \ge -D(\rho||\sigma) \tag{11}$$

which was given by Ogawa and Nagaoka [14]. A direct proof of the equality (8) was also given by Hayashi [6] using the information spectrum approach in quantum setting [10, 12], and a considerably simple proof of the converse part (11) was given in [11].

In this lecture, the asymptotic behavior of the error exponent $\frac{1}{n} \log \beta_n(A_n)$ under the exponential constraint

$$\alpha_n(A_n) \le e^{-nr}, \qquad r > 0$$

is studied, and a non-commutative analogue of the Hoeffding bound [9] similar to (3) is given as follows.

Theorem 278 (Ogawa and Hayashi 2004, [13]) *For all $r > 0$, there exists a sequence of tests $\{A_n\}$ which satisfies*

$$\limsup_{n\to\infty} \frac{1}{n} \log \alpha_n(A_n) \le -r, \tag{12}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \beta_n(A_n) \le -\max_{0<s\le 1} \frac{\overline{\psi}(s) - (1-s)r}{s} \tag{13}$$

where

$$\overline{\psi}(s) \triangleq -\log \mathrm{Tr}\left[\rho\sigma^{\frac{s}{2}}\rho^{-s}\sigma^{\frac{s}{2}}\right]. \tag{14}$$

We will prove the theorem in 4. If ρ and σ commutate, $\overline{\psi}(s)$ is identical with the classical counterpart $\Psi(s)$ defined in (4), and (13) coincides with the Hoeffding bound (3), which is optimal in classical hypothesis testing.

This lecture is organized as follows. In 3, upper bounds on the error probabilities are shown based on a key operator inequality [6]. Using the upper bounds, we will prove Theorem 278 in 4. In 5, we will make some remarks toward further investigations.

Section 7 is devoted to the definition of pinching (see, e.g., [2], p. 50), which is known as a special notion of the conditional expectation in literature on the operator

algebra and is used effectively in 3. In 8, the key operator inequality used in 3 is summarized along with another proof of it for readers' convenience.

3 Bounds on Error Probabilities

In the sequel, let $\mathcal{E}_{\sigma_n}(\rho_n)$ be the conditional expectation of ρ_n to the commutant of the $*$-subalgebra generated by σ_n, which we call pinching (see 7) and denote it as $\overline{\rho_n}$ for simplicity. Let $v(\sigma_n)$ be the number of eigenvalues of σ_n mutually different from others as defined in 7. Then a key operator inequality[1] follows from Lemma 285 in 8, which originally appeared in [6]

$$\rho_n \leq v(\sigma_n)\overline{\rho}n. \tag{15}$$

Note that the type counting argument provides

$$v(\sigma_n) \leq (n+1)^d \tag{16}$$

where $d \triangleq \dim \mathcal{H}$. Following [6], let us apply the operator monotonicity of the function $x \mapsto -x^{-s}$, $0 \leq s \leq 1$ (see, e.g, [2, Sec. V.1]) to (15) so that we have

$$\overline{\rho_n}^{-s} \leq v(\sigma_n)^s \rho_n^{-s} \leq (n+1)^{sd} \rho_n^{-s}. \tag{17}$$

Following the notation used in [10, 12], let us define the projection $\{X > 0\}$ for a Hermitian operator $X = \sum_i x_i E_i$ as

$$\{X > 0\} \triangleq \sum_{i:x_i>0} E_i \tag{18}$$

where E_i is the projection onto the eigenspace corresponding to an eigenvalue x_i. In the sequel, we will focus on a test defined by

$$\overline{S}_n(a) \triangleq \{\overline{\rho_n} - e^{na}\sigma_n > 0\} \tag{19}$$

where a is a real parameter, and derive the upper bounds on the error probabilities for the test $\overline{S}_n(a)$ as follows.

Theorem 279 (Ogawa and Hayashi 2004, [13])

$$\alpha_n\left(\overline{S}_n(a)\right) \leq (n+1)^d e^{-n\overline{\varphi}(a)}, \tag{20}$$

$$\beta_n\left(\overline{S}_n(a)\right) \leq (n+1)^d e^{-n[\overline{\varphi}(a)+a]} \tag{21}$$

[1] Although the way to derive the operator inequality and the definition of $v(\sigma_n)$ are different from those of [6], it results in the same one as [6] in the case that both of ρ_n and σ_n are tensored states.

where $\overline{\varphi}(a)$ is defined by $\overline{\psi}(s)$ given in (14) as

$$\overline{\varphi}(a) \stackrel{def}{=} \max_{0\le s\le 1} \left\{\overline{\psi}(s) - as\right\}. \tag{22}$$

Proof The definition of $\overline{S}_n(a)$ and commutativity of operators $\overline{\rho_n}$ and σ_n lead to

$$\left(\overline{\rho_n}^{1-s} - e^{na(1-s)}\sigma_n^{1-s}\right)\overline{S}_n(a) \ge 0 \tag{23}$$

$$\left(\overline{\rho_n} - e^{nas}\sigma_n^s\right)\left(I_n - \overline{S}_n(a)\right) \le 0 \tag{24}$$

for all $0 \le s \le 1$. Note that $\overline{S}_n(a)$ also commutes with σ_n. Therefore, the inequality (24), with the property of pinching (63) in 7, provides

$$\begin{aligned}
\alpha_n(\overline{S}_n(a)) &= \mathrm{Tr}[\rho_n(I_n - \overline{S}_n(a))] \\
&= \mathrm{Tr}[\overline{\rho_n}(I_n - \overline{S}_n(a))] \\
&= \mathrm{Tr}[\overline{\rho_n}^{1-s}\overline{\rho_n}^{s}(I_n - \overline{S}_n(a))] \\
&\le e^{nas}\,\mathrm{Tr}[\overline{\rho_n}^{1-s}\sigma_n^s(I_n - \overline{S}_n(a))] \\
&\le e^{nas}\,\mathrm{Tr}[\overline{\rho_n}^{1-s}\sigma_n^s].
\end{aligned} \tag{25}$$

In the same way, (23) yields

$$\begin{aligned}
\beta_n(\overline{S}_n(a)) &= \mathrm{Tr}[\sigma_n\overline{S}_n(a)] \\
&= \mathrm{Tr}[\sigma_n^s\sigma_n^1\overline{S}_n(a)] \\
&\le e^{-na(1-s)}\,\mathrm{Tr}[\sigma_n^s\overline{\rho_n}^{1-s}\overline{S}_n(a)] \\
&\le e^{-na}e^{nas}\,\mathrm{Tr}[\overline{\rho_n}^{1-s}\sigma_n^s.
\end{aligned} \tag{26}$$

It follows from (63) and (17) that

$$\begin{aligned}
\mathrm{Tr}[\overline{\rho_n}^{1-s}\sigma_n^s] &= \mathrm{Tr}\left[\overline{\rho_n}\sigma_n^{\frac{s}{2}}\overline{\rho_n}^{-s}\sigma_n^{\frac{s}{2}}\right] \\
&= \mathrm{Tr}\left[\rho_n\sigma_n^{\frac{s}{2}}\overline{\rho_n}^{-s}\sigma_n^{\frac{s}{2}}\right] \\
&\le (n+1)^{sd}\,\mathrm{Tr}\left[\rho_n\sigma_n^{\frac{s}{2}}\rho_n^{-s}\sigma_n^{\frac{s}{2}}\right]
\end{aligned}$$

$$= (n+1)^{sd} \left(\mathrm{Tr} \left[\rho \sigma^{\frac{s}{2}} \rho^{-s} \sigma^{\frac{s}{2}} \right] \right)^n$$

$$= (n+1)^{sd} e^{-n\overline{\psi}(s)} \tag{27}$$

for all $0 \le s \le 1$. Combining (25)–(27), we have

$$\alpha_n \left(\overline{S}_n(a) \right) \le (n+1)^{sd} e^{-n\left[\overline{\psi}(s) - as\right]}$$

$$\le (n+1)^{d} e^{-n\left[\overline{\psi}(s) - as\right]}, \tag{28}$$

$$\beta_n \left(\overline{S}_n(a) \right) \le (n+1)^{sd} e^{-n\left[\overline{\psi}(s) - as + a\right]}$$

$$\le (n+1)^{d} e^{-n\left[\overline{\psi}(s) - as + a\right]}, \tag{29}$$

which lead to (20) and (21) by taking the maximum in the exponents. □

4 Proof of Theorem 278 and the Quantum Stein's Lemma

In this section, we will prove Theorem 278 by using Theorem 279. To this end, the behavior of $\overline{\varphi}(a)$ in the error exponents (20) and (21) is investigated in the following lemmas. We will also show that Theorem 279 provides a simple proof of the direct part of the quantum Stein's lemma (10).

Lemma 280 *$\overline{\varphi}(a)$ is convex and monotonically nonincreasing.*

Proof The assertion immediately follows from the definition of $\overline{\varphi}(a)$. Actually, we have for all $0 \le t \le 1$

$$\overline{\varphi}(ta + (1-t)b) = \max_{0 \le s \le 1} \{\overline{\psi}(s) - (ta + (1-t)b)s\}$$

$$\le t \max_{0 \le s \le 1} \{\overline{\psi}(s) - as\} + (1-t) \max_{0 \le s \le 1} \{\overline{\psi}(s) - bs\}$$

$$= t\overline{\varphi}(a) + (1-t)\overline{\varphi}(b). \tag{30}$$

Next, let $a \le b$ and $s_b \triangleq \arg\max_{0 \le s \le 1} \{\overline{\psi}(s) - bs\}$. Then we have

$$\overline{\varphi}(b) = \overline{\psi}(s_b) - bs_b$$

$$\le \overline{\psi}(s_b) - as_b$$

$$\leq \max_{0\leq s\leq 1}\{\overline{\psi}(s) - as\}$$

$$= \overline{\varphi}(a). \tag{31}$$

□

Lemma 281 *$\overline{\varphi}(a)$ ranges from 0 to infinity.*

Proof Since we can calculate the derivative of $\overline{\psi}(s)$ explicitly, $\overline{\psi}(s)$ is continuous and differentiable. Therefore, it follows from the mean value theorem that for $s > 0$ there exists $0 \leq t \leq s$ such that

$$\overline{\psi}(t) = \frac{\overline{\psi}(s) - \overline{\psi}(0)}{s - 0}. \tag{32}$$

Let $a \leq \max_{0\leq t\leq 1} \overline{\psi}'(t)$, then we have

$$a \geq \frac{\overline{\psi}(s) - \overline{\psi}(0)}{s - 0}. \tag{33}$$

and hence,

$$\overline{\psi}(0) \geq \overline{\psi}(s) - as \tag{34}$$

which yields

$$0 = \overline{\psi}(0) = \max_{0\leq s\leq 1}\{\overline{\psi}(s) - as\} = \overline{\varphi}(a). \tag{35}$$

On the other hand, it is obvious that

$$\lim_{a\to-\infty} \overline{\varphi}(a) = \infty. \tag{36}$$

Since $\overline{\varphi}(a)$ is continuous, which follows from convexity by Lemma 280, the assertion follows from (35) and (36). □

Combined with the above lemma, Theorem 279 leads to Theorem 278 as follows.

Proof of Theorem 278 For all $r > 0$, there exists $a_r \in \mathbb{R}$ such that $r = \overline{\varphi}(a_r)$ from Lemma 281. Let $\overline{u}(r) \triangleq \overline{\varphi}(a_r) + a_r$, then it follows from Theorem 279 that

$$\limsup_{n\to\infty} \frac{1}{n} \log \alpha_n(\overline{S}n(a_r)) \leq -r \tag{37}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \beta_n \overline{S}_n(a_r)) \leq -\overline{u}(r). \tag{38}$$

Therefore, it suffices to show that

$$\overline{u}(r) = \max_{0 \le s \le 1} \frac{\overline{\psi} - (1-s)r}{s} \tag{39}$$

For all $0 \le s \le 1$, we have from the definition of $\overline{\varphi}(a)$

$$r = \overline{\varphi}(a_r) \ge \overline{\psi}(s) - a_r s \tag{40}$$

and there exists a number s_0, $0 < s_0 \le 1$, achieving the equality since $r = \overline{\varphi}(a_r) > 0$. On the other hand, the definitions of $\overline{u}(r)$ and a_r lead to

$$\overline{u}(r) = \overline{\varphi}(a_r) + a_r = r + a_r. \tag{41}$$

Eliminating a_r from (40) and (41), we have

$$\overline{u}(r) \ge \frac{\overline{\psi}(s) - (1-s)r}{s}) \tag{42}$$

and s_0 achieves the equality in (42) as well. Thus, we have shown (39), and Theorem 278 has been proved. □

Next, observing that $\overline{\psi}(0) = 0$ and $\overline{\psi}'(0) = D(\rho||\sigma)$, we have

$$\overline{\varphi}(a) > 0 \qquad \text{for all } a < D(\rho||\sigma) \tag{43}$$

which leads to the following theorem combined with Theorem 279.

Theorem 282 (Ogawa and Hayashi 2004, [13]) *For all $a < D(\rho||\sigma)$, we have*

$$\lim_{n\to\infty} \alpha_n(\overline{S}_n(a)) = 0 \tag{44}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \beta_n(\overline{S}_n(a)) \le -a. \tag{45}$$

Since $a < D(\rho||\sigma)$ can be arbitrarily near $D(\rho||\sigma)$, we have shown the direct part of the quantum Stein's lemma (10).

5 Toward Further Investigations

The error exponents derived here do not seem to be natural, since $\overline{\psi}(s)$ lacks symmetry between ρ and σ that the original hypothesis testing problem has. We need further investigation to determine the error exponents in quantum hypothesis testing. In this section, we make a few remarks on some candidates for the

alternative to $\overline{\psi}(s)$ in the expectation that the error exponents would be written in the form of Theorem 278.

Among many candidates, let us consider the following functions:

$$\psi_1(s) \triangleq \max\left\{\overline{\psi}(s), \tilde{\psi}(s)\right\} \tag{46}$$

$$\psi_2(s) \triangleq -\log \mathrm{Tr}\left[\rho^{1-s}\sigma^s\right] \tag{47}$$

$$\psi_3(s) \triangleq -\log \mathrm{Tr}\left[e^{(1-s)\log\rho + s\log\sigma}\right] \tag{48}$$

where

$$\tilde{\psi}(s) \triangleq -\log \mathrm{Tr}\left[\sigma\rho^{\frac{1-s}{2}}\sigma^{-(1-s)}\rho^{1-s}2\right] \tag{49}$$

and define the corresponding functions

$$u_i(r) \triangleq \max_{0<2\le 1} \frac{\psi_i(s) - (1-s)r}{s} \qquad i = 1, 2, 3. \tag{50}$$

The reason to consider these functions is as follows. First $\psi_1(s)$ is a symmetrized version of $\overline{\psi}(s)$, and Theorem 278 still holds with $\overline{\psi}(s)$ replaced by $\psi_1(s)$, since similar upper bounds to Theorem 279 using $\tilde{\psi}(s)$ are valid by exchanging ρ and σ and replacing s with $1-s$. On the other hand, $\psi_2(s)$ for $-1 \le s \le 0$ appeared in [14] to show the strong converse property in quantum hypothesis testing. Concerning $\psi_3(s)$, $u_3(r)$ is a quantum analogue of (2). Actually, we can show that

$$u_3(r) = \min_{\rho' : D(\rho'||\rho) \le r} D(\rho'||\rho) \tag{51}$$

by the same way as [14, Sec. VI]. At present it is not clear whether $u_2(r)$ and $u_3(r)$ are achievable exponents in quantum hypothesis testing. It should be noted, however, that $\psi_i(s)$, $i = 1, 2, 3$, are reduced to the classical one (4) if ρ and σ commute, and they have desirable properties

$$\begin{aligned} &\psi_i(0) = \psi_i(1) = 0 \\ &\psi_i'(0) = D(\rho||\sigma), \\ &\psi_i'(1) = D(\rho||\sigma) \qquad i = 1, 2, 3 \end{aligned} \tag{52}$$

which are consistent with the quantum Stein's lemma. The above properties of $\psi_2(s)$ and $\psi_3(s)$ are verified by the direct calculations while those of $\psi_1(s)$ follow from

the following fact:

$$\psi_1(s) = \overline{\psi}(s) \geq \tilde{\psi}(s), \qquad \text{if } s \text{ is sufficiently near 0} \tag{53}$$

$$\psi_1(s) = \tilde{\psi}(s) \geq \overline{\psi}(s), \qquad \text{if } s \text{ is sufficiently near 1} \tag{54}$$

which is a consequence of $\overline{\psi}(0) = \psi_2(0)$, $\tilde{\psi}(1) = \psi_2(1)$, and the following lemma.

Lemma 283 *For all* $0 \leq s \leq 1$*, we have*

$$\overline{\psi}(s) \leq \psi_2(s) \tag{55}$$

$$\tilde{\psi}(s) \leq \psi_2(s) \tag{56}$$

Proof Let us apply the monotonicity property of the quantum quasi-entropy [17, 18] to $\mathrm{Tr}[\rho^{1-s}\sigma^s]$, $0 \leq s \leq 1$,[2] so that we have

$$\begin{aligned} e^{-n\psi_2(s)} &= \left(\mathrm{Tr}[\rho^{1-s}\sigma^s]\right)^n \\ &= \mathrm{Tr}[\rho_n^{1-s}\sigma_n^s] \\ &\leq \mathrm{Tr}[\overline{\rho_n}^{1-s}\sigma_n^s] \\ &\leq (n+1)^{sd} e^{-n\overline{\psi}(s)} \end{aligned} \tag{57}$$

where we used (27) in the last inequality. Thus, we obtain

$$\overline{\psi}(s) \leq \psi_2(s) + \frac{sd}{n}\log(n+1) \tag{58}$$

for any natural number n, and we have (55) by letting n go to infinity. Exchanging ρ and σ and replacing s with $1-s$ in (55), we obtain (56). □

It follows immediately from Lemma 283 that $\psi_1(s) \leq \psi_2(s)$, and it was pointed out in [14] that we have $\psi_2(s) \leq \psi_3(s)$ as a consequence of the Golden-Thompson inequality (see, e.g., [16, p. 128])

$$\mathrm{Tr}\left[e^{A+B}\right] \leq \mathrm{Tr}\left[e^A e^B\right] \tag{59}$$

[2] Comprehensible explanations of the monotonicity property are found in [1, Sec. 7.2] and [14].

for Hermitian operators A and B with the equality if and only if A and B commute. These facts are stated as the following proposition

Proposition 284 *It holds that*

$$\psi_1(s) \leq \psi_2(s) \leq \psi_3(s) \qquad \forall\, 0 \leq s \leq 1 \tag{60}$$

$$u_1(r) \leq u_2(r) \leq u_3(r) \qquad \forall\, r > 0 \tag{61}$$

Especially, if ρ and σ do not commute, we have $\psi_2(s) < \psi_2(s)$ and $u_2(r) < u_3(r)$.

As mentioned above, $u_1(r)$ is an achievable exponent in quantum hypothesis testing, while it is not known whether $u_2(r)$ and $u_3(r)$ are achievable or not. It is interesting to study the achievability of these functions, especially that of $u_2(r)$, and the problem is left open.

6 Concluding Remarks

In the quantum hypothesis problem, we have presented upper bounds on the error probabilities of the first and the second kind, based on a key operator inequality satisfied by a density operator and pinching of it. The upper bounds are regarded as a noncommutative analogue of the Hoeffding bound [9], which is the optimal bound in classical hypothesis testing, and the upper bounds provide a simple proof of the direct part of the quantum Stein's lemma. Compared with [6], the proof is considerably simple and leads to the exponential convergence of the error probability of the first kind.

7 Definition of Pinching

In this section, we summarize the definition of pinching (see, e.g., [2, p. 50]) for readers' convenience. Pinching is known as a special notion of the conditional expectation in the field of operator algebra.

Given a Hermitian operator $A \in \mathcal{L}(\mathcal{H})$, let $A = \sum_{i=1}^{v(A)} a_i E_i$ be its spectral decomposition, where $v(A)$ is the number of eigenvalues of A mutually different from others, and each E_i is the projection corresponding to an eigenvalue a_i. The following map defined by using the PVM $E = \{Ei\}_{i=1}^{v(A)}$ is called pinching:

$$\mathcal{E}_A : B \in \mathcal{L}(\mathcal{H}) \to \mathcal{E}_A(B) = \triangleq \sum_{i=1}^{v(A)} E_i B E_i \in \mathcal{L}(\mathcal{H}). \tag{62}$$

The operator $\mathcal{E}_A(B)$ is also called pinching when no confusion is likely to arise, and it is sometimes denoted as $\mathcal{E}_E(B)$. It should be noted here that pinching is the conditional expectation (with respect to the tracial state) to the commutant of the $*$-subalgebra generated by A or PVM E, since $\mathcal{E}_A(B)$ is the one and only operator which satisfies

$$\mathrm{Tr}[BC] = \mathrm{Tr}[\mathcal{E}_A(B)C] \tag{63}$$

for any operator $C \in \mathcal{L}(\mathcal{H})$ commuting with A.

8 Key Operator Inequality

The following lemma has played an important role in this lecture. Although the lemma for a two-valued PVM has been widely used, it appeared in [6] for the general case. Here, we will show another proof of it for readers' convenience.

Lemma 285 (Hayashi 2002, [6]) *Given a PVM $M = \{M_i\}_{i=1}^{v(M)}$ on $\mathcal{H}$, we have for all $\rho \in \mathcal{S}(\mathcal{H})$*

$$\rho \le v(M)\mathcal{E}_M(\rho) \tag{64}$$

where $\mathcal{E}_M(\rho)$ is the pinching defined in 7.

Proof First, note that the following map, defined with respect to a non-negative operator $A \in \mathcal{L}(\mathcal{H})$, is operator convex

$$f_A : X \in \mathcal{L}(\mathcal{H}) \to X^*AX \in \mathcal{L}(\mathcal{H}) \tag{65}$$

which is shown by a direct calculation

$$tf_A(X)+(1-t)f_A(Y)-f_A(tX+(1-t)Y) = t(1-t)(X-Y)^*A(X-Y) \ge 0 \tag{66}$$

for $0 \le t \le 1$. Using the convexity, the lemma is verified as follows:

$$\begin{aligned}\frac{1}{v(M)^2}\rho &= \left(\frac{1}{v(M)}\sum_{i=1}^{v(M)} M_i\right)\rho\left(\frac{1}{v(M)}\sum_{i=1}^{v(M)} M_i\right)\\ &\le \frac{1}{v(M)}\sum_{i=1}^{v(M)} M_i\rho M_i\\ &= \frac{1}{v(M)}\mathcal{E}_M(\rho).\end{aligned} \tag{67}$$

□

References

1. S. Amari, H. Nagaoka, *Methods of Information Geometry* (AMS/Oxford University, Oxford, 1993)
2. R. Bhatia, *Matrix Analysis* (Springer, New York, 1997)
3. R.E. Blahut, Hypothesis testing and information theory. IEEE Trans. Inf. Theory **IT-20**, 405–417 (1974)
4. R.E. Blahut, *Principles and Practice of Information Theory* (Addison-Wesley, Reading, 1991)
5. M. Hayashi, Exponents of quantums fixed-length pure state source coding. Phys. Rev. A **66**(3), 032321 (2002)
6. M. Hayashi, Optimal sequence of POVM's in the sense of Stein's lemma in quantum hypothesis testing. J. Phys. A Math. Gen. **35**, 10759–10773 (2002)
7. M. Hayashi, H. Nagaoka, General formulas for capacity of classical-quantum channels. IEEE Trans. Inf. Theory **49**(7), 1753–1768 (2003)
8. F. Hiai, D. Petz, The proper formula for relative entropy and its asymptotics in quantum probability. Commun. Math. Phys. **143**, 99–114 (1991)
9. W. Hoeffding, On probabilities of large deviations, in *Proceedings of the 5th Berkeley Symposium Mathematical Statistics and Probability, Berkeley, CA* (1965), pp. 203–219
10. H. Nagaoka, On asymptotic theory of quantum hypothesis testing, in *Proceedings of the Symposium Statistical Inference Theory and its Information Theoretical Aspect* (1998), pp. 49–52
11. H. Nagaoka, Strong converse theorems in quantum information theory, in *Proceedings of the ERATO Workshop in Quantum Information Science* (2001)
12. H. Nagaoka, M. Hayashi, *An Information-Spectrum Approach to Classical and Quantum Hypothesis Testing for Simple Hypotheses* (2002)
13. T. Ogawa, M. Hayashi, On error exponents in quantum hypothesis testing. IEEE Trans. Inf. Theory **50**(6), 1368–1372 (2004)
14. T. Ogawa, H. Nagaoka, Strong converse and Stein's lemma in quantum hypothesis testing. IEEE Trans. Inf. Theory **46**, 2428–2433 (2000)
15. T. Ogawa, H. Nagaoka, A new proof of the channel coding theorem via hypothesis testing in quantum information theory, in *Proceedings of the 2002 IEEE International Symposium Information Theory, Lausanne, Switzerland* (2002)
16. M. Ohya, D. Petz, *Quantum Entropy and its Use, Berlin/Heidelberg* (Springer, Germany, 1993)
17. D. Petz, *Quasi-entropies for States of a von Neumann Algebra* (RIMS, Kyoto University, Kyoto, 1985), pp. 787–800
18. D. Petz, Quasi-entropies for finite quantum systems. Rep. Math. Phys. **23**, 57–65 (1986)

Part V
Identification and Statistics

Identification via Compressed Data

We introduce and analyze a new coding problem for a correlated source $(X^n, Y^n)_{n=1}^{\infty}$. The observer of X^n can transmit data depending on X^n at a prescribed rate R. Based on these data the observer of Y^n tries to identify whether for some distortion measure ρ (like the Hamming distance) $\frac{1}{n}\rho(X^n, Y^n) \leq d$, a prescribed fidelity criterion. We investigate as functions of R and d the exponents of two error probabilities, the probabilities for misacceptance and the probabilities for misrejection. Our analysis has led to a new method for proving converses. Its basis is "The Inherently Typical Subset Lemma". It goes considerably beyond the "Entropy Characterisation of [2], the "Image Size Characterisation of [3], and its extensions in [7]. It is conceivable that it has a strong impact on Multi-user Information Theory.

1 Introduction and Formulation of the Problem

Introduction

In this lecture (see [4]), we consider a new model: identification via compressed data. To put it in perspective, let us first review the traditional problems in source coding theory. Consider the following diagram, where $\{X_n\}_{n=1}^{\infty}$ is an i.i.d. source taking values in a finite alphabet $\mathcal{X}$. The encoder output is a binary sequence which appears at a rate of R bits per symbol. The decoder output is a sequence $\{\hat{X}_n\}_{n=1}^{\infty}$ which takes values in a finite reproduction alphabet $\mathcal{Y}$. In traditional source coding theory, the decoder is required to recover $\{X_n\}_{n=1}^{\infty}$ either completely or with some allowable distortion. That is, the output sequence $\{\hat{X}_n\}_{n=1}^{\infty}$ of the decoder must

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_26

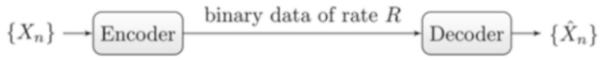


Fig. 1 Source coding

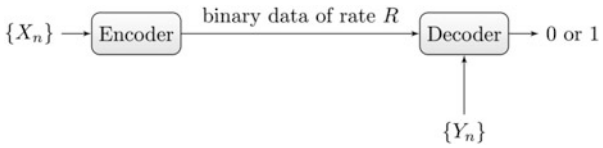


Fig. 2 Joint source coding and identification

satisfy

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\rho(X_i, \hat{X}_i) \le d \ , \tag{1}$$

for sufficiently large n, where $\mathbb{E}$ devotes the expected value,

$$\rho : \mathcal{X} \times \mathcal{Y} \to [0, +\infty) \tag{2}$$

is a distortion measure, and d is the allowable distortion between the source sequence and the reproduction sequence. The problem is then to determine the infimum of the rate R such that the system shown in Fig. 1 can operate in such a way that (1) is satisfied. It is known from rate distortion theory [5], that the infimum is given by the rate distortion function of the source $\{X_n\}_1^\infty$.

Let us now consider the system shown in Fig. 2. The sequence $\{Y_n\}_1^\infty$ is a sequence of i.i.d RV's taking values from $\mathcal{Y}$. Knowing Y^n, the decoder is now required to be able to identify whether or not the source sequence X^n and the sequence Y^n have some prescribed relation F in such a way that two kinds of error probabilities satisfy some prescribed conditions. In parallel with rate distortion theory, we consider in this lecture the following relation F defined by

$$n^{-1}\sum_{i=1}^{n}\rho(X_i, Y_i) \le d \ . \tag{3}$$

That is, the values X^n and Y^n are said to have relation F if (3) is satisfied. The problem we are interested in is to determine the infimum of the rate R such that the system shown in Fig. 2 can operate so that the error probability of misrejection, that is the decoder votes for 0 even though F holds and the error probability of misacceptance, that is the decoder votes for 1 even though F does not hold, satisfy certain constraints. So the goal of the decoder is to identify whether X^n is close to Y^n (in the sense of relation F) or not. The encoder is cooperative.

Formal Statement of Problem

First, we present some notation used throughout the lecture. Script capitals $\mathcal{X}$, $\mathcal{Y}$, $\cdots$, denote finite sets. The cardinality of a set $\mathcal{A}$ is denoted by $|\mathcal{A}|$. The letters P, Q, always stand for probability distributions on finite sets. X, Y, $\cdots$, denote RV's. The distributions of RV's X and Y are denoted by P_X and P_Y, respectively. $\mathcal{P}(\mathcal{X})$ stands for the set of all probability distributions on $\mathcal{X}$. The functions log and exp are understood to be to the base 2. If $\mathcal{A}$ is a finite set, then $\mathcal{A}^n$ denotes the set of all n-tuples $a^n = (a_1, \cdots, a_n)$ from $\mathcal{A}$. If $a = (a_i)$ is a finite or infinite sequence of letters from $\mathcal{A}$, let $a_m^n = (a_m, \cdots, a_n)$ and, for simplicity, write a_1^n as a^n. A similar convention also applies to RV's.

Let $\{(X_n, Y_n)\}_{n=1}^{\infty}$ be a sequence of independent drawings of a pair (X, Y) of RV's with joint distribution P_{XY} taking values in $\mathcal{X} \times \mathcal{Y}$. Let $\rho : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ be a distortion measure. Let $\{\rho_n : n = 1, 2, \cdots\}$ be a single-letter fidelity criterion generated by ρ, where

$$\rho_n : \mathcal{X}^n \times \mathcal{Y}^n \to [0, +\infty)$$

is a mapping defined by

$$\rho_n(x^n, y^n) = \frac{1}{n} \sum_{i=1}^{n} \rho(x_i, y_i),$$

for any $x^n \in \mathcal{X}^n$ and any $y^n \in \mathcal{Y}^n$. Without loss of generality, we assume that the distortion measure ρ satisfies

$$\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} \rho(x, y) = 0 \ . \tag{4}$$

Let $d \geq 0$ satisfy

$$d < \mathbb{E}\rho(X, Y) \ . \tag{5}$$

An nth order identification source (IDS) code $\mathcal{C}_n$ is defined as a triple $\mathcal{C}_n = (f_n, B_n, g_n)$ where $B_n \subset \{0, 1\}^*$ is a prefix-free set, f_n (called an encoder) is a mapping from $\mathcal{X}^n$ to B_n, and g_n (called a decoder) is a mapping from $\mathcal{Y}^n \times B_n$ to $\{0, 1\}$. Note that, in this definition, the encoder f_n can be of variable-length. The correspondence between an identification source code as defined here and the system shown in Figure 2 should be clear. When an identification source code $\mathcal{C}_n = (f_n, B_n, g_n)$ is used in the system shown in Figure 2, the performance can be measured by three quantities: the resulting average rate per symbol $r_n(\mathcal{C}_n)$, the first kind of error probability $P_{e1}(\mathcal{C}_n)$, and the second kind of error probability $P_{e2}(\mathcal{C}_n)$, where

$$r_n(\mathcal{C}_n) = \frac{1}{n}\mathbb{E}(\text{ the length of } f_n(X^n)) \ , \tag{6}$$

$$P_{e1}(\mathcal{C}_n) = \Pr\{g_n(Y^n, f_n(X^n)) = 0 | \rho_n(X^n, Y^n) \leq d\} \ , \tag{7}$$

and

$$P_{e2}(\mathcal{C}_n) = \Pr\{g_n(Y^n, f_n(X^n)) = 1 | \rho_n(X^n, Y^n) > d\} . \tag{8}$$

Clearly, $P_{e1}(\mathcal{C}_n)$ and $P_{e2}(\mathcal{C}_n)$ can be interpreted as the probability of misrejection and the probability of misacceptance (or false identification), respectively.

Let $R \in [0, +\infty)$, $\alpha \in (0, +\infty]$, and $\beta \in (0, +\infty]$. A triple (R, α, β) is said to be achievable with respect to a given d, if for any $\epsilon > 0$ there exists a sequence $\{\mathcal{C}_n\}_{n=1}^{\infty}$ nth order IDS codes $\mathcal{C}_n$, such that for sufficiently large n,

$$r_n(\mathcal{C}_n) \le R + \epsilon , \tag{9}$$

$$P_{e1}(\mathcal{C}_n) \le 2^{-n(\alpha-\epsilon)} , \tag{10}$$

and

$$P_{e2}(\mathcal{C}_n) \le 2^{-n(\beta-\epsilon)} , \tag{11}$$

where as a convention, $\alpha = +\infty (\beta = +\infty$, resp.) means that the probability of misrejection (false identification, resp.) is zero. Let $\mathcal{R}(d)$ be the set of all achievable triples. Let $\bar{\mathcal{R}}(d)$ denote the closure of $\mathcal{R}(d)$ with respect to the usual topology under which $a_n \to +\infty$ means that a_n is equal to ∞ for all but finitely many integers n. In this lecture, we are interested in determining the region $\bar{\mathcal{R}}(d)$. Specifically, we define for each pair $(\alpha, \beta) \in [0, +\infty]^2$,

$$R^*_{XY}(\alpha, \beta, d) = \inf\{R : (R, \alpha, \beta) \in \bar{\mathcal{R}}(d)\} . \tag{12}$$

Our main problem is the determination of this function.

Note that since $\bar{\mathcal{R}}(d)$ is closed, the infimum in (12) is actually a minimum. It is easy to see that $R^*_{XY}(\alpha, \beta, d) \ge R^*_{XY}(\alpha, 0, d)$ for any $\beta \ge 0$. On the other hand, since $\bar{\mathcal{R}}(d)$ is closed, it follows from (12) that

$$R^*_{XY}(\alpha, 0, d) = \lim_{\beta \to 0} R^*_{XY}(\alpha, \beta, d).$$

Therefore, $R^*_{XY}(\alpha, \beta, d)$ is continuous at $\beta = 0$. A similar result holds for $\alpha = 0$.

Discussion

In the last subsection we formulated the problem we are interested in as investigating the trade-off between the rate R and the error exponents α and β. A natural question to ask at this point is why the problem should be set up in this way. To answer this question, we first note that since $d < \mathbb{E}\rho(X, Y)$, it follows immediately that $\Pr(\rho_n(X^n, Y^n) \le d) \to 0$ as n goes to infinity. Therefore, if instead of the two

kinds of error probabilities, we use the error probability

$$P_e(\mathcal{C}_n) = \Pr(\rho_n(X^n, Y^n) \le d) P_{e1}(\mathcal{C}_n) + \Pr(\rho_n(X^n, Y^n) > d) P_{e2}(\mathcal{C}_n)$$

as a criterion, as studied by Ahlswede and Csizar in their 1-Bit Theorem [1], then the present problem becomes trivial and no information needs to be sent. This leads us to consider the two kinds of error probabilities. Second, let us see what happens if the two kinds of error probabilities are only required to vanish as n goes to infinity. The following theorem tells us that in this case the minimum achievable rate is always equal to zero.

Theorem 286 *For any distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$,*

$$R^0_{XY} = 0 \,, \tag{13}$$

where R^0_{XY} is the infimum of all positive real numbers R such that there exists for any $\epsilon > 0$ a sequence $\{\mathcal{C}_n\}$ of IDS codes, where $\mathcal{C}_n$ is an nth order IDS code, such that for sufficiently large n, $r_n(\mathcal{C}_n) \le R + \epsilon$, and

$$\lim_{n\to\infty} P_{e1}(\mathcal{C}_n) = 0 \quad \textit{and} \quad \lim_{n\to\infty} P_{e2}(\mathcal{C}_n) = 0 \,.$$

Proof of Theorem 286 To prove $R^o_{XY} = 0$, we construct, for sufficiently large n, an nth order ID source code $\mathcal{C}_n = (f_n, B_n, g_n)$ as follows. For each $x^n \in \mathcal{X}^n$, the encoder sends x^k completely to the decoder. This needs $n^{-1}\lceil k \log |\mathcal{X}| \rceil$ bits per source symbol. Observing $y^n \in \mathcal{Y}^n$ and receiving x^k, the decoder outputs 1 if $\rho_k(x^k, y^k) \le d + \delta$ and 0 otherwise, where $\delta > 0$ is selected so that $d + \delta < \mathbb{E}\rho(X, Y)$. The probability of false identification is given by

$$P_{e2}(\mathcal{C}_n) = \Pr(\rho_k(X^k, Y^k) \le d + \delta | \rho_n(X^n, Y^n) > d).$$

Since $d < \mathbb{E}\rho(X, Y)$, it is easy to see that for sufficiently large n

$$P_{e2}(\mathcal{C}_n) \le 2 \Pr(\rho_k(X^k, Y^k) \le d + \delta).$$

Let $\mathcal{P}^d$ denote the set of all $Q \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ such that

$$\mathbb{E}_Q \rho(X_0, Y_0) \le d$$

Clearly, $\mathcal{P}^d$ is a convex and closed set. Let Q^* be the unique element of $\mathcal{P}^d$ such that

$$D(Q^* || P_{XY}) = \min_{Q \in \mathcal{P}^d} D(Q || P_{XY}).$$

For any $\epsilon > 0$, select k so large that

$$P_{e2}(\mathcal{C}_n) \leq \epsilon$$

and

$$\Pr(\rho_k(\hat{X}^k, \hat{Y}^k) > d + \delta) < \epsilon.,$$

where $(\hat{X}^k, \hat{Y}^k)$ is the sequence of k-times independent drawings of a pair of RV's $(\hat{X}, \hat{Y})$ taking values on $\mathcal{X} \times \mathcal{Y}$ with joint distribution of $P_{\hat{X}\hat{Y}} = Q^*$. Fix such a k. All remaining is to prove that for sufficiently large n, the probability of misidentification $P_{e1}(\mathcal{C}_n)$ will be less than ϵ. To see this is true, note that

$$P_{e1}(\mathcal{C}_n) = \Pr(\rho_k(X^k, Y^k) > d + \delta | \rho_n(X^n, Y^n) \leq d).$$

By virtue of the conditional limit theorem [6, 9], it is not hard to prove that

$$\lim_{n\to\infty} P_{e1}(\mathcal{C}_n) = \Pr(\rho_k(\hat{X}^k, \hat{Y}^k) > d + \delta) < \epsilon.$$

This completes the proof of Theorem 286. □

Therefore the only interesting problem left is to investigate the trade-off of the rate R and the two error exponents. Indeed, the results we obtained in this lecture show that the problem proposed in the last subsection is really very interesting and even led us to develop a new powerful method for proving converses in information theory.

2 Statement and Discussion of the Main Results

As before, let (X, Y) be a pair of RV's with probability distribution P_{XY} taking values in $\mathcal{X} \times \mathcal{Y}$. Let $d < \mathbb{E}\rho(X, Y)$ and define

$$\beta(d) = \inf_{\mu \in \mathcal{P}(d)} D(\mu || P_{XY}) \,, \tag{14}$$

where

$$\mathcal{P}(d) = \{\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mu(x, y)\rho(x, y) \leq d\}$$

and $D(.||.)$ stands for the relative entropy function. It is not hard to see that

$$\lim_{n\to\infty} -\frac{1}{n} \log \Pr\{\rho_n(X^n, Y^n) \leq d\} = \beta(d) \,. \tag{15}$$

We distinguish two cases: (i) X and Y are independent; (ii) X and Y are correlated. To build up ideas we begin with the easier case (i), in which we have conclusive results.

Independent Case
In this subsection, we assume that X and Y are independent, that is $P_{XY} = P_X \times P_Y$. Without loss of generality we further assume that $P_X(x) > 0$ and $P_Y(y) > 0$ for every $x \in \mathcal{X}$ and every $y \in \mathcal{Y}$.

Let U be a RV taking values in some finite set $\mathcal{U}$. Let P_{XU} be the joint distribution of X and U on $\mathcal{X} \times \mathcal{U}$. Define

$$\mathcal{E}(P_{XU}, d) = \inf[D(P_{\tilde{Y}}||P_Y) + I(U \wedge \tilde{Y})] \,, \tag{16}$$

where the infimum is taken over all RV's $\tilde{Y}$ taking values in $\mathcal{Y}$ and being jointly distributed with X, U such that $\mathbb{E}\rho(X, \tilde{Y}) \leq d$. By using the same argument as in the proof of Lemma 2.2. of [5, pp. 124], it is not hard to prove that $\mathcal{E}(P_{XU}, d)$ has the following property.

Lemma 287 *$\mathcal{E}(P_{XU}, d)$ is non-increasing and convex as a function of d and continuous as a function of the pair (P_{XU}, d) where P_{XU} ranges over the set $\mathcal{P}(\mathcal{X} \times \mathcal{U})$.*

For any $\beta > 0$, we next define,

$$R(P_X, P_Y, \beta, d) = \inf\{I(X \wedge U) : U \text{ is a finite valued RV with } \mathcal{E}(P_{XU}, d) \geq \beta\} \,. \tag{17}$$

Let

$$R(P_X, P_Y, 0, d) = \lim_{\beta \to 0^+} R(P_X, P_Y, \beta, d) \,. \tag{18}$$

Clearly

$$R(P_X, P_Y, 0, d) =$$

$$\inf\{I(X \wedge U) : U \text{ is a finite valued RV with } \mathcal{E}(P_{XU}, d) > 0\} \,. \tag{19}$$

Define

$$\bar{R}(P_X, P_Y, \beta, d) = \lim_{\beta' \to \beta^-} R(P_X, P_Y, \beta', d) \,. \tag{20}$$

This is well defined since $R(P_X, P_Y, \beta, d)$ as a function of β is non-decreasing.

The following theorem gives a formula for $R^*(+\infty, \beta, d)$.

Theorem 288 *Assume X and Y are independent. Then for any $0 < d < \mathbb{E}\rho(X, Y)$ and $0 \leq \beta \leq \beta(d)$,*

$$R^*_{XY}(+\infty, \beta, d) = \bar{R}(P_X, P_Y, \beta, d) \ . \tag{21}$$

Remark At this point, let us pause to give a few comments on the issue concerning the computation of $R(P_X, P_Y, \beta, d)$. In the following subsection, we shall compute $R(P_X, P_Y, \beta, d)$ in the binary symmetric case. It turns out that in this special case, $R(P_X, P_Y, \beta, d)$ can be expressed in terms of the rate-distortion function of the source X. In general, however, the computation of this function may be very difficult. It seems to the authors that there is no easy way to apply the support lemma [2, 7, Chapter 3] to upper bound the cardinality of the set $\mathcal{U}$. Instead, we shall take an alternative approach to the problem. We define for each integer $k \geq 1$, and any $\beta > 0$

$$R_k(P_X, P_Y, \beta, d) = \inf\{I(X \wedge U) \tag{22}$$

where U is a RV taking at most k values with $\mathcal{E}(P_{XU}, d) \geq \beta$.

For $\beta = 0$, $R_k(P_X, P_Y, 0, d)$ is defined similarly. Clearly, $R_k(P_X, P_Y, \beta, d)$ as a function of k is non-increasing and converges to $R(P_X, P_Y, \beta, d)$ as k goes to infinity. Later on, we shall estimate the rate at which $R_k(P_X, P_Y, \beta, d)$ converges to $R(P_X, P_Y, \beta, d)$ to provide a partial solution to the problem of the computation of $R(P_X, P_Y, \beta, d)$.

To give a general formula for the function $R^*_{XY}(\alpha, \beta, d)$, we next modify the definition of the quantities $\mathcal{E}(P_{XU}, d)$ and $R(P_X, P_Y, \beta, d)$ as follows. For any $\gamma \geq 0$ and any $\alpha \geq 0$, define

$$\mathcal{E}(P_{XU}, \alpha, \gamma, d) = \inf[D(P_{\tilde{Y}}||P_Y) + I(U \wedge \tilde{Y})] \ , \tag{23}$$

where the infimum is taken over all RV's $\tilde{Y}$ taking values in $\mathcal{Y}$ and being jointly distributed with X, U such that $\mathbb{E}\rho(X, \tilde{Y}) \leq d$ and

$$D(P_{\tilde{Y}}||P_Y) + I(XU \wedge \tilde{Y}) \leq \gamma + \alpha \ . \tag{24}$$

Here we make use of the convention that the infimum taken over an empty set is $+\infty$. Let

$$\beta(P_X, d) = \inf_{\mathbb{E}\rho(X,\tilde{Y}) \leq d} D(P_{\tilde{Y}}||P_Y) + I(X \wedge \tilde{Y}) \ , \tag{25}$$

where the infimum is taken over all RV's $\tilde{Y}$ taking values in $\mathcal{Y}$ such that $\mathbb{E}\rho(X, \tilde{Y}) \leq d$. Then it is easy to see that in case $\gamma + \alpha < \beta(P_X, d)$ the following holds

$$\mathcal{E}(P_{XU}, \alpha, \gamma, d) = +\infty \tag{26}$$

for any RV U. In analogy to Lemma 287, it is not hard to see that $\mathcal{E}(P_{XU}, \alpha, \gamma, d)$ has the following property.

Lemma 289 *$\mathcal{E}(P_{XU}, \alpha, \gamma, d)$ is non-increasing and convex as a function of α (γ or d resp.) and continuous as a function of the quadruple $(P_{XU}, \alpha, \gamma, d)$, where the quadruple $(P_{XU}, \alpha, \gamma, d)$ ranges over all quadruples satisfying $\gamma + \alpha > \beta(P_X, d)$, $\alpha > 0$ and $d > 0$.*

Similarly to (17) and (18), we define for any $\beta > 0$

$$R(P_X, P_Y, \alpha, \gamma, \beta, d) = \inf\{I(X \wedge U) \tag{27}$$

where U is a finite valued RV with $\mathcal{E}(P_{XU}, \alpha, \gamma, d) \geq \beta$.
Let

$$R(P_X, P_Y, \alpha, \gamma, 0, d) = \lim_{\beta \to 0^+} R(P_X, P_Y, \alpha, \gamma, \beta, d) \ . \tag{28}$$

Define

$$\bar{R}(P_X, P_Y, \alpha, \gamma, \beta, d) = \lim_{\beta' \to \beta^-} R(P_X, P_Y, \alpha, \gamma, \beta', d) \ . \tag{29}$$

The following theorem gives a general formula for $R^*_{XY}(\alpha, \beta, d)$.

Theorem 290 *Assume that X and Y are independent, then for any $0 < d < \mathbb{E}\rho(X, Y)$, $0 < \alpha \neq \beta(P_X, d) - \beta(d)$, and $0 \leq \beta \leq \beta(d)$, the following holds*

$$R^*_{XY}(\alpha, \beta, d) = \bar{R}(P_X, P_Y, \alpha, \beta(d), \beta, d) \ . \tag{30}$$

Remark Obviously, $\beta(d) \leq \beta(P_X, d)$. If $\beta(d) < \beta(P_X, d)$, then it follows from (26) and (27) that for any $\alpha < \beta(P_X, d) - \beta(d)$ and $\beta \geq 0$

$$R(P_X, P_Y, \alpha, \beta(d), \beta, d) = 0 \ . \tag{31}$$

On the other hand, it is easy to see that in this special case, $R^*_{XY}(\alpha, \beta, d) = 0$ for any $\beta \in [0, +\infty]$. (This will become clear when we prove the direct part of Theorem 290.) Furthermore, it follows from the definition of $R^*_{XY}(\alpha, \beta, d)$ that as a function of α it is left continuous. Thus it will suffice for us to prove Theorem 290 for $\alpha > \beta(P_X, d) - \beta(d)$.

Note that Theorem 288 is actually a special case of Theorem 290, because for $\alpha = +\infty$

$$R(P_X, P_Y, \alpha, \beta(d), \beta, d) = R(P_X, P_Y, \beta, d) \ . \tag{32}$$

The reasons why we state Theorems 288 and 290 separately can be seen in the following sections. Similarly to (22), we can also define $R_k(P_X, P_Y, \alpha, \gamma, \beta, d)$ for

each $k \geq 1$. We conclude this section with pointing out the following facts on $R_k(P_X, P_Y, \beta, d)$ and $R_k(P_X, P_Y, \alpha, \gamma, \beta, d)$.

1. $R_k(P_X, P_Y, \beta, d)$ as a function of the triple (P_X, β, d) is lower semi-continuous.
2. $R_k(P_X, P_Y, \beta, d)$ as a function of β is left continuous.
3. $R_k(P_X, P_Y, \alpha, \gamma, \beta, d)$ as a function of the quintuple $(P_X, \alpha, \gamma, \beta, d)$ is lower semi-continuous in the range $\gamma + \alpha > \beta(P_X, d), \alpha > 0$ and $d > 0$.
4. $R_k(P_X, P_Y, \alpha, \gamma, \beta, d)$ as a function of β is left continuous if $\gamma + \alpha > \beta(P_X, d)$.

Example: The Binary Symmetric Case

In this subsection, we consider the binary symmetric case where $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, X and Y are independent and uniformly distributed over $\{0, 1\}$, and the distortion measure ρ is the Hamming distance.

We first evaluate $R^*_{XY}(+\infty, \beta, d)$ from Theorem 288 and then show as an example how to prove Theorem 288 in this special case. Note that in this case, $\beta(d) = 1 - h(d)$ where $h(\cdot)$ is the binary entropy function.

Theorem 291 *For any* $0 \leq d < \frac{1}{2}$ *and* $0 \leq \beta \leq \beta(d)$,

$$R^*_{XY}(+\infty, \beta, d) = 1 - h(d_\beta - d), \tag{33}$$

where $d_\beta \leq \frac{1}{2}$ *satisfies* $h(d_\beta) = 1 - \beta$.

Proof It is easy to see that $1 - h(d_\beta - d)$ is a continuous function of β. In view of Theorem 288, it suffices to prove that for any $0 < \beta < \beta(d)$

$$R(P_X, P_Y, \beta, d) = 1 - h(d_\beta - d)\ . \tag{34}$$

Let U be a RV taking values uniformly in $\{0, 1\}$ and such that

$$I(X \wedge U) = 1 - h(d_\beta - d) \quad \textit{and} \quad \mathbb{E}\rho(X, U) \leq d_\beta - d\ . \tag{35}$$

Since X takes values uniformly in $\{0, 1\}$, such a RV exists [6]. It is easy to verify that

$$\begin{aligned}
\mathcal{E}(P_{XU}, d) &= \inf_{\mathbb{E}\rho(X,\tilde{Y}) \leq d} [D(P_{\tilde{Y}} || P_Y) + I(U \wedge \tilde{Y})] \\
&= \inf_{\mathbb{E}\rho(X,\tilde{Y}) \leq d} [1 - H(\tilde{Y}|U)] \\
&\geq \inf_{\mathbb{E}\rho(X,\tilde{Y}) \leq d} I(U \wedge \tilde{Y}) \\
&\geq \inf_{\mathbb{E}\rho(U,\tilde{Y}) \leq d_\beta} I(U \wedge \tilde{Y}) \\
&= 1 - h(d_\beta) \\
&= \beta
\end{aligned} \tag{36}$$

where the last inequality is due to (35). Thus it follows from (17) that

$$R(P_X, P_Y, \beta, d) \leq I(X \wedge U) = 1 - h(d_\beta - d) \ .$$

To prove the reverse inequality of (37), let U be any RV taking values in some finite set $\mathcal{U}$ such that $\mathcal{E}(P_{XU}, d) \geq \beta$. Since X takes values uniformly in $\{0, 1\}$, it suffices to prove that

$$H(X|U) \leq h(d_\beta - d) \ . \tag{37}$$

To this end, we solve the following optimization problem

$$\inf_{\mathbb{E}\rho(X,\tilde{Y}) \leq d} [1 - H(\tilde{Y}|U)] \ . \tag{38}$$

For each $u \in \mathcal{U}$, let x_u be an element of $\{0, 1\}$ such that $P_{X|U}(x_u|u) \leq \frac{1}{2}$, where $P_{X|U}(x_u|u)$ denotes the conditional probability of $X = x_u$ given $U = u$. Since

$$\mathcal{E}(P_{XU}, d) \geq \beta > 0 \ , \tag{39}$$

it is not hard to see that

$$\sum_{u \in \mathcal{U}} P_U(u)(\frac{1}{2} - P_{X|U}(x_u|u)) > d \ . \tag{40}$$

From (40), it follows that the optimization problem (38) is equivalent to the following optimization problem

$$\inf_{\mathbb{E}\rho(X,\tilde{Y}) = d} [1 - H(\tilde{Y}|U)] \ . \tag{41}$$

Since the objective function $1 - H(\tilde{Y}|U)$ depends only on $P_{\tilde{Y}|U}(x_u|u)$, it is not hard to see that the optimization problem (41) can be reformulated as maximizing

$$\sum_u P_U(u) h(P_{\tilde{Y}|U}(x_u|u)) \tag{42}$$

subject to

$$\sum_u P_U(u) |P_{X|U}(x_u|u)) - P_{\tilde{Y}|U}(x_u|u))| = d \tag{43}$$

and

$$0 \leq P_{\tilde{Y}|U}(x_u|u)) \leq 1 \ . \tag{44}$$

Since $P_{X|U}(x_u|u)) \leq \frac{1}{2}$, conditions (43) and (44) can be replaced by

$$\sum_u P_U(u)(P_{\tilde{Y}|U}(x_u|u)) - P_{X|U}(x_u|u))) = d \tag{45}$$

and

$$P_{X|U}(x_u|u)) \leq P_{\tilde{Y}|U}(x_u|u)) \leq 1 \ . \tag{46}$$

The standard Lagrange Multiplier Method can be used to show that the maximum of the optimization problem given by (42), (45), and (46) is achieved at the point $\{P_{\tilde{Y}|U}(x_u|u))\}$ for which there exists a $\lambda > 0$ such that

(i) $P_{\tilde{Y}|U}(x_u|u)) = \lambda$ for any $u \in \mathcal{U}$ satisfying $P_{X|U}(x_u|u)) \leq \lambda$;
(ii) $P_{\tilde{Y}|U}(x_u|u)) = P_{X|U}(x_u|u))$ for any $u \in \mathcal{U}$ satisfying $P_{X|U}(x_u|u)) > \lambda$;
(iii)

$$\sum_{u:P_{X|U}(x_u|u)\leq\lambda} P_U(u)[\lambda - P_{X|U}(x_u|u))] = d \ . \tag{47}$$

Clearly, this is something like water-filling. Since $\mathcal{E}(P_{XU}, d) \geq \beta$, it follows that

$$\sum_{u:P_{X|U}(x_u|u)\leq\lambda} P_U(u)h(\lambda) + \sum_{u:P_{X|U}(x_u|u)>\lambda} P_U(u)h(P_{X|U}(x_u|u)) \leq 1 - \beta \ . \tag{48}$$

We now claim that (37) holds. Otherwise, say,

$$H(X|U) = \sum_u P_U(u)h(P_{X|U}(x_u|u)) > h(d_\beta - d) \ . \tag{49}$$

Then, in view of (40), (47) and the fact that the derivative of the function $h(s)$ is strictly decreasing over the interval $s \in (0, \frac{1}{2}]$, we can see that

$$\sum_{u:P_{X|U}(x_u|u)\leq\lambda} P_U(u)h(\lambda) + \sum_{u:P_{X|U}(x_u|u)>\lambda} P_U(u)h(P_{X|U}(x_u|u)) > h(d_\beta)$$
$$= 1 - \beta \ . \tag{50}$$

This contradicts (48). From (37) and (34) follows immediately. This completes the proof of Theorem 291. □

Note that $1 - h(d_\beta - d)$ is the rate distortion function of the source X evaluated at the point $d_\beta - d$. In some sense, therefore, Theorem 291 shows that there exists a close relationship between the rate $R^*_{XY}(+\infty, \beta, d)$ and the rate distortion function of X.

Next we outline the proof of Theorem 288 in the binary symmetric case. The direct part is easy. For any $d' < d_\beta - d$, roughly speaking, $2^{n(1-h(d'))}$ balls of radius nd' can almost cover the whole space. For each $x^n \in \{0,1\}^n$, we send simply the center of the ball in which x^n lies. Upon receiving this center, the decoder first calculates the Hamming distance between y^n and the center, and then outputs 1 if the distance is $\leq n(d'+d)$ and 0 otherwise. It is not hard to see that the misrejection probability is guaranteed to be zero, and the misacceptance probability is upper bounded by $2^{-n(1-h(d'+d))}$, which is less than or equal to $2^{-n\beta}$. This implies $R^*_{XY}(+\infty,\beta,d) \leq 1 - h(d_\beta - d)$.

To prove the converse part, let $(R,+\infty,\beta)$ be achievable, where $\beta > 0$. By definition, there exists for any $\epsilon > 0$ and sufficiently large n an nth-order IDS code $\mathcal{C}_n = (f_n, B_n, g_n)$ such that

$$r_n(\mathcal{C}_n) \leq R + \epsilon \ , \quad P_{e1}(\mathcal{C}_n) = 0 \ , \quad \text{and} \quad P_{e2}(\mathcal{C}_n) \leq 2^{-n(\beta-\epsilon)} \ . \tag{51}$$

For any $b^n \in B_n$, let

$$S(b^n) = \{x^n \in \mathcal{X}^n | f_n(x^n) = b^n\}$$

and

$$S^d(b^n) = \{y^n \in \mathcal{Y}^n | \text{ there exist } x^n \in S(b^n) : \rho_n(x^n, y^n) \leq d \ \} \ .$$

From (51) and the Markov inequality, it follows that with very high probability $b^n \in B_n$ satisfies

$$\Pr\{Y^n \in S^d(b^n)\} \leq 2^{-n(\beta-2\epsilon)} \ .$$

To continue our derivation, we use at this point an isoperimetric theorem in combinatorial extremal theory [10] which says roughly that for any subset $A \subset \{0,1\}^n$ with $|A| = \sum_{i=0}^{k} \binom{n}{i}$ for some k, the cardinality of the Hamming l-neighbourhood $\Gamma^l A$ of A is minimized when A is a sphere, where for any $l \geq 0$

$$\Gamma^l A = \{y^n \in \{0,1\}^n | \text{ there exist } x^n \in A : n\rho_n(x^n, y^n) \leq l\} \ .$$

Using this result, one can prove that with very high probability $b^n \in B_n$ satisfies $|S(b^n)| \leq 2^{nh(d_{\beta,\epsilon}-d)}$, where $d_{\beta,\epsilon} \leq 1/2$ satisfies $h(d_{\beta,\epsilon}) = 1 - \beta + 2\epsilon$. This implies the converse part of Theorem 288.

The above argument is typical. It will be generalized to the general case in the subsequent sections. What makes the proof of the converse part easy is the solution of the isopermetric problem. Unfortunately, the solution of the isopermetric problem is very difficult in general. For the simplest distortion measure—Hamming distance—the solution is known only in the binary case and even for general alphabets, an asymptotically optimal solution can be derived using the image—size characteristization of [3] (see also [7]) In Sect. 3, we develop a new method which

yields in particular the asymptotic solution of the general isopermetric problem for arbitrary finite alphabets and arbitrary distortion measures. Although the exact solution of the problem is extremly difficult, the asymptotic solution is good enough for our identification problem at hand.

Correlated Case

In this subsection, X and Y may be correlated. Unlike the independent case, only partial results on $R^*_{XY}(+\infty, 0, d)$ are obtained in this general case. First note that when X and Y are independent, $R(P_X, P_Y, 0, d)$ given by (19) can be rewritten as

$$R(P_X, P_Y, 0, d) = \inf_U I(X \wedge U) \tag{52}$$

where the infimum is taken over all RV's U taking values in some finite set such that

$$\mathbb{E}\bar{\rho}(P_{X|U}(\cdot|U), P_Y) > d \tag{53}$$

where $\bar{\rho}(P_{X|U}(\cdot|U), P_Y)$ denotes the $\bar{\rho}$—distance between the conditional distribution $P_{X|U}(\cdot|U)$ and the distribution P_Y of Y. (For the definition of $\bar{\rho}$ distance, we refer to [8]). The expression (52) of $R(P_X, P_Y, 0, d)$ will be extended to the general case.

Let $W(\cdot|\cdot) : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be a stochastic matrix such that for any $x \in \mathcal{X}$ and any $y \in \mathcal{Y}$, $W(y|x)$ is the conditional probability of $Y = y$ given $X = x$. A stochastic matrix $\hat{W}(\cdot|\cdot) : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ is said to be absolutely continuous with respect to W if for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $W(y|x) = 0$ implies $\hat{W}(y|x) = 0$. Let $X_0(Y_0$, resp.) denote the projection of $\mathcal{X} \times \mathcal{Y}$ onto $\mathcal{X}$ ($\mathcal{Y}$, resp.). For any $P \in \mathcal{P}(\mathcal{X})$, define

$$\bar{\rho}_e(P) = \inf_Q \mathbb{E}_Q \, \rho(X_0, Y_0), \tag{54}$$

where the infimum is taken over all $Q \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ such that

(i) The marginal of Q on $\mathcal{X}$ is P;
(ii) The marginal of Q on $\mathcal{Y}$ is PW, where $PW \in \mathcal{P}(\mathcal{Y})$ is given by

$$PW(y) = \sum_{x \in \mathcal{X}} P(x)W(y|x), \quad y \in \mathcal{Y} \tag{55}$$

(iii) The transition probability matrix from X_0 to Y_0 under the distribution Q is absolutely continuous with respect to W.

Clearly, if $W(y|x) > 0$ for any $x \in \mathcal{X}$ and any $y \in \mathcal{Y}$, then $\bar{\rho}_e(P)$ is just the $\bar{\rho}$-distance between P and PW. For any $0 < d < \mathbb{E}\rho(X, Y)$, we next define

$$\tilde{R}(P_{XY}, 0, d) = \inf_U [I(X \wedge U) - I(Y \wedge U)], \tag{56}$$

where the infimum is taken over all RV's U taking values in some finite set $\mathcal{U}$ such that (i) $U \to X \to Y$ forms a Markov chain, and (ii) $\mathbf{E}\bar{\rho}_e(P_{X|U}(\cdot|U), P_Y) > d$. We will prove that $\tilde{R}(P_{XY}, 0, d)$ has the following property.

Lemma 292 *$\tilde{R}(P_{XY}, 0, d)$ as a function of d is convex over the interval $0 < d < \mathbb{E}\rho(X, Y)$. Moreover, in evaluating $\tilde{R}(P_{XY}, 0, d)$ from (56), it suffices to consider only sets $\mathcal{U}$ with $|\mathcal{U}| \le |\mathcal{X}| + 2$.*

Proof of Lemma 292 For $i = 1, 2$, let (U_i, X_i, Y_i) be a random vector such that

(i) $P_{X_iY_i} = P_{XY}$;
(ii) $U_i \to X_i \to Y_i$ forms a Markov chain;
(iii) $\mathbb{E}\bar{\rho}_e(P_{X_i|U_i}(\cdot|U_i)) > d_i$.

Let I be a RV taking values in $\{1, 2\}$ with $\Pr(I = 1) = \lambda$. I is assumed to be independent of (U_i, X_i, Y_i) for $i = 1, 2$. Define

$$\tilde{X} = X_I, \tilde{Y} = Y_I, \text{ and } U = (U_I, I).$$

Clearly, $P_{\tilde{X}\tilde{Y}} = P_{XY}$ and $U \to \tilde{X} \to \tilde{Y}$ forms a Markov chain. Furthermore, it is not hard to see that

$$\mathbb{E}\bar{\rho}_e(P_{\tilde{X}|U}(\cdot|U)) = \lambda\mathbb{E}\bar{\rho}_e(P_{X_1|U_1}(\cdot|U_1)) + (1-\lambda)\mathbb{E}\bar{\rho}_e(P_{X_2|U_2}(\cdot|U_2)) > \lambda d_1 + (1-\lambda)d_2$$

and

$$I(\tilde{X}\wedge U) - I(\tilde{Y}\wedge U) = \lambda(I(X_1\wedge U_1) - I(Y_1\wedge U_1)) + (1-\lambda)(I(X_2\wedge U_2) - I(Y_2\wedge U_2)).$$

From this and the definition of $\tilde{R}(P_{XY}, 0, d)$, it follows that $\tilde{R}(P_{XY}, 0, d)$ as a function of d is convex.

To prove the second part of Lemma 292, first note that $\bar{\rho}_e(P)$ is convex as a function of P over $\mathcal{P}(\mathcal{X})$. Since $\mathcal{P}(\mathcal{X})$ is a convex polytope, it follows from [11] that $\bar{\rho}_e(P)$ is upper semicontinuous on $\mathcal{P}(\mathcal{X})$. On the other hand, from the definition of $\bar{\rho}_e(P)$, it is easy to prove that $\bar{\rho}_e(P)$ is lower semicontinuous on $\mathcal{P}(\mathcal{X})$. Therefore, $\bar{\rho}_e(P)$ is continuous on $\mathcal{P}(\mathcal{X})$. Applying the support lemma to the definition of $\tilde{R}(P_{XY}, 0, d)$ yields immediately the second result of Lemma 292. □

Similarly to (56), we define for any $0 < d < \mathbb{E}(\rho(X, Y))$

$$R(P_{XY}, 0, d) = \inf_U I(X \wedge U) \tag{57}$$

where the infimum is taken over all RV's U taking values in some finite set $\mathcal{U}$ such that (i) $U \to X \to Y$ forms a Markov chain, (ii) $\mathbb{E}\bar{\rho}(P_{X|U}(\cdot|U), P_{Y|U}(\cdot|U)) > d$. Obviously, (57) is the extension of (52) to the general case. It is easy to see that a similar result to Lemma 292 holds also for $R(P_{XY}, 0, d)$. The following theorem gives an upper and a lower bound for $R^*_{XY}(+\infty, 0, d)$ in the general case that X and Y may be correlated.

Theorem 293 *For any* $0 < d < \mathbb{E}\rho(X, Y)$,

$$\tilde{R}(P_{XY}, 0, d) \leq R^*_{XY}(+\infty, 0, d) \leq R(P_{XY}, 0, d). \tag{58}$$

Note that when X and Y are independent, the lower and the upper bounds are the same in Theorem 293 and equal to $R(P_X, P_Y, 0, d)$. Considering the expression given by (56), a natural question to ask at this point is whether the lower bound is always tight. Unfortunately, the following example shows that this is not true in general.

Example Let X, Y, Z be three RV's taking values in finite sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ respectively, such that

(i) Y is independent of X, Z,
(ii) $P_{XZ}(x, z) > 0$ for any pair $(x, z) \in \mathcal{X} \times \mathcal{Z}$.

Assume the decoder in the system shown in Figure 2 now knows (Y^n, Z^n) and wants to identify whether $\rho_n(X^n, Y^n) \leq d$. In another word, in addition to Y^n, the decoder knows side information Z^n which is correlated with X^n. Since $P_{XZ}(x, z) > 0$ for all pairs (x, z) and the distortion measure is irrelevant to Z, it is not hard to see that the side information is of no use and the minimum rate in bits per source symbol required to guarantee the zero probability of misrejection is still equal to $R^*_{XY}(+\infty, 0, d) = R(P_{XY}, 0, d)$. On the other hand, if we think of (Y, Z) as one RV defined on $\mathcal{Y} \times \mathcal{Z}$ and extend ρ from $\mathcal{X} \times \mathcal{Y}$ to $\mathcal{X} \times (\mathcal{Y} \times \mathcal{Z})$ by letting $\rho(x, (y, z)) = \rho(x, y)$ for all triples (x, y, z), then we have

$$\begin{aligned} \tilde{R}(P_{X(YZ)}, 0, d) &= \inf_U [I(X \wedge U) - I(YZ \wedge U))] \\ &= \inf_U [I(X \wedge U) - I(Z \wedge U)], \end{aligned} \tag{59}$$

where the infimum is taken over all RV's U taking values in some finite set $\mathcal{U}$ such that (i) $U \to X \to (YZ)$ forms a Markov chain, or equivalently $U \to X \to Z$ forms a Markov chain, (ii) $\mathbb{E}\bar{\rho}(P_{X|U}(\cdot|U), P_{YZ|U}(\cdot|U)) > d$, or equivalently $\mathbb{E}\bar{\rho}(P_{X|U}(\cdot|U), P_Y) > d$. From (52), (57), and (59), it follows that if X and Z are highly correlated, then in general,

$$\tilde{R}(P_{X(YZ)}, 0, d) < R(P_{XY}, 0, d) = R(P_{X(YZ)}, 0, d). \tag{60}$$

This shows that, in this case, the upper bound $R(P_{X(YZ)}, 0, d)$ is tight, but the lower bound $\tilde{R}(P_{X(YZ)}, 0, d)$ is not. ▲

The example shows a case where side information is of no use to reduce the transmission rate R in the system shown in Figure 2. Let us now look at a case where side information does help in reducing the transmission rate R.

Let X, Y, Z be three RV's taking values on finite sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ respectively, such that $X \to Z \to Y$ form a Markov chain. Let $\{X^n, Z^n, Y^n\}$ be n independent

drawings of the triple X, Z, Y. Assume that both the encoder and the decoder now know the side information Z^n. The decoder is still required to identify whether $\rho_n(X^n, Y^n) \leq d$ with zero probability of misrejection. Clearly, this is a special case of the situation we considered in Theorem 293, if we think of (X, Z) and (Y, Z) as two RV's, and extend $\rho(x, y)$ to $\rho((x, z), (y, z'))$ accordingly. Interestingly enough, in this special case the side information does help in reducing the transmission rate.

Theorem 294 *If $X \to Z \to Y$ form a Markov chain, then for any $0 < d < \mathbb{E}\rho(X, Y)$,*

$$R^*_{(XZ),(YZ)}(+\infty, 0, d) = \tilde{R}(P_{(XZ)(YZ)}, 0, d). \tag{61}$$

In contrast to the example, Theorem 294 gives us another example for which the lower bound of (58) is tight, but the corresponding upper bound is not.

We conclude this subsection with pointing out that if $X \to Z \to Y$ forms a Markov chain, then $\tilde{R}(P_{(XZ)(YZ)}, 0, d)$ can be rewritten as

$$\tilde{R}(P_{(XZ)(YZ)}, 0, d) = \inf_U I(X \wedge U|Z)\ , \tag{62}$$

where the infimum is taken over all RV's U taking values in some finite set $\mathcal{U}$ such that (i) $U \to (X, Z) \to Y$ forms a Markov chain; and (ii) $\mathbb{E}\bar{\rho}(P_{X|UZ}(\cdot|UZ), P_{Y|Z}(\cdot|Z)) > d$.

3 Inherently Typical Subset Lemma

This section is devoted to develop a new method for proving converses, which can be used to prove the converse parts of Theorem 288 and 3 and to solve the general isoperimetric problem (a subject to which we intend to return in another paper). The main idea of this method is embodied in what we call inherently typical subset lemma.

For each integer $m > 0$, let $\mathcal{P}_m(\mathcal{X})$ denote the set of all m-ED's on $\mathcal{X}$, that is

$$\mathcal{P}_m(\mathcal{X}) = \{P \in \mathcal{P}(\mathcal{X}) : P(x) \in \{0, \frac{1}{m}, \frac{2}{m}, \cdots, \frac{m-1}{m}, 1\}\ \forall\, x \in \mathcal{X}\}. \tag{63}$$

Let $\mathcal{U}_m = \{u_1, \cdots, u_{|\mathcal{P}_m(\mathcal{X})|}\}$ be an arbitrary set. Since $|\mathcal{U}_m| = |\mathcal{P}_m(\mathcal{X})|$, we can associate with each $P \in \mathcal{P}_m(\mathcal{X})$ an element $u \in \mathcal{U}_m$ so that elements of $\mathcal{U}_m$ associated with distinct m-ED's are distinct. If $u \in \mathcal{U}_m$ is associated with $P \in \mathcal{P}_m(\mathcal{X})$, for convenience, we shall write P as $P(\cdot|u)$. In terms of this notation, we have

$$\mathcal{P}_m(\mathcal{X}) = \{P(\cdot|u) : u \in \mathcal{U}_m\}. \tag{64}$$

Let A be any subset of $\mathcal{X}^n$. For any $0 \le i \le n-1$, define

$$A_i = \{x^i \in \mathcal{X}^i : x^i \text{is a prefix of some element of} A\}. \tag{65}$$

Here, we make use of the convention that $A_0 = \{\Lambda\}$, where Λ is the empty string. Assume that the integer m is greater than or equal to $2^{16|\mathcal{X}|^2}$.

Definition 295 $A \subset \mathcal{X}^n$ is called m-inherently typical if there exists a mapping

$$\phi : \cup_{i=0}^{n-1} A_i \to \mathcal{U}_m \tag{66}$$

such that the following hold:

(i) There exists an n-ED $Q \in \mathcal{P}_n(\mathcal{X} \times \mathcal{U}_m)$ such that for any $x^n \in A$,

$$P_{x^n u^n}(x,u) = Q(x,u), \quad x \in \mathcal{X}, \ \ u \in \mathcal{U}_m \tag{67}$$

where $u^n = (u_1, u_2, \cdots, u_n) \in \mathcal{U}_m^n$ is a sequence defined by $u_i = \phi(x^{i-1})$ for all $i : 1 \le i \le n$, (Such a sequence is called a sequence associated with x^n through ϕ) and for any $x \in \mathcal{X}$ and any $u \in \mathcal{U}_m$,

$$P_{x^n u^n}(x,u) = \frac{1}{n}|\{i : (x_i, u_i) = (x,u)\}|. \tag{68}$$

(ii) If $(\hat{X}, \hat{U})$ is a pair of RV's taking values in $\mathcal{X} \times \mathcal{U}_m$ with joint distribution Q, then

$$\frac{1}{n}\log|A| \le H(\hat{X}|\hat{U}) \le \frac{1}{n}\log|A| + \frac{\log^2 m}{m}. \tag{69}$$

Let $A \subset \mathcal{X}^n$ be m-inherently typical. Let ϕ be the corresponding mapping such that (67) and (69) are satisfied. For any random vector $\tilde{X}^n = (\tilde{X}_1, \cdots, \tilde{X}_n)$ taking values in A, we define another random vector $\tilde{U}^n = (\tilde{U}_1, \cdots, \tilde{U}_n)$ by letting $\tilde{U}_i = \phi(\tilde{X}^{i-1})$ for all $i : 1 \le i \le n$. Clearly (67) implies that with probability one, the following holds:

$$P_{\tilde{X}^n \tilde{U}^n}(x,u) = \frac{1}{n}\sum_{i=1}^{n} \Pr\{\tilde{X}_i = x, \tilde{U}_i = u\} \ \ x \in \mathcal{X}, \ \ u \in \mathcal{U}_m. \tag{70}$$

Note that the left hand side of (70) is the frequency, i.e. the average over time, and the right hand side is the average probability over ensemble. Intuitively, therefore, (70) just says that with probability one, the average over time is equal to the average over the ensemble. This is where the word "inherently typical" comes from. In typical applications, (see the following sections), the random vector $\tilde{X}^n$ is

often assumed to be uniformly distributed on A. In this case

$$\frac{1}{n}\log|A| = \frac{1}{n}H(\tilde{X}^n) = \frac{1}{n}\sum_{i=1}^{n} H(\tilde{X}_i|\tilde{X}^{i-1}). \tag{71}$$

Let I be a RV taking values uniformly in $\{1, \cdots, n\}$ and independent of $\tilde{X}^n$. Let $\tilde{X} = \tilde{X}_I$ and $U = (\tilde{X}^{I-1}, I)$, then

$$\frac{1}{n}\log|A| = H(\tilde{X}|U). \tag{72}$$

If we extend the mapping ϕ in the obvious way so that $\phi(U) = \phi(\tilde{X}^{I-1})$ whenever $U = (\tilde{X}^{I-1}, I)$, then it is not hard to see that $\tilde{X}$ and $\tilde{U}$ have the joint distribution $P_{\tilde{X}\tilde{U}} = Q$ where $\tilde{U} = \phi(U)$. Therefore, (69) just says that

$$H(\tilde{X}|U) \le H(\tilde{X}|\tilde{U}) \le H(\tilde{X}|U) + \frac{\log^2 m}{m}\,. \tag{73}$$

Lemma 296 (Inherently Typical Subset Lemma) *For any $m \ge 2^{16|\mathcal{X}|^2}$, n satisfying $((m+1)^{5|\mathcal{X}|+4}\ln(n+1))/n \le 1$, and any $A \subset \mathcal{X}^n$, there exists an m-inherently typical subset $\tilde{A} \subset A$ such that*

$$\frac{1}{n}\log\frac{|A|}{|\tilde{A}|} \le |\mathcal{X}|(m+1)^{|\mathcal{X}|}\frac{\log(n+1)}{n}\,. \tag{74}$$

Before proving Lemma 296, we remind the reader of the following two basic inequalities.

Lemma 297 (Pinsker Inequality, [11]) *For any two distributions $P_1, P_2 \in \mathcal{P}(\mathcal{X})$,*

$$D(P_1||P_2) \ge \frac{1}{2\ln 2}||P_1 - P_2||^2\,.$$

Lemma 298 (Folklore, Lemma 1.2.7 of [7]) *If P_1 and P_2 are two distributions on $\mathcal{X}$ such that*

$$||P_1 - P_2|| \le \Theta \le \frac{1}{2}\,,$$

then

$$|H(P_1) - H(P_2)| \le -\Theta \log \frac{\Theta}{|\mathcal{X}|} .$$

Proof of Lemma 296 Let A be any subset of $\mathcal{X}^n$. Let $\tilde{X}^n = (\tilde{X}_1, \cdots, \tilde{X}_n)$ be a random vector taking values uniformly in A. Let p denote the distribution of $\tilde{X}^n$ on A. Define a mapping ϕ from $\cup_{i=0}^{n-1} A_i$ to $\mathcal{U}_m$ so that for any $x^i \in A_i$, $0 \le i \le n-1$,

$$||p(\cdot|x^i) - P(\cdot|\phi(x^i))|| \le \frac{2|\mathcal{X}|}{m} , \tag{75}$$

where $p(\cdot|x^i)$ is the conditional distribution of $\tilde{X}_{i+1}$ given $\tilde{X}^i = x^i$, $P(\cdot|\phi(x^i))$ is the distribution in $\mathcal{P}_m(\mathcal{X})$ corresponding to $u = \phi(x^i)$(see (64)), and $||\cdot||$ denotes the variational distance between distributions. It is easy to see that such a mapping exists. (Essentially, this says that we use m-ED's to quantize distributions $p(\cdot|x^i)$, $x^i \in A_i$ and $0 \le i \le n-1$.) For each n-ED $\bar{Q} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{U}_m)$, let $A_{\bar{Q}} \subset A$ consist of all sequences $x^n \in A$ such that

$$P_{x^n u^n}(x, u) = \bar{Q}(x, u), \quad x \in \mathcal{X} \ \ and \ \ u \in \mathcal{U}_m ,$$

where $u^n \in \mathcal{U}_m^n$ is the sequence associated with x^n through the mapping ϕ and

$$P_{x^n u^n}(x, u) = \frac{1}{n} |\{i : (x_i, u_i) = (x, u)\}|.$$

Clearly, $\{A_{\bar{Q}} : \bar{Q} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{U}_m)\}$ is a partition of A. Let $Q \in \mathcal{P}_n(\mathcal{X} \times \mathcal{U}_m)$ satisfy

$$|A_Q| = \max\{|A_{\bar{Q}}| : \bar{Q} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{U}_m)\} . \tag{76}$$

We claim that $\tilde{A} = A_Q$ is the desired subset in Lemma 296. That is, that $\tilde{A}$ satisfies (74) and is an m-inherently typical subset under the mapping ϕ. To see this, first note that

$$|A| = \sum_{\bar{Q} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{U}_m)} |A_{\bar{Q}}| \le |\mathcal{P}_n(\mathcal{X} \times \mathcal{U}_m)||A_Q| \le (n+1)^{|\mathcal{X}||\mathcal{U}_m|} |\tilde{A}| . \tag{77}$$

This, together with $|\mathcal{U}_m| \le (m+1)^{|\mathcal{X}|}$, implies immediately (74). On the other hand, by the definition of A_Q, it follows that for any $x^n \in \tilde{A}$

$$P_{x^n u^n}(x, u) = Q(x, u), \quad x \in \mathcal{X}, \quad u \in \mathcal{U}_m, \tag{78}$$

where u^n is the sequence associated with x^n through ϕ. Therefore, all remaining to be proved is that if $(\hat{X}, \hat{U})$ is a random vector taking values on $\mathcal{X} \times \mathcal{U}_m$ with joint

distribution $P_{\hat{X}\hat{U}} = Q$, then

$$\frac{1}{n}\log|\tilde{A}| \le H(\hat{X}|\hat{U}) \le \frac{1}{n}\log|\tilde{A}| + \frac{\log^2 m}{m}\,. \tag{79}$$

To prove (79), let $\bar{X}^n = (\bar{X}_1, \cdots, \bar{X}_n)$ be a random vector taking values uniformly in $\tilde{A}$. Let $\tilde{p}$ denotes the distribution of $\bar{X}^n$ on $\tilde{A}$. As in the analysis following definition 1, let I be a RV taking values uniformly on $\{1, \cdots, n\}$. Let $\bar{X} = \bar{X}_I$, $U = (\bar{X}^{I-1}, I)$ and $\bar{U} = \phi(U)$ where $\phi(U) = \phi(\bar{X}^{I-1})$ whenever $U = (\bar{X}^{I-1}, I)$. Then

$$\frac{1}{n}\log|\tilde{A}| = H(\bar{X}|U) \le H(\bar{X}|\bar{U}). \tag{80}$$

In view of (78), it is easy to see that $\bar{X}$ and $\bar{U}$ have the joint distribution $P_{\bar{X}\bar{U}} = Q$. Consequently, in the following it suffices to prove

$$H(\bar{X}|\bar{U}) \le H(\bar{X}|U) + \frac{\log^2 m}{m}\,. \tag{81}$$

To this end, note that for any $x \in \mathcal{X}$ and $u \in \mathcal{U}_m$,

$$\begin{aligned} Q(x,u) &= P_{\bar{X}\bar{U}}(x,u) \\ &= \frac{1}{n}\sum_{i=1}^{n} \Pr(\bar{X}_i = x, \phi(\bar{X}^{i-1}) = u) \\ &= \sum_{x^n \in \tilde{A}} \tilde{p}(x^n)\frac{1}{n}\sum_{i=1}^{n} \delta(x_i, x)\delta(\phi(x^{i-1}), u) \\ &= \sum_{x^n \in \tilde{A}} \tilde{p}(x^n)\frac{1}{n}\sum_{i=1}^{n} \tilde{p}(x|x^{i-1})\delta(\phi(x^{i-1}), u)\,, \end{aligned} \tag{82}$$

where $\delta(\cdot,\cdot)$ is the Kronecker Delta function, that is

$$\delta(z, z') = \begin{cases} 1 & \text{if } z = z' \\ 0 & \text{otherwise} \end{cases}$$

and $\tilde{p}(x|x^{i-1})$ is the conditional probability of $\bar{X}_i = x$ given $\bar{X}^{i-1} = x^{i-1}$. Since for any $x^n \in \tilde{A}$,

$$P_{\bar{U}}(u) = \frac{1}{n}\sum_{i=1}^{n} \delta(\phi(x^{i-1}), u), \quad u \in \mathcal{U}_m$$

it follows from (82) that

$$P_{\bar{X}\bar{U}}(x,u) - P_{\bar{U}}(u)P(x|u) = \sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} \Big(\tilde{p}(x|x^{i-1}) - P(x|u) \Big) \delta(\phi(x^{i-1}, u). \tag{83}$$

This implies

$$\sum_{x \in \mathcal{X}} |P_{\bar{X}\bar{U}}(x,u) - P_{\bar{U}}(u)P(x|u)| \leq \sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} ||\tilde{p}(\cdot|x^{i-1}) - P(\cdot|u)|| \delta(\phi(x^{i-1}), u). \tag{84}$$

On the other hand, from (74) it follows that

$$\begin{aligned} \frac{1}{n} \sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \log \frac{\tilde{p}(x^n)}{p(x^n)} &= \sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} \log \frac{\tilde{p}(x_i|x^{i-1})}{p(x_i|x^{i-1})} \\ &= \sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} D(\tilde{p}(\cdot|x^{i-1})||p(\cdot|x^{i-1})) \\ &\leq |\mathcal{X}|(m+1)^{|\mathcal{X}|} \frac{\log(n+1)}{n} . \end{aligned} \tag{85}$$

Using Pinsker's inequality, we get

$$\sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} ||\tilde{p}(\cdot|x^{i-1}) - p(\cdot|x^{i-1})||^2 \leq 2|\mathcal{X}|(m+1)^{|\mathcal{X}|} \frac{\ln(n+1)}{n} . \tag{86}$$

Applying Schwartz inequality to (86) yields

$$\sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} ||\tilde{p}(\cdot|x^{i-1}) - p(\cdot|x^{i-1})|| \leq \left[2|\mathcal{X}|(m+1)^{|\mathcal{X}|} \frac{\ln(n+1)}{n} \right]^{\frac{1}{2}} . \tag{87}$$

Going back to (84), we get

$$\begin{aligned} &\sum_{x \in \mathcal{X}} |P_{\bar{X}\bar{U}}(x,u) - P_{\bar{U}}(u)P(x|u)| \\ &\leq \sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} ||\tilde{p}(\cdot|x^{i-1}) - p(\cdot|x^{i-1})|| \delta(\phi(x^{i-1}), u) \end{aligned}$$

$$+ \sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} ||p(\cdot|x^{i-1}) - P(\cdot|u)|| \delta(\phi(x^{i-1}), u)$$

$$\leq \left[2|\mathcal{X}|(m+1)^{|\mathcal{X}|} \frac{\ln(n+1)}{n} \right]^{\frac{1}{2}} + \frac{2|\mathcal{X}|}{m} P_{\bar{U}}(u) \ , \tag{88}$$

where the last inequality is due to (87) and (75). Therefore, if

$$P_{\bar{U}}(u) \geq \left[2|\mathcal{X}|(m+1)^{|\mathcal{X}|} \frac{\ln(n+1)}{n} \right]^{\frac{1}{4}} \ ,$$

then

$$\begin{aligned} ||P_{\bar{X}|\bar{U}}(\cdot|u) - P(\cdot|u)|| &\leq \left[2|\mathcal{X}|(m+1)^{|\mathcal{X}|} \frac{\ln(n+1)}{n} \right]^{\frac{1}{4}} + \frac{2|\mathcal{X}|}{m} \\ &\leq \frac{4|\mathcal{X}|}{m} \ , \end{aligned} \tag{89}$$

where $P_{\bar{X}\bar{U}}(\cdot|u)$ is the conditional probability distribution of $\bar{X}$ given $\bar{U} = u$ and the last inequality is due to the assumption that $[(m+1)^{5|\mathcal{X}|+4} \ln(n+1)]/n \leq 1$. From (89) and Lemma 298, we have

$$\begin{aligned} H(\bar{X}|\bar{U}) &= \sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} H(P_{\bar{X}|\bar{U}}(\cdot|\phi(x^{i-1}))) \\ &\leq \sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} H(P(\cdot|\phi(x^{i-1}))) + 4|\mathcal{X}| \frac{\log m}{m} \\ &\quad + |\mathcal{U}_m| \left[2|\mathcal{X}|(m+1)^{|\mathcal{X}|} \frac{\log(n+1)}{n} \right]^{\frac{1}{4}} \log |\mathcal{X}| \\ &\leq^{*} \sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} H(P(\cdot|\phi(x^{i-1}))) + (4|\mathcal{X}|+1) \frac{\log m}{m} \\ &\leq^{**} \sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} H(p(\cdot|x^{i-1})) + (6|\mathcal{X}|+1) \frac{\log m}{m} \ , \end{aligned} \tag{90}$$

where the inequality * is due to the assumption that $[(m+1)^{5|\mathcal{X}|+4} \ln(n+1)]/n \leq 1$ and $m \geq 2^{16|\mathcal{X}|^2}$, and inequality ** is due to (75) and Lemma 298. To continue (90),

we next compare $H(p(\cdot|x^{i-1}))$ with $H(\tilde{p}(\cdot|x^{i-1}))$. Let

$$F = \left\{ (i, x^{i-1}) : 1 \le i \le n, x^n \in \tilde{A}, \right.$$
$$\left. \|\tilde{p}(\cdot|x^{i-1}) - p(\cdot|x^{i-1})\| \ge \left[2|\mathcal{X}|(m+1)^{|\mathcal{X}|} \frac{\ln(n+1)}{n} \right]^{\frac{1}{4}} \right\}.$$

From (87) and the Markov inequality,

$$\sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} 1_F(i, x^n) \le [2|\mathcal{X}|(m+1)^{|\mathcal{X}|} \frac{\log(n+1)}{n}]^{\frac{1}{4}}, \tag{91}$$

where 1_F denotes the indicator function of F. From (91) and Lemma 298, it is not hard to verify that

$$\begin{aligned} \sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} H(p(\cdot|x^{i-1})) &\le \sum_{x^n \in \tilde{A}} \tilde{p}(x^n) \frac{1}{n} \sum_{i=1}^{n} H(\tilde{p}(\cdot|x^{i-1})) \\ &\quad + \frac{2|\mathcal{X}|^2}{m} + 2|\mathcal{X}| \frac{\log m}{m} \\ &= H(\bar{X}|U) + \frac{2|\mathcal{X}|^2}{m} + 2|\mathcal{X}| \frac{\log m}{m} \\ &\le H(\bar{X}|U) + (2|\mathcal{X}| + 1) \frac{\log m}{m}. \end{aligned} \tag{92}$$

Combining (92) with (90) yields

$$\begin{aligned} H(\bar{X}|\bar{U}) &\le H(\bar{X}|U) + (8|\mathcal{X}| + 2) \frac{\log m}{m} \\ &\le H(\bar{X}|U) + \frac{\log^2 m}{m}, \end{aligned} \tag{93}$$

where the last inequality is due to the assumption that $m > 2^{16|\mathcal{X}|^2}$. This completes the proof of (81) and hence the proof of Lemma 296. □

Note that Lemma 296 is proved by estimating the variational distance between the distribution $P_{\bar{X}|\bar{U}}(\cdot|\phi(x^{i-1}))$ and $\tilde{p}(\cdot|x^{i-1})$ where $x^{i-1} \in \tilde{A}_{i-1}$. Roughly speaking, the attempt we have made in the proof of Lemma 296 is to show that the variational distance between these two distributions is roughly upper bounded by $\frac{\log^2 m}{m}$. In fact, this is just what the second condition of the definition of m-inherently

typical subset implies. To see this, let us go back to (73), where $A \subset \mathcal{X}^n$ is assumed to be m-inherently typical. It is not hard to see that (73) can be rewritten as

$$H(\tilde{X}|\tilde{U}) - H(\tilde{X}|U) = \sum_{x^n \in A} p(x^n) \frac{1}{n} \sum_{i=1}^{n} D(p(\cdot|x^{i-1}) || P_{\tilde{X}|\tilde{U}}(\cdot|\phi(x^{i-1})))$$
$$\leq \frac{\log^2 m}{m} . \tag{94}$$

Using Pinsker's inequality once again, we get

$$\sum_{x^n \in A} p(x^n) \frac{1}{n} \sum_{i=1}^{n} ||p(\cdot|x^{i-1}) - P_{\tilde{X}|\tilde{U}}(\cdot|\phi(x^{i-1}))||^2 \leq 2\frac{\log^2 m}{m} \tag{95}$$

which, together with Schwartz inequality, implies

$$\sum_{x^n \in A} p(x^n) \frac{1}{n} \sum_{i=1}^{n} ||p(\cdot|x^{i-1}) - P_{\tilde{X}|\tilde{U}}(\cdot|\phi(x^{i-1}))|| \leq \sqrt{2\frac{\log^2 m}{m}} . \tag{96}$$

This means that the average variational distance between $P_{\tilde{X}|\tilde{U}}(\cdot|\phi(x^{i-1}))$ and $p(\cdot|x^{i-1})$ is upper bounded by $\sqrt{2\frac{\log^2 m}{m}}$. Therefore, the second condition in the definition of m-inherently typical subsets is also stringent.

4 Proofs of Theorems 288 and 290

In this section, we assume X and Y are independent. First of all, let us review some basic facts about ED's and typical sequences. Let U be a RV taking values on some finite set $\mathcal{U}$. Let $\{\gamma_n\}$ be a sequence of positive numbers such that $\gamma_n \to 0$ and $\sqrt{n}\gamma_n \to \infty$ as $n \to +\infty$. Recall that $\mathcal{P}_n(\mathcal{U})$ denotes the set of all n-ED's on $\mathcal{U}$. For each $u^n \in \mathcal{U}^n$, the ED P_{u^n} of u^n is defined by

$$P_{u^n}(u) = \frac{1}{n}|\{i : u_i = u\}|, \quad u \in \mathcal{U}.$$

For each $P \in \mathcal{P}_n(\mathcal{U})$, let

$$T_P^n(\mathcal{U}) = \{u^n \in \mathcal{U}^n : P_{u^n} = P\}$$

and denote by $\mathcal{V}_n(P, \mathcal{U}\times\mathcal{X})$ the set of all stochastic matrices $V = (V(x|u))_{x\in\mathcal{X},u\in\mathcal{U}}$ such that

$$V(x|u) \in \{0, \frac{1}{nP(u)}, \frac{2}{nP(u)}, \cdots, 1\} \text{ for all } x \in \mathcal{X},\ u \in \mathcal{U}.$$

Given $u^n \in \mathcal{U}^n$ and $V \in \mathcal{V}_n(P_{u^n}, \mathcal{U} \times \mathcal{X})$, a sequence x^n is said to be V-generated by u^n if for all $x \in \mathcal{X}$ and all $u \in \mathcal{U}$,

$$P_{u^n x^n}(u, x) = P_{u^n}(u)V(x|u).$$

Denote by $T_V^n(u^n, \mathcal{X})$ the set of all sequences x^n V-generated by u^n.
An n-ED $P \in \mathcal{P}_n(\mathcal{U})$ is said to be (U, γ_n)-essential if

$$|P(u) - P_U(u)| \le \gamma_n$$

and $P(u) = 0$ whenever $P_U(u) = 0$. A sequence $u^n \in \mathcal{U}^n$ is called (U, γ_n)-typical if P_{u^n} is (U, γ_n)-essential. Denote by T^n_{U,γ_n} the set of all (U, γ_n)-typical sequences. Similarly, for $u^n \in \mathcal{U}^n$, we call $V \in \mathcal{V}_n(P_{u^n}, \mathcal{U} \times \mathcal{X})$ $(u^n, X|U, \gamma_n)$-essential if

$$|P_{u^n}(u)V(x|u) - P_{u^n}(u)P_{X|U}(x|u)| \le \gamma_n$$

and $V(x|u) = 0$ whenever $P_{X|U}(x|u) = 0$ where $P_{X|U}(x|u)$ is the conditional probability of $X = x$ given $U = u$. A sequence $x^n \in \mathcal{X}^n$ is called $(u^n, X|U, \gamma_n)$-typical if there exists a $(u^n, X|U, \gamma_n)$-essential stochastic matrix $V \in \mathcal{V}_n(P_{u^n}, \mathcal{U} \times \mathcal{X})$ such that x^n is V-generated by u^n. Denote by $T^n_{X|U,\gamma_n}(u^n)$ the set of all $(u^n, X|U, \gamma_n)$-typical sequences x^n.

Although the above notation is introduced for RV's X and U and for finite sets $\mathcal{X}$ and $\mathcal{U}$, in the following we shall use freely these notation and terminology for other RV's and finite sets. Note that if u^n is (U, γ_n)-typical, and x^n is $(u^n, X|U, \gamma_n)$-typical, then $u^n x^n$ is $(UX, 2\gamma_n)$-typical and x^n is $(X, 2|\mathcal{U}|\gamma_n)$-typical. The following facts will be used. For all $P \in \mathcal{P}_n(\mathcal{U})$, and $V \in \mathcal{V}_n(P, \mathcal{U} \times \mathcal{X})$, $u^n \in T_P^n$

$$|\mathcal{V}_n(P, \mathcal{U} \times \mathcal{X})| \le (n + 1)^{|\mathcal{U}||\mathcal{X}|} \tag{97}$$

$$(n + 1)^{-|\mathcal{U}|}2^{nH(P)} \le |T_P^n(\mathcal{U})| \le 2^{nH(P)} \tag{98}$$

$$(n + 1)^{-|\mathcal{U}||\mathcal{X}|}2^{nH(V|P)} \le |T_V^n(u^n, \mathcal{X})| \le 2^{nH(V|P)} \tag{99}$$

where

$$\begin{aligned} H(V|P) &= \sum_{u\in\mathcal{U}} P(u)H(V(\cdot|u)) \\ &= \sum_{u\in\mathcal{U}}\sum_{x\in\mathcal{X}} -P(u)V(x|u)\log V(x|u)\ . \end{aligned} \tag{100}$$

Furthermore,

$$\Pr(U^n \in T^n_{U,\gamma_n}) \geq 1 - \frac{|\mathcal{U}|}{4n\gamma_n^2} \tag{101}$$

and if $\Pr(U^n = u^n) > 0$, then

$$\Pr(X^n \in T^n_{X|U,\gamma_n}(u^n)|U^n = u^n) \geq 1 - \frac{|\mathcal{X}||\mathcal{U}|}{4n\gamma_n^2}\,, \tag{102}$$

where (X^n, U^n) are n independent drawings of (X, U).

Proof of Theorem 288
In view of the fact that $R^*_{XY}(\alpha, \beta, d)$ is continuous at $\beta = 0$, it suffices to prove Theorems 288 and 290 for $0 < \beta < \beta(d)$.

Proof of Theorem 288 We first prove the direct part, that is,

$$R^*_{XY}(+\infty, \beta, d) \leq \bar{R}(P_X, P_Y, \beta, d).$$

By the definition of $R^*_{XY}(+\infty, \beta, d)$, it suffices to prove that $(R, +\infty, \beta)$ is achievable for any $R > \bar{R}(P_X, P_Y, \beta, d)$. To this end, let us fix $R > \bar{R}(P_X, P_Y, \beta, d)$ below and prove $(R, +\infty, \beta)$ is achievable.

In view of the definition of $\bar{R}(P_X, P_Y, \beta, d)$ and $R(P_X, P_Y, \beta, d)$, it is not hard to see that for any $\delta > 0$, there exists a RV U taking values on some finite set $\mathcal{U}$ such that

$$I(X \wedge U) < R, \quad and \quad \mathcal{E}(P_{XU}, d) \geq \beta - \delta\,. \tag{103}$$

Based on the pair (X, U), the standard technique of [2] (see also [7]) can be used to show that there exists for sufficiently large n a system $\{(u^n(i), \mathcal{S}_i) : 1 \leq i \leq M\}$ which has the following properties:

(i) $\log M \leq n(I(X \wedge U) + \delta)$,
(ii) For $1 \leq i \leq M$, $u^n(i) \in T^n_{U,\gamma_n}$, $\mathcal{S}_i \subset T^n_{X|U,\gamma_n}(u^n(i))$ and

$$\Pr(X^n \in \mathcal{S}_i|U^n = u^n(i)) \geq \frac{\delta}{2},$$

where (X^n, U^n) are n-independent drawings of (X, U).
(iii) $\mathcal{S}_i : 1 \leq i \leq M$ are disjoint and

$$\Pr(X^n \in \cup_{i=1}^M \mathcal{S}_i) \geq 1 - \delta.$$

Based on this system, we construct an nth-order ID source code $\mathcal{C}_n = (f_n, B_n, g_n)$ as follows. For each $x^n \notin \cup_{i=1}^M \mathcal{S}_i$, the encoder simply sends the sequence x^n itself to the decoder. After receiving x^n, the decoder outputs 1 if $\rho_n(x^n, y^n) \leq d$

and 0 otherwise. The number of bits needed for the lossless transmission of x^n is $\lceil n \log |\mathcal{X}| \rceil$ plus one bit flag indicating $x^n \notin \cup_{i=1}^M \mathcal{S}_i$. For each $x^n \in \cup_{i=1}^M \mathcal{S}_i$, the encoder first finds the integer i such that $x^n \in \mathcal{S}_i$ and then transmits i to the decoder. Upon receiving i, the decoder outputs 1 if $\rho_n(\mathcal{S}_i, y^n) \le d$, and 0 otherwise, where

$$\rho_n(\mathcal{S}_i, y^n) = \min\{\rho_n(\tilde{x}^n, y^n) : \tilde{x}^n \in \mathcal{S}_i\}.$$

The number of bits needed for the transmission of the integer i is $\lceil \log M \rceil$ plus one bit flag indicating $x^n \in \cup_{i=1}^M \mathcal{S}_i$. Therefore the average rate of the IDS code described above is upper bounded by

$$\begin{aligned} r_n(\mathcal{C}_n) &= \frac{1}{n} \Pr(X^n \in \cup_{i=1}^M \mathcal{S}_i)(\lceil \log M \rceil + 1) + \frac{1}{n} \Pr(X^n \notin \cup_{i=1}^M \mathcal{S}_i)(\lceil n \log |\mathcal{X}| \rceil + 1) \\ &\le I(X \wedge U) + \delta + \delta \log |\mathcal{X}| + \frac{2}{n} \\ &\le R + (1 + \log |\mathcal{X}|)\delta + \frac{2}{n}, \end{aligned} \tag{104}$$

where the last inequality is due to (103). From the construction of $\mathcal{C}_n$, it is clear that the probability of misrejection is zero and the probability of false identification is upper bounded by

$$\begin{aligned} P_{e2}(\mathcal{C}_n) &\le \frac{1}{\Pr(\rho_n(X^n, Y^n) > d)} \sum_{i=1}^M \Pr(X^n \in \mathcal{S}_i) \Pr(Y^n \in \mathcal{S}_i^d) \\ &\le 2 \sum_{i=1}^M \Pr(X^n \in \mathcal{S}_i) \Pr(Y^n \in \mathcal{S}_i^d) \end{aligned} \tag{105}$$

for sufficiently large n, where

$$\mathcal{S}_i^d = \{y^n \in \mathcal{Y}^n : \rho_n(\mathcal{S}_i, y^n) \le d\} \tag{106}$$

and the last inequality is due to the fact that $d < \mathbb{E}\rho(X, Y)$. To continue (105), note that for $1 \le i \le M$ and $x^n \in \mathcal{S}_i$, $(u^n(i), x^n)$ is $(UX, 2\gamma_n)$-typical. Since $\mathcal{E}(P_{XU}, d) \ge \beta - \delta$, it follows from Lemma 287 that for sufficiently large n and any $x^n \in \mathcal{S}_i$,

$$\mathcal{E}(P_{x^n u^n(i)}, d) \ge \beta - 2\delta. \tag{107}$$

Clearly, for each $1 \le i \le M$,

$$\mathcal{S}_i^d = \bigcup_{V \in \mathcal{V}_n(P_{u^n(i)}, \mathcal{U} \times \mathcal{Y})} \mathcal{S}_i^d \cap T_V^n(u^n(i), \mathcal{Y}). \tag{108}$$

It is easy to see that if $\mathcal{S}_i^d \cap T_V^n(u^n(i), \mathcal{Y})$ is not empty, then there exists $x^n \in \mathcal{S}_i$ and $Q \in \mathcal{P}_n(\mathcal{U} \times \mathcal{X} \times \mathcal{Y})$ such that

(i) the marginal of Q on $\mathcal{U} \times \mathcal{X}$ is $P_{u^n(i)x^n}$;
(ii) the marginal of Q on $\mathcal{U} \times \mathcal{Y}$ is given by

$$P_{u^n(i)}(u)V(y|u), \quad u \in \mathcal{U} \ \text{ and } \ y \in \mathcal{Y};$$

(iii) under the distribution Q, $\mathbb{E}\rho(X_0, Y_0) \le d$.

In view of (16) and (107), this implies

$$\sum_{u\in\mathcal{U}} P_{u^n(i)}(u) D(V(\cdot|u)||P_Y) = \sum_{u\in\mathcal{U}, y\in\mathcal{Y}} P_{u^n(i)}(u)V(y|u) \log \frac{V(y|u)}{P_Y(y)} > \beta - 2\delta \ . \tag{109}$$

Therefore, if $\mathcal{S}_i^d \cap T_V^n(u^n(i), \mathcal{Y})$ is not empty, then

$$\begin{aligned}
&\Pr(Y^n \in \mathcal{S}_i^d \cap T_V^n(u^n(i), \mathcal{Y})) \\
&\le \Pr\{Y^n \in T_V^n(u^n(i), \mathcal{Y})\} \\
&= |T_V^n(u^n(i), \mathcal{Y})| 2^{-n[H(V|P_{u^n(i)}) + \sum_{u\in\mathcal{U}} P_{u^n(i)}(u) D(V(\cdot|u)||P_Y)]} \\
&\le^* 2^{-n \sum_{u\in\mathcal{U}} P_{u^n(i)}(u) D(V(\cdot|u)||P_Y)} \\
&\le 2^{-n(\beta - 2\delta)},
\end{aligned} \tag{110}$$

where the inequality * is due to (99) and the last inequality is due to (109). Going back to (108), we get for sufficiently large n

$$\begin{aligned}
\Pr(Y^n \in \mathcal{S}_i^d) &\le |\mathcal{V}_n(P_{u^n(i)}, \mathcal{U} \times \mathcal{Y})| 2^{-n(\beta - 2\delta)} \\
&\le 2^{-n(\beta - 3\delta)},
\end{aligned} \tag{111}$$

where the last inequality is due to (97). Substituting (111) into (105) yields

$$P_{e2}(\mathcal{C}_n) \le 2^{-n(\beta - 3\delta)}. \tag{112}$$

Since $\delta > 0$ is arbitrary, by definition, (104) and (112) imply that $(R, +\infty, \beta)$ is achievable. This completes the proof of the direct part of Theorem 288.

We next turn to the converse part, that is

$$R_{XY}^*(+\infty, \beta, d) \ge \bar{R}(P_X, P_Y, \beta, d).$$

Clearly, it is enough to prove that for any achievable triple $(R, +\infty, \beta)$

$$R \geq \bar{R}(P_X, P_Y, \beta, d).$$

To this end, let us below fix an achievable triple $(R, +\infty, \beta)$. By definition, there exists for any $\epsilon > 0$ a sequence of ID source codes $\mathcal{C}_n = (f_n, B_n, g_n)$ such that for sufficiently large n,

$$r_n(\mathcal{C}_n) \leq R + \epsilon,\ P_{e1}(\mathcal{C}_n) = 0, \quad and \quad P_{e2}(\mathcal{C}_n) \leq 2^{-n(\beta-\epsilon)}. \tag{113}$$

As in the binary symmetric case, we define for each $b^n \in B_n$

$$\mathcal{S}(b^n) = \{x^n \in \mathcal{X}^n : f_n(x^n) = b^n\}$$

and

$$\mathcal{S}^d(b^n) = \{y^n \in \mathcal{Y}^n : \rho_n(x^n, y^n) \leq d \text{ for some } x^n \in S(b^n)\}\ .$$

Since $P_{e1}(\mathcal{C}_n) = 0$, we must have

$$\mathcal{S}^d(b^n) \subset \{y^n \in \mathcal{Y}^n : g_n(y^n, b^n) = 1\}.$$

Therefore, the inequality $P_{e2}(\mathcal{C}_n) \leq 2^{-n(\beta-\epsilon)}$ implies

$$\sum_{b^n \in B_n} \Pr(X^n \in \mathcal{S}(b^n)) \Pr(Y^n \in \mathcal{S}^d(b^n)) \leq 2^{-n(\beta-\epsilon)} + \Pr(\rho_n(X^n, Y^n) \leq d).$$

In view of (15) and the fact that $\beta < \beta(d)$, it follows that for sufficiently large n

$$\sum_{b^n \in B_n} \Pr(X^n \in \mathcal{S}(b^n)) \Pr(Y^n \in \mathcal{S}^d(b^n)) \leq 2^{-n(\beta-\epsilon)}.$$

From Markov' s inequality, one gets that

$$\sum_{b^n \in B_n'} \Pr(X^n \in \mathcal{S}(b^n)) \geq 1 - \epsilon, \tag{114}$$

where

$$B_n' = \{b^n \in B_n : \Pr\{Y^n \in S^d(b^n)\} \leq 2^{-n(\beta-\epsilon-n^{-1}\log(2/\epsilon))}\}\ .$$

Let m be a sufficiently large positive integer to be specified later. Fix a $b^n \in B_n^{'}$. Applying the inherently typical subset lemma (i.e., Lemma 296) to $S(b^n) \cap T_{X,r_n}^n$,

where T^n_{X,r_n} is the set of all (X, r_n)-typical sequences x^n, we get an m-inherently typical subset $A \subset S(b^n) \cap T^n_{X,r_n}$ such that

$$\frac{1}{n}\log\frac{|S(b^n)\cap T^n_{X,r_n}|}{|A|} \le |\mathcal{X}|(m+1)^{|\mathcal{X}|}\frac{\log(n+1)}{n}\,. \tag{115}$$

By the definition of m-inherently typical subsets, there exists a mapping ϕ from $\cup_{i=0}^{n-1} A_i$ to $\mathcal{U}_m$, where $A_i = \{x^i \in \mathcal{X}^i : x^i \text{ is a prefix of some element in } A\}$, such that the following hold:

(i) There exists an n-ED $Q \in \mathcal{P}_n(\mathcal{X}\times\mathcal{U}_m)$ such that for any $x^n \in A$,

$$P_{x^n u^n}(x,u) = Q(x,u), \qquad x\in\mathcal{X},\ u\in\mathcal{U}_m\,, \tag{116}$$

where $u^n \in \mathcal{U}_m^n$ is the sequence associated with x^n through ϕ.

(ii) If $(\hat{X},\hat{U})$ is a random vector taking values on $\mathcal{X}\times\mathcal{U}_m$ with joint distribution Q, then

$$\frac{1}{n}\log|A| \le H(\hat{X}|\hat{U}) \le \frac{1}{n}\log|A| + \frac{\log^2 m}{m}\,. \tag{117}$$

As what we did in the analysis following Definition 1, let $\tilde{X}^n = (\tilde{X}_1,\cdots,\tilde{X}_n)$ be a random vector taking values uniformly on A. Define a random vector $\tilde{U}^n = (\tilde{U}_1,\cdots,\tilde{U}_n)$ by letting $\tilde{U}_i = \phi(\tilde{X}^{i-1})$, $1 \le i \le n$. Let I be a RV taking values uniformly on $\{1,2,\cdots,n\}$ and independent of $\tilde{X}^n$. Let

$$\tilde{X} = \tilde{X}_I,\ \tilde{U} = \tilde{U}_I,\ U = (\tilde{X}^{I-1}, I). \tag{118}$$

Clearly, if we extend the mapping ϕ in the obvious way so that $\phi(U) = \phi(\tilde{X}^{I-1})$ whenever $U = (\tilde{X}^{I-1}, I)$, then $\tilde{U} = \phi(U)$. As pointed out in the analysis following Definition 1, $\tilde{X}$ and $\tilde{U}$ have the joint distribution $P_{\tilde{X}\tilde{U}} = Q$. Furthermore, $n^{-1}\log|A| = H(\tilde{X}|U)$ and (117) can be rewritten as

$$H(\tilde{X}|U) \le H(\tilde{X}|\tilde{U}) \le H(\tilde{X}|U) + \frac{\log^2 m}{m}. \tag{119}$$

Having defined the random pair $(\tilde{X},\tilde{U})$ taking values on $\mathcal{X}\times\mathcal{U}_m$, we next lower bound $\Pr(Y^n \in A^d)$ by a function of $\mathcal{E}(P_{\tilde{X}\tilde{U}}, d-\epsilon)$, where

$$A^d = \{y^n \in \mathcal{Y}^n | \text{there exist } x^n \in A : \rho_n(x^n, y^n) \le d\}.$$

In view of the definition of $\mathcal{E}(P_{\tilde{X}\tilde{U}}, d-\epsilon)$, let $\tilde{Y}$ be a RV taking values on $\mathcal{Y}$ such that $\mathbb{E}\rho(\tilde{X},\tilde{Y}) \le d-\epsilon$. Let $V = (V(y|xu))_{x\in\mathcal{X},u\in\mathcal{U}_m,y\in\mathcal{Y}}$ be a stochastic matrix so that $V(y|xu)$ is the conditional probability of $\tilde{Y} = y$ given $\tilde{X} = x$ and $\tilde{U} = u$. Let

$\tilde{Y}^n = (\tilde{Y}_1, \cdots, \tilde{Y}_n)$ be the random vector resulting from passing $(\tilde{X}^n, \tilde{U}^n)$ through the channel V. From (102), it follows that for any $x^n \in A$,

$$\Pr(\tilde{Y}^n \in T^n_{\tilde{Y}|\tilde{X}\tilde{U},\gamma_n}(x^n u^n)|\tilde{X}^n = x^n, \tilde{U}^n = u^n) > 1 - \frac{|\mathcal{X}||\mathcal{Y}||\mathcal{U}_m|}{4n\gamma_n^2}, \tag{120}$$

where $u^n \in \mathcal{U}_m^n$ is the sequence associated with x^n through ϕ. From (116), it is not hard to see that for any $x^n \in A$,

$$T^n_{\tilde{Y}|\tilde{X}\tilde{U},\gamma_n}(x^n, u^n) \subset T^n_{\tilde{Y},|\tilde{X}||\tilde{U}|\gamma_n}. \tag{121}$$

Furthermore, since $\mathbb{E}\rho(\tilde{X}, \tilde{Y}) \leq d - \epsilon$, it follows that for sufficiently large n and any $y^n \in T^n_{\tilde{Y}|\tilde{X}\tilde{U},\gamma_n}(x^n u^n)$,

$$\rho_n(x^n, y^n) \leq d.$$

Therefore, if we let

$$F = \cup_{x^n \in A} T^n_{\tilde{Y}|\tilde{X}\tilde{U},\gamma_n}(x^n u^n),$$

where $u^n \in \mathcal{U}_m^n$ is the sequence associated with x^n through ϕ, then $F \subset A^d$ and for any $x^n \in A$,

$$\Pr(\tilde{Y}^n \in F|\tilde{X}^n = x^n, \tilde{U}^n = u^n) > 1 - \frac{|\mathcal{X}||\mathcal{Y}||\mathcal{U}_m|}{4n\gamma_n^2}.$$

This implies

$$\Pr(\tilde{Y}^n \in F) > 1 - \frac{|\mathcal{X}||\mathcal{Y}||\mathcal{U}_m|}{4n\gamma_n^2}. \tag{122}$$

For convenience, let

$$\delta_n = \frac{|\mathcal{X}||\mathcal{Y}||\mathcal{U}_m|}{4n\gamma_n^2}.$$

From (122), we have

$$H(\tilde{Y}^n) \leq h(\delta_n) + \log|F| + n\delta_n \log|\mathcal{Y}|,$$

where $h(\cdot)$ represents the binary entropy function. From this, it is not hard to verify that

$$\begin{aligned}\log|F| &\geq H(\tilde{Y}^n) - h(\delta_n) - n\delta_n \log|\mathcal{Y}| \\ &= \sum_{i=1}^{n} H(\tilde{Y}_i|\tilde{Y}^{i-1}) - h(\delta_n) - n\delta_n \log|\mathcal{Y}| \\ &\geq \sum_{i=1}^{n} H(\tilde{Y}_i|\tilde{Y}^{i-1}\tilde{X}^{i-1}\tilde{U}^{i-1}) - h(\delta_n) - n\delta_n \log|\mathcal{Y}| \\ &=^{*} \sum_{i=1}^{n} H(\tilde{Y}_i|\tilde{X}^{i-1}\tilde{U}^{i-1}) - h(\delta_n) - n\delta_n \log|\mathcal{Y}| \\ &=^{**} \sum_{i=1}^{n} H(\tilde{Y}_i|\tilde{X}^{i-1}) - h(\delta_n) - n\delta_n \log|\mathcal{Y}|, \end{aligned} \tag{123}$$

where equality * is due to the fact that given $(\tilde{X}^{i-1}, \tilde{U}^{i-1})$, $\tilde{Y}^{i-1}$ and $\tilde{Y}_i$ are conditionally independent, and the equality ** follows from the fact that $\tilde{U}^{i-1} = \phi(\tilde{X}^{i-2})$. Recall that I is the RV which takes values uniformly on $\{1, \cdots, n\}$ and is independent of $\tilde{X}^n$ and $\tilde{Y}^n$. Let $\bar{Y} = \tilde{Y}_I$. In view of (118) and (123) continues as follows

$$\begin{aligned}\log|F| &\geq nH(\bar{Y}|U)) - h(\delta_n) - n\delta_n \log|\mathcal{Y}| \\ &= nH(\bar{Y}|\tilde{U}) + n(H(\bar{Y}|U) - H(\bar{Y}|\tilde{U})) - h(\delta_n) - n\delta_n \log|\mathcal{Y}| \\ &= nH(\tilde{Y}|\tilde{U}) + n(H(\bar{Y}|U) - H(\bar{Y}|\tilde{U})) - h(\delta_n) - n\delta_n \log|\mathcal{Y}| \end{aligned} \tag{124}$$

where the last equality follows from the observation that $(\tilde{X}, \tilde{U}, \bar{Y})$ has the same joint distribution as that of $(\tilde{X}, \tilde{U}, \tilde{Y})$. To continue (124) further, we next estimate the difference $H(\bar{Y}|\tilde{U}) - H(\bar{Y}|U)$. Since $\tilde{U} = \phi(U)$, it is not hard to verify that

$$H(\bar{Y}|\tilde{U}) - H(\bar{Y}|U) = \frac{1}{n}\sum_{i=1}^{n} \sum_{\substack{x^{i-1} \\ \in A_{i-1}}} \Pr(\tilde{X}^{i-1} = x^{i-1}) \sum_{y\in\mathcal{Y}} \Pr(\tilde{Y}_i = y|\tilde{X}^{i-1} = x^{i-1}) \cdot$$

$$\cdot \log \frac{\Pr(\tilde{Y}_i = y|\tilde{X}^{i-1} = x^{i-1})}{P_{\tilde{Y}|\tilde{U}}(y|\phi(x^{i-1}))}, \tag{125}$$

where $P_{\bar{Y}|\tilde{U}}(y|\phi(x^{i-1}))$ is the conditional probability of $\bar{Y} = y$ given $\tilde{U} = \phi(x^{i-1})$. By construction, it is not hard to see that

$$\Pr(\tilde{Y}_i = y|\tilde{X}^{i-1} = x^{i-1}) = \sum_{x\in\mathcal{X}} \Pr(\tilde{X}_i = x|\tilde{X}^{i-1} = x^{i-1}) V(y|x\phi(x^{i-1}))$$

and

$$P_{\tilde{Y}|\tilde{U}}(y|\phi(x^{i-1})) = \sum_{x\in\mathcal{X}} P_{\tilde{X}|\tilde{U}}(x|\phi(x^{i-1})) V(y|x\phi(x^{i-1})).$$

Using the log-sum inequality, one gets that

$$\sum_{y\in\mathcal{Y}} \Pr(\tilde{Y}_i = y|\tilde{X}^{i-1} = x^{i-1}) \log \frac{\Pr(\tilde{Y}_i = y|\tilde{X}^{i-1} = x^{i-1})}{P_{\tilde{Y}|\tilde{U}}(y|\phi(x^{i-1}))}$$

$$\leq \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \Pr(\tilde{X}_i = x|\tilde{X}^{i-1} = x^{i-1}) V(y|x\phi(x^{i-1})) \log \frac{\Pr(\tilde{X}_i = x|\tilde{X}^{i-1} = x^{i-1})}{P_{\tilde{X}|\tilde{U}}(x|\phi(x^{i-1}))}$$

$$\leq \sum_{x\in\mathcal{X}} \Pr(\tilde{X}_i = x|\tilde{X}^{i-1} = x^{i-1}) \log \frac{\Pr(\tilde{X}_i = x|\tilde{X}^{i-1} = x^{i-1})}{P_{\tilde{X}|\tilde{U}}(x|\phi(x^{i-1}))}. \tag{126}$$

Substituting (126) into (125) yields

$$H(\bar{Y}|\tilde{U}) - H(\bar{Y}|U) \leq H(\tilde{X}|\tilde{U}) - H(\tilde{X}|U) \leq \frac{\log^2 m}{m}, \tag{127}$$

where the last inequality is due to (119). Combining (124) and (127) yields

$$\log|F| \geq nH(\tilde{Y}|\tilde{U}) - n\frac{\log^2 m}{m} - -h(\delta_n) - n\delta_n \log|\mathcal{Y}|.$$

From (121),

$$F \subset T^n_{\tilde{Y},|\mathcal{X}||\mathcal{U}_m|\gamma_n}.$$

Thus,

$$\Pr(Y^n \in F) \geq |F| 2^{-n(H(\tilde{Y})+D(P_{\tilde{Y}}||P_Y)+o(1))}$$

$$\geq \exp\{-n(I(\tilde{U} \wedge \tilde{Y}) + D(P_{\tilde{Y}}||P_Y) + \frac{\log^2 m}{m} + \epsilon_n)\}\,, \tag{128}$$

where $\epsilon_n \to 0$ as n goes to infinity. Since $F \subset A^d$, (128) implies

$$\Pr(Y^n \in A^d) \geq \exp\{-n(I(\tilde{U} \wedge \tilde{Y}) + D(P_{\tilde{Y}}||P_Y) + \frac{\log^2 m}{m} + \epsilon_n)\}. \tag{129}$$

Note that (129) holds for any RV $\tilde{Y}$ taking values on $\mathcal{Y}$ such that $\mathbb{E}\rho(\tilde{X}, \tilde{Y}) \leq d - \epsilon$. This, together with the definition of $\mathcal{E}(P_{\tilde{X}\tilde{U}}, d - \epsilon)$, implies

$$\Pr(Y^n \in A^d) \geq \exp\{-n(\mathcal{E}(P_{\tilde{X}\tilde{U}}, d - \epsilon) + \frac{\log^2 m}{m} + \epsilon_n)\}. \tag{130}$$

Let us go back to (114) and (115). We next want to estimate the probability $\Pr\{X^n \in \mathcal{S}(b^n) \cap T^n_{X,\gamma_n}\}$, where $b^n \in B'_n$. Since $A \subset \mathcal{S}(b^n) \cap T^n_{X,\gamma_n}$, it is easy to see from (116) and (118) that $P_{\tilde{X}}$ is (X, γ_n)-essential, that is

$$|P_{\tilde{X}}(x) - P_X(x)| \leq \gamma_n, \quad x \in \mathcal{X} \tag{131}$$

and $P_{\tilde{X}}(x)$ whenever $P_X(x) = 0$. For convenience, let

$$a_n = |\mathcal{X}|(m+1)^{|\mathcal{X}|}\frac{\log(n+1)}{n}.$$

It is not hard to see that

$$\begin{aligned}
\Pr(X^n \in \mathcal{S}(b^n) \cap T^n_{X,\gamma_n}) &\leq |\mathcal{S}(b^n) \cap T^n_{X,\gamma_n}| 2^{-n(H(\tilde{X})-o(1))} \\
&\leq^* 2^{-n(H(\tilde{X})-n^{-1}\log|A|-a_n-o(1))} \\
&\leq^{**} 2^{-n(H(\tilde{X})-H(\tilde{X}|\tilde{U})-a_n-o(1))} \\
&= 2^{-n(I(\tilde{X}\wedge\tilde{U})-\epsilon'_n)}\,,
\end{aligned} \tag{132}$$

where the inequality * is due to (115), the inequality ** is due to the fact that $n^{-1}\log|A| \leq H(\tilde{X}|\tilde{U})$ and ϵ'_n goes to zero as n goes to infinity. Since $b^n \in B'_n$, it follows that

$$\Pr(Y^n \in A^d) \leq \Pr(Y^n \in \mathcal{S}^d(b^n)) \leq 2^{-n(\beta-\epsilon-\frac{1}{n}\log\frac{2}{\epsilon})}. \tag{133}$$

Comparing (133) with (130) yields

$$\mathcal{E}(P_{\tilde{X}\tilde{U}}, d - \epsilon) \geq \beta - \epsilon - \frac{\log^2 m}{m} - \epsilon_n - \frac{1}{n}\log\frac{2}{\epsilon}.$$

Note that $\tilde{U}$ takes values on $\mathcal{U}_m$. From the definition of $R_k(P_X, P_Y, \beta, d)$, it follows that

$$I(\tilde{X} \wedge \tilde{U}) \geq R_{|\mathcal{U}_m|}(P_{\tilde{X}}, P_Y, \beta - \epsilon - \frac{\log^2 m}{m} - \epsilon_n - \frac{1}{n}\log\frac{2}{\epsilon}, d - \epsilon)$$

which, combined with (132), implies

$$\begin{aligned} &-\frac{1}{n}\log \Pr(X^n \in \mathcal{S}(b^n) \cap T^n_{X,\gamma_n}) \\ &\geq R_{|\mathcal{U}_m|}(P_{\tilde{X}}, P_Y, \beta - \epsilon - \frac{\log^2 m}{m} - \epsilon_n - \frac{1}{n}\log\frac{2}{\epsilon}, d - \epsilon) - \epsilon'_n. \end{aligned} \tag{134}$$

In view of Fact 1 in the first subsection of 2 and (131), (134) continues as follows.

$$-\frac{1}{n}\log \Pr(X^n \in \mathcal{S}(b^n) \cap T^n_{X,\gamma_n}) \geq R_{|\mathcal{U}_m|}(P_X, P_Y, \beta - \epsilon - \frac{\log^2 m}{m}, d - \epsilon) - \epsilon''_n\,, \tag{135}$$

where ϵ''_n goes to zero as n goes to infinity.

Note that (135) holds for any $b^n \in B'_n$. Now it is not hard to verify that

$$\begin{aligned} R + \epsilon \geq r_n(\mathcal{C}_n) &\geq \frac{1}{n}H(f_n(X^n)) \\ &\geq \sum_{b^n \in B'_n} -\frac{1}{n}\Pr(X^n \in \mathcal{S}(b^n) \cap T^n_{X,\gamma_n}) \log \Pr(X^n \in \mathcal{S}(b^n) \cap T^n_{X,\gamma_n}) - \frac{1}{n} \\ &\geq (1 - \epsilon - \frac{|\mathcal{X}|}{4n\gamma_n^2})(R_{|\mathcal{U}_m|}(P_X, P_Y, \beta - \epsilon - \frac{\log^2 m}{m}, d - \epsilon) - \epsilon''_n) - \frac{1}{n}, \end{aligned} \tag{136}$$

where the last inequality is due to (114) and the following inequality

$$\Pr(X^n \in T^n_{X,\gamma_n}) \geq 1 - \frac{|\mathcal{X}|}{4n\gamma_n^2}.$$

In view of Fact 1, letting $n \to +\infty$ and then letting $\epsilon \to 0$ in (136) yield

$$R \geq R_{|\mathcal{U}_m|}(P_X, P_Y, \beta - \frac{\log^2 m}{m}, d).$$

Since $(R, +\infty, \beta)$ is an arbitrary achievable triple, this implies

$$R^*_{XY}(+\infty, \beta, d) \geq R_{|\mathcal{U}_m|}(P_X, P_Y, \beta - \frac{\log^2 m}{m}, d). \tag{137}$$

Thus, for sufficiently large m,

$$R^*_{XY}(+\infty, \beta, d) \geq R(P_X, P_Y, \beta - \frac{\log^2 m}{m}, d).$$

Letting m go to infinity yields

$$R^*_{XY}(+\infty, \beta, d) \geq \bar{R}(P_X, P_Y, \beta, d).$$

This completes the proof of the converse part and hence the proof of Theorem 288. □

Remark In response to the remark following Theorem 288, let us note that in view of Lemma 296, the inequality (137) actually holds for any $m \geq 2^{16|\mathcal{X}|^2}$. From the proof of Theorem 288, therefore we obtain that for any $m \geq 2^{16|\mathcal{X}|^2}$,

$$\begin{aligned} R_{|\mathcal{U}_m|}(P_X, P_Y, \beta - \frac{\log^2 m}{m}, d) &\leq R^*_{XY}(+\infty, \beta, d) \\ &= \bar{R}(P_X, P_Y, \beta, d) \\ &\leq R(P_X, P_Y, \beta, d) \\ &\leq R_{|\mathcal{U}_m|}(P_X, P_Y, \beta, d). \end{aligned} \tag{138}$$

This gives us in a sense how accurate the value obtained could be if we approximate $\bar{R}(P_X, P_Y, \beta, d)$ by $R_{|\mathcal{U}_m|}(P_X, P_Y, \beta, d)$. If some regular conditions are satisfied, hopefully this approximation could be as accurate as $O(\log^2 m/m)$.

Proof of Theorem 290

We next turn to the proof of Theorem 290. Although the proof of Theorem 290 is more complicated than that of Theorem 288, the basic idea is the same and in fact, most parts of the proof are just the translation of the corresponding parts in the proof of Theorem 288 to the present case. This is why we stated separately Theorems 288 and 290. Hope this will help the reader understand the proofs more easily.

Proof of Theorem 290 In view of the remark following Theorem 290, it suffices to prove Theorem 290 for $\alpha > \beta(P_X, d) - \beta(d)$ and $0 < \beta < \beta(d)$. We first prove the direct part, that is

$$R^*_{XY}(\alpha, \beta, d) \leq \bar{R}(P_X, P_Y, \alpha, \beta(d), \beta, d).$$

By the definition of $R^*_{XY}(\alpha, \beta, d)$, it is enough to prove that for any R satisfying

$$R > \bar{R}(P_X, P_Y, \alpha, \beta(d), \beta, d),$$

(R, α, β) is achievable. To this end, let us fix below $R > \bar{R}(P_X, P_Y, \alpha, \beta(d), \beta, d)$. As in the proof of Theorem 288, it is not hard to see that for any $\delta > 0$, there exists a random variable U taking values on some finite set $\mathcal{U}$ such that

$$I(X \wedge U) < R \qquad \text{and} \qquad \mathcal{E}(P_{XU}, \alpha, \beta(d), d) \geq \beta - \delta. \tag{139}$$

Corresponding to the random pair (X, U), there exists for sufficiently large n a system $\{(u^n(i), \mathcal{S}_i) | 1 \leq i \leq M\}$ which satisfies properties (i)–(iii). Based on this system, we construct an nth order ID source code $\mathcal{C}_n = (f_n, B_n, g_n)$ as follows. For each $x^n \notin \cup_{i=1}^M \mathcal{S}_i$, the encoder simply sends the sequence x^n itself to the decoder. After receiving x^n, the decoder outputs 1 if $\rho_n(x^n, y^n) \leq d$ and 0 otherwise. For each $x^n \in \cup_{i=1}^M \mathcal{S}_i$, the encoder first finds the integer i such that $x^n \in \mathcal{S}_i$ and then transmits i to the decoder. Upon receiving i, the decoder outputs 1 if $y^n \in \mathcal{Y}^n$ satisfies that there exists some $x^n \in S_i$ such that $\rho_n(x^n, y^n) \leq d$ and

$$\sum_{x \in \mathcal{X}, u \in \mathcal{U}} P_{x^n u^n(i)}(x, u) D(P_{y^n | x^n u^n(i)}(\cdot | xu) || P_Y) \leq \beta(d) + \alpha, \tag{140}$$

where

$$P_{y^n | x^n u^n(i)}(\cdot | xu) \in \mathcal{V}_n(P_{x^n u^n(i)}, (\mathcal{X} \times \mathcal{U}) \times \mathcal{Y})$$

is defined by $P_{x^n u^n(i)}(x, u) P_{y^n | x^n u^n(i)}(y | xu) = P_{x^n u^n(i) y^n}(x, u, y)$ for all $y \in \mathcal{Y}$; and otherwise outputs 0. Clearly, the encoder f_n defined here is the same as in the proof of Theorem 288. From (104), therefore, the average rate $r_n(\mathcal{C}_n)$ is also upper bounded by

$$r_n(\mathcal{C}_n) \leq R + (1 + \log |\mathcal{X}|)\delta + \frac{2}{n}. \tag{141}$$

For each $1 \leq i \leq M$, let

$$\mathcal{S}_i^d = \{y^n \in \mathcal{Y}^n | \text{there exist } x^n \in S_i : \rho_n(x^n, y^n) \leq d\} \tag{142}$$

$$\hat{\mathcal{S}}_i^d = \{y^n \in \mathcal{Y}^n | g_n(y^n, i) = 1\} \tag{143}$$

and

$$\bar{\mathcal{S}}_i^d = \mathcal{Y}^n - \hat{\mathcal{S}}_i^d. \tag{144}$$

Obviously, $\hat{S}_i^d \subset S_i^d$ for $1 \leq i \leq M$. For each $x^n \in \mathcal{S}_i$, let $B^i(x^n)$ denote the set of all $y^n \in \mathcal{Y}^n$ such that $\rho_n(x^n, y^n) \leq d$ and

$$\sum_{x \in \mathcal{X}, u \in \mathcal{U}} P_{x^n u^n(i)}(x,u) D(P_{y^n|x^n u^n(i)}(\cdot|xu)||P_Y) > \beta(d) + \alpha. \tag{145}$$

From (145), it is not hard to see that

$$\Pr(Y^n \in B^i(x^n)) \leq (n+1)^{|\mathcal{X}||\mathcal{U}||\mathcal{Y}|} 2^{-n(\beta(d)+\alpha)} . \tag{146}$$

By the construction of the ID source code $\mathcal{C}_n$, we can now verify that

$$\begin{aligned} P_{e1}(\mathcal{C}_n) &= \frac{\Pr\{(X^n, Y^n) \in \cup_{i=1}^M \mathcal{S}_i \times \bar{\mathcal{S}}_i^d, \& \rho_n(X^n, Y^n) \leq d\}}{\Pr\{\rho_n(X^n, Y^n) \leq d\}} \\ &\leq \frac{1}{\Pr\{\rho_n(X^n, Y^n) \leq d\}} \sum_{i=1}^M \sum_{x^n \in \mathcal{S}_i} \Pr(X^n = x^n) \Pr(Y^n \in B^i(x^n)) \\ &\leq^* \frac{(n+1)^{|\mathcal{X}||\mathcal{U}||\mathcal{Y}|}}{\Pr(\rho_n(X^n, Y^n) \leq d)} 2^{-n(\beta(d)+\alpha)} \\ &\leq^{**} 2^{-n(\alpha-\delta)} \end{aligned} \tag{147}$$

for sufficiently large n, where the inequality * follows from (146) and the inequality ** follows from (15). As in the proof of the direct part of Theorem 288, it is clear that the probability $P_{e2}(\mathcal{C}_n)$ of false identification of the ID source code $\mathcal{C}_n$ is upper bounded by

$$\begin{aligned} P_{e2}(\mathcal{C}_n) &\leq \frac{1}{\Pr\{\rho_n(X^n, Y^n) > d\}} \sum_{i=1}^M \Pr(X^n \in \mathcal{S}_i)\ Pr(Y^n \in \hat{\mathcal{S}}_i^d) \\ &\leq 2 \sum_{i=1}^M \Pr(X^n \in \mathcal{S}_i) \Pr(Y^n \in \hat{\mathcal{S}}_i^d) \end{aligned} \tag{148}$$

for sufficiently large n. To continue (148), we do the same thing as we did before. First note that for any $1 \leq i \leq M$ and $x^n \in \mathcal{S}_i$, $(u^n(i), x^n)$ is $(UX, 2\gamma_n)$-typical. In view of Lemma 289 and (139), therefore, it follows that for sufficiently large n and any $x^n \in \mathcal{S}_i$,

$$\mathcal{E}(P_{x^n u^n(i)}, \alpha, \beta(d), d) \geq \beta - 2\delta . \tag{149}$$

Let us now look at $\hat{\mathcal{S}}_i^d \cap T_V^n(u^n(i), \mathcal{Y})$, where $V \in \mathcal{V}_n(P_{u^n(i)}, \mathcal{U} \times \mathcal{Y})$. Clearly, if $\hat{\mathcal{S}}_i^d \cap T_V^n(u^n(i), \mathcal{Y})$ is not empty, then from the definition of $\hat{\mathcal{S}}_i^d$, there exist a $x^n \in \mathcal{S}_i$ and a $Q \in \mathcal{P}_n(\mathcal{U} \times \mathcal{X} \times \mathcal{Y})$ such that

(i) the marginal of Q on $\mathcal{U} \times \mathcal{X}$ is $P_{u^n(i)x^n}$;
(ii) the marginal of Q on $\mathcal{U} \times \mathcal{Y}$ is given by

$$P_{u^n(i)}(u)V(y|u), \quad u \in \mathcal{U} \text{ and } y \in \mathcal{Y};$$

(iii) if $(\tilde{U}, \tilde{X}, \tilde{Y})$ is a random vector taking values on $\mathcal{U} \times \mathcal{X} \times \mathcal{Y}$ with joint distribution $P_{\tilde{X}\tilde{U}\tilde{Y}} = Q$, then $\mathbb{E}\rho(X_0, Y_0) \leq d$ and $D(P_{\tilde{Y}}||P_Y) + I(\tilde{X}\tilde{U} \wedge \tilde{Y}) \leq \beta(d) + \alpha$.

In view of the definition of $\mathcal{E}(P_{x^n u^n(i)}, \alpha, \beta(d), d)$ and the inequality (149), this implies

$$\sum_{u \in \mathcal{U}} P_{u^n(i)}(u) D(V(\cdot|u)||P_Y) > \beta - 2\delta \, . \tag{150}$$

Therefore, if $\hat{\mathcal{S}}_i^d \cap T_V^n(u^n(i), \mathcal{Y})$ is not empty, then

$$\begin{aligned} \Pr\{Y^n \in \hat{\mathcal{S}}_i^d \cap T_V^n(u^n(i), \mathcal{Y})\} &\leq \Pr\{Y^n \in T_V^n(u^n(i), \mathcal{Y})\} \\ &\leq 2^{-n \sum_{u \in \mathcal{U}} P_{u^n(i)}(u) D(V(\cdot|u)||P_Y))} \\ &\leq 2^{-n(\beta - 2\delta)}, \end{aligned} \tag{151}$$

which implies

$$Pr(Y^n \in \hat{\mathcal{S}}_i^d) \leq |\mathcal{V}_n(P_{u^n(i)}, \mathcal{U} \times \mathcal{Y})| 2^{-n(\beta - 2\delta)} \leq 2^{-n(\beta - 3\delta)} \tag{152}$$

for sufficiently large n. Substituting (152) into (148) yields

$$P_{e2}(\mathcal{C}_n) \leq 2 2^{-n(\beta - 3\delta)} \, . \tag{153}$$

Since $\delta > 0$ is arbitrary, (141), (147), and (153) implies that (R, α, β) is achievable. This completes the proof of the direct part of Theorem 290.

We next turn to the converse part. Clearly, it is enough to prove that for any achievable triple (R, α, β)

$$R \geq \bar{R}(P_X, P_Y, \alpha, \beta(d)\beta, d).$$

To this end, let us below fix an achievable triple (R, α, β). By definition, there exists for any $\epsilon > 0$ a sequence of ID source codes $\mathcal{C}_n = (f_n, B_n, g_n)$ such that for

sufficiently large n,

$$r_n(\mathcal{C}_n) \le R + \epsilon,\ P_{e1}(\mathcal{C}_n) \le 2^{-n(\alpha-\epsilon)},\ \ and\ \ P_{e2}(\mathcal{C}_n) \le 2^{-n(\beta-\epsilon)}\ . \tag{154}$$

For each $b^n \in B_n$, let

$$\mathcal{S}(b^n) = \{x^n \in \mathcal{X}^n : f_n(x^n) = b^n\}$$

and

$$G(b^n) = \{y^n \in \mathcal{Y}^n : g_n(y^n, b^n) = 1\}.$$

For each $x^n \in \mathcal{X}^n$, denote by $B(x^n)$ the set of sequences $y^n \in \mathcal{Y}^n$ such that $\rho_n(x^n, y^n) \le d$ and $y^n \notin G(f_n(x^n))$. It is not hard to see that

$$P_{e1}(\mathcal{C}_n) = \frac{1}{\Pr(\rho_n(X^n, Y^n) \le d)} \sum_{x^n \in \mathcal{X}^n} \Pr(X^n = x^n) \Pr(Y^n \in B(x^n)).$$

By virtue of (15) and (154), we have for sufficiently large n

$$\sum_{x^n \in \mathcal{X}^n} \Pr(X^n = x^n) \Pr(Y^n \in B(x^n)) \le 2^{-n(\alpha-2\epsilon+\beta(d))}\ . \tag{155}$$

Let

$$F_n = \{x^n \in \mathcal{X}^n : \Pr(Y^n \in B(x^n)) \le \epsilon^{-1} 2^{-n(\alpha-2\epsilon+\beta(d))}\}.$$

From (155) and the Markov inequality,

$$\Pr(X^n \in F_n) \ge 1 - \epsilon\ . \tag{156}$$

As in the proof of the converse part of Theorem 288, it is not hard to prove that the inequality $P_{e2}(\mathcal{C}_n) \le 2^{-n(\beta-\epsilon)}$ implies that for sufficiently large n,

$$\sum_{b^n \in B_n} \Pr(X^n \in \mathcal{S}(b^n)) \Pr(Y^n \in G(b^n)) \le 2 2^{-n(\beta-\epsilon)}.$$

Using Markov inequality once again, one gets that

$$\sum_{b^n \in B'_n} \Pr(X^n \in \mathcal{S}(b^n)) \ge 1 - \epsilon,$$

where

$$B_n' = \{b^n \in B_n | \Pr(Y^n \in G(b^n)) \leq 2^{-n(\beta-\epsilon-\frac{1}{n}\log\frac{2}{\epsilon})}\}.$$

We are now in a position to apply the inherently typical subset lemma (Lemma 296). Fix a $b^n \in B_n'$. Applying the lemma to $\mathcal{S}(b^n) \cap T^n_{X,\gamma_n} \cap F_n$, we get an m-inherently typical subset $A \subset \mathcal{S}(b^n) \cap T^n_{X,\gamma_n} \cap F_n$ such that

$$\frac{1}{n}\log\frac{|\mathcal{S}(b^n) \cap F_n \cap T^n_{X,\gamma_n}|}{|A|} \leq |\mathcal{X}|(m+1)^{|\mathcal{X}|}\frac{\log(n+1)}{n}\ . \tag{157}$$

The remaining proof is much the same as that shown in the proof of the converse part of Theorem 288. In what follows, therefore, we only point out places where changes are needed. (Unless otherwise specified, all notation below is the same as in the proof of the converse part of Theorem 288).

Having defined the random pair $(\tilde{X}, \tilde{U})$ taking values on $\mathcal{X} \times \mathcal{U}_m$, we, instead of lower bounding $\Pr(Y^n \in A^d)$ by a function of $\mathcal{E}(P_{\tilde{X}\tilde{U}}, d-\epsilon)$, lower bound $\Pr(Y^n \in G(b^n))$ by a function of $\mathcal{E}(P_{\tilde{X}\tilde{U}}, , \alpha - 4\epsilon, \beta(d), d-\epsilon)$. In view of the definition of $\mathcal{E}(P_{\tilde{X}\tilde{U}}, , \alpha - 4\epsilon, \beta(d), d-\epsilon)$, let $\tilde{Y}$ be a RV taking values on $\mathcal{Y}$ such that

$$\mathbb{E}\rho(\tilde{X}, \tilde{Y}) \leq d-\epsilon \text{ and } D(P_{\tilde{Y}}||P_Y) + I(\tilde{X}\tilde{U} \wedge \tilde{Y}) \leq \beta(d) + \alpha - 4\epsilon\ . \tag{158}$$

Let $V = (V(y|xu))_{x\in\mathcal{X},u\in\mathcal{U}_m,y\in\mathcal{Y}}$ be a stochastic matrix so that $V(y|xu)$ is the conditional probability of $\tilde{Y} = y$ given $\tilde{X} = x, \tilde{U} = u$. Let $\tilde{Y}^n = (\tilde{Y}_1, \cdots, \tilde{Y}_n)$ be a random vector resulting from passing $(\tilde{X}^n, \tilde{U}^n)$ through the channel V. From each $x^n \in A$, consider $T^n_{\tilde{Y}|\tilde{X}\tilde{U},r_n}(x^n, u^n)$, where $u^n \in \mathcal{U}_m^n$ is the sequence associated with x^n through ϕ. In view of (158), it is not hard to see that for sufficiently large n and for any $y^n \in T^n_{\tilde{Y}|\tilde{X}\tilde{U},r_n}(x^n, u^n)$,

$$\rho_n(x^n, y^n) \leq d \tag{159}$$

and

$$\sum_{x\in\mathcal{X},u\in\mathcal{U}_m} P_{x^nu^n}(x,u)D(P_{y^n|x^nu^n}(\cdot|xu)||P_Y) \leq \beta(d) + \alpha - 3\epsilon. \tag{160}$$

Let $\hat{V} \in \mathcal{V}_n(P_{x^nu^n}, (\mathcal{X} \times \mathcal{U}_m) \times \mathcal{Y})$ be $(x^nu^n, \tilde{Y}|\tilde{X}\tilde{U}, \gamma_n)$-essential, then (160) implies

$$\sum_{x\in\mathcal{X},u\in\mathcal{U}_m} P_{x^nu^n}(x,u)D(\hat{V}(\cdot|xu)||P_Y) \leq \beta(d) + \alpha - 3\epsilon. \tag{161}$$

Since $A \subset \mathcal{S}(b^n) \cap F_n \cap T^n_{X,\gamma_n}$, $x^n \in A$ implies $x^n \in F_n$. By definition of F_n, therefore, it follows that

$$\Pr(Y^n \in B(x^n)) \le \epsilon^{-1} 2^{-n(\alpha+\beta(d)-2\epsilon)}. \tag{162}$$

By comparing (162) with (161), we can obtain that

$$\begin{aligned} |T^n_{\hat{V}}(x^n u^n, \mathcal{Y}) \cap B(x^n)| &\le^* \epsilon^{-1} 2^{-n\epsilon} (n+1)^{|\mathcal{X}||\mathcal{Y}||\mathcal{U}_m|} |T^n_{\hat{V}}(x^n u^n, \mathcal{Y})| \\ &\le 2^{-\frac{n\epsilon}{2}} |T^n_{\hat{V}}(x^n u^n, \mathcal{Y})| \end{aligned} \tag{163}$$

for sufficiently large n, where in derivation of *, the following inequality was used:

$$|T^n_{\hat{V}}(x^n u^n, \mathcal{Y})| \ge (n+1)^{-|\mathcal{X}||\mathcal{Y}||\mathcal{U}_m|} \exp\{n \sum_{x \in \mathcal{X}, u \in \mathcal{U}_m} P_{x^n u^n}(x,u) H(\hat{V}(\cdot|x,u))\}.$$

From (163), it is now easy to check that

$$\begin{aligned} &\Pr(\tilde{Y}^n \in T^n_{\tilde{Y}|\tilde{X}\tilde{U},\gamma_n}(x^n u^n) \cap B(x^n) | \tilde{X}^n = x^n, \tilde{U}^n = u^n) \\ &= \sum_{\hat{V}} \Pr(\tilde{Y} \in T^n_{\hat{V}}(x^n u^n, \mathcal{Y}) \cap B(x^n) | \tilde{X}^n = x^n, \tilde{U}^n = u^n) \\ &\le 2^{-\frac{n\epsilon}{2}} \sum_{\hat{V}} \Pr(\tilde{Y} \in T^n_{\hat{V}}(x^n u^n, \mathcal{Y}) | \tilde{X}^n = x^n, \tilde{U}^n = u^n) \\ &= 2^{-\frac{n\epsilon}{2}} \Pr(\tilde{Y} \in T^n_{\tilde{Y}|\tilde{X}\tilde{U},\gamma_n}(x^n u^n) | \tilde{X}^n = x^n, \tilde{U}^n = u^n) \\ &\le 2^{-\frac{n\epsilon}{2}}, \end{aligned} \tag{164}$$

where the summation is taken over all $\hat{V}$ such that $\hat{V}$ is $(x^n u^n, \tilde{Y}|\tilde{X}\tilde{U}, \gamma_n)$-essential. In view of (120), (159), and (164), it follows that

$$\begin{aligned} &\Pr(\tilde{Y}^n \in T^n_{\tilde{Y}|\tilde{X}\tilde{U},\gamma_n}(x^n u^n) \cap G(b^n) | \tilde{X}^n = x^n, \tilde{U}^n = u^n) \\ &\ge 1 - \frac{|\mathcal{U}_m||\mathcal{X}||\mathcal{Y}|}{4n\gamma_n^2} - 2^{-\frac{n\epsilon}{2}} \\ &\ge 1 - \delta'_n, \end{aligned} \tag{165}$$

where $\delta'_n \to 0$ as $n \to \infty$. Let

$$G = \cup_{x^n \in A} T^n_{\tilde{Y}|\tilde{X}\tilde{U},\gamma_n}(x^n u^n) \cap G(b^n).$$

(165) implies that for any $x^n \in A$,

$$\Pr(\tilde{Y}^n \in G | \tilde{X}^n = x^n, \tilde{U}^n = u^n) \geq 1 - \delta'_n,$$

which in turn implies

$$\Pr(\tilde{Y}^n \in G) \geq 1 - \delta'_n. \tag{166}$$

Note that (166) is in parallel with (122). A similar argument to the derivation of (128) can be used to show that

$$\Pr(Y^n \in G) \geq \exp\{-n(I(\tilde{U} \wedge \tilde{Y}) + D(P_{\tilde{Y}} || P_Y) + \frac{\log^2 m}{m} + \epsilon_n)\}, \tag{167}$$

where $\epsilon_n \to 0$ as n goes to infinity. Since $G \subset G(b^n)$, (167) implies

$$\Pr(Y^n \in G(b^n)) \geq \exp\{-n(I(\tilde{U} \wedge \tilde{Y}) + D(P_{\tilde{Y}} || P_Y) + \frac{\log^2 m}{m} + \epsilon_n)\}. \tag{168}$$

Note that (168) holds for any RV $\tilde{Y}$ taking values on $\mathcal{Y}$ such that (158) is satisfied. This, together with the definition of $\mathcal{E}(P_{\tilde{X}\tilde{U}}, \alpha - 4\epsilon, \beta(d), d - \epsilon)$, implies that

$$\Pr(Y^n \in G(b^n)) \geq \exp\{-n(\mathcal{E}(P_{\tilde{X}\tilde{U}}, \alpha - 4\epsilon, \beta(d), d - \epsilon) + \frac{\log^2 m}{m} + \epsilon_n)\}. \tag{169}$$

On the other hand, since $b^n \in B'_n$,

$$\Pr(Y^n \in G(b^n)) \leq 2^{-n(\beta - \epsilon - \frac{1}{n}\log\frac{2}{\epsilon})}$$

which, combined with (169), yields

$$\mathcal{E}(P_{\tilde{X}\tilde{U}}, \alpha - 4\epsilon, \beta(d), d - \epsilon) \geq \beta - \epsilon - \frac{\log^2 m}{m} - \epsilon'_n \tag{170}$$

where ϵ'_n goes to zero as n goes to infinity. A similar argument to the derivation of (132) can be used to show that

$$\Pr(X^n \in \mathcal{S}(b^n) \cap F_n \cap T^n_{X,\gamma_n}) \leq 2^{-n(I(\tilde{X} \wedge \tilde{U}) - \epsilon''_n)},$$

that is

$$-\frac{1}{n}\log \Pr(X^n \in \mathcal{S}(b^n) \cap F_n \cap T^n_{X,\gamma_n}) \geq I(\tilde{X} \wedge \tilde{U}) - \epsilon''_n, \tag{171}$$

where ϵ_n'' goes to zero as n goes to infinity. In view of (170) and the definition of $R_{|\mathcal{U}_m|}(P_{\tilde{X}}, P_Y, \alpha, \gamma, \beta, d)$, (171) continues as follows

$$-\frac{1}{n}\log \Pr(X^n \in \mathcal{S}(b^n) \cap F_n \cap T^n_{X,\gamma_n})$$

$$\geq R_{|\mathcal{U}_m|}(P_{\tilde{X}}, P_Y, \alpha - 4\epsilon, \beta(d), \beta - \epsilon - \frac{\log^2 m}{m} - \epsilon_n', d - \epsilon) - \epsilon_n''. \tag{172}$$

Using Fact 3 in the first subsection of 2, we get

$$-\frac{1}{n}\log \Pr(X^n \in \mathcal{S}(b^n) \cap F_n \cap T^n_{X,\gamma_n})$$

$$\geq R_{|\mathcal{U}_m|}(P_X, P_Y, \alpha - 4\epsilon, \beta(d), \beta - \epsilon - \frac{\log^2 m}{m}, d - \epsilon) - \bar{\epsilon}_n\ , \tag{173}$$

where $\bar{\epsilon}_n$ goes to zero as n goes to infinity. Note that (173) holds for any $b^n \in B_n'$. In parallel with (136), we now have

$$R + \epsilon \geq r_n(\mathcal{C}_n) \geq \frac{1}{n}H(f_n(X^n))$$

$$\geq \sum_{b^n \in B_n'} -\frac{1}{n}\Pr(X^n \in \mathcal{S}(b^n) \cap F_n \cap T^n_{X,\gamma_n}) \log \Pr(X^n \in \mathcal{S}(b^n) \cap F_n \cap T^n_{X,\gamma_n}) - \frac{2}{n}$$

$$\geq (1 - 2\epsilon - \frac{|\mathcal{X}|}{4n\gamma_n^2})(R_{|\mathcal{U}_m|}(P_X, P_Y, \alpha - 4\epsilon, \beta(d), \beta - \epsilon - \frac{\log^2 m}{m}, d - \epsilon) - \bar{\epsilon}_n) - \frac{2}{n}. \tag{174}$$

In view of Fact 3 in the first subsection of 2 once again, letting $n \to \infty$ and then letting $\epsilon \to 0$ yields

$$R \geq R_{|\mathcal{U}_m|}(P_X, P_Y, \alpha, \beta(d), \beta - \frac{\log^2 m}{m}, d),$$

which implies

$$R^*_{XY}(\alpha, \beta, d) \geq R_{|\mathcal{U}_m|}(P_X, P_Y, \alpha, \beta(d), \beta - \frac{\log^2 m}{m}, d). \tag{175}$$

Letting $m \to \infty$ in(175) yields

$$R^*_{XY}(\alpha, \beta, d) \geq \bar{R}(P_X, P_Y, \alpha, \beta(d), \beta, d),$$

which completes the proof of the converse part and hence the proof of Theorem 290. □

5 Proofs of Theorems 5 and 6

In this section, X and Y may be correlated. As in the third subsection of 2, let $W = (W(y|x))_{x\in\mathcal{X},y\in\mathcal{Y}}$ denote the transition probability matrix from X to Y.

Proof of Theorem 293 We begin with proving

$$R^*_{XY}(+\infty, 0, d) \le R(P_{XY}, 0, d). \tag{176}$$

To prove (176), it suffices to prove that for any $R > R(P_{XY}, 0, d)$, there exists a $\delta_0 > 0$ such that $(R, +\infty, \delta_0)$ is achievable. To this end, we fix below $R > R(P_{XY}, 0, d)$. By the definition of $R(P_{XY}, 0, d)$, there exists a RV U taking values on some finite set $\mathcal{U}$ such that

(i) $U \to X \to Y$ form a Markov chain;
(ii) $I(X \wedge U) < R$ and $\mathbb{E}\bar{\rho}(P_{X|U}(\cdot|U), P_{Y|U}(\cdot|U)) > d$.

Without loss of generality, in what follows, we shall assume $P_U(u) > 0$ for any $u \in \mathcal{U}$. Let δ be a positive real to be specified later. As in the proof of the direct part of Theorem 288, corresponding to the random pair (X, U), there exists for sufficiently large n a system $\{(u^n(i), \mathcal{S}_i)|1 \le i \le M\}$ which has the properties (i) to (iii). Let $\mathcal{C}_n = (f_n, B_n, g_n)$ be the nth order IDS code which is based on the system we just defined and constructed as in the proof of the direct part of Theorem 288. From the proof of the direct part of Theorem 288, the probability of misidentification of $\mathcal{C}_n$ is zero and the average rate in bits per symbol of $\mathcal{C}_n$ is upper bounded by

$$r_n(\mathcal{C}_n) \le I(X \wedge U) + (1 + \log|\mathcal{X}|)\delta + \frac{2}{n}. \tag{177}$$

Furthermore, the probability of false identification of $\mathcal{C}_n$ is now upper bounded by

$$\begin{aligned} P_{e2}(\mathcal{C}_n) &\le \frac{1}{\Pr(\rho_n(X^n, Y^n) > d)} \sum_{i=1}^{M} \Pr((X^n, Y^n) \in \mathcal{S}_i \times \mathcal{S}_i^d) \\ &\le 2\sum_{i=1}^{M} \sum_{y^n\in\mathcal{S}_i^d} \sum_{x^n\in\mathcal{S}_i} \Pr\{X^n = x^n\} \Pr\{Y^n = y^n|X^n = x^n\} \end{aligned} \tag{178}$$

for sufficiently large n. Since $u^n(i) \in T^n_{U,\gamma_n}$ and $\mathcal{S}_i \subset T^n_{X|U,\gamma_n}(u^n(i))$ for each $1 \le i \le M$, it follows that for sufficiently large n and for any $x^n \in \mathcal{S}_i$,

$$\Pr(X^n = x^n) = 2^{-n(I(X\wedge U)+o(1))} \Pr(X^n = x^n|U^n = u^n(i))$$

which, together with (178) and the fact that $U \to X \to Y$ form a Markov chain, implies

$$P_{e2}(\mathcal{C}_n) \le 2^{2^{-n(I(X\wedge U)+o(1))}} \sum_{i=1}^{M} \sum_{y^n \in \mathcal{S}_i^d} \Pr(Y^n = y^n | U^n = u^n(i))$$

$$= 2^{-n(I(X\wedge U)+o(1))} \sum_{i=1}^{M} \Pr(Y^n \in \mathcal{S}_i^d | U^n = u^n(i)) \,, \tag{179}$$

where (U^n, X^n, Y^n) is n independent drawings of (U, X, Y). Let

$$d_0 = \mathbb{E}\bar{\rho}(P_{X|U}(\cdot|U), P_{Y|U}(\cdot|U)).$$

For convenience, we think of $\mathbb{E}\bar{\rho}(P_{X|U}(\cdot|U), P_{Y|U}(\cdot|U))$ as a function of $(P_U, P_{X|U}, P_{Y|U})$ which is denoted by $F(P_U, P_{X|U}, P_{Y|U})$. It is not hard to prove that this function is continuous. Since $d_0 > d$, there exists a $\sigma > 0$ such that for any $P \in \mathcal{P}(\mathcal{U})$ and any stochastic matrix $V = V(x|u)_{u\in\mathcal{U}, x\in\mathcal{X}}$,

$$||P - P_u|| \le \sigma, ||V - P_{X|U}|| \le \sigma \quad \Longrightarrow \quad F(P, V, P_{Y|U}) > \frac{d_0 + d}{2}, \tag{180}$$

where

$$||V - P_{X|U}|| = \sum_{u\in\mathcal{U}} ||V(\cdot|u) - P_{X|U}(\cdot|u)||.$$

Particularly, for sufficiently large n and for any $x^n \in \mathcal{S}_i$,

$$F(P_{u^n(i)}, P_{x^n|u^n(i)}, P_{Y|U}) > \frac{d_0 + d}{2}, \tag{181}$$

where $P_{x^n|u^n(i)} \in \mathcal{V}_n(P_{u^n(i)}, \mathcal{U} \times \mathcal{X})$ is the stochastic matrix so that x^n is $P_{x^n|u^n(i)}$-generated by $u^n(i)$, since $(u^n(i), x^n)$ is $(UX, 2\gamma_n)$-typical. To continue (179), let us note that if $\mathcal{S}_i^d \cap T_V^n(u^n(i), \mathcal{Y})$ is not empty, where $V \in \mathcal{V}_n(P_{u^n(i)}, \mathcal{U} \times \mathcal{Y})$, then there exists $x^n \in \mathcal{S}_i$ and $Q \in \mathcal{P}_n(\mathcal{U} \times \mathcal{X} \times \mathcal{Y})$ such that

(i) the marginal of Q on $\mathcal{U} \times \mathcal{X}$ is $P_{u^n(i)x^n}$;
(ii) the marginal of Q on $\mathcal{U} \times \mathcal{Y}$ is given by

$$P_{u^n(i)}(u)V(y|u), \quad u \in \mathcal{U} \ \text{and} \ y \in \mathcal{Y};$$

(iii) under the distribution Q, $\mathbb{E}\rho(X_0, Y_0) \le d$.

This implies

$$F(P_{u^n(i)}, P_{x^n|u^n(i)}, V) \leq d. \tag{182}$$

In view of (180), (181), and (182) implies

$$\sum_{u\in\mathcal{U}} P_{u^n(i)}(u) D(V(\cdot|u)||P_{Y|U}(\cdot|u)) > 3\delta_0, \tag{183}$$

where $\delta_0 > 0$ is a constant independent of i, $u^n(i)$ and V. Therefore, if $\mathcal{S}_i^d \cap T_V^n(u^n(i), \mathcal{Y})$ is not empty, then

$$\begin{aligned} &\Pr\{Y^n \in \mathcal{S}_i^d \cap T_V^n(u^n(i), \mathcal{Y})|U^n = u^n(i)\} \\ &\leq \Pr\{Y^n \in T_V^n(u^n(i), \mathcal{Y})|U^n = u^n(i)\} \leq 2^{-3n\delta_0} \end{aligned} \tag{184}$$

which in turn implies

$$\Pr\{Y^n \in \mathcal{S}_i^d|U^n = u^n(i)\} \leq |\mathcal{V}_n(P_{u^n(i)}, \mathcal{U}\times\mathcal{Y})|2^{-3n\delta_0} \leq 2^{-2n\delta_0} \tag{185}$$

for sufficiently large n. Substituting (185) into (179) yields,

$$P_{e2}(\mathcal{C}_n) \leq 2^{-n(2\delta_0-\delta+o(1))}.$$

Selecting $\delta < \delta_0$ so small that the right hand side of (177) is less than R. Accordingly,

$$P_{e2}(\mathcal{C}_n) \leq 2^{-n\delta_0} \tag{186}$$

for sufficiently large n. This shows that $(R, +\infty, \delta_0)$ is achievable and hence completes the proof of (176).

We next turn to proving

$$R_{XY}^*(+\infty, 0, d) \geq \tilde{R}(P_{XY}, 0, d). \tag{187}$$

By the definition of $R_{XY}^*(+\infty, 0, d)$, it suffices to prove that for any achievable triple $(R, +\infty, \beta)$,

$$R \geq \tilde{R}(P_{XY}, 0, d). \tag{188}$$

To this end, let us fix below an achievable triple $(R, +\infty, \beta)$. By definition, there exists for any $\epsilon > 0$ a sequence $\{\mathcal{C}_n\}$ of IDS codes, where $\mathcal{C}_n = (f_n, B_n, g_n)$ is an nth order IDS code, such that for sufficiently large n,

$$r_n(\mathcal{C}_n) \leq R + \epsilon,\ P_{e1}(\mathcal{C}_n) = 0,\ \ and\ \ P_{e2}(\mathcal{C}_n) \leq 2^{-n(\beta-\epsilon)}. \tag{189}$$

As what we did before, for each $b^n \in B_n$, let

$$\mathcal{S}(b^n) = \{x^n \in \mathcal{X}^n : f_n(x^n) = b^n\}.$$

Let $\hat{\mathcal{S}}^d(b^n)$ denote the set of all sequences $y^n \in \mathcal{Y}^n$ such that

$$\Pr(X^n \in \mathcal{S}(b^n) \cap B_d(y^n) | Y^n = y^n) > 0. \tag{190}$$

where

$$B_d(y^n) = \{x^n \in \mathcal{X}^n : \rho_n(x^n, y^n) \le d\}.$$

Clearly, $P_{e1}(\mathcal{C}_n) = 0$ implies

$$\hat{\mathcal{S}}^d(b^n) \subset \{y^n \in \mathcal{Y}^n : g_n(y^n, b^n) = 1\}.$$

From (189), therefore, it is not hard to see that

$$\sum_{b^n \in B_n} \Pr\{(X^n, Y^n) \in \mathcal{S}(b^n) \times \hat{\mathcal{S}}^d(b^n)\} \le 2^{-n(\beta-\epsilon)} + \Pr\{\rho_n(X^n, Y^n) \le d\}$$

which in turn implies

$$\sum_{b^n \in B_n} \Pr\{X^n \in \mathcal{S}(b^n)\} \sum_{x^n \in \mathcal{S}(b^n)} \frac{\Pr\{X^n = x^n\}}{\Pr\{X^n \in \mathcal{S}(b^n)\}} \Pr\{Y^n \in \hat{\mathcal{S}}^d(b^n) | X^n = x^n\} \to 0 \tag{191}$$

as n goes to infinity. Let B'_n consist of all $b^n \in B_n$ such that

$$\sum_{x^n \in \mathcal{S}(b^n)} \frac{\Pr(X^n = x^n)}{\Pr(X^n \in \mathcal{S}(b^n))} \Pr(Y^n \in \hat{\mathcal{S}}^d(b^n) | X^n = x^n) < \frac{\epsilon}{d}. \tag{192}$$

From (191) and the Markov inequality, for sufficiently large,

$$\sum_{b^n \in B'_n} \Pr(X^n \in \mathcal{S}(b^n)) > 1 - \epsilon.$$

Fix $b^n \in B'_n$ and consider $\mathcal{S}(b^n) \cap T^n_{X,\gamma_n}$. It is easy to see that there exists $A \subset \mathcal{S}(b^n) \cap T^n_{X,\gamma_n}$ such that $A \subset T^n_P(\mathcal{X})$ for some (X, γ_n)-essential P and

$$\frac{1}{n} \log \frac{|\mathcal{S}(b^n) \cap T^n_{X,\gamma_n}|}{|A|} \le |\mathcal{X}| \frac{\log n + 1}{n}. \tag{193}$$

From (192),

$$\sum_{x^n \in A} \frac{1}{|A|} \Pr(Y^n \in \hat{A}^d | X^n = x^n) \le \frac{\epsilon}{d} \tag{194}$$

where $\hat{A}^d$ is defined in the same way as $\hat{\mathcal{S}}^d(b^n)$was. Focusing on A, we define a random vector $\tilde{X}^n = (\tilde{X}_1, \cdots, \tilde{X}_n)$ taking values uniformly on A. Let $\tilde{Y}^n = (\tilde{Y}_1, \cdots, \tilde{Y}_n)$ be the output of the memoryless channel W resulting from passing $\tilde{X}^n$ through W. It is easy to verify that

$$\begin{aligned}
\frac{1}{n}\log|A| = \frac{1}{n}H(\tilde{X}^n) &= \frac{1}{n}H(\tilde{X}^n|\tilde{Y}^n) + \frac{1}{n}I(\tilde{X}^n;\tilde{Y}^n) \\
&= \frac{1}{n}\sum_{i=1}^{n} H(\tilde{X}_i|\tilde{X}^{i-1},\tilde{Y}^n) + \frac{1}{n}I(\tilde{X}^n;\tilde{Y}^n) \\
&\le \frac{1}{n}\sum_{i=1}^{n} H(\tilde{X}_i|\tilde{X}^{i-1},\tilde{Y}_i,\tilde{Y}_{i+1}^n) + \frac{1}{n}H(\tilde{Y}^n) - \frac{1}{n}\sum_{i=1}^{n} H(\tilde{Y}_i|\tilde{X}_i).
\end{aligned} \tag{195}$$

Let I be a RV taking values uniformly on $\{1, \cdots, n\}$ and independent of $\tilde{X}^n$ and $\tilde{Y}^n$. Let

$$\tilde{X} = \tilde{X}_I, \tilde{Y} = \tilde{Y}_I \text{ and } U = (\tilde{X}^{I-1}, \tilde{Y}_{I+1}^n, I). \tag{196}$$

Then (195) continues as follows

$$\frac{1}{n}\log|A| \le H(\tilde{X}|\tilde{Y},U) + I(\tilde{X}\wedge\tilde{Y}) = H(\tilde{X}|U) + I(U\wedge\tilde{Y}), \tag{197}$$

where the last step follows from the fact that $U \to \tilde{X} \to \tilde{Y}$ form a Markov chain. From (193) and (197), we now have

$$\begin{aligned}
\Pr(X^n \in \mathcal{S}(b^n) \cap T_{X,\gamma_n}^n) &\le |\mathcal{S}(b^n) \cap T_{X,\gamma_n}^n| 2^{-n(H(\tilde{X})+o(1))} \\
&\le \exp\{-n[I(\tilde{X}\wedge U) - I(\tilde{Y}\wedge U) - o(1)]\}.
\end{aligned} \tag{198}$$

Next we show that $U, \tilde{X}, \tilde{Y}$ satisfy

$$\mathbb{E}\bar{\rho}_e(P_{\tilde{X}|U}(\cdot|U)) > d - \epsilon. \tag{199}$$

To this end, Let $\mathcal{U}$ be the finite set on which U takes values, that is

$$\mathcal{U} = \{(x^{i-1}, y_{i+1}^n, i) : x^{i-1} \in \mathcal{X}^{i-1}, y_{i+1}^n \in \mathcal{Y}^{n-i}, 1 \le i \le n\}.$$

For each $u \in \mathcal{U}$, let $\hat{W}_u = (\hat{W}_u(y|x))$ be a stochastic matrix such that

(i) $P_{\tilde{X}|U}(\cdot|u)\hat{W}_u = P_{\tilde{X}|U}(\cdot|u)W$;
(ii) $\hat{W}_u$ is absolutely continuous with respect to W;
(iii)

$$\bar{\rho}_e(P_{\tilde{X}|U}(\cdot|u)) = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} P_{\tilde{X}|U}(x|u)\hat{W}_u(y|x)\rho(x, y).$$

Therefore,

$$\mathbb{E}\bar{\rho}_e(P_{\tilde{X}|U}(\cdot|U)) = \sum_{u\in\mathcal{U}, x\in\mathcal{X}, y\in\mathcal{Y}} P_U(u)P_{\tilde{X}|U}(x|u)\hat{W}_u(y|x)\rho(x, y).$$

We now write $\hat{W}_u(y|x)$ as $\hat{W}_i(y|x^{i-1}, y_{i+1}^n, x)$ whenever $u = (x^{i-1}, y_{i+1}^n, i)$. Think of $\hat{W}_i(\cdot|\cdot)$ as a channel $\mathcal{X}^{i-1} \times \mathcal{Y}^{n-i} \times \mathcal{X} \to \mathcal{Y}$ and construct a random vector $\hat{Y}^n = (\hat{Y}_1, \cdots, \hat{Y}_n)$ as follows:

Step 1. For $i = n$, $\hat{W}_n$ is from $\mathcal{X}^{n-1} \times \mathcal{X}$ to $\mathcal{Y}$. Pass $\tilde{X}^n$ (viewed as $(\tilde{X}^{n-1}, \tilde{X}_n)$) through $\hat{W}_n$ and denote the output by $\hat{Y}_n$;
Step 2. Pass $(\tilde{X}^{n-2}, \hat{Y}_n, \tilde{X}_{n-1})$ through $\hat{W}_{n-1}$, and denote the output by $\hat{Y}_{n-1}$;
Step i. So far, $\hat{Y}_{n-j}$ for $j = 0, \cdots, i-2$ have been constructed. Pass $(\tilde{X}_{n-i}, \hat{Y}_{n-i+2}^n, \tilde{X}_{n-i+1})$ through channel $\hat{W}_{n-i+1}$ and denote the output by $\hat{Y}_{n-i+1}$.
Continue this procedure until
Step n. Pass $(\hat{Y}_2^n, \tilde{X}_1)$ through the channel $\hat{W}_1$ and denote the output by $\hat{Y}_1$.

Since $P_{\tilde{X}|U}(\cdot|u)\hat{W}_u = P_{\tilde{X}|U}(\cdot|u)W$, from the above construction, we can see that for any $i : 1 \le i \le n$, $(\tilde{X}^{n-i}, \hat{Y}_{n-i+1}^n)$ has the same distribution as that of $(\tilde{X}^{n-i}, \tilde{Y}_{n-i+1}^n)$. From this, we obtain

$$\mathbb{E}\bar{\rho}_e(P_{\tilde{X}|U}(\cdot|U)) = \mathbb{E}\rho_n(\tilde{X}^n, \hat{Y}^n) = \mathbb{E}[\mathbb{E}(\rho_n(\tilde{X}^n, \hat{Y}^n)|\hat{Y}^n)], \tag{200}$$

where $\mathbb{E}(\cdot|\hat{Y}^n)$ denote the conditional expectation with respect to $\hat{Y}^n$. Since $\hat{W}_u$ is absolutely continuous with respect to W for any $u \in \mathcal{U}$, it follows from the construction of $\hat{Y}^n$ that $P_{\tilde{X}^n\hat{Y}^n}$ is also absolutely continuous with respect to $P_{\tilde{X}^n\tilde{Y}^n}$. Therefore, for any $y^n \notin \hat{A}^d$, if $\Pr(\hat{Y}^n = y^n) > 0$, or equivalently, $\Pr(\tilde{Y}^n = y^n) > 0$, then from the definition of $\hat{A}^d$, we have

$$\Pr(\rho_n(\tilde{X}^n, y^n) \le d|\hat{Y}^n = y^n) = 0$$

which implies

$$\mathbb{E}(\rho_n(\tilde{X}^n, \hat{Y}^n)|\hat{Y}^n = y^n) \ge d. \tag{201}$$

Note that (194) can be rewritten as

$$\Pr(\tilde{Y}^n \in \hat{A}^d) < \frac{\epsilon}{d}.$$

This, combined with (200) and (201), yields

$$\begin{aligned}\mathbb{E}\tilde{\rho}_e(P_{\tilde{X}|U}(\cdot|U)) &= \mathbb{E}[\mathbb{E}(\rho_n(\tilde{X}^n, \hat{Y}^n)|\hat{Y}^n)] \\ &\geq d\Pr(\hat{Y}^n \notin \hat{A}^d) \\ &= d\Pr(\tilde{Y}^n \notin \hat{A}^d) \\ &> d - \epsilon.\end{aligned}$$

Finally, let us go back to (198). In view of the definition of $\tilde{R}(P_{\tilde{X}\tilde{Y}}, 0, d)$, we have

$$\begin{aligned}\Pr(X^n \in \mathcal{S}(b^n) \cap T_{X,\gamma_n}) &\leq \exp\{-n\tilde{R}(P_{\tilde{X}\tilde{Y}}, 0, d-\epsilon) + o(n)\} \\ &\leq \exp\{-n\tilde{R}(P_{XY}, 0, d-\epsilon) + o(n)\}\,, \end{aligned} \tag{202}$$

where the last inequality is due to the fact that $P_{\tilde{X}}$ is (X, γ_n)-essential. Note that (202) holds for any $b^n \in B_n'$. In view of (189), we have

$$\begin{aligned}R + \epsilon &\geq r_n(\mathcal{C}_n) \\ &\geq \sum_{b^n \in B_n'} -\frac{1}{n}\Pr(X^n \in \mathcal{S}(b^n) \cap T_{X,\gamma}^n)\log\Pr(X^n \in \mathcal{S}(b^n) \cap T_{X,\gamma}^n) - \frac{1}{n} \\ &\geq (1 - \epsilon - \frac{|\mathcal{X}|}{4n\gamma_n^2})(\tilde{R}(P_{XY}, 0, d-\epsilon) - o(1)) - \frac{1}{n}.\end{aligned}$$

In view of Lemma 292, letting $n \to \infty$ and then letting $\epsilon \to 0$ yield

$$R \geq \tilde{R}(P_{XY}, 0, d).$$

This completes the proof of (188) and hence the proof of Theorem 293. □

Remark At this point, we point out the reason why the method used in Sect. 4 to prove Theorems 288 and 290 can not be generalized to the general case in which X and Y may be correlated. The main difficulty lies in the fact that even in the simplest case of $\alpha = +\infty$ and $\beta = 0$, the auxiliary RV U introduced in the proof of the lower bound of Theorem 293 involves both sets $\mathcal{X}$ and $\mathcal{Y}$.

Proof of Theorem 294 Clearly, we need to prove only

$$R^*_{(XZ)(YZ)}(+\infty, 0, d) \leq \tilde{R}(P_{(XZ)(YZ)}, 0, d). \tag{203}$$

By using the formula (62) for $\tilde{R}(P_{(XZ)(YZ)}, 0, d)$, an argument similar to the derivation of (176) can be used to show (203). □

6 Open Problems

The following problems remain open:

1. When X and Y are independent, Theorem 290 gives $R^*_{XY}(\alpha, \beta, d)$ for $0 \leq \beta < \beta(d)$. What happens if $\beta \geq \beta(d)$?
2. What is the counterpart of Theorem 290 in the general case in which X and Y may be correlated? This problem may be too difficult to solve.
3. An easier problem is the following: what is $R^*_{XY}(+\infty, 0, d)$ in the general case?
4. In this lecture, we considered the case when $d < \mathbb{E}\rho(X, Y)$. What happens if $d > \mathbb{E}\rho(X, Y)$? In the binary symmetric case, of course, the problem associated with $d > \mathbb{E}\rho(X, Y)$ is equivalent to that associated with $d < \mathbb{E}\rho(X, Y)$. In general, however, this is not true.

References

1. R. Ahlswede, I. Csiszar, To get a bit of information may be as hard as to get full information. IEEE Trans. Inf. Theory **IT-27**, 389–408 (1981)
2. R. Ahlswede, J. Körner, Source coding with side information and a converse for degrades broadcast channels. IEEE Trans. Inf. Theory, **IT-21**, 629–637 (1975)
3. R. Ahlswede, P. Gács, J. Körner, Bounds on conditional probabilities with applications in multiuser communication. Z. Wahrscheinlichkeitstheorie und verw. Geb. **34**, 157–177 (1976)
4. R. Ahlswede, E. Yang, Z. Zhang, Identification via compressed data. IEEE Trans. Inf. Theory **43**(1), 48–70 (1997)
5. T. Berger, *Rate Distortion Theory* (Prentice-Hall, Englewood Cliffs, 1971)
6. T.M. Cover, J.A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991)
7. I. Csiszar, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems* (Academic Press, New York, 1981)
8. R.M. Gray, D.L. Neuhoff, P.C. Shields, A generalization of Ornstein's $\bar{d}$-distance with application to information theory. Ann. Probab. **3**, 315–328 (1975)
9. B. Grunbaum, *Convex Polytopes* (Intersciences, New York, 1967)
10. L.H. Harper, Optimal numbering and isoperimetric problems on graphs. J. Comb. Theory **1**(3), 385–394 (1966)
11. M.S. Pinsker, *Information and Information Stability of Random Variables and Processes* (Holden-Day, San Francisco, 1964)

Part VI
Recent Results

New Results in Identification Theory

Holger Boche, Christian Deppe(✉), Wafa Labidi, and Roberto Ferrara

Machine-to-machine and human-to-machine communications are essential aspects incorporated in the framework of fifth generation wireless connectivity.

These new applications demand a strict adherence to end-to-end latency and robustness/reliability of a communication link. For a comprehensive discussion of these applications and their requirements concerning end-to-end latency, see [48].

As is shown in [48], the security for these applications and their necessary latency requirements must be embedded in the physical domain. Furthermore, for many of the applications discussed, the message transmission problem, as has been defined by Shannon [70], is too limiting. For this kind of communication, the receiver must be in a position to successfully decode all the messages from the sender.

To the contrary, as was discussed in [48], it is the communication task of identification best depicting the communication task in the new applications. The task of identification was introduced in [8] by R. Ahlswede and G. Dueck.

This survey is a supplement to the Lecture Notes on Identification Theory. Rudolf Ahlswede work on this lecture notes until his death in 2010. Since then, some new results and more applications of identification theory have emerged. The BMBF even supports one project titled "Post Shannon Communication". This project is about new communication models that are not implemented in the sense of the Shannon approach. Identification theory is a first example that new communication models lead to significant increases in performance. As shown in the book, in this scenario the receiver only wants to decide if the sender has sent a relevant message or not. Of course, the sender has no prior information about the messages that the recipient considers important. The relevance of certain messages to the recipient may be changed during the application. Ahlswede and Dueck have shown that there are identification codes with double exponential size in the block length of

The original version of this chapter was revised. A correction to this chapter can be found at https://doi.org/10.1007/978-3-030-65072-8_28

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_27

codewords, while in Shannon's transmission scheme the best possible codes have exponential size. The benefits offered by communication networks are limited by the ability of non-intended users (eavesdropping) or the manipulation of the sensitive data (jamming). These effects are different in the theory of identification than in Shannon's approach.

In this section we will be looking at various capacities defined following the same template. Namely, given a definition for a channel model and a channel W in this models, and given a definition of (n, M, ε) codes for these channels (where n is the blocklength, M the size and ε the error), then the pessimistic capacity of the channels under these codes is defined as

$$\begin{aligned}&\inf_{\varepsilon>0} \liminf_{n\to\infty} \frac{1}{n} \log M(W, n, \varepsilon)\\ &= \sup\left\{R : \forall\, \varepsilon > 0,\ \exists\, N,\ \forall\, n \geq N : \frac{1}{n} \log M(\mathcal{W}, n, \varepsilon) \geq R - \varepsilon\right\},\end{aligned}$$

where $M(W, n, \varepsilon)$ is the maximum M for which an (n, M, ε) code (as per the given definition) for W exists. The optimistic capacity is defined as

$$\begin{aligned}&\inf_{\varepsilon>0} \limsup_{n\to\infty} \frac{1}{n} \log M(W, n, \varepsilon)\\ &= \sup\left\{R : \forall\, \varepsilon > 0,\ \forall\, N,\ \exists\, n \geq N : \frac{1}{n} \log M(\mathcal{W}, n, \varepsilon) \geq R - \varepsilon\right\},\end{aligned}$$

The only exceptions are codes marked as ID codes, for which the pessimistic and optimistic capacity are achievable rates that scale double exponentially and thus are defined as

$$\inf_{\varepsilon>0} \liminf_{n\to\infty} \frac{1}{n} \log M(W, n, \varepsilon)$$

and

$$\inf_{\varepsilon>0} \limsup_{n\to\infty} \frac{1}{n} \log M(W, n, \varepsilon)$$

respectively.

Unless otherwise stated, the capacity is always implicitly taken to be the pessimistic capacity. Most importantly to remember, we will leave these steps implicit each time we define denote a capacity after defining a certain set codes.

1 Secure and Robust Identification Against Eavesdropping and Jamming Attacks

We present the main results established in [23–25]. Namely, we consider identification over robust (protection against systematic random errors), secure (protection against eavesdropping) channels, and the case if a jammer has influence to the channel. In this work, the identification capacity is compared to the transmission capacity of different channels and also in this case we have the double exponential advantage.

First, the effects are discussed by the example of a simple discrete memoryless channel. The channel model is then generalized to be include robustness. In information theory, robustness is modeled with a compound channel or, more generally, with an arbitrarily varying channel. Both models retain the effect of double exponential identification capacity. After that, security is considered and modeled with the help of a wiretapper. Here it turns out that there are even greater advantages of identification compared to transmission. While using a transmission code the capacity of a secure code depends directly on the channel to the wiretapper, for an identification code it is sufficient to know just that secure information can be transmitted. Once this is guaranteed, the channel can be used to identify the messages with the same capacity as the channel without wiretapper.

1.1 Compound Channels

In this section, the compound channel (CC) is discussed. In this model, the channel uncertainty is modeled by a given set of channels. The communication participants know this set, but they do not know which channel of this set describes the channel actually used. This compound channel model was introduced in [19] by Blackwell, Breiman, and Thomasian. It can be considered as a channel with state selector, choosing t and fixing it for all channel uses, but the sender and the receiver do not know his selection (see Fig. 1).

For the transmission capacity, the following result is obtained.

Theorem 299 ([19, 81]) *The transmission capacity of the CC $\mathcal{W}$ is*

$$C(\mathcal{W}) = \max_{P} \min_{W \in \mathcal{W}} I(P; W). \tag{1}$$

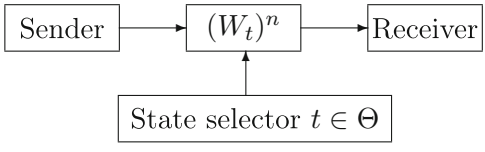


Fig. 1 The compound channel. Once t is selected W_t is used for the whole bolck

As in the case of the DMC, the CC has the same identification capacity as the transmission capacity.

Theorem 300 *The ID capacity of the CC $\mathcal{W}$ is*

$$C_{\mathrm{ID}}(\mathcal{W}) = C(\mathcal{W}). \tag{2}$$

1.2 Arbitrarily-Varying Channels

Another possibility to model the robustness in terms of information theory is with the arbitrarily-varying channel. This setting is also modeled by a given set of channels. The communication participants are aware of the state set but not of the actual realization of each state, which is determined probabilistically. The choice of state is arbitrary and can be thought as being controlled by a jammer, where at every time step the state of the channel can be changed by the jammer, and so that the sender and the receiver do not know the jamming strategy. Such a model is called an arbitrarily-varying channel (AVC) (Fig. 2).

It is to note that for an AVC there is the possibility that the channel is symmetrizable. The intuitive meaning of this is that the jammer can choose the state of the channel such that any two codewords, x and x', may be confused by the receiver.

Definition 301 An AVC $\mathcal{W} : \mathcal{X} \times \mathcal{S} \to \mathcal{P}(\mathcal{Y})$ is symmetrizable if there exists a channel $U : \mathcal{X} \to \mathcal{P}(\mathcal{S})$, such that for all $x, x' \in \mathcal{X}$ $y \in \mathcal{Y}$

$$\sum_s U(s|x')W(y|x,s) = \sum_s U(s|x)W(y|x',s). \tag{3}$$

In this situation, the decoder will be unable to tell if the transmitted codeword was x or x'. When a channel is symmetrizable, it is not possible to transmit or identify a message.

Furthermore, to give the capacity of an AVC we need the concept of a correlated code, where the sender and the receiver have access to some source with correlated (or common) randomness. The capacity of the AVC $\mathcal{W}$ using a correlated random

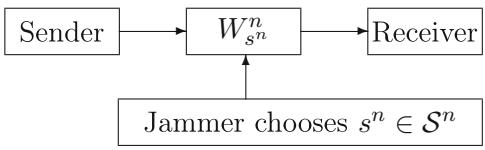


Fig. 2 The arbitrarily-varying channel

code is called correlated random coding capacity and denoted by

$$C_{\text{ran}}(\mathcal{W}) = \max_P \min_S I(P; S\mathcal{W}) = \min_S \max_P I(P; S\mathcal{W})$$

where we generalize the notation $PW \in \mathcal{P}(\mathcal{Y})$ to the case $S\mathcal{W} : \mathcal{X} \to \mathcal{P}(Y)$, where the probability distribution is only over part of the input of the channel and produces the channel $S\mathcal{W}(y|x) = \sum_s P_S(s)\mathcal{W}(y|x, s)$. The correlated randomness codes can be derandomized whenever the normal capacity is non-zero. Therefore the normal capacity C, without correlated randomness, of the AVC is discontinuous, as it is either zero or exactly equal to C_{ran}. In [47], Ericson showed that for positivity of the capacity it is necessary that $\mathcal{W}$ be non-symmetrizable, and in [43] it was proven to also be sufficient. Namely, the from all the above the following theorem holds.

Theorem 302 ([43]) *The transmission capacity of the AVC $\mathcal{W}$ is*

$$C(\mathcal{W}) = \begin{cases} 0 & \text{if } \mathcal{W} \text{ is symmetrizable} \\ C_{ran}(\mathcal{W}) & \text{otherwise.} \end{cases} \tag{4}$$

Notice that examples of symmetrizabe AVCs with positive C_{ran} exist and can be found for example in [42].

Again, it could be shown that in this case the identification capacity is equal to the transmission capacity.

Theorem 303 ([7]) *The ID capacity of the AVC $\mathcal{W}$ is*

$$C_{ID}(\mathcal{W}) = C(\mathcal{W}) \tag{5}$$

1.3 Compound Wiretap Channels

The two robust models from the previous section are considered again and a wiretapper is added to the model. First, the case of the compound channel is discussed. As before, a state selector chooses t, and the sender and the receiver do not know his selection, but the wiretapper does. This corresponds to an active attack by the state selector and passive wiretapping. The previously introduced model allows for joining state selector and wiretapper. In this view the wiretapper chooses t and the code has to work independently of the choice of t. For that reason, this is called an active attack.

Definition 304 Let $\Theta = \{1, \ldots, T\}$ and $\Sigma = \{1, \ldots, S\}$ be finite index sets. A discrete memoryless compound wiretap channel (CWC) is a quintuple $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{W}, \mathcal{V})$, where $\mathcal{X}$ is the finite input alphabet, $\mathcal{Y}$ is the finite output alphabet for the legitimate receiver, $\mathcal{Z}$ is the finite output alphabet for the wiretapper, $\mathcal{W} = (W_t)_{t\in\Theta}$ is the CC whose output is available to the legitimate receiver, and

$\mathcal{V} = (V_s)_{t\in\Sigma}$ the CC whose output is available to the wiretapper. The channel is assumed to be memoryless.

Theorem 305 ([17]) *The secrecy capacity C_S of a CWC $(\mathcal{W}, \mathcal{V})$ is given by*

$$C_S(\mathcal{W}, \mathcal{V}) = \lim_{n\to\infty} \frac{1}{n} \max_{U\to X^n\to Y_t^n Z_s^n} \left(\min_{t\in\Theta} I(U \wedge Y_t^n) - \max_{s\in\Sigma} I(U \wedge Z_s^n) \right), \qquad (6)$$

where Y_t is the resulting random variable at the output of the intended receiver channel and Z_s is the resulting random variable at the output of the wiretap channel, if the channel is (W_t, V_s). The maximum is taken over all random variables U that satisfy the Markov chain relationships $U \to X^n \to Y_t^n Z_s^n$.

Theorem 306 ([24]) *The ID capacity of the CWC is*

$$C_{SID}(\mathcal{W}, \mathcal{V}) = \begin{cases} C(\mathcal{W}) & \text{if } C_S(\mathcal{W}, \mathcal{V}) > 0, \\ 0 & \text{if } C_S(\mathcal{W}, \mathcal{V}) = 0. \end{cases}$$

1.4 Arbitrarily-Varying Wiretap Channels

In the case of the AVWC, symmetrizability also plays a role in the secure transmission and ID capacity, denoted C_S and C_{SID}.

The secret transmission capacity of codes using common randomness is given by

$$C_{S,ran}(\mathcal{W}, \mathcal{V}) = \lim_{n\to\infty} \frac{1}{n} \sup_{U\to X^n\to Y_t^n Z_{s^n}^n} \left(\min_{t\in\mathcal{P}(\Theta)} I(U \wedge Y_t^n) - \max_{s^n\in\mathcal{S}^n} I(U \wedge Z_{s^n}^n) \right)$$

This capacity gives the secret transmission capacity in case of non-symmetrizable AVC at the intended receiver.

Theorem 307 ([62]) *The secret transmission capacity of the AVWC is*

$$C_S(\mathcal{W}, \mathcal{V}) = \begin{cases} 0 & \text{if } \mathcal{W} \text{ is symmetrizable} \\ C_{S,ran}(\mathcal{W}, \mathcal{V}) & \text{otherwise.} \end{cases}$$

The following dichotomy (introduced in Theorem 87) result is obtained.

Theorem 308 ([25]) *The secret ID capacity of the AVWC is*

$$C_{SID}(\mathcal{W}, \mathcal{V}) = \begin{cases} C_{ran}(\mathcal{W}) & \text{if } C_{S,ran}(\mathcal{W}, \mathcal{V}) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

2 Classical-Quantum Channels

In this section, we present the most important results cited in [26].

Let $\mathcal{H}$ be a Hilbert space, we denote by $\mathcal{B}(\mathcal{H})$ the set of bounded operators, by $\mathcal{S}(\mathcal{H})$ the set of quantum states (bounded positive operators of trace one). All Hilbert spaces here will be finite dimensional, we can then denote by $\mathcal{P}(\mathcal{H})$ the set of pure quantum states (rank-1/projector states).

A POVM defines the most general measurement map from a quantum state to a classical random variable. As such POVMs also describe the most general form of decoder of classical information from a quantum state.

Definition 309 Let $\mathcal{H}$ be a finite dimensional Hilbert space. A *POVM (positive operator valued measure)* on $\mathcal{H}$ is a collection $(D_i)_{i=1}^N$ of positive semidefinite operators D_i on $\mathcal{H}$ such that $\sum_{i=1}^N D_i = \mathbb{1}_{\mathcal{H}}$, where $\mathbb{1}_{\mathcal{H}}$ denotes the identity operator on $\mathcal{H}$.

In identification theory the goal for Bob is changed: Assume that he "only" wants to know if the transmitted message is equal to some j. Instead of having a single decoder POVM, we have now a binary POVM for each message.

Definition 310 A *set of measurement operators* $\{D_i\}$, i.e. $0 \leq D_i \leq \mathbb{1}$ but not necessarily a POVM, is called *simultaneous* if there exist a POVM, called *generating POVM,* such that each D_i can be written as a sum of some the POVM's elements.

Each operator D_i implicitly define a binary POVM, namely $\{D_i, \mathbb{1} - D_i\}$ for each message i. A simultaneous set of measurement operators $\{D_i\}_i$ means that for each i the POVM $\{D - i, \mathbb{1} - D_i\}$ can be performed as classical post-processing of the generating POVM.

2.1 Classical-Quantum Channels

In [26], the identification capacity of classical-quantum channels ("cq-channels") under channel uncertainty and privacy constraints is studied. To be precise, the compound memoryless cq-channels are first considered and their identification capacity is determined; then an eavesdropper is added by considering compound memoryless wiretap cqq-channels, and their secret identification capacity is determined. In the first case (without privacy), Boche, Deppe and Winter find that the identification capacity always equal to the transmission capacity. In the second case, they find the same dichotomy of classical channels: either the secrecy capacity (also known as private capacity) of the channel is zero, and then the secrecy identification capacity is also zero, or the secrecy capacity is positive and then the secrecy identification capacity equals the transmission capacity of the main channel without the wiretapper.

First some basic definitions related to cq-channels are introduced. Cq-channels have a classical sender, having access to an input alphabet $\mathcal{X}$, but their output is quantum, being described by a Hilbert space $\mathcal{H}$.

Definition 311 A *discrete classical-quantum channel (cq-channel)* is a map $W : \mathcal{X} \longrightarrow \mathcal{S}(\mathcal{H})$ where $\mathcal{X}$ is a finite set and $\mathcal{S}(\mathcal{H})$ is the set of quantum states of the complex Hilbert space $\mathcal{H}$, which we assume to be finite dimensional. Furthermore, we denote $a = |\mathcal{X}|$ the cardinality of $\mathcal{X}$, and $d = |\mathcal{H}|$ the dimension of $\mathcal{H}$.

Given a probability distribution $P \in \mathcal{P}(\mathcal{X})$ we define the state:

$$PW \equiv W(P) \triangleq \sum_{x \in \mathcal{X}} P(x)W(x).$$

This is the basic definition upon which all the other models of classical-quantum channels are built upon.

2.2 Wiretap Classical-Quantum Channels

An important aspect in information theory is security, or privacy. Wyner [82] introduced the classical wiretap channel, which he solved in the degraded case, and later Csiszár and Körner [41] in the general case. It can be described by two channels from the sender ("Alice") to the intended receiver ("Bob") and to the eavesdropper ("Eve"), respectively. The wiretap channel was generalized to the setting of quantum information theory in [38, 45]. Formally, in contrast to the classical case, quantumly the channel has to be described by a single quantum operation T, from Alice to the joint system of Bob and Eve together: then the intended channel $W = \mathrm{Tr}_B \circ T$ and the wiretapper channel $V = \mathrm{Tr}_E \circ T$ are defined in [26]. Here, only one case of the cq-channel is considered, where Alice's input is described by a letter $x \in \mathcal{X}$ from a finite alphabet. Then the classical-quantum wiretap channel is defined in a simple way.

Definition 312 A *classical-quantum wiretap channel (wiretap cqq-channel)* is a pair (W, V) of two discrete memoryless cq-channels $W : \mathcal{X} \longrightarrow \mathcal{S}(B)$ and $V : \mathcal{X} \longrightarrow \mathcal{S}(E)$. When Alice sends a classical input $x^n \in \mathcal{X}^n$, Bob (intended receiver) and Eve (eavesdropper) receive the states $W^{\otimes n}(x^n)$ and $V^{\otimes n}(x^n)$, respectively.

Definition 313 An (n, M, ε, μ) *wiretap transmission code* for the wiretap cqq-channel (W, V) is a collection $\{(P_i, D_i) : i \in [M]\}$ of pairs consisting of probability distributions P_i on $\mathcal{X}^n$ and a POVM $(D_i)_{i=1}^N$ on B^n such that

$$\forall\, i \in [M] \qquad \mathrm{Tr}\, W^{\otimes n}(P_i)D_i \geq 1 - \varepsilon,$$

$$\forall\, i, j \in [M] \qquad \frac{1}{2}\|V^{\otimes n}(P_i) - V^{\otimes n}(P_j)\|_1 \leq \mu.$$

We denote by $C_S(W, V)$ the capacity of the wiretap cq-channel achieved by the wiretap transmision codes.

Definition 314 An (n, M, ε) *wiretap (simultaneous) ID code* for the wiretap cqq-channel (W, V) is a collection $\{(P_i, D_i) : i \in [M]\}$ of pairs consisting of probability distributions P_i on $\mathcal{X}^n$ and (simultaneous) measurement operators D_i on B^n such that

$$\begin{aligned} \forall\, i \in [M] \qquad & \operatorname{Tr} W^{\otimes n}(P_i) D_i \geq 1 - \varepsilon, \\ \forall\, i \neq j \in [M] \qquad & \operatorname{Tr} W^{\otimes n}(P_i) D_j \leq \varepsilon, \\ \forall\, i, j \in [M] \qquad & \frac{1}{2} \| V^{\otimes n}(P_i) - V^{\otimes n}(P_j) \|_1 \leq \varepsilon. \end{aligned}$$

We denote by $C_{\mathrm{SID}}(W, V)$ ($C_{\mathrm{SID}}^{\mathrm{sim}}(W, V)$) the capacity of the wiretap cqq-channel achieved by the wiretap (simultaneous) ID codes.

Theorem 315 ([38]) *The secrecy capacity of a wiretap cqq-channel is given by*

$$C_S(W, V) = \lim_{n \to \infty} \max_{U \to X^n \to B^n E^n} \frac{1}{n} \Big(I(U : B^n) - I(U : E^n) \Big),$$

where the maximum is taken over all random variables that satisfy the Markov chain relationships $U \to X^n \to B^n E^n$.

The wiretap cqq-channel is considered and a multi-letter formula for its secure identification capacity is derived. The idea is similar to the classical case. A combination of two codes is used. For the converse, inequalities of [10] and [49] are generalized.

Theorem 316 ([24], Dichotomy Theorem) *Let $C(W)$ be the capacity of the cq-channel W and let $C_S(W, V)$ be the secrecy capacity of the wiretap cqq-channel. Then,*

$$C_{SID}(W, V) = C_{SID}^{sim}(W, V) = \begin{cases} C(W) & \text{if } C_S(W, V) > 0, \\ 0 & \text{if } C_S(W, V) = 0. \end{cases}$$

2.3 Compound Classical-Quantum Channels

We will consider robust codes against compound cq-channels. The results for cq-channels follow as special cases.

Definition 317 Let Θ be an index set, $\mathcal{X}$ a finite set and $\mathcal{H}$ a finite-dimensional Hilbert space. Let $W_t : \mathcal{X} \longrightarrow \mathcal{S}(\mathcal{H})$ be a cq-channel for every $t \in \Theta$:

$$W_t : x \mapsto W_t(x) \in \mathcal{S}(\mathcal{B}).$$

If the memoryless extension of the cq-channel W_t is given by $W_t(x^n) = W_t^{\otimes n}(x^n) = W_t(x_1) \otimes \ldots \otimes W_t(x_n)$ for $x^n \in \mathcal{X}^n$ then we call $\mathcal{W} = \{W_t\}_{t\in\Theta}$ a *compound cq-channel*.

In this case t model a channel that can change from one session to another (between different blocks of length n). The special case of regular cq-channels is recovered for $|\Theta| = 1$.

In transmission, Alice uses the classical-quantum channel to transmit messages from the set $\mathcal{X}$ to Bob. Bob tries to find out the transmitted messages by measuring with a POVM.

Definition 318 An (n, M, ε) *transmission code* for the compound cq-channel $\mathcal{W}$ is a family $\mathcal{C} := \{(P_m, D_m) : m \in [M]\}$ consisting of stochastic encodings given by probability distributions P_m over $\mathcal{X}^n$ and a POVM $\{D_i\} \subset \mathcal{B}(B^{\otimes n})$, such that

$$\forall\, t \in \Theta,\ \forall\, i \in [M] \qquad\qquad \operatorname{Tr} W_t^{\otimes n}(P_i) D_i \geq 1 - \varepsilon,$$

where M is the *size* of the code and ε the error probability.

We denote $C(\mathcal{W})$ the capacity of the compound cq-channel achieved by the transmission codes.

In identification, Alice uses the classical-quantum channel to encode from the set $\mathcal{X}$ to Bob. Bob test for the message he is interested in with a binary measurement, i.e a POVM $\{D, \mathbb{1} - D\}$. Thus, the collection of all measurement operators used to test for each identification message does not need to be a POVM.

Definition 319 An (n, N, ε) *(simultaneous) ID code* for the compound channel

$$\mathcal{W} = \{W_t : \mathcal{X} \to \mathcal{S}(\mathcal{H})\}_{t\in\Theta}$$

is a set of pairs $\{(P_i, D_i) : i \in [N]\}$ where the P_i are probability distributions on $\mathcal{X}^n$ and the D_i are (simultaneous) measurement operators on $\mathcal{H}^{\otimes n}$, such that

$$\forall\, t \in \Theta,\ i \in [M] \qquad\qquad \operatorname{Tr} W_t^{\otimes n}(P_i) D_i \geq 1 - \varepsilon$$

$$\forall\, t \in \Theta,\ i \neq j \in [M] \qquad\qquad \operatorname{Tr} W_t^{\otimes n}(P_i) D_j \leq \varepsilon.$$

We denote $C_{\mathrm{ID}}(\mathcal{W})$ ($C_{\mathrm{ID}}^{\mathrm{sim}}(\mathcal{W})$) the capacity of the compound cq-channel achieved by the (simultaneous) ID codes.

Theorem 320 ([16]) *Let $\mathcal{W}$ be a compound cq-channel. Then,*

$$C(\mathcal{W}) = \inf_{t\in\Theta} \sup_{Q} I(Q; W_t).$$

Theorem 321 ([26]) *Let $\mathcal{W}$ be a compound cq-channel. Then,*

$$C_{\mathrm{ID}}(\mathcal{W}) = C_{\mathrm{ID}}^{sim}(\mathcal{W}) = C(\mathcal{W}).$$

The weak converse holds also for the optimistic ID capacity:

$$\overline{C}_{\mathrm{ID}}(\mathcal{W}) \triangleq \inf_{\varepsilon>0} \limsup_{n\to\infty} \frac{1}{n} \log N(n, \varepsilon) = C_{\mathrm{ID}}(\mathcal{W}).$$

2.4 Compound Wiretap Classical-Quantum Channels

In Theorem 326, the construction for the more general compound model is described. The same idea is used here to show the direct part: Alice and Bob first create shared randomness at a rate equal to the channel capacity. A code with an arbitrary small positive rate is then sufficient to use the method of Ahlswede and Dueck by sending and decoding the function values.

Definition 322 Let Θ and Σ be an index sets and let $\mathcal{W} = \{W_t : \mathcal{X} \to \mathcal{S}(\mathcal{B}) : t \in \Theta\}$ and $\mathcal{V} = \{V_s : \mathcal{X} \to \mathcal{S}(E) : s \in \Sigma\}$ be compound cq-channels. We call the pair $(\mathcal{W}, \mathcal{V})$ a compound wiretap cqq-channel. The channel output of $\mathcal{W}$ is available to the legitimate receiver (Bob) and the channel output of $\mathcal{V}$ is available to the wiretapper (Eve).

Definition 323 An (n, M, ε) wiretap transmission code for the compound wiretap cqq-channel $(W_t, V_s)_{t\in\Theta, s\in\Sigma}$ consists of a family $\mathcal{C} = (P_i, D_i)_{i\in[M]}$ where the P_i are probability distributions on $\mathcal{X}^n$ and $(D_i)_{i\in[M]}$ a POVM on $\mathcal{B}^{\otimes n}$ such that

$$\forall\, t \in \Theta,\ i \in [M] \qquad \mathrm{Tr}\, W^{\otimes n}(P_i) D_i \geq 1 - \varepsilon,$$

$$\forall\, s \in \Sigma,\ i, j \in [M] \qquad \frac{1}{2} \| V_s^{\otimes n}(P_i) - V_s^{\otimes n}(P_j) \|_1 \leq \mu.$$

The capacity is defined as before.

Definition 324 An (n, N, ε) *wiretap (simultaneous) ID code* for the compound wiretap cqq-channel $(\mathcal{W}, \mathcal{V})$ is a set of pairs $\{(P_i, D_i) : i \in [N]\}$ where the P_i are probability distributions on $\mathcal{X}^n$ and the D_i are (simultaneous) measurement

operators on $B^{\otimes n}$ such that,

$$\begin{aligned}
&\forall\, t \in \Theta,\ i \in [M] && \operatorname{Tr} W_t^{\otimes n}(Q_i) D_i \geq 1-\varepsilon,\\
&\forall\, t \in \Theta,\ i \neq j \in [M] && \operatorname{Tr} W_t^{\otimes n}(Q_j) D_i \leq \varepsilon,\\
&\forall\, s \in \Sigma,\ i, j \in [M] && \frac{1}{2}\| V_s^{\otimes n}(Q_j) - V_s^{\otimes n}(Q_i)\|_1 \leq \varepsilon. \qquad (7)
\end{aligned}$$

We denote by $C_{\mathrm{SID}}(\mathcal{W},\mathcal{V})$ ($C_{\mathrm{SID}}^{\mathrm{sim}}(\mathcal{W},\mathcal{V})$) the capacity of the compound wiretap cqq-channel achieved by the wiretap (simultaneous) ID codes.

Again a dichotomy result is obtained.

Theorem 325 ([21]) *Let $(\mathcal{W},\mathcal{V})$ be a compound wiretap cqq-channel. Then*

$$C_S(\mathcal{W},\mathcal{V}) = \lim_{n\to\infty} \sup_{U\to X^n\to(B_t^n E_s^n)} \frac{1}{n}\left(\inf_{t\in\Theta} I(U;B_t^n) - \sup_{s\in\Sigma} I(U;E_s^n)\right).$$

Theorem 326 ([26]) *Let $(\mathcal{W},\mathcal{V})$ be a compound wiretap cqq-channel. Then,*

$$C_{SID}(\mathcal{W},\mathcal{V}) = C_{SID}^{sim}(\mathcal{W},\mathcal{V}) = \begin{cases} C(\mathcal{W}) & \text{if } C_S(\mathcal{W},\mathcal{V}) > 0,\\ 0 & \text{if } C_S(\mathcal{W},\mathcal{V}) = 0. \end{cases}$$

2.5 Arbitrarily-Varying Classical-Quantum Channels

Now, the same analysis is performed for the case of arbitrarily-varying cq-channels, with analogous findings.

Definition 327 We say that the arbitrarily-varying cq-channel $\mathcal{W} = \{W_t : t \in \Theta\}$ is *symmetrizable* if there exists a parametrized set of distributions $\{\tau(\cdot|x) : x \in \mathcal{X}\}$, on Θ also known as a channel τ from $\mathcal{X}$ to Θ, such that for all $x, x' \in \mathcal{X}$,

$$\sum_{t\in\Theta} \tau(t|x) W_t(x') = \sum_{t\in\Theta} \tau(t|x') W_t(x).$$

It is worth noting that from now only *finite* index sets Θ are considered.

Definition 328 Let Θ be a finite index set, $\mathcal{X}$ a finite set and B a finite-dimensional Hilbert space. Let $W_t : \mathcal{X} \longrightarrow \mathcal{S}(B)$ be a cq-channel for every $t \in \Theta$:

$$W_t : \mathcal{X} \ni x \mapsto W_t(x) \in \mathcal{S}(B), \quad t \in \Theta.$$

Let $t^n \in \Theta^n$ be a state sequence. If the memoryless extension of the cq-channel W_{t^n} is given by $W_{t^n}(x^n) = W_{t_1}(x_1) \otimes \ldots \otimes W_{t_n}(x_n)$ for $x^n \in \mathcal{X}^n$, then we call $\mathcal{W} = \{W_t\}_{t\in\Theta}$ *an arbitrarily-varying cq-channel*.

In this case t^n models a jammer that can change the channel during the transmission. Again, non-compound or non-arbitrarily-varying channels are recovered for $|\Theta| = 1$. Like in the compound case, here we allow explicitly stochastic encoders.

Definition 329 An (n, M, λ) *transmission code* for the arbitrarily-varying cq-channel $\mathcal{W}$ is a family $((P_m, D_m) : m \in [M])$ consisting of probability distributions P_m over $\mathcal{X}^n$ and a POVM $\{D_i\}$ over $B^{\otimes n}$, such that for all

$$\forall\, t^n \in \Theta^n,\ \forall\, i \in [M] \qquad\qquad \operatorname{Tr} W_{t^n}^{\otimes n}(P_i) D_i \geq 1 - \varepsilon.$$

We denote by $C(\mathcal{W})$ the capacity of the arbitrarily-varying cq-channel achieved by the transmission codes.

Definition 330 An (n, M, ε) *(simultaneous) ID code* for the arbitrarily-varying cq-channel $\mathcal{W}$ is a family $((P_m, D_m) : m \in [M])$ where P_m are probability distributions over $\mathcal{X}^n$ and the D_i are (simultaneous) measurement operators on $B^{\otimes n}$, such that for all

$$\forall\, t^n \in \Theta^n,\ \forall\, i \in [M] \qquad\qquad \operatorname{Tr} W_{t^n}^{\otimes n}(P_i) D_i \geq 1 - \varepsilon,$$
$$\forall\, t^n \in \Theta^n,\ \forall\, i \neq j \in [M] \qquad\qquad \operatorname{Tr} W_{t^n}^{\otimes n}(P_i) D_j \leq \varepsilon.$$

We denote by $C_{\mathrm{ID}}(\mathcal{W})$ ($C_{\mathrm{ID}}^{\mathrm{sim}}(\mathcal{W})$) the capacity of the arbitrarily-varying cq-channel achieved by the (simultaneous) ID codes.

Furthermore, we set

$$C_{\mathrm{ran}}(\mathcal{W}) := \max_{P\in\mathcal{P}(\mathcal{X})} \min_{T\in\mathcal{P}(\Theta)} I(P; T\mathcal{W}),$$

where T is the probability distribution of the jammer input. This is called the random coding capacity of the channel. Under this notion, the encoding with a stochastic encoder is generalized to a common-randomness code. It is assumed that the sender and the receiver have access to some source of common randomness, *which, however, is secret from the jammer*. Examples of symmetrizable channels with non-zero C_{ran} can be found in [27].

Theorem 331 ([4]) *Let $\mathcal{W}$ be an arbitrarily-varying cq-channel. Then its capacity $C(\mathcal{W})$ is given by*

$$C(\mathcal{W}) = \begin{cases} 0 & \text{if } \mathcal{W} \text{ is symmetrizable,} \\ C_{ran}(\mathcal{W}) & \text{otherwise.} \end{cases}$$

With the help of the method from Theorem 321, it is shown in [26] that the transmission capacity of the channel corresponds to the identification capacity. To do this, in the proof of the direct part a code for an arbitrary varying cq-channel is simply used instead of the code for the compound cq-channel. To show the converse, it is shown that the error of the first type in the identification can not be arbitrarily small if the channel is symmetrizable. Therefore, the following theorem is obtained.

Theorem 332 ([26]) *Let $\mathcal{W}$ be an arbitrarily-varying cq-channel. Then its ID capacity is given by*

$$C_{\mathrm{ID}}^{sim}(\mathcal{W}) = C_{\mathrm{ID}}(\mathcal{W}) = C(\mathcal{W})$$

2.6 Arbitrarily-Varying Wiretap Classical-Quantum Channels

Now, a wiretapper is added to the arbitrarily-varying cq-channel. First the transmission codes are defined.

Definition 333 Let Θ and Σ be finite index sets, and let $\mathcal{W} = \{W_t : \mathcal{X} \to \mathcal{S}(B) : t \in \Theta\}$ and $\mathcal{V} = \{V_s : \mathcal{X} \to \mathcal{S}(E) : s \in \Sigma\}$ be arbitrarily-varying cq-channels. We call the pair $(\mathcal{W}, \mathcal{V})$ an *arbitrarily-varying wiretap cqq-channel*. The channel output of $\mathcal{W}$ is available to the legitimate receiver (Bob) and the channel output of $\mathcal{V}$ is available to the wiretapper (Eve). We may sometimes write the channel as a family of pairs $(\mathcal{W}, \mathcal{V}) = (W_t, V_s)_{t\in\Theta, s\in\Sigma}$.

Then the secure transmission capacity is determined. Using this result, the secure identification capacity of the arbitrarily-varying wiretap cq-channel could be then calculated.

Definition 334 An (n, M, λ) wiretap transmission code for the arbitrarily-varying wiretap cqq-channel $(W_t, V_s)_{t\in\Theta, s\in\Sigma}$ consists of a family $\{P_i, D_i\}_{i\in[M]}$, where the P_i are probability distributions on $\mathcal{X}^n$ and $\{D_i\}_{i\in[M]}$ a POVM on $B^{\otimes n}$ such that

$$\forall\, t^n \in \Theta^n,\ \forall\, i \in [M] \qquad \operatorname{Tr} W_{t^n}^{\otimes n}(P_i) D_i \geq 1 - \lambda, \tag{8}$$

$$\forall\, s^n \in \Sigma^n,\ \forall\, i, j \in [M] \qquad \frac{1}{2}\| V_{s^n}^{\otimes n}(P_i) - V_{s^n}^{\otimes n}(P_j)\|_1 \leq \varepsilon. \tag{9}$$

We denote by $C(\mathcal{W}, \mathcal{V})$ the capacity of the arbitrarily-varying wiretap cq-channel achieved by the wiretap transmission codes.

Definition 335 An (n, M, λ) wiretap (simultaneous) ID code for the arbitrarily-varying wiretap cqq-channel $(W_t, V_s)_{t\in\Theta, s\in\Sigma}$ consists of a family $\{P_i, D_i\}_{i\in[M]}$, where P_m are probability distributions over $\mathcal{X}^n$ and the D_i are (simultaneous)

measurement operators on $B^{\otimes n}$ such that

$$\forall\, t^n \in \Theta^n,\ \forall\, i \in [M] \qquad \operatorname{Tr} W_{t^n}^{\otimes n}(P_i) D_i \geq 1 - \lambda, \tag{10}$$

$$\forall\, t^n \in \Theta^n,\ \forall\, i \neq j \in [M] \qquad \operatorname{Tr} W_{t^n}^{\otimes n}(P_i) D_j \leq \varepsilon. \tag{11}$$

$$\forall\, s^n \in \Sigma^n,\ \forall\, i, j \in [M] \qquad \frac{1}{2} \| V_{s^n}^{\otimes n}(P_i) - V_{s^n}^{\otimes n}(P_j) \|_1 \leq \varepsilon. \tag{12}$$

We denote by $C_{\mathrm{ID}}(\mathcal{W}, \mathcal{V})$ ($C_{\mathrm{ID}}^{\mathrm{sim}}(\mathcal{W}, \mathcal{V})$) the capacity of the arbitrarily-varying wiretap cq-channel achieved by the (simultaneous) wiretap ID codes.

To state the result of [22] we again introduce the random coding capacity,

$$C_{\mathrm{S,ran}}(\mathcal{W}, \mathcal{V}) \triangleq \lim_{n \to \infty} \frac{1}{n} \max_{U \to X^n \to B_{t^n}^n E_{s^n}^n} \left(\min_{\widehat{W} \in \operatorname{conv}\{W_{t^n}^{\otimes n}\}} I(P; \widehat{W}) - \max_{s^n \in \Sigma^n} I(P; V_{s^n}^{\otimes n}) \right).$$

Here, $B_{t^n}^n$ are the resulting quantum states at the output of the legitimate receiver's channels. $E_{s^n}^n$ are the resulting quantum states at the output of the wiretap channels.

In [22] the following dichotomy is shown.

Theorem 336 ([22]) *Let $(\mathcal{W}, \mathcal{V})$ be an arbitrarily-varying wiretap cqq-channel. Then,*

$$C_{\mathrm{S}}(\mathcal{W}, \mathcal{V}) = \begin{cases} 0 & \text{if } \mathcal{W} \text{ is symmetrizable,} \\ C_{S,ran}(\mathcal{W}, \mathcal{V}) & \text{otherwise.} \end{cases}$$

Again a dichotomy result is shown. We use the idea of Theorem 326. As fundamental codes we use a code C' for the arbitrarily-varying cq-channel and a code C'' for the arbitrarily-varying wiretap cqq-channel, both reaching the capacity. If the transmission capacity for $\mathcal{W}$ is positive, one gets as an identification capacity the transmission capacity of $\mathcal{W}$. The security follows by the strong secrecy condition like in Theorem 326. Also the converse follows the same idea. Therefore following theorem is obtained.

Theorem 337 (Dichotomy) *Let $C(\mathcal{W})$ be the capacity of the arbitrarily-varying cq-channel $\mathcal{W}$ and let $C_{\mathrm{S}}(\mathcal{W}, \mathcal{V})$ be the secrecy capacity of the arbitrarily-varying wiretap cq-channel $(\mathcal{W}, \mathcal{V})$. Then,*

$$C_{SID}(\mathcal{W}, \mathcal{V}) = C_{SID}^{sim}(\mathcal{W}, \mathcal{V}) = \begin{cases} C(\mathcal{W}) & \text{if } C_{\mathrm{S}}(\mathcal{W}, \mathcal{V}) > 0, \\ 0 & \text{if } C_{\mathrm{S}}(\mathcal{W}, \mathcal{V}) = 0. \end{cases}$$

3 Quantum Channels

Here, we report the results reviewed in [80].

As is often the case, in quantum probability, there is not just one but several quantizations: we know at least two different concepts of identification of classical information via quantum channels, and three different identification capacities for quantum information. In this section, we concentrate on conceptual points and open problems presented by Andreas Winter in [79]. Let $\mathcal{P}(K)$ be the space of pure states of a finite dimensional Hilbert space K.

Definition 338 (Winter [79]) A *quantum ID code for the channel $\mathcal{N}$ with error ϵ*, for the Hilbert space K, is a pair of maps $\mathcal{E} : \mathcal{P}(K) \longrightarrow \mathcal{S}(A)$ and $\mathcal{D} : \mathcal{P}(K) \longrightarrow \mathcal{L}(B)$ with $0 \leq \mathcal{D}_\varphi \leq \mathbb{1}$ for all $\varphi = |\varphi\rangle\langle\varphi| \in \mathcal{P}(K)$, such that for all pure states/rank-one projectors $\psi, \varphi \in \mathcal{P}(K)$,

$$\left|\operatorname{Tr}\psi\varphi - \operatorname{Tr}\mathcal{N}(\mathcal{E}_\psi)\mathcal{D}_\varphi\right| \leq \epsilon.$$

If the encoding $\mathcal{E}$ is a quantum channel we speak of a *blind* code, otherwise we call it *visible*.

For the case of an iid channel $\mathcal{N}^{\otimes n}$, we denote the maximum dimension of a blind (visible) quantum ID code by $M(n, \epsilon)$ ($M_{\mathrm{v}}(n, \epsilon)$).

This notion can be motivated as follows: In quantum transmission, the objective for the receiver is to recover the state ψ by means of a suitable decoding (cptp) map $\widetilde{\mathcal{D}} : \mathcal{L}(B) \longrightarrow \mathcal{L}(K)$, with high accuracy. Of course then the receiver can perform any measurement on the decoded state, effectively simulating an arbitrary measurement on the original input state, in the sense that for any state ρ and POVM $M = (M_i)_i$ on K, there exists another POVM $M' = (M'_i)_i$ on B such that the measurement statistics of ρ under M is approximately that of $\mathcal{N}(\mathcal{E}(\rho))$ under M' ($\mathcal{E}$ in this case is a quantum channel). (M' can be written down directly via the adjoint $\widetilde{\mathcal{D}}^\dagger : \mathcal{L}(K) \longrightarrow \mathcal{L}(B)$ of the decoding map, which maps measurement POVMs on K to POVMs on B: $M'_i = \widetilde{\mathcal{D}}^\dagger(M_i)$.) The converse is also true: If the receiver can simulate sufficiently general measurements on the input state by suitable measurements on the channel output, then he can actually decode the state by a cptp map $\widetilde{\mathcal{D}}$ [66].

This allows us to relax the task of quantum information transmission to requiring only that the receiver be able to simulate the statistics of certain restricted measurements. In the case of quantum identification, these are $(\varphi, \mathbb{1} - \varphi)$ for arbitrary rank-one projectors $\varphi = |\varphi\rangle\langle\varphi| \in \mathcal{P}(K)$. They are the measurements which allow the receiver to ask the (quantum) question: "Is the state equal to φ or orthogonal to it?". Obviously, in quantum theory this question cannot be answered with certainty, but for each test state it yields a characteristic distribution and $\operatorname{Tr}\psi\varphi$ is the probability of answering yes when measuring on the state ψ. The quantum-ID task above is about reproducing this distribution by substituting ψ with the encoded state followed by the channel action $\mathcal{N}(\mathcal{E}_\psi)$, and φ by $\mathcal{D}_\varphi$.

Note that we can always concatenate a blind or visible quantum ID code for the Hilbert space K with a fingerprinting set of pure states [37] in K, to obtain a classical ID code. This is because in fingerprinting the encodings are pure states ψ_i and the tests precisely the POVMs $(\psi_i, \mathbb{1} - \psi_i)$. Hence, as the cardinality of the fingerprinting set is exponential in the dimension $|K|$, $M(n, \epsilon)$ and $M_v(n, \epsilon)$ can be at most exponential in n.

Definition 339 For a quantum channel $\mathcal{N}$, the *blind, respectively visible, quantum ID capacity* is defined as

$$Q_{\mathrm{ID}}(\mathcal{N}) := \inf_{\epsilon>0} \liminf_{n\to\infty} \frac{1}{n} \log M(n, \epsilon),$$

$$Q_{\mathrm{ID,v}}(\mathcal{N}) := \inf_{\epsilon>0} \liminf_{n\to\infty} \frac{1}{n} \log M_{\mathrm{v}}(n, \epsilon).$$

If we leave out the qualifier, the quantum ID capacity is by default the blind variety.

Note that by definition and the above remark,

$$Q_{\mathrm{ID}}(\mathcal{N}) \leq Q_{\mathrm{ID,v}}(\mathcal{N}) \leq C_{\mathrm{ID}}(\mathcal{N}). \tag{13}$$

The first quantum ID capacity that had been determined was for the ideal qubit channel:

Theorem 340 (Winter [79]) *For the noiseless channel* id_A *on Hilbert space A, there exists a (blind) quantum ID code with error* ϵ *and encoding a space K of dimension* $|K| \geq C(\epsilon)|A|^2$*, for some universal function* $C(\epsilon) > 0$.

As a consequence, $Q_{\mathrm{ID}}(\mathrm{id}_2) = Q_{ID,v}(\mathrm{id}_2) = 2$*, twice the quantum transmission capacity.*

In view of this theorem, we gain at least 2 in capacity for each noiseless qubit we use additionally to the given channel. This motivates the following definition.

Definition 341 (Hayden/Winter [52]) For a quantum channel $\mathcal{N}$, the *amortized (blind/visible) quantum ID capacity* is defined as

$$Q_{\mathrm{ID}}^{\mathrm{am}}(\mathcal{N}) := \sup_k \; Q_{\mathrm{ID}}(\mathcal{N} \otimes \mathrm{id}_k) - 2\log k,$$

$$Q_{\mathrm{ID,v}}^{\mathrm{am}}(\mathcal{N}) := \sup_k \; Q_{\mathrm{ID,v}}(\mathcal{N} \otimes \mathrm{id}_k) - 2\log k,$$

respectively.

The blind quantum ID-capacities are among the best understood, thanks to recently made conceptual progress, which we review in the next section. We will then also ask the question *how much* amortization is required. This is formalized in the usual way: Namely, for a rate $Q \leq Q_{\mathrm{ID}}^{\mathrm{am}}(\mathcal{N})$, we say that A is an *achievable*

amortization rate if there exist k_n for all n, such that

$$\liminf_{n\to\infty} \frac{1}{n}\left(Q_{\mathrm{ID}}(\mathcal{N}^{\otimes n}\otimes \mathrm{id}_{k_n}) - 2\log k_n\right) \geq Q \quad \text{and} \quad \limsup_{n\to\infty} \frac{1}{n}\log k_n \leq A,$$

giving rise to an achievable quantum ID-rate/amortization region, *viz.* a tradeoff between Q and A. Similarly of course for the visible variant. To state the following conceptual points about blind(!) quantum ID codes, it is useful to fix an encoding cptp map $\mathcal{E} : \mathcal{L}(K) \longrightarrow \mathcal{L}(A)$ and to combine it with the noisy channel, $\mathcal{N}' = \mathcal{N}\circ\mathcal{E}$, for which we choose a Stinespring dilation $V : K \hookrightarrow B\otimes F$. The quantum ID code is now the entire input space K of this effective new channel, together with the previously given operators $\mathcal{D}_\varphi$ on B. The next result states that just as quantum error correctability of $\mathcal{N}'$ is equivalent to $\widehat{\mathcal{N}'}$ being decoupling [46], quantum identification is essentially equivalent to weak decoupling from the environment:

Theorem 342 (Hayden/Winter [52]) *If K is a ϵ quantum ID code for the channel $\mathcal{N}'$ with Stinespring dilation $V : K \hookrightarrow B\otimes F$, then the complementary channel $\widehat{\mathcal{N}'}$ is approximately forgetful:*

$$\forall|\varphi\rangle, |\psi\rangle \in K \quad \frac{1}{2}\left\|\widehat{\mathcal{N}'}(\varphi) - \widehat{\mathcal{N}'}(\psi)\right\|_1 \leq \delta := 7\sqrt[4]{\epsilon}.$$

Conversely, if $\widehat{\mathcal{N}'}$ is approximately forgetful with error δ, then the trace-norm geometry is approximately preserved by $\mathcal{N}'$:

$$\forall|\varphi\rangle, |\psi\rangle \in S \quad 0 \leq \left\|\varphi-\psi\right\|_1 - \left\|\mathcal{N}'(\varphi) - \mathcal{N}'(\psi)\right\|_1 \leq \epsilon := 4\sqrt{2\delta}.$$

If, in addition, the nonzero eigenvalues of the environment's states $\widehat{\mathcal{N}'}(\varphi)$ lie in the interval $[\mu,\lambda]$ for all $|\varphi\rangle \in K$, then one can construct an η quantum ID code for $\mathcal{N}'$ (i.e. a set of operators D_φ for all $|\varphi\rangle \in K$ as in Definition 338), with $\eta := 7\delta^{1/8}\sqrt{\lambda/\mu}$.

Remark 343 While it would be desirable to eliminate the eigenvalue condition at the end of the theorem, the condition is fairly natural in this context. If the environment's states $\widehat{\mathcal{N}'}(\varphi)$ are very close to a single state σ^F for all $|\varphi\rangle \in K$, then all the $V|\varphi\rangle$ are very close to being purifications of σ^F, meaning that they differ from one another only by a unitary plus a small perturbation. If σ^F is the maximally mixed state or close to it, then the assumption will be satisfied. In the asymptotic iid setting we are looking at this turns to be the case.

This characterization of quantum ID codes (albeit "only" blind ones) allows the determination of capacities by a random coding argument, for which only the weak decoupling has to be verified. The above duality theorem is not only the basis for the direct but also for the converse part(s) of the following capacity theorem.

Theorem 344 (Hayden/Winter [52]) *For a quantum channel $\mathcal{N}$, its (blind) quantum ID capacity is given by*

$$Q_{\mathrm{ID}}(\mathcal{N}) = \lim_{n\to\infty} \frac{1}{n} Q_{\mathrm{ID}}^{(1)}(\mathcal{N}^{\otimes n}),$$

where

$$Q_{\mathrm{ID}}^{(1)}(\mathcal{N}) = \sup_{|\phi\rangle} \left\{ I(A:B)_\rho \text{ s.t. } I(A\rangle B)_\rho > 0 \right\},$$

$|\phi\rangle$ is the purification of an input state to $\mathcal{N}$, $\rho^{AB} = (\mathrm{id}\otimes\mathcal{N})(\phi)$, and $I(A:B)_\rho = S(\rho^A) + S(\rho^B) - S(\rho^{AB})$ is the mutual information, and $I(A\rangle B)_\rho = S(\rho^B) - S(\rho^{AB})$ the coherent information. We declare the sup *to be* 0 *if the set above is empty. In particular, $Q_{\mathrm{ID}}(\mathcal{N}) = 0$ if and only if $Q(\mathcal{N}) = 0$.*

Furthermore, the amortized quantum ID capacity equals

$$Q_{\mathrm{ID}}^{\mathrm{am}}(\mathcal{N}) = \sup_{|\phi\rangle} I(A:B)_\rho = C_E(\mathcal{N}),$$

the entanglement-assisted classical capacity of $\mathcal{N}$ [14].

4 Classical Gaussian Channels

In this section, we want to recall some results about identification over classical non-discrete Channels, e.g., Gaussian channels. We then deal with secure identification over Gaussian wiretap channels. Burnashev considered in [35] discrete-time channels with independent additive noise:

- $y_i = x_i + \xi_i, \quad \forall\, i \in \{1,\dots,n\}$
- ξ_i are iid and $\xi_i \sim \mathrm{f} \in \mathbb{R}^1, \quad \forall\, i \in \{1,\dots,n\}$. The noise function f should satisfy some regularity conditions. There exist some constants K, K_1, γ, α such that:

$$\int_{-\infty}^{\infty} \left(\max_{|t-x|\le u} \sqrt{\mathrm{f}(t)} - \min_{|t-x|\le u} \sqrt{\mathrm{f}(t)} \right)^2 \mathrm{d}x \le K u^\gamma, \qquad u > 0,\, 1 < \gamma \le 2 \tag{14}$$

$$\int_{|x|\ge z} \mathrm{f}(x)\mathrm{d}x \le K_1 z^{-\alpha}, \quad z > 0,\ \alpha > 2 \tag{15}$$

$$1/\alpha + 1/\gamma < 1 \tag{16}$$

- Average power constraint: $\frac{1}{n}\sum_{i=1}^n x_i^2 \le a^2, \quad a > 0$
- The output set is infinite $\mathcal{Y} = \mathbb{R}$, the input set $\mathcal{X} = [-a\sqrt{n}, a\sqrt{n}]$.

$\Longrightarrow$ We denote such channel by $W(\mathrm{f}, a)$ and its Shannon capacity by $C(\mathrm{f}, a)$.

Burnashev conjectured that the direct part of the identification coding theorem for the channels described above is similar to the discrete case.

We assume the existence of an (M, n, λ) ID code $\{(Q(\cdot|i), \mathcal{D}_i),\ i = 1, \ldots, M\}$ for the channel $W(\mathrm{f}, a)$. The maximal cardinality M such that an (M, n, λ) is denoted by $M_{\mathrm{ID}}(n, \delta)$. Burnashev proved in [36] the following theorem:

Theorem 345 *For any output measure Q_π on $\mathbb{R}^n$, there exist input blocks $x^n(i) \in \mathcal{L}^n,\ i = 1, \ldots, N$ with $\frac{1}{n} \ln N \leq C(f, a) + \delta$ such that their generated output measures $\{W(\cdot|x^n(i))\}$ satisfy the following inequality:*

$$|Q_\pi(W(f, a)) - \mathrm{conv}\{W(\cdot|x^n(i)), i = 1, \ldots, N\}| \leq \delta, \quad \delta > 0 \tag{17}$$

In other words, each output measure $Q_i \in \mathcal{P}(\mathbb{R})$ can be δ-approximated by another output measure $Q_i' \triangleq \mathrm{conv}\{W(\cdot|x^n(i)), i = 1, \ldots, N\}$ generated by $N \approx e^{n(C(f,a)+\delta)}$ input blocks in $\mathcal{L}^n$. We fix $\delta > 0$ such that $2(\delta + \lambda) < 1$.

Definition 346 A collection $\{P_i, i = 1, \ldots, M\}$ with $\forall\, i,\ P_i \in \mathcal{P}(\mathbb{R}^n)$ is called an (M, n, δ, W) pairwise-separated collection if:

$$|W^n P_i - W^n P_j| \geq 2(1 - \delta), \quad i \neq j$$

The maximal possible cardinality of pairwise-separated collection for the channel W is denoted by $M_p(\delta, W)$.

Definition 347 A collection $\{P_i, i = 1, \ldots, M\}$ with $\forall\, i,\ P_i \in \mathcal{P}(\mathbb{R}^n)$ is called an (M, n, δ, W) completely-separated collection if:

$$|W^n P_i - \mathrm{conv}\{W^n P_j,\ j \neq i\}| \geq 2(1 - \delta), \quad i \neq j$$

The maximal possible cardinality of completely-separated collection for the channel W is denoted by $M_c(\delta, W)$.

Theorem 345 implies the following result:

$$C_{\mathrm{ID}}(f, a) \leq C(f, a) + \delta, \quad 0 < \delta < \frac{1}{2} \tag{18}$$

where $C_{\mathrm{ID}}(f, a)$ denotes the identification capacity of the channel $W(f, a)$.

4.1 Classical Gaussian Wiretap Channels

Now, we consider information theoretical security. Existing cryptographic approaches commonly used for wireless local area networks can be overridden sufficient computing power. In contrast, information theory provides a tool for designing codes that are proven to be unbreakable. In our coding scheme the

authorized sender wants to transmit a secure identification message to the authorized receiver so that the receiver is able to identify his message. The unauthorized party is a wiretapper who can wiretap the transmission. He tries to identify an unknown message. We develop a coding scheme so that secure identification over an Gaussian channel is possible. The receiver can identify a message with high probability. Furthermore, the wiretapper is not able to identify a message with high probability. We consider the following Gaussian wiretap channel:

$$\begin{aligned} y_i &= x_i + \xi_i, && \forall\, i \in \{1, \ldots, n\} \\ z_i &= x_i + \phi_i, && \forall\, i \in \{1, \ldots, n\} \end{aligned} \tag{19}$$

where $x^n = (x_1, x_2, \ldots, x_n)$ is the channel input sequence. $y^n = (y_1, y_2, \ldots, y_n)$ and $z^n = (z_1, z_2, \ldots, z_n)$ are Bob and Eve's observations, respectively. $W_{Y|X}$ is the main channel, while $V_{Z|X}$ is the wiretapper's channel. $\xi^n = (\xi_1, \xi_2, \ldots, \xi_n)$ and $\phi^n = (\phi_1, \phi_2, \ldots, \phi_n)$ are the noise sequences of the main channel and the wiretapper's channel, respectively. ξ_i are identically and independently distributed (i.i.d) with and drawn from a normal distribution with variance σ^2 denoted by g. ϕ_i are i.i.d and drawn from a normal distribution with variance σ'^2 denoted by g$'$. The channel input fulfills the following energy constraint.

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq P \tag{20}$$

This implies that the input set is reduced to $\mathcal{E} = [-\sqrt{nP}, \sqrt{nP}]$. The output sets are infinite $\mathcal{Y} = \mathcal{Z} = \mathbb{R}$. In the sequel, we consider only strong secrecy requirement $I(U, Z^n) \leq \lambda$. We denote this channel by $(W, V, \mathrm{g}, \mathrm{g}', P)$ with secrecy capacity $C(\mathrm{g}, \mathrm{g}', P)$. We denote by $W(\mathrm{g}, P)$ and $V(\mathrm{g}', P)$ the channels to the legitimate receiver and to the wiretapper, respectively. A secrecy or a wiretap code for is defined as follows. In [78] and [73], the following theorem about the strong secrecy capacity of the Gaussian channel was shown:

Theorem 348 ([78]) *Let $C_{\mathrm{S}}(g, g', P)$ be the secrecy capacity of the channel (W, V, g, g', P) then:*

$$C_{\mathrm{S}}(g, g', P) = \begin{cases} \frac{1}{2} \log \left(\frac{1+\frac{P}{\sigma^2}}{1+\frac{P}{\sigma'^2}} \right) & \text{if } \sigma'^2 \geq \sigma^2 \\ 0 & \text{else} \end{cases}$$

Based on the definitions of transmission wiretap codes in [82] and identification wiretap codes in [10], we introduce wiretap codes for the Gaussian wiretap channel described above (19).

Definition 349 A randomized (n, M, λ) transmission-code for the Gaussian wiretap channel $(V, W, \mathrm{g}, \mathrm{g}', P)$ is a family of pairs $\{(Q(\cdot|i), \mathcal{D}_i),\ i = 1, \ldots, M\}$ with

$$Q(\cdot|i) \in \mathcal{P}\left(\mathcal{X}^n\right),\ \mathcal{D}_i \subset \mathcal{Y}^n, \qquad \forall\, i \in \{1, \ldots, N\} \tag{21}$$

$$\sum_{l=1}^{n} x_l^2 \leq n \cdot P, \qquad \forall\, x^n \in \mathcal{X}^n \tag{22}$$

such that for all $i \in \{1, \ldots, M\}$ and $i \neq j$

$$\int_{x^n \in \mathcal{X}^n} Q(x^n|i) W^n(\mathcal{D}_i^c|x^n) d^n x^n \leq \lambda \tag{23}$$

$$\mathcal{D}_i \cap \mathcal{D}_j = \emptyset \tag{24}$$

$$I(U, Z^n) \leq \lambda \tag{25}$$

where U is a uniformly distributed RV on $\{1, \ldots, M\}$ and Z^n is the output of the channel V i.e., the wiretapper's observation.

Definition 350 A randomized $(n, N, \lambda_1, \lambda_2)$ identification code for the Gaussian wiretap channel $(V, W, \mathrm{g}, \mathrm{g}', P)$ is a family of pairs $\{(Q(\cdot|i), \mathcal{D}_i),\ i = 1, \ldots, N\}$ with

$$Q(\cdot|i) \in \mathcal{P}\left(\mathcal{X}^n\right), \mathcal{D}_i \subset \mathcal{Y}^n, \qquad \forall\, i \in \{1, \ldots, N\} \tag{26}$$

$$\sum_{l=1}^{n} x_l^2 \leq n \cdot P, \qquad \forall\, x^n \in \mathcal{X}^n \tag{27}$$

such that for all $i, j \in \{1, \ldots, N\}$, $i \neq j$ and some $\mathcal{E} \in \mathcal{Z}^n$

$$\int_{x^n \in \mathcal{X}^n} Q(x^n|i) W^n(\mathcal{D}_i^c|x^n) d^n x^n \leq \lambda_1 \tag{28}$$

$$\int_{x^n \in \mathcal{X}^n} Q(x^n|j) W^n(\mathcal{D}_i|x^n) d^n x^n \leq \lambda_2 \tag{29}$$

$$\int_{x^n \in \mathcal{X}^n} Q(x^n|j) V^n(\mathcal{E}|x^n) d^n x^n + \int_{x^n \in \mathcal{X}^n} Q(x^n|i) V^n(\mathcal{E}^c|x^n) d^n x^n \geq 1 - \lambda \tag{30}$$

The following theorem is proved in [56].

Theorem 351 *Let $C_{SID}(g, g', P)$ be the secure identification capacity of the wiretap channel (W, V, g, g', P) and then:*

$$C_{SID}(g, g', P) = \begin{cases} C(g, P) & \text{if } C_S(g, g', P) > 0 \\ 0 & \text{if } C_S(g, g', P) = 0 \end{cases} \tag{31}$$

$C(g, P)$ defines the identification capacity of the main Gaussian $W(g, P)$.

5 Identification and Continuity

Here, we review the results of [25]. A similar work for channels with feedback has been done in [31].

In this section, we investigate an important performance criterion for the identification and secure identification capacity. We analyze the dependence of the capacity on its channel parameters. A communication system is easier to handle if the capacity continuously depends on the channel parameters, because otherwise, minor changes in the parameters can lead to dramatic changes in the performance. Shannon [71] assumed in 1956 that the zero-error capacity is additive. This was refuted in [11] by Alon. The property shown there is called super-additivity. It is generally not known which channels are super-additive and which are not. We have shown that for the AVC, the identification capacity is equal to the transmission capacity. Thus the analytical properties of the identification capacity are completely determined by the work [28] and [29]. We first introduce the necessary definitions for the investigation of continuity and additivity, and then list the known results.

5.1 Basic Definitions and Results

We denote the set of channels from $\mathcal{X}$ to $\mathcal{Y}$ as $CH(\mathcal{X};\mathcal{Y})$. In the case of multiple inputs or outputs sets we separate them with a comma, e.g. $CH(\mathcal{X},\mathcal{X}';\mathcal{Y},\mathcal{Y}')$. For the analysis that follows, we introduce the function $F(\mathcal{W}) : CH(\mathcal{X},\mathcal{S};\mathcal{Y}) \to \mathbb{R}_+$

$$F(\mathcal{W}) = \inf_{\sigma\in CH(\mathcal{X};\mathcal{S})} \max_{\substack{x\neq\hat{x}\\ x,\hat{x}\in X}} \sum_{y\in\mathcal{Y}} \left| \sum_{s\in\mathcal{S}} \left[W(y|x,s)\sigma(s|\hat{x}) - W(y|\hat{x},s)\sigma(s|\hat{x}) \right] \right|. \tag{32}$$

Since $CH(\mathcal{X};\mathcal{S})$ is a bounded and closed set, there exists for any AVC $\mathcal{W}$ a channel $\sigma^* \in CH(\mathcal{X};S)$ such that the infimum above is achieved and F can be expressed as a minimum. Further, we have $F(\mathcal{W}) \geq 0$ with equality if and only if $\mathcal{W}$ is symmetrizable. We also need a concept of distance. For two DMCs $\mathcal{W}_1, \mathcal{W}_2 \in CH(\mathcal{X};\mathcal{Y})$ we define the distance between $\mathcal{W}_1$ and $\mathcal{W}_2$ based on the total variation distance as

$$d(\mathcal{W}_1,\mathcal{W}_2) := \max_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} |\mathcal{W}_1(y|x) - \mathcal{W}_2(y|x)|. \tag{33}$$

To extend this concept to AVCs, we consider two AVCs $\mathcal{W}_1 = \{W_1(\cdot|\cdot,s_1)\}_{s_1\in\mathcal{S}_1}$ and $\mathcal{W}_2 = \{W_2(\cdot|\cdot,s_2)\}_{s_2\in\mathcal{S}_2}$ with $W_i(\cdot|\cdot,s_i) \in CH(X,S_i;Y)$, $i = 1,2$, and define

$$G(\mathcal{W}_1,\mathcal{W}_2) := \max_{s_1\in\mathcal{S}_1} \min_{s_2\in\mathcal{S}_2} d(W_1(\cdot|\cdot,s_1), W_2(\cdot|\cdot,s_2)) \tag{34}$$

which describes how well one AVC can be approximated by the other one. Note that the function G is not symmetric. Accordingly, we define the distance between $\mathcal{W}_1$ and $\mathcal{W}_2$ as

$$D(\mathcal{W}_1, \mathcal{W}_2) := \max\{G(\mathcal{W}_1, \mathcal{W}_2), G(\mathcal{W}_2, \mathcal{W}_1)\}. \tag{35}$$

Note that S_1 and S_2 can be arbitrary finite state sets and we do not necessarily need to have $|S_1| = |S_2|$. In [29] the following basis properties are shown.

Lemma 352 *Let $\mathcal{W}_1$ and $\mathcal{W}_2$ be two finite AVCs. Then the following inequalities hold:*

$$F(\mathcal{W}_2) \leq 2G(\mathcal{W}_1, \mathcal{W}_2) + F(\mathcal{W}_1) \tag{36}$$

$$F(\mathcal{W}_1) \leq 2G(\mathcal{W}_2, \mathcal{W}_1) + F(\mathcal{W}_2) \tag{37}$$

$$|F(\mathcal{W}_1) - F(\mathcal{W}_2)| \leq 2D(\mathcal{W}_1, \mathcal{W}_2). \tag{38}$$

Furthermore, the following is shown in [29].

Lemma 353 *Let $\tilde{\mathcal{W}}$ be an arbitrary finite AVC and let $\{\mathcal{W}_n\}_{n=1}^{\infty}$ be a sequence of finite AVCs such that*

$$\lim_{n\to\infty} D(\mathcal{W}_n, \tilde{\mathcal{W}}) = 0. \tag{39}$$

Then

$$\lim_{n\to\infty} F(\mathcal{W}_n) = F(\tilde{\mathcal{W}}). \tag{40}$$

For the next results, we need the concept of orthogonal (or parallel) AVCs. For two AVCs $\mathcal{W}_1$ and $\mathcal{W}_2$ we define the AVC $\mathcal{W}$ as

$$\mathcal{W} = \mathcal{W}_1 \times \mathcal{W}_2 = \{W_1(\cdot|\cdot, s_1)\}_{s_1\in\mathcal{S}_1} \times \{W_2(\cdot|\cdot, s_2)\}_{s_2\in\mathcal{S}_2} \tag{41}$$

which means that the underlying channel law is

$$W(y_1, y_2|x_1, x_2, s_1, s_2) = W_1(y_1|x_1, s_1)W_2(y_2|x_2, s_2) \tag{42}$$

for all $x_i \in \mathcal{X}_i$, $y_i \in \mathcal{Y}_i$, and $s_i \in \mathcal{S}_i$, $i = 1, 2$.

Definition 354 Let $\mathcal{W}$ be a finite AVC. A capacity $C(\mathcal{W})$ is said to be continuous in all finite AVCs $\mathcal{W}$, if for all sequences of finite AVCs $\{\mathcal{W}_n\}_{n=1}^{\infty}$ with

$$\lim_{n\to\infty} D(\mathcal{W}_n, \mathcal{W}) = 0 \tag{43}$$

we have

$$\lim_{n\to\infty} C(\mathcal{W}_n) = C(\mathcal{W}). \tag{44}$$

Note that similarly as in Lemma 353, the only restriction on the state set is $|\mathcal{S}_n| < \infty$, but we can have $\lim_{n\to\infty} |\mathcal{S}_n| = \infty$. Based on this definition, the capacity $C(\mathcal{W})$ is discontinuous in $\mathcal{W}$ if and only if there is a sequence $\{\mathcal{W}_n\}_{n=1}^{\infty}$ of finite AVCs satisfying (43) but

$$\limsup_{n\to\infty} C(\mathcal{W}_n) > \liminf_{n\to\infty} C(\mathcal{W}_n) \tag{45}$$

is satisfied.

5.2 Continuity and Discontinuity Behavior of C_{ID}

We have shown that $C_{\mathrm{ID}}(\mathcal{W}) = C(\mathcal{W})$. Therefore, we can use the results from [29] where C was considered, to give analytical properties of C_{ID}.

First we give a complete characterization of the discontinuity points of the capacity.

Theorem 355 *The capacity $C_{\mathrm{ID}}(\mathcal{W})$ is discontinuous in the finite AVC $\mathcal{W}$ if and only if the following conditions hold:*

1. $C_{ran}(\mathcal{W}) > 0$
2. $F(\mathcal{W}) = 0$ *and for every* $\epsilon > 0$ *there exists a finite AVC* $\tilde{\mathcal{W}}$ *with* $D(W, \tilde{W}) < \epsilon$ *and* $F(\tilde{\mathcal{W}}) > 0$.

The following result establishes certain robustness properties of the capacity.

Theorem 356 *Let $\mathcal{W}$ be a finite AVC with $F(\mathcal{W}) > 0$ (non symmetrizable). Then there exists an $\tilde{\epsilon} > 0$ such that all finite AVCs $\tilde{\mathcal{W}}$ with*

$$D(\tilde{\mathcal{W}}, \mathcal{W}) < \tilde{\epsilon} \tag{46}$$

are continuity points of C_{ID}.

Next, we want to further analyze the discontinuous behavior for finite AVCs. For this purpose, let

$$\mathcal{N} = \{\mathcal{W} : F(\mathcal{W}) = 0\} \tag{47}$$

be the set of symmetrizable channels (the kernel of F) and $\partial\mathcal{N}$ be the boundary of $\mathcal{N}$ under the distance D, which is given by

$$\partial\mathcal{N} = \{\mathcal{W} \in \mathcal{N} : \forall\, \epsilon > 0,\ \exists\, \hat{\mathcal{W}} \notin \mathcal{N} \text{ such that } D(\hat{\mathcal{W}}, \mathcal{W}) < \epsilon\}.$$

Let $\{W_n\} = \{W_n\}_{n=1}^{\infty}$ be a sequence of finite AVCs. We define the variance of C_{ID} over the sequence as

$$V(\{W_n\}) \triangleq \limsup_{n\to\infty} C_{\mathrm{ID}}(\mathcal{W}_n) - \liminf_{n\to\infty} C_{\mathrm{ID}}(\mathcal{W}_n). \tag{48}$$

Let now $\mathcal{W}$ be a finite AVC. We define the maximum variance around $\mathcal{W}$ as

$$\bar{V}(\mathcal{W}) = \sup V(\{\mathcal{W}_n\}) \tag{49}$$

where the sup is taken over all sequences of finite AVCs $\{\mathcal{W}_n\}_{n=1}^{\infty}$ that satisfy

$$\lim_{n\to\infty} D(\mathcal{W}_n, \mathcal{W}) = 0. \tag{50}$$

With this, $\bar{V}(\mathcal{W})$ describes the maximal variation of $C_{\mathrm{ID}}(\mathcal{W})$ in the neighborhood of a certain AVC $\mathcal{W}$. Then, we define

$$\bar{V} \triangleq \sup_{\mathcal{W}\in CH(X,S;Y)} \bar{V}(\mathcal{W}) = \bar{V}(\mathcal{X}, \mathcal{Y}) \tag{51}$$

as the maximal variation for all AVCs $\mathcal{W}$ on the given legitimate-input and output alphabets. Furthermore we set

$$\mathcal{N}_{\mathrm{ran}} \triangleq \{\mathcal{W} : C_{\mathrm{ran}}(\mathcal{W}) = 0\}$$

as the set of channels that cannot transmit even in the presence of common randomness (the kernel of C_{ran}). A rephrasing of the theorems above is thus the following.

Theorem 357 *For finite AVCs $\mathcal{W}$ the following assertions hold:*

1. $\bar{V}(\mathcal{W}) = 0$ *for* $\mathcal{W} \notin \mathcal{N} \setminus \mathcal{N}_{ran}$
2. $\bar{V}(\mathcal{W}) = C_{ran}(\mathcal{W})$ *for* $\mathcal{W} \in \mathcal{N} \setminus \mathcal{N}_{ran}$
3. $\bar{V} = \sup_{\mathcal{W}\in D} C_{ran}(\mathcal{W})$.

As noted before, $\mathcal{N} \setminus \mathcal{N}_{\mathrm{ran}}$ is not empty and examples can be found in [42].

5.3 Additivity and Super-Additivity of C_{ID}

In this section, we now examine the additivity of the capacity function. Alon was the first who showed the phenoma of super-additivity for the zero-error capacity [11].

Definition 358 Then, a capacity is said to be super-additive if there exist two finite AVCs $\mathcal{W}_1$ and $\mathcal{W}_2$ such that

$$C(\mathcal{W}_1 \times \mathcal{W}_2) > C(\mathcal{W}_1) + C(\mathcal{W}_2), \tag{52}$$

i.e., a joint use of both channels yields a higher capacity than the sum of their individual uses.

Theorem 359 *Let $\mathcal{W}_1$ and $\mathcal{W}_2$ be two AVCs. Then*

$$C_{\mathrm{ID}}(\mathcal{W}_1 \times \mathcal{W}_2) = 0 \tag{53}$$

if and only if

$$C_{\mathrm{ID}}(\mathcal{W}_1) = C_{\mathrm{ID}}(\mathcal{W}_2) = 0. \tag{54}$$

The theorem above shows that super additivity cannot happen when both channels have zero capacity, which otherwise would be called super-activation of the channels. We will see that super-activation can occur for secret identification.

The next result shows that the ID capacity is super-additive.

Theorem 360 *Let $\mathcal{W}_1$ and $\mathcal{W}_2$ be two AVCs. Then*

$$C_{\mathrm{ID}}(\mathcal{W}_1 \times \mathcal{W}_2) > C_{\mathrm{ID}}(\mathcal{W}_1) + C_{\mathrm{ID}}(\mathcal{W}_2) \tag{55}$$

if and only if

$$\begin{aligned} \min\{F(\mathcal{W}_1), F(\mathcal{W}_2)\} &= 0, \\ \max\{F(\mathcal{W}_1), F(\mathcal{W}_2)\} &> 0, \\ C_{ran}(\mathcal{W}_1), C_{ran}(\mathcal{W}_2) &> 0, \end{aligned}$$

namely if one of the channels is symmetrizable but both can transmit using common randomness.

5.4 Continuity of C_{SID} for AVWCs

In Theorem 308 we showed:

$$C_{\mathrm{SID}}(\mathcal{W}, \mathcal{V}) = \begin{cases} C(\mathcal{W}) & \text{if } C_{\mathrm{S}}(\mathcal{W}, \mathcal{V}) > 0 \\ 0 & \text{if } C_{\mathrm{S}}(\mathcal{W}, \mathcal{V}) = 0 \end{cases} \tag{56}$$

We shall now use this representation to fully characterize the continuity behavior and the discontinuity behavior of C_{SID}.

Therefore, we will need a suitable measure of distance between AVWCs. Recall Definition 34 of the G function, which describes how well one AVC can be approximated by the other one.

Definition 361 Let $(\mathcal{W}, \mathcal{V})$ and $(\mathcal{W}', \mathcal{V}')$ be two AVWCs, then we set

$$D((\mathcal{W}, \mathcal{V}), (\mathcal{W}', \mathcal{V}')) := \max\{G(\mathcal{W} \times \mathcal{V}, \mathcal{W}' \times \mathcal{V}'), G(\mathcal{W}' \times \mathcal{V}', \mathcal{W} \times \mathcal{V})\}. \tag{57}$$

Definition 362 Let $(\mathcal{W}, \mathcal{V})$ be a finite AVWC. The capacity $C(\mathcal{W}, \mathcal{V})$ is said to be continuous in $(\mathcal{W}, \mathcal{V})$, if for all sequences of finite AVWCs $\{(\mathcal{W}_n, \mathcal{V}_n)\}_{n=1}^{\infty}$ with

$$\lim_{n\to\infty} D((\mathcal{W}_n, \mathcal{V}_n), (\mathcal{W}, \mathcal{V})) = 0 \tag{58}$$

we have

$$\lim_{n\to\infty} C(\mathcal{W}_n, \mathcal{V}_n) = C(\mathcal{W}, \mathcal{V}). \tag{59}$$

The capacity is said to be discontinuous in $(\mathcal{W}, \mathcal{V})$ otherwise.

Let us now fully characterize the discontinuity of C_{SID}. We split the characterisation in two cases:

1. $C_{\mathrm{S,ran}}(\mathcal{W}, \mathcal{V}) > 0$
2. $C_{\mathrm{S,ran}}(\mathcal{W}, \mathcal{V}) = 0$

We start with the first case.

Theorem 363 *Let $C_{S,ran}(\mathcal{W}, \mathcal{V}) > 0$. $(\mathcal{W}, \mathcal{V})$ is a discontinuity point of C_{SID} if and only if $F(\mathcal{W}) = 0$ and for all $\epsilon > 0$ there exists a finite AVC $\mathcal{W}_\epsilon$ with $D(\mathcal{W}, \mathcal{W}_\epsilon) < \epsilon$ such that $F(\mathcal{W}_\epsilon) > 0$.*

We now examine the second case.

Theorem 364 *Let $C_{S,ran}(\mathcal{W}, \mathcal{V}) = 0$. $(\mathcal{W}, \mathcal{V})$ is a point of discontinuity of C_{SID} if and only if $C_{ran}(\mathcal{W}) > 0$ and for all $\epsilon > 0$ there exits a finite AVWC $(\mathcal{W}_\epsilon, \mathcal{V}_\epsilon)$ with $D(\mathcal{W}, \mathcal{W}_\epsilon) < \epsilon$ such that $F(\mathcal{W}_\epsilon) > 0$ and $C_{S,ran}(\mathcal{W}_\epsilon, \mathcal{V}_\epsilon) > 0$.*

Therefore the following corrollary follows from Theorems 363 and 364.

Corollary 365 *$(\mathcal{W}, \mathcal{V})$ is a point of discontinuity of C_{SID} if and only if $C_{ran}(\mathcal{W}) > 0$ and for all $\epsilon > 0$ there exits a finite AVWC $(\mathcal{W}_\epsilon, \mathcal{V}_\epsilon)$ with $D(\mathcal{W}, \mathcal{W}_\epsilon) < \epsilon$ such that $F(\mathcal{W}_\epsilon) > 0$ and $C_{S,ran}(\mathcal{W}_\epsilon, \mathcal{V}_\epsilon) > 0$.*

5.5 Super-Additivity and Super-Activation for C_{SID}

In this section, we will fully characterize the occurrence of super-activation and super-additivity for C_{SID}. Of course, super-activation is the most powerful form of super-additivity, in this case two channels with capacity zero add up to one with positive capacity. We first prove the complete characterization of super-activation. Subsequently we prove the cases of super-additivity, which is not caused by super-activation. First we start with the formal definition.

Definition 366 We say that two AVWCs $(\mathcal{W}, \mathcal{V})$ and $(\widetilde{\mathcal{W}}, \widetilde{\mathcal{V}})$ are superactivating if

$$C(\mathcal{W} \times \widetilde{\mathcal{W}}, \mathcal{V} \times \widetilde{\mathcal{V}}) > 0 \tag{60}$$

despite

$$C(\mathcal{W}, \mathcal{V}) = C(\widetilde{\mathcal{W}}, \widetilde{\mathcal{V}}) = 0. \tag{61}$$

Super-activation is the extreme case of super-additivity for channels with zero capacity.

In [24] we showed that C_{ID} does not have any super-activation. For C_{SID} this is different.

Theorem 367 *Let $(\mathcal{W}_1, \mathcal{V}_1)$, $(\mathcal{W}_2, \mathcal{V}_2)$ be two AVWCs. Then*

1. *Let* $\max\{C_{S,ran}((\mathcal{W}_1, \mathcal{V}_1), C_{S,ran}((\mathcal{W}_2, \mathcal{V}_2)\} > 0$ *then the AVWCs are super-activating if and only if*

$$\min\{C_{S,ran}(\mathcal{W}_1, \mathcal{V}_1), C_{S,ran}(\mathcal{W}_2, \mathcal{V}_2)\} = 0. \tag{62}$$

   Let w.l.o.g. $C_{S,ran}(\mathcal{W}_1, \mathcal{V}_1) > 0$, *then* $C_{S,ran}(\mathcal{W}_2, \mathcal{V}_2) = 0$ *and* $F(\mathcal{W}_1) = F(\mathcal{W}_2) = 0$.
2. *Let* $C_{S,ran}(\mathcal{W}_1, \mathcal{V}_1) = C_{S,ran}(\mathcal{W}_2, \mathcal{V}_2) = 0$ *then the AVWCs are super-activating if and only if*

$$F(\mathcal{W}_1 \times \mathcal{W}_2) > 0 \tag{63}$$

   (The condition $F(\mathcal{W}_1 \times \mathcal{W}_2) > 0$ *is equivalent to* $\max\{F(\mathcal{W}_1), F(\mathcal{W}_2)\} > 0$.*)*

We now want to examine the case in which we observe super-additivity, but in which no super-activation occurs. This characterization provides the following theorem.

Theorem 368 *Let $(\mathcal{W}_1, \mathcal{V}_1)$ and $(\mathcal{W}_2, \mathcal{V}_2)$ be two finite AVWCs for which no super-activation occurs. Then C_{SID} is superadditive for these channels if and only if*

$$C_{ran}(\mathcal{W}_1), C_{ran}(\mathcal{W}_2) > 0.$$

and exactly one channel has non-zero C_{SID}, i.e.

$$\max\{C_{SID}(\mathcal{W}_1, \mathcal{V}_1), C_{SID}(\mathcal{W}_2, \mathcal{V}_2)\} > 0,$$
$$\min\{C_{SID}(\mathcal{W}_1, \mathcal{V}_1), C_{SID}(\mathcal{W}_2, \mathcal{V}_2)\} = 0.$$

The meaning of the statements above is that super-additivity only happens whenever the secure ID capacity of at least one of the channels is zero, namely at the discontinuity points of the capacity, but this channel can be used to produce common randomness with the help of the key produced by the other channel. The condition the channels are not super-activating is essentially excluding the case of $C_{\mathrm{SID}}(\mathcal{W}_1, \mathcal{V}_1) = C_{\mathrm{SID}}(\mathcal{W}_2, \mathcal{V}_2) = 0$

6 Identification and Computability

In this section, we give a summary of [30]. In [30], the problem of identification over correlation-assisted DMCs is considered. In this scenario, the transmitter and the receiver have further access to correlated source observations as visualized in Fig. 3. Based on this resource, the encoder and decoder can be chosen. The corresponding identification capacity of this communication scenario remains unknown to date and in [30] analytical properties and representations of the identification capacity are studied.

Definition 369 A Turing machine is a mathematical model of an abstract machine that manipulates symbols on a strip of tape according to certain given rules. It can simulate any given algorithm and therewith provides a simple but very powerful model of computation. Turing machines have no limitations on computational complexity, unlimited computing capacity and storage, and execute programs completely error-free. Accordingly they provide fundamental performance limits for today's digital computers. Turing machines account for all those problems and tasks that are algorithmically solvable on a classical (i.e., non-quantum) machine. They

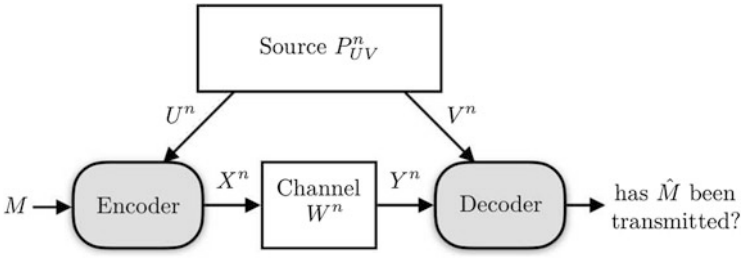


Fig. 3 Identification via a correlation-assisted DMC W. Transmitter and receiver have access to a correlated source observations U^n and V^n and can adapt their encoder and decoder accordingly

are further equivalent to the von Neumann-architecture without hardware limitations and to the theory of recursive functions.

Definition 370 In correlation-assisted identification, the transmitter and the receiver have access to a correlated source $P_{UV} \in \mathcal{P}(\mathcal{U} \times \mathcal{V})$. Similarly, the encoding can be adapted according to the received source sequence $u^n \in \mathcal{U}^n$, i.e., $Q(\cdot|i, u^n) \in \mathcal{P}(\mathcal{X}^n)$, and the decoding according to the received source sequence $v^n \in \mathcal{V}^n$, i.e., $\mathcal{D}_i(v^n) \subset \mathcal{Y}^n$. The definitions of a code, an achievable rate, and the identification capacity follow accordingly.

The identification capacity for DMCs with correlated sources is given by the following theorem.

Theorem 371 *There is no natural number $n_0 \in \mathcal{N}$ such that the identification capacity can be expressed as*

$$C_{\mathrm{ID}}(W, P_{UV}) = \max_{p \in \mathcal{P}} F(p, W, P_{UV}) \tag{64}$$

with $\mathcal{P} \subset \mathcal{R}^{n_0}$ a compact set (i.e. closed and bounded) and $F : \mathcal{P} \times \mathrm{CH} \times \mathcal{P}(\mathcal{U} \times \mathcal{V}) \to \mathcal{R}$ a continuous function.

Remark 372 Theorem 371 shows that the correlation-assisted identification capacity possesses a completely different behavior than the task of message transmission. In the case of message transmission, correlated sources can help to increase the capacity or to stabilize the transmission. For example, for arbitrarily-varying channels and arbitrarily-varying wiretap channels, the correlation-assisted capacities are always continuous and the capacity expressions can be expressed as optimization problems of the structure (64).

The following behavior for identification over DMCs is noted. The identification capacity without correlation is continuous and can be expressed as an optimization problem. In the case of correlation, this is not longer true.

Remark 373 Theorem 371 further immediately implies that the correlation-assisted identification capacity C_{ID} cannot be expressed by a finite multi-letter formula. As a consequence, if C_{ID} can be described by entropic quantities, then this must be done via a corresponding sequence.

Now, the question is: Whether or not the identification capacity is algorithmically computable. The concept of computability and computable reals goes back to Turing [74, 75]. A sequence of rational numbers $\{r_n\}_{n \in \mathcal{N}}$ is called a *computable sequence* if there exist recursive functions $a, b, s : \mathcal{N} \to \mathcal{N}$ with $b(n) \neq 0$ for all $n \in \mathcal{N}$ and

$$r_n = (-1)^{s(n)} \frac{a(n)}{b(n)}, \qquad n \in \mathcal{N}, \tag{65}$$

cf. [72, Def. 2.1 and 2.2] for a detailed treatment. A real number x is said to be *computable* if there exists a computable sequence of rational numbers $\{r_n\}_{n\in\mathcal{N}}$ such that

$$|x - r_n| < 2^{-n} \tag{66}$$

for all $n \in \mathcal{N}$. We denote the set of computable real numbers by $\mathcal{R}_c$. Based on this, the set of computable probability distributions $\mathcal{P}_c(\mathcal{X})$ is defined as the set of all probability distributions $P \in \mathcal{P}(\mathcal{X})$ such that $P(x) \in \mathcal{R}_c$, $x \in \mathcal{X}$. Further, let CH_c be the set of all computable channels, i.e., for a channel $W : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$, $W(\cdot|x) \in \mathcal{P}_c(\mathcal{Y})$ for every $x \in \mathcal{X}$. This is important since a Turing machine can only operate on computable real numbers.

Definition 374 A function $f : \mathcal{R}_c \to \mathcal{R}_c$ is called *Borel computable* if there is an algorithm that transforms each given computable sequence of a computable real x into a corresponding representation for $f(x)$.

It is to note that Turing's definition of computability conforms to the definition of Borel computability above. This captures the engineering intuition of the computability of functions. Intuitively, a function $f : \mathcal{R} \to \mathcal{R}$ is computable if every algorithm that computes the computable input $x \in \mathcal{R}$ can be algorithmically transformed into an algorithm that computes the output $f(x)$.

There are weaker forms of computability known as *Markov computability* and *Banach-Mazur computability*, of which the latter one is the weakest form of computability.

Definition 375 A function $f : \mathcal{R}_c \to \mathcal{R}_c$ is called *Banach-Mazur computable* if f maps any given computable sequence $\{x_n\}_{n=1}^{\infty}$ of real numbers into a computable sequence $\{f(x_n)\}_{n=1}^{\infty}$ of real numbers.

Theorem 376 *The identification capacity C_{ID} is not Banach-Mazur computable and therewith also not Turing computable.*

The previous results do not exclude that the correlation-assisted identification capacity can be expressed as

$$C_{\mathrm{ID}}(W, P_{UV}) = \lim_{n\to\infty} \max_{p\in\mathcal{P}^n} F_n(p, W, P_{UV}) \tag{67}$$

with $\mathcal{P}^n \subset \mathcal{R}^{l_n}$ a compact set, $l_n \in \mathcal{N}$, and $F_n : \mathcal{P}^n \times \mathrm{CH} \times \mathcal{P}(\mathcal{U} \times \mathcal{V}) \to \mathcal{R}$ a continuous function. From a practical point of view, an expression of the form (67) would not yield any particular problems, if the convergence of this expression is effective, i.e., algorithmically computable, and the function F_n is Borel computable according to Definition 374.

7 Converse Coding Theorems for Identification via Channels

In [64], Oohama deals with the ID channel for general noisy channels. He derives a function, which serves as an upper bound of the quantity $1 - (\mu_n + \lambda_n)$. This function has a property that it tends to zero as for noisy channels satisfying the strong converse property. Hence, this result on upper bound of $1 - (\mu_n + \lambda_n)$ together with this property yields the result obtained by Han and Verdú [51]. In particular, for DMCs, Oohama shows that $1 - (\mu_n + \lambda_n)$ tends to zero exponentially as $n \to \infty$ at transmission rates above the ID capacity, deriving an explicit form of the lower bound of this exponent. For derivation of the results, the channel resolvability problem formulated by Han and Verdú [51] was considered. a stronger result on the direct coding theorem for this problem is first established by deriving an exponential lower bound for the approximation error of channel outputs to tend to zero as n goes to infinity. Next, the converse coding theorem for the ID channel is derived based on an idea of converting the direct coding theorem for the channel resolvability problem into the converse coding theorem of the ID channel.

7.1 Main Results

A noisy channel with input set $\mathcal{X}^n$ and output set $\mathcal{Y}^n$ is defined a stochastic matrix $W^n : \mathcal{X}^n \longrightarrow \mathcal{Y}^n$. Formal definition of W^n is $W^n \triangleq \{W^n(\cdot|\boldsymbol{x})\}_{\boldsymbol{x}\in\mathcal{X}^n}$, where $W^n(\cdot|\boldsymbol{x}) \in \mathcal{P}(\mathcal{Y}^n)$ is a conditional distribution on $\mathcal{Y}^n$ given $\boldsymbol{x} \in \mathcal{X}^n$.

Proposition 377 *For any $(n, N_n, \mu_n, \lambda_n)$ ID code with $\mu_n + \lambda_n < 1$, if the rate $r_n = \frac{1}{n}\log\log N_n$ satisfies*

$$r_n \geq R + \frac{\log n}{n} + \frac{1}{n}\log(2\log|\mathcal{X}|)^2$$

then, for any $\gamma \geq 0$, the sum $\mu_n + \lambda_n$ of two error probabilities satisfies the following.

$$1 - \mu_n - \lambda_n \leq \Omega_{n,\gamma}(R|W^n)$$

From this proposition, the following corollary is obtained.

Corollary 378 *For any sequence of ID codes $\{(n, N_1, N_2, \mu_n, \lambda_n)\}_{n=1}^{\infty}$ satisfying $\mu_n + \lambda_n < 1$, $n = 1, 2, \ldots$, if*

$$\liminf_{n\to\infty} r_n \geq R$$

then,for any $\delta \geq 0$, *there exists* $n_0 = n_0(\delta)$

$$1 - \mu_n - \lambda_n \leq \Omega_{n,\gamma}(R - \delta | W^n)$$

Next, the speed of the convergence for the sum of two types of error probabilities to tend to one is discussed.

Theorem 379 *For any sequence of ID codes* $\{(n, N_1, N_2, \mu_n, \lambda_n)\}_{n=1}^{\infty}$ *satisfying* $\mu_n + \lambda_n < 1,\ n = 1, 2, \ldots,$ *if*

$$\liminf_{n\to\infty} r_n \geq R$$

then the sum $\lambda_n + \mu_n$ *of two error probabilities satisfies the following.*

$$\liminf_{n\to\infty} \left(-\frac{1}{n}\right) \log(1 - \mu_n - \lambda_n) \geq \frac{1}{2}\sigma_1(R|\mathbf{W})$$

In particular, if $\mathbf{W}$ *is a DMC with* $W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$, *we have*

$$\liminf_{n\to\infty} \left(-\frac{1}{n}\right) \log(1 - \mu_n - \lambda_n) \geq \frac{1}{2}G(R|W)$$

For more details about the used functions, we refer the reader to [64].

7.2 Average Error Criterion

the following average error criterion is considered.

$$\overline{\mu}_n \triangleq \frac{1}{N_n} \sum_{1 \leq i \leq N_n} \mu_{i,n}, \quad \overline{\lambda}_n \triangleq \frac{1}{N_n} \sum_{1 \leq i \leq N_n} \lambda_{i,n}$$

For $0 \leq \mu, \lambda \leq 1$, let $C_{\mathrm{ID,a}}(\mu, \lambda|\mathbf{W})$ be denoted by the identification capacity defined by replacing the maximum error probability criterion by the above average error probability criterion. Since $\overline{\mu}_n \leq \mu_n$ and $\overline{\lambda}_n \leq \lambda_n$, it is obvious that for any $\mu, \lambda \geq 0$

$$C_{\mathrm{ID}}(\mu, \lambda|\mathbf{W}) \leq C_{\mathrm{ID,a}}(\mu, \lambda|\mathbf{W})$$

It should be shown that $C_{\mathrm{ID,a}}(\mu, \lambda|\mathbf{W})$ has the same upper bound as $C_{\mathrm{ID}}(\mu, \lambda|\mathbf{W})$. An important key result in the case of the average error criterion is given in the following proposition.

Proposition 380 *Fix $\tau > 0$ arbitrarily. For any $(n, N_1, N_2, \overline{\mu}_n, \overline{\lambda}_n)$ code with $\overline{\mu}_n + \overline{\lambda}_n < 1$ if the rate $r_n = \frac{1}{n} \log \log N_n$, satisfies*

$$r_n \geq R + \tau + \frac{\log n}{n} + \frac{1}{n} \log(2 \log |\mathcal{X}|)$$

then, for any $\gamma \geq 0$, the sum $\overline{\mu}_n + \overline{\lambda}_n$ of two average error probabilities satisfies the following.

$$1 - \overline{\mu}_n - \overline{\lambda}_n \leq \Omega_{n,\gamma}(R|W^n) + \nu_{n,\tau}(R, |\mathcal{X}|) \tag{68}$$

where

$$\nu_{n,\tau}(R, |\mathcal{X}|) \triangleq |\mathcal{X}|^{-2n(e^{n\tau}-1)e^{nR}}$$

Since $e^{n\tau} - 1 \geq n\tau$, we have

$$0 \leq \nu_{n,\tau}(R, |\mathcal{X}|) \leq |\mathcal{X}|^{-2n^2\tau e^{nR}}$$

which implies that for each fixed $\tau > 0$, $\nu_{n,\tau}(R, |\mathcal{X}|)$ decays double exponentially as n tends to infinity.

From the above proposition, we obtain the following corollary.

Corollary 381 *For any sequence of ID codes $\{(n, N_1, N_2, \mu_n, \lambda_n)\}_{n=1}^{\infty}$ satisfying $\mu_n + \lambda_n < 1,\ n = 1, 2, \ldots,$ if*

$$\liminf_{n\to\infty} r_n \geq R$$

then,for any $\delta \geq 0$, there exists $n_0 = n_0(\delta)$

$$1 - \mu_n - \lambda_n \leq \nu_{n,\tau}(R - \delta, |\mathcal{X}|) + \Omega_{n,\gamma}(R - \delta|W^n)$$

Now, the following theorem is established from previous results.

Theorem 382 *For any sequence of ID codes $\{(n, N_1, N_2, \mu_n, \lambda_n)\}_{n=1}^{\infty}$ satisfying $\mu_n + \lambda_n < 1,\ n = 1, 2, \ldots,$ if*

$$\liminf_{n\to\infty} r_n > \overline{C}(\mathbf{W})$$

then,

$$\liminf_{n\to\infty} \{\overline{\mu}_n + \overline{\lambda}_n\} = 1,$$

which implies that for any $\mu \geq 0, \lambda \geq 0, \mu + \lambda < 1$ *and any noisy channel* $\mathbf{W}$*,*

$$\underline{C}(\mathbf{W}) \leq C_{\mathrm{ID}}(\mu, \lambda|\mathbf{W}) \leq C_{ID,a}(\mu, \lambda|\mathbf{W}) \leq \overline{C}(\mathbf{W}).$$

In particular, if

$$\underline{C}(\mathbf{W}) = \overline{C}(\mathbf{W})$$

then for any $\mu \geq 0,\ \lambda \geq 0$*, and* $\mu + \lambda < 1$

$$\underline{C}(\mathbf{W}) = C_{ID,a}(\mu, \lambda|\mathbf{W}) = \overline{C}(\mathbf{W})$$

Furthermore, $\overline{\mu}_n + \overline{\lambda}_n$ *converges to one as* $n \to \infty$ *at rates above the ID capacity. This implies that the strong converse property holds with respect to the sum of two types of average error probabilities.*

Now, the following theorem is established from previous results.

Theorem 383 *For any sequence of ID codes* $\{(n, N_1, N_2, \mu_n, \lambda_n)\}_{n=1}^{\infty}$ *satisfying* $\mu_n + \lambda_n < 1$*, if*

$$\liminf_{n\to\infty} r_n \geq R$$

8 Converse Coding Theorems for Identification via Multiple Access Channels

In [63], Oohama deals with the problem of identification via multiple access channels (MACs) for general noisy channels with two inputs and one output finite sets and channel transition probabilities that may be arbitrary for every block length n. First, he established a stronger result on the direct coding theorem for this problem by deriving an upper bound for the approximation error of channel outputs to tend to zero as n goes to infinity. Next, he proved the converse coding theorem by converting the direct coding theorem for the MAC resolvability problem into the converse coding theorem for the ID via MACs.

8.1 Identification via Multiple Access Channels

Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be finite sets. Let $\mathcal{P}(\mathcal{X}^n)$ and $\mathcal{P}(\mathcal{Y}^n)$ be sets of probability distributions on $\mathcal{X}^n$ and $\mathcal{Y}^n$, respectively. A source $\mathbf{X}$ with alphabet $\mathcal{X}$ is the sequence $\{P_X^n : P_X^n \in \mathcal{P}(\mathcal{X}^n)\}_{n=1}^{\infty}$ and a source $\mathbf{Y}$ with alphabet $\mathcal{Y}$ is the sequence

$\{P_Y^{\cdot} P_Y^n \in \mathcal{P}(\mathcal{Y}^n)\}_{n=1}^{\infty}$. Similarly, a noisy channel $\mathbf{W}$ with two inputs alphabets $\mathcal{X}$ and $\mathcal{Y}$ and one output alphabet $\mathcal{Z}$ is a sequence of conditional distributions $\{W^n(\cdot|\cdot,\cdot)\}_{n=1}^{\infty}$, where $W^n(\cdot|\cdot,\cdot) = \{W^n(\cdot|\boldsymbol{x},\boldsymbol{y}) \in \mathcal{P}(\mathcal{Z}^n)\}_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}^n\times\mathcal{Y}^n}$. Next, for $P_{X^n} \in \mathcal{P}(\mathcal{X}^n)$, $P_{Y^n} \in \mathcal{P}(\mathcal{Y}^n)$ and $\boldsymbol{z} \in \mathcal{Z}^n$, set

$$P_{X^n} P_{Y^n} W^n \boldsymbol{z} = \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}^n\times\mathcal{Y}^n} P_{X^n}(\boldsymbol{x}) P_{Y^n}(\boldsymbol{y}) W^n(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}), \tag{69}$$

which becomes a probability distribution on $\mathcal{Z}^n$. We denote it by $P_{X^n} P_{Y^n} W^n = \{P_{X^n} P_{Y^n} W^n(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{Z}^n}$. Set $P_{Z^n} = P_{X^n} P_{Y^n} W^n$ and call P_{Z^n} the response of (P_{X^n}, P_{Y^n}) through noisy channel W^n.

Definition 384 An $(n, N_1, N_2, \mu_n, \lambda_n)$ ID code for W^n is a collection

$$\{(P_{X^n|i}, (P_{Y^n|j}, D_{i,j}),\ i = 1, 2, \ldots, N_1,\ j = 1, 2, \ldots, N_2\}$$

such that

1. $P_{X^n|i} \in \mathcal{P}(\mathcal{X}^n)$, $P_{Y^n|j} \in \mathcal{P}(\mathcal{Y}^n)$,
2. $D_{i,j} \in \mathcal{Z}^n$,
3. $P_{Z^n|i,j}$ is the response of $(P_{X^n|i}(P_{Y^n|j})$,
4. $\mu_{n,ij} = P_{Z^n|i,j}(D_{i,j}^c)$, $\mu_n = \max_{\substack{1\le i\le N_1\\ 1\le j\le N_2}} \mu_{n,ij}$
5. $\lambda_{n,ij} = \max_{(k,l)\neq(i,j)} P_{Z^n|k,l}(D_{i,j})$, $\max_{\substack{1\le i\le N_1\\ 1\le j\le N_2}} \lambda_n = \lambda_{n,ij}$

The rate of an $(n, N_1, N_2, \mu_n, \lambda_n)$ ID code is defined by

$$r_{i,n} \triangleq \frac{1}{n} \log\log N_i, \quad i = 1, 2 \tag{70}$$

A rate pair (R_1, R_2) is said to be (μ, λ)-achievable ID rate pair if there exists an $(n, N_1, N_2, \mu_n, \lambda_n)$ code such that

$$\limsup_{n\to\infty} \mu_n \le \mu, \tag{71}$$

$$\limsup_{n\to\infty} \lambda_n \le \lambda, \tag{72}$$

$$\liminf_{n\to\infty} r_{i,n} \ge R_i, \qquad i = 1, 2 \tag{73}$$

The set of all (μ, λ)-achievable ID rate pairs for W is denoted by $C_{\mathrm{ID}}(\mu, \lambda|W)$, which we call the (μ, λ) ID capacity region.

8.2 Main Results

The main result of [63] for identification via MACs is the following.

Proposition 385 *For any $(n, N_1, N_2, \mu_n, \lambda_n)$ code with $\mu_n + \lambda_n < 1$, if the rate $r_{i,n} = \frac{1}{n} \log\log N_i$ satisfies*

$$r_{1,n} \geq R_1 + \frac{\log n}{n} + \frac{1}{n} \log\log(3|\mathcal{X}|)^2, \tag{74}$$

$$r_{2,n} \geq R_2 + \frac{\log n}{n} + \frac{1}{n} \log\log(3|\mathcal{Y}|)^2, \tag{75}$$

then, for any $\gamma \geq 0$, the sum $\mu_n + \lambda_n$ of two error probabilities satisfies the following.

$$1 - \mu_n - \lambda_n \leq \Omega_{n,\gamma}(R_1, R_2|W^n)$$

For more details, we refer to [63].

Theorem 386 *For any sequence of ID codes $\{(n, N_1, N_2, \mu_n, \lambda_n)\}_{n=1}^{\infty}$ satisfying $\mu_n + \lambda_n < 1,\ n = 1, 2, \ldots,$ if*

$$\liminf_{n\to\infty} r_{i,n} \geq R_i,\ i = 1, 2, \quad (R_1, R_2) \notin \overline{\mathcal{C}}'(\mathbf{W})$$

then,

$$\liminf_{n\to\infty} \{\mu_n + \lambda_n\} = 1,$$

which implies that for any $\mu \geq 0, \lambda \geq 0, \mu + \lambda < 1$ and any noisy channel $\mathbf{W}$,

$$\underline{\mathcal{C}}(\mathbf{W}) \subset \mathcal{C}_{\mathrm{ID}}(\mu, \lambda|\mathbf{W}) \subset \overline{\mathcal{C}}'(\mathbf{W}).$$

Furthermore, $\mu_n + \lambda_n$ converges to one as $n \to \infty$ at rates above the ID capacity. This implies that the strong converse property holds with respect to the sum of two types of error probabilities.

Proposition 387 *Fix $\tau > 0$ arbitrarily. For any $(n, N_1, N_2, \overline{\mu}_n, \overline{\lambda}_n)$ code with $\overline{\mu}_n + \overline{\lambda}_n < 1$ if the rate $r_{i,n} = \frac{1}{n} \log\log N_i, \quad i = 1, 2$, satisfy*

$$r_{1,n} \geq R_1 + \tau + \frac{\log n}{n} + \frac{1}{n} \log\log(|\mathcal{X}|)^2, \tag{76}$$

$$r_{2,n} \geq R_2 + \tau + \frac{\log n}{n} + \frac{1}{n} \log\log(|\mathcal{Y}|)^2, \tag{77}$$

then, for any $\gamma \geq 0$, the sum $\mu_n + \lambda_n$ of two average error probabilities satisfies the following.

$$1 - \overline{\mu}_n - \overline{\lambda}_n \leq \Omega_{n,\gamma}(R_1, R_2|W^n) + \nu_{n,\tau}(R_1, R_2, |\mathcal{X}|, |\mathcal{Y}|)$$

where

$$\begin{aligned}\nu_{n,\tau}(R_1, R_2, |\mathcal{X}|, |\mathcal{Y}|) \triangleq\ & |\mathcal{X}|^{-2n(e^{n\tau}-1)e^{nR_1}} + |\mathcal{Y}|^{-2n(e^{n\tau}-1)e^{nR_2}} \\ & + |\mathcal{X}|^{-2n(e^{n\tau}-1)e^{nR_1}} \cdot |\mathcal{Y}|^{-2n(e^{n\tau}-1)e^{nR_2}} \end{aligned} \tag{78}$$

Since $e^{n\tau} - 1 \geq n\tau$, we have

$$0 \leq \nu_{n,\tau}(R_1, R_2, |\mathcal{X}|, |\mathcal{Y}|) \tag{79}$$

$$\leq |\mathcal{X}|^{-2n(e^{n\tau}-1)e^{nR_1}} + |\mathcal{Y}|^{-2n(e^{n\tau}-1)e^{nR_2}} + |\mathcal{X}|^{-2n(e^{n\tau}-1)e^{nR_1}} \cdot |\mathcal{Y}|^{-2n(e^{n\tau}-1)e^{nR_2}} \tag{80}$$

$$\leq 3|\mathcal{X}|^{-2n(e^{n\tau}-1)e^{nR_1}} \cdot |\mathcal{Y}|^{-2n(e^{n\tau}-1)e^{nR_2}} \tag{81}$$

which implies for each fixed $\tau > 0$, $\nu_{n,\tau}(R_1, R_2, |\mathcal{X}|, |\mathcal{Y}|)$ decays double exponentially as n tends to infinity.

Theorem 388 *For any sequence of ID codes $\{(n, N_1, N_2, \mu_n, \lambda_n)\}_{n=1}^{\infty}$ satisfying $\mu_n + \lambda_n < 1,\ n = 1, 2, \ldots,$ if*

$$\liminf_{n\to\infty} r_{i,n} \geq R_i,\ i = 1, 2, \quad (R_1, R_2) \notin \overline{\mathcal{C}}'(\mathbf{W})$$

then,

$$\liminf_{n\to\infty} \{\mu_n + \lambda_n\} = 1,$$

which implies that for any $\mu \geq 0, \lambda \geq 0, \mu + \lambda < 1$ and any noisy channel $\mathbf{W}$,

$$\underline{\mathcal{C}}(\mathbf{W}) = \mathcal{C}_{\mathrm{ID}}(\mu, \lambda|\mathbf{W}) = \mathcal{C}_{ID,a}(\mu, \lambda|\mathbf{W}) = \overline{\mathcal{C}}'(\mathbf{W}).$$

Furthermore, $\mu_n + \lambda_n$ converges to one as $n \to \infty$ at rates above the ID capacity. This implies that the strong converse property holds with respect to the sum of two types of error probabilities.

9 Explicit Constructions for Identification

In this section we review the results of [76] and [44]. In particular we report the results of [76] with the observations and reformulations of [44]. These include a capacity-achieving explicit construction of identification codes based on Reed-Solomon codes.

A common way of constructing identification codes is, as first done both in chapters "Identification via Channels" and "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas", by having all the channel errors corrected by a transmission code and performing identification for the almost-noiseless channel provided by the transmission code. In the noiseless channel then, as done in chapter "Identification in the Presence of Feedback: A Discovery of New Capacity Formulas", identification can be done by associating to each identity i a function T_i and by sending a single input-output pair $(j, T_i(j))$ for verification through the channel, with j picked uniformly at random using local randomness. The receiver, given another identity i', can verify whether the two chosen identities are the same by computing its own output $T_{i'}(j')$ on the input part j' of the received input-output pair (j', t'), and verifying that the received and computed output are the same ($T_{i'}(j') = t'$?). If the channel is noiseless, then $(j', t') = (j, T_i(j))$ this scheme has zero missed identification error (if $i = i'$ then the computed outputs will match) and false identification bounded by the fraction of inputs any two identities map to the same output (if $i \neq i'$ then an incorrect accept will happen only on those j such that $T_{i'}(j) = T_i(j)$) [9, 76]. In case of a noisy channel, the error probability of the transmission code (the probability that $(j', t') \neq (j, T_i(j))$) adds to the missed- and false-identification error probabilities of the identification code [8, 76].

The important observation is that these functions are none other than error correction codes seen and used with a different approach than error correction. In order to make this observation we first recall the definition of an error correction code.

Definition 389 (Error-Correction Codes) Let n, N, d, s be integers and let $\mathcal{X}$ be without loss of generality the alphabet $\mathcal{X} = [s]$. An error-correction code of length n, size N and distance d over an alphabet of size s, or simply an $(n, N, d)_s$ error-correction code, is a subset of words $\mathcal{C} = \{\mathbf{c}_i\}_{i\in[N]} \subseteq \mathcal{X}^n$ such that the hamming distance between any two codewords is at least d. If $N = s^k$ for some integer k, then the codes can be use to encode the elements of $\mathcal{X}^k$ and the codes are then called $(n, k, d)_s$ block codes.

Observation 390 *A set of N functions $\{T_i\}_{i=1,\ldots,N}$ from inputs of size M to outputs of size q such that, for any two of these functions their outputs are equal on at most $M - d$ inputs, corresponds to an error correction code of size N, blocklength M, alphabet of size q and distance d, and vice versa. To each function corresponds a*

Arguably, in such cases the transmission code should have the error probability comparable to the error probability of the identification code.

codeword, obtained by concatenating all the possible function outputs in order as $c_i = T_i(1) \dots T_i(M)$ *Similarly, given any correction code, each codeword defines a function, namely the function* $T_i(j) = c_{ij}$ *that map positions to the symbols of the codeword.*

For the sake of clarity, we repeat the identification pre- and post-processing using error-correction codes. Given an error-correction code of size N, blocklength n, alphabet of size q and minimum Hamming distance d, namely a $(n, N, d)_q$ error-correction code, an identification code is constructed as follows. The error-correction code is not used to correct error, but instead is used in a different manner. The codewords c_i of the error-correction code are associated each to an identity i. The identification sender of identity i randomly and uniformly chooses a position j from $1, 2, \dots, n$, and then sends j and the jth letter c_{ij} to the receiver, using a transmission code if the channel is noisy. The receiver must make a choice on what identity he is interested, say i'. Upon receiving j, c_{ij}, the receiver checks whether jth letter of the codeword associated with his interested message i'th is c_i. Namely, it checks whether $c_{i'j} = c_{ij}$ and says "yes/accept i'" if so, or "no/reject i'" otherwise. In the noiseless case, the only possible error is the false identification error (error of the second kind), which occurs only when the receiver is interested in a different message than the one sent, and two codewords c_i and $c_{i'}$ have the same letter as in the ith position. Thus, the probability of false identification is bounded by

$$\lambda_2 \leq 1 - \frac{d}{n}. \tag{82}$$

Again, we highlight that this is a bound on the false identification error only in the absence of a transmission error.

In order to uniquely refer to the input-output pairs $(j, T_i(j))$ produced in the pre-processing we will call j the *randomness* and $T_i(j)$ the *tag*, since $T_i(j)$ takes the role of a small label as will become clear next. For convenience we may then call the functions T_i *tagging functions*. Traditionally, the identities have been called messages [8, 9]. However, the identities are not the messages that are sent through the transmission code. Furthermore, transmission can be performed in parallel to identification without trade-off, meaning that the capacity of both can be achieved at the same time [50]. This is because the goal of the pre-processing is to use all the capacity of the channel to send the local randomness j and produce common randomness between the sender and receiver. A small tag size (the size of $T_i(j)$ or c_{ij}) that asymptotically does not use any capacity [9] is then enough to allow a rate of identities that grows doubly exponentially in the blocklength. The intuition is that identification is performed by verifying the two tags, the senders and the receivers, of a random challenge. In other works [24–26], the randomness has been called a colouring number, the tag a colour and the identity a colouring.

Observation 391 *In [55, 76] error-correction codes are used to construct constant-weight binary codes that are then used to construct identification codes following the method in [76]. This at first glance might seem like a different way of constructing identification codes, however it is implicitly the same use of error-correction codes presented above. The way the constant-weight binary codes are further encoded in [76] (via the use of the incidence matrix of the binary code), its an implicit encoding of the randomness-tag pair. Both in [55, 76] and in the scheme above, the information that is sent through the channel via the transmission code is a randomness-tag pair.*

9.1 Conditions for Achieving Identification Capacity

As just explained, there is no difference between error correction codes and sets of mapping functions. However, for the sake of clarity, it is more convenient to keep the discussion in terms of mapping functions. Thus, to aid the exposition we will make the following definitions.

Definition 392 (Tag/Coloring Code) Let I, M, H be integers and let $\varepsilon \in [0, 1]$. A tag/coloring code of I identities, from messages of size M to tags of length H, with error ε, or simply a (I, M, H, ε) tag code, is a collection of maps $\mathcal{T} = \{T_i : [M] \to [H]\}_{i\in[I]}$, such that the pairwise collisions satisfy $\frac{1}{M}\sum_{m\in[M]} \delta_{T_i(m),T_{i'}(m)} \leq \varepsilon$ for all $i \neq i'$.

If the messages and the tags are strings over the same alphabets, namely if for some integer q we have $M = q^m$ and $H = q^h$ then we denote the tag code by $[\log_q \log_q I, m, h, \varepsilon]_q$, and with $[\log\log I, m, h, \varepsilon]$ if $q = 2$.

The previous observations then simply mean that an $(n, N, d)_q$ error-correction code $\{\mathbf{c}_i\}$ defines an $(N, n, q, 1 - \frac{d}{n})$ tag code $\{T_i\}$ via $T_i(m) = c_{im}$, and vice versa.

In particular, a $[q^n, q^k, d]_q$ error-correction code defines a $[k, n, 1, 1 - \frac{d}{n}]_q$ tag code.

The following is a rephrasing of the main achievability theorems of identification that reduce the problem of identification over a noisy channel to the problem of identification over a noiseless channel.

Construction 393 ([8, 76]) *Let $\{(x_m, D_m)\}_{m\in[M]}$ be a (n, M, ε) transmission code, let $\{(x'_m, D'_m)\}_{m\in[M']}$ be a (n', M', ε') transmission code, and let $\mathcal{T}$ be a (I, M, M', μ) tag code. Then*

$$\left\{\left(x^{\mathrm{TI}}_{mi} = x_m \cdot x'_{T_i(m)},\ D^{\mathrm{TI}}_{mi} = D_m \times D'_{T_i(m)}\right)\right\}_{m\in[M],i\in[I]}$$

is an $(n+n', M, I, \varepsilon+\varepsilon', \varepsilon+\varepsilon'+\mu)$ *transmission–identification code as defined in [50]. Let* $\{U_i\}_{i\in[I]}$ *be i.i.d. uniform random variables over* $[M]$*, then*

$$\left\{\left(P_i^{\mathrm{ID}} = x_{U_i i}^{\mathrm{TI}} = x_{U_i}\cdot x'_{T_i(U_i)},\ D_i^{\mathrm{ID}} = \bigcup\nolimits_{m\in[M]} D_m \times D'_{T_i(m)}\right)\right\}_{i\in[I]}$$

is an $(n+n', I, \varepsilon+\varepsilon', \varepsilon+\varepsilon'+\mu)$ *identification code.*

We will now from the behaviour of the rates of these codes, derive conditions required for tag codes to achieve capacity as done in [76].

Remark 394 In [76] the term "optimal" is used in the meaning of capacity-achieving. However, since optimal is also used to mean optimal parameters at finite blocklengths (which is an independent condition from being capacity achieving), here we will avoid the term "optimal"

The rates of the new codes are easily computed as follows. Let us first compute the transmission rate achieved by using the two transmission codes in parallel as

$$R := \frac{\log M + \log M'}{n+n'}. \tag{83}$$

The transmission–identification rate tuple for the above construction, where the second rate is also the rate of the identification code, is then

$$(R_{\mathrm{T}}, R_{\mathrm{ID}}) := \left(\frac{\log M}{n+n'}, \frac{\log\log I}{n+n'}\right) \tag{84}$$

$$= \left(\frac{\log M}{\log M + \log M'}\cdot R,\ \frac{\log\log I}{\log M + \log M'}\cdot R\right) \tag{85}$$

Notice now that the number of identities I is limited by the number of all functions M'^{M} from messages to tags and thus

$$\log\log I \le \log M + \log\log M' \le \log M + \log M',$$

giving

$$(R_{\mathrm{T}}, R_{\mathrm{ID}}) \le \left(\frac{\log M}{\log M + \log M'}\cdot R,\ \frac{\log M + \log\log M'}{\log M + \log M'}\cdot R\right) \le (R,\ R)\,. \tag{86}$$

In particular, this is also a specialized but simple converse proof for such constructions, namely that the transmission–identification and identification capacities they can achieve cannot exceed the transmission capacity of the channel. With a more detailed analysis we can gain some insight in what property of tag codes we should look for(beyond the obvious requirement that μ goes to zero asymptotically) in order to achieve the best rates with tagged transmission codes. A refinement of

Eq. (86) actually shows that (asymptotically, namely if we now think of sequence of codes of increasing blocklength) the achieved R_{ID} are always lower than the achieved R_{T}. Indeed we have more precisely that

$$R_{\mathrm{T}} = \frac{1}{1+\frac{\log M'}{\log M}} \cdot R, \tag{87}$$

$$R_{\mathrm{ID}} \leq \frac{1}{1+\frac{\log M'}{\log M}} \cdot R + \frac{\log\log M'}{\log M + \log M'} \cdot R \tag{88}$$

$$\leq \frac{1}{1+\frac{\log M'}{\log M}} \cdot R + \frac{\log\log M'}{\log M'} \cdot R. \tag{89}$$

Since $\log\log x/\log x \in o(1)$, namely it tends to zero as x goes to infinity, when taking asymptotically larger codes we see that the achieved identification rates are always smaller than the achieved transmission rates. In particular, if we the identification rate achieves the transmission capacity, it also means that both the transmission rate R_{T} and the identification rate R_{ID} achieve the transmission capacity simultaneously. We will see that asymptotically large M' are needed to send the error μ to zero, thus we can restrict to this case.

Remark 395 Notice that rates $R_{\mathrm{ID}} > R_{\mathrm{T}}$ are of course achievable with general codes. The reason why we obtain that $R_{\mathrm{ID}} \leq R_{\mathrm{T}}$ for the codes above, is because by construction they do not use any randomness at the encoder, but they can be easily generalized to codes where the random input to the tag code is generated partly by messages and partly a uniform randomness. This will use some of the transmission capacity to produce and send the randomness, thus artificially lowering R_{T} below R_{ID}.

We can now extract sufficient conditions for the identification and transmission rates to achieve the transmission capacity. In order to do this we can rewrite the transmission–identification rate as

$$(R_{\mathrm{T}}, R_{\mathrm{ID}}) = \left(\frac{1}{1+\frac{\log M'}{\log M}} \cdot R,\ \frac{\log\log I}{\log M} \cdot \frac{1}{1+\frac{\log M'}{\log M}} \cdot R \right)$$

$$= \left(\frac{1}{1+\frac{\log M'}{\log M}} \cdot R,\ \frac{\log\log I}{\log M} \cdot R_{\mathcal{T}} \right)$$

We thus have that for R_{ID} to achieve the identification capacity with increasingly larger codes we firstly need R to achieve capacity, namely capacity achieving transmission codes, and then we need $\frac{\log M'}{\log M}$ to approach zero and $\frac{\log\log I}{\log M}$ to approach one. All this while at the same time having the transmission errors ε and

ε' going to zero, which is guaranteed by capacity achieving transmission codes, and the tag-code error μ also to go to zero. This justifies the following definition.

Definition 396 A sequence of $(I_n, M_n, H_n, \varepsilon_n)$ tag codes is achieving identification capacity if

$$\begin{aligned} \frac{\log\log I_n}{\log M_n} &\to 1 \\ \frac{\log H_n}{\log M_n} &\to 0 \\ \varepsilon_n &\to 0. \end{aligned} \tag{90}$$

In case of $[\pi_n, m_n, h_n, \varepsilon_n]_{q_n}$ tag codes the conditions become

$$\begin{aligned} \frac{\pi_n}{m_n} &\to 1 \\ \frac{h_n}{m_n} &\to 0 \\ \varepsilon_n &\to 1. \end{aligned} \tag{91}$$

The following is a reformulation of the main result in [76] with the error correction code used to compute the tags extracted. The original formulation uses these code to create a identification-capacity achieving constant-weight binary code.

Proposition 397 ([76]) *Let q_n be a sequence of prime powers, and k_n and δ_n integer sequences satisfying*

$$1 \preceq \delta \prec k \prec q$$

(in big-O notation: $\delta_n \in \Omega(1)$, $k_n \in \omega(\delta_n)$, and $q_n \in \omega(k_n)$).

Let $\mathcal{C}$ be a $[q_n, k_n]$ and $\mathcal{C}'$ a $[q_n^{k_n}, q_n^{k_n-\delta_n}]$ Reed-Solomon code, then $\mathcal{C} \circ \mathcal{C}'$ is an error-correction code achieving identification capacity.

9.2 A Simple Achievability Proof of Identification

The existence of identification-capacity achieving codes exist was first proven using a counting argument in the first appearance of identification [8]. However, in general these do not need to be tag codes, but rather collections of subsets over which to take the uniform distribution. In such codes, there is thus no distinction between the message/randomness and the tag. This is inconvenient if one is interested in adding security requirement on the identities. In the absence of messages, since the identification rate depends on how much randomness is

available, performing identification with tag codes allows to achieve higher rates by sending the randomness in plain while sending the tag secretly. Such proofs use a random coding argument to prove that tag codes achieving identification capacity exist and can be chosen randomly with high probability [9]. The explicit identification codes constructed in [55, 76] actually implicitly construct tag codes. Because of the correspondence to error correction codes, an analogous to the counting argument from [8] for the case of tagging function already exists and this is the Gilbert–Varshamov bound, which proves the existence of tag codes in any input-output sizes without any random coding. The Gilbert–Varshamov bound had already been used to get an achievability proof for identification in [10] were it was used to construct constant-weight binary codes.

Lemma 398 (Gilbert–Varshamov Bound) *Let $M \geq 2$ and $H \geq 3$ be finite integers, and let $\Omega = 0, \ldots, M-1$.*

Then there exist $(I, M, H, \Omega/M)$ tag codes with

$$I > \left(\frac{H}{2}\right)^{M} \frac{H-2}{(H-1)^{M-\Omega}-1} > \frac{H^{\Omega}}{2^{M}}. \tag{92}$$

In particular, let $c > 0$, then for $\log H \geq c+1$ there always exist

$$\left(2^{cM},\ M,\ H,\ \frac{c+1}{\log H}\right)$$

tag codes (by choosing $\Omega = \lfloor (c+1)M/\log H \rfloor$ in the first lower bound).

Corollary 399 *For and $M \geq 2$ and $H \geq 4$ there always exist*

$$\left(2^{M},\ M,\ H,\ \frac{2}{\log H}\right)$$

tag codes.

Since there exist $\left(2^{M}, M, H, 2/\log H\right)$ tag codes, by choosing a sequence of tag codes such that $1 \prec \log H_n \prec \log M_n$, we automatically get a sequence achieving identification capacity and plugging it in Construction 393 we obtain a capacity achieving identification code.

10 Secure Storage for Identification

In this section, we present the main contributions in [13].

In [10] the authors consider identification for the wiretap channel. They show that the number of messages that can reliably be identified in this case grows doubly exponentially with the block length. The secure identification capacity

even equals the Shannon capacity of the main channel. This result is generalized in [23, 24] where the authors consider robust identification for wiretap channels. In [53] and [58] the authors interpret the discrete memoryless source from the source model as a biometric source. That is why they consider the privacy leakage of the protocols for secret key generation.

In [13] the source model from [6] is considered for generating common randomness. But in contrast to [7] the privacy leakage of the corresponding protocols is considered while interpreting the source as a biometric source as in [53] and [58]. This common randomness is then used for identification. So the contribution is twofold. We characterize the capacity for common randomness generation from a discrete memoryless source while considering privacy leakage. Then protocols for identification using a discrete memoryless source are constructed. In contrast to [7] and [3] the helper message is assumed to be stored on a public database. The protocols for identification are constructed such that they provide secrecy. So these protocols allow for secure storage for identification. The privacy leakage of these protocols are also considered. Secure storage on a public database for identification is considered. The authors make use of a physical unclonable function (PUF) source. A PUF source is essentially equivalent to a biometric source. The output of a biometric source is assumed to uniquely characterize a person whereas the output of a PUF source uniquely characterizes a device. This allows to use the output of a PUF source for secure storage as described in this work.

10.1 Storage for Identification Model

Consider the secure storage for identification process depicted in Fig. 4. The process consists of two phases. In the first phase the system gets the message d that should be stored on the database consisting of k storage cells which can each store a value from the alphabet $\mathcal{U}$. The system reads X^n from the PUF source. The system then generates U^k from X^n using an encoder depending on d and stores U^k on the public database. In the second phase the system reads U^k from the database and Y^n from

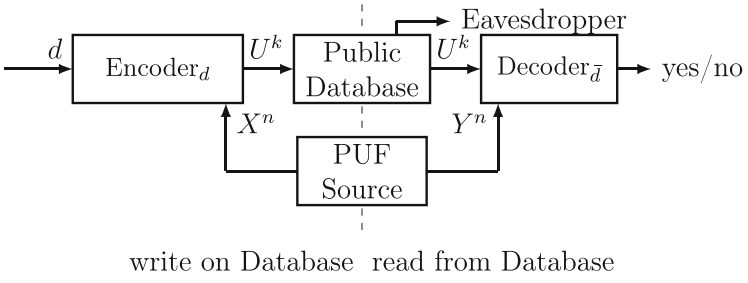


Fig. 4 Secure storage for identification process

the PUF source. The system then uses a decoder, depending on the message of interest $\bar{d}$, to decide whether $\bar{d}$ is the message stored on the database, making use of Y^n and U^k.

An information theoretic model of the storage process for identification is now defined.

Definition 400 Let $k, n \in \mathbb{N}$. The *storage for identification model* consists of the alphabet $\mathcal{U}$, a discrete memoryless multiple source (DMMS) P_{XY} on the alphabet $\mathcal{X} \times \mathcal{Y}$, a set of (possibly randomized) encoders $\{\Phi_d\}_{d\in\mathcal{D}}$, $\Phi_d \colon \mathcal{X}^n \to \mathcal{U}^k$ for all $d \in \mathcal{D}$ and a set of (possibly randomized) decoders $\{\Psi_d\}_{d\in\mathcal{D}}$, $\Psi_d \colon \mathcal{U}^k \times \mathcal{Y}^n \to \{0, 1\}$ for all $d \in \mathcal{D}$. Let X^n and Y^n be the random variables (RVs) generated from P_{XY}.We call $(\{\Phi_d\}_{d\in\mathcal{D}}, \{\Psi_d\}_{d\in\mathcal{D}})$ a *storage for identification protocol*.

The authors consider an eavesdropper who reads from the public database and assume that he wants to identify a specific message. The eavesdropper knows the protocol used and. T could be even assumed that he knows the message the decoder wants to identify. The authors want that the sum of the probability that the eavesdropper makes an error of the first kind and the probability that the eavesdropper makes an error of the second kind is close to one.

The output of the PUF source uniquely characterizes a device, so the authors possibly want to reuse parts of it. That is why the attacker should not have a lot of information about the PUF source output X^n.

This motivates the following definition of achievability for the storage for identification model.

Definition 401 Let $B > 0$. We call the tuple $(R_{\mathrm{ID}}, R_{\mathrm{PL}})$ $R_{\mathrm{ID}}, R_{\mathrm{PL}} \geq 0$ an *achievable rate pair* for the storage for identification model if for every $\delta > 0$ there is a $k_0 = k_0(\delta)$ such that for all $k \geq k_0$ and $n = \lceil B \cdot k \rceil$ there exists a storage for identification protocol such that for all $d, \bar{d} \in \mathcal{D}$, $d \neq \bar{d}$

$$\begin{aligned} \Pr(\Psi_d(\Phi_d(X^n), Y^n) = 0) &\leq \delta \\ \Pr(\Psi_d(\Phi_{\bar{d}}(X^n), Y^n) = 1) &\leq \delta \\ \Pr(\Psi_d^E(\Phi_d(X^n)) = 0) + \Pr(\Psi_d^E(\Phi_{\bar{d}}(X^n)) = 1) &\geq 1 - \delta \qquad (93) \\ \frac{1}{k} \log\log|\mathcal{D}| &\geq R_{\mathrm{ID}} - \delta \\ \frac{1}{k} I(\Phi_d(X^n); X^n) &\leq R_{\mathrm{PL}} + \delta \end{aligned}$$

for all strategies $\{\Psi_d^E\}_{d\in\mathcal{D}}$ of the eavesdropper. We denote the corresponding storage for identification protocols by *secure storage protocols*. We call the set of all rate pairs that are achievable using such storage for identification protocols *capacity region* $\mathcal{R}_{\mathrm{ID}}(B)$.

Lemma 402 *Let $B > 0$. $R_{\mathrm{ID}}(B)$ is a closed set.*

10.2 Results on Common Randomness and Secret Key Generation

Some results concerning common randomness (CR) and secret key (SK) generation from a DMMS are needed now. The following information theoretic model is considered.

Definition 403 Let $n \in \mathbb{N}$. The *source model* consists of a DMMS P_{XY}, a (possibly randomized) encoder $F\colon \mathcal{X}^n \to \mathcal{K} \times \mathcal{M}$ and a (possibly randomized) decoder $G\colon \mathcal{Y}^n \times \mathcal{M} \to \hat{\mathcal{K}}$. Let X^n and Y^n be the output of the DMMS. The RVs (K, M) are generated from X^n using F and the RV $\hat{K}$ is generated from (Y^n, M) using G. We call (F, G) a *CR/SK generation protocol*.

Generation of common randomness [7] is now considered.

Definition 404 Let $L \geq 0$. We call $R(L) \geq 0$ an *achievable CR generation rate with forward communication rate constraint* L for the source model if for every $\delta > 0$ there is an $n_0 = n_0(\delta)$ such that for all $n \geq n_0$ there is a CR/SK generation protocol such that

$$\frac{1}{n} \log |\mathcal{M}| \leq L + \delta$$

$$\Pr(K = \hat{K}) \geq 1 - \delta$$

$$\frac{1}{n} H(K) \geq R - \delta \tag{94}$$

$$\frac{1}{n} \log |\mathcal{K}| \leq c \tag{95}$$

for a $c > 0$. We denote the corresponding CR/SK generation protocols by *CR generation protocols with rate constraint*. We denote the supremum of all achievable CR generation rates with forward communication rate constraint L by *CR capacity* $C_{\mathrm{CR}}(L)$.

We now also consider privacy leakage for the source model. This makes sense when we assume that the DMMS, that is part of the source model, models a PUF source.

Definition 405 We call the triple $(R_{\mathrm{CR}}, R_{\mathrm{FC}}, R_{\mathrm{PL}})$, $R_{\mathrm{CR}}, R_{\mathrm{FC}}, R_{\mathrm{PL}} \geq 0$ an *achievable CR generation rate versus forward communication rate versus privacy leakage rate triple* for the source model if for every $\delta > 0$ there is an $n_0 = n_0(\delta)$ such that for all $n \geq n_0$ there is a CR/SK generation protocol such that

$$\Pr(K = \hat{K}) \geq 1 - \delta \tag{96}$$

$$\frac{1}{n} \log |\mathcal{M}| \leq R_{\mathrm{FC}} + \delta$$

$$\frac{1}{n} I(M; X^n) \leq R_{\mathrm{PL}} + \delta$$
$$\frac{1}{n} H(K) \geq R_{\mathrm{CR}} - \delta$$
$$\frac{1}{n} \log |\mathcal{K}| \leq c,$$

for a $c > 0$. We denote the corresponding CR/SK generation protocols by *private CR generation protocols*. The set of all rate triples that are achievable using private CR generation protocols is denoted by the *CR capacity region* $\mathcal{R}_{\mathrm{CR}}$.

In a first approach private CR generation protocols with deterministic encoders and decoders (f, g) are considered. The corresponding CR capacity region is denoted by $\mathcal{R}^{\mathrm{d}}_{\mathrm{CR}}$. In [7] the authors also consider deterministic CR generation protocols with rate constraint and characterize the corresponding capacity, which is denoted in [13] by $C^{\mathrm{d}}_{\mathrm{CR}}(L)$. The following result is proved in [7].

Theorem 406 *It holds that*

$$C^{\mathrm{d}}_{\mathrm{CR}}(L) = \max_V I(V; X),$$

where the maximization is over all RVs V such that $V - X - Y$ and $I(V; X) - I(V; Y) \leq L$. We also only have to consider RVs V with $|\mathcal{V}| \leq |\mathcal{X}|$.

The authors in [13] use this result to characterize $\mathcal{R}^{\mathrm{d}}_{\mathrm{CR}}$.

Theorem 407 *It holds that*

$$\mathcal{R}^{\mathrm{d}}_{\mathrm{CR}} = \bigcup_{V:\, V-X-Y} \left\{ (R_{\mathrm{CR}}, R_{\mathrm{FC}}, R_{\mathrm{PL}}) : \begin{array}{r} 0 \leq R_{\mathrm{CR}} \leq I(V; X) \\ R_{\mathrm{FC}} \geq I(V; X|Y) \\ R_{\mathrm{PL}} \geq I(V; X|Y) \end{array} \right\} \tag{97}$$

and the authors in [13] only have to consider RVs V with $|\mathcal{V}| \leq |\mathcal{X}| + 1$.

Secret key generation [6] with perfect secrecy is also considered.

Definition 408 $R \geq 0$ is called an *achievable SK generation rate* for the source model if for every $\delta > 0$ there is an $n_0 = n_0(\delta)$ such that for all $n \geq n_0$ there is a CR/SK generation protocol such that

$$\Pr(K = \hat{K}) \geq 1 - \delta \tag{98}$$
$$I(K; M) = 0$$
$$H(K) = \log |\mathcal{K}|$$
$$\frac{1}{n} \log |\mathcal{K}| \geq R - \delta.$$

The corresponding CR/SK generation protocols are denoted by *perfect SK generation protocols*. We call the supremum of all achievable SK generation rates *SK capacity* C_{SK}.

In [6] the authors prove the following result.

Theorem 409 *It holds that*

$$C_{\mathrm{SK}} = I(X;Y).$$

10.3 Achievability Result for Secure Storage for Identification

Now $\mathcal{R}_{\mathrm{ID}}(B)$ is characterized. In order to do so the authors in [13] make use of results for CR and SK generation while considering the privacy leakage. Firstly consider deterministic secure storage for identification protocols $(\{\phi_d\}_{d\in\mathcal{D}}, \{\psi_d\}_{d\in\mathcal{D}})$. The corresponding capacity region is denoted by $\mathcal{R}_{\mathrm{ID}}^{\mathrm{d}}(B)$. The following achievability result is obtained.

Theorem 410 *It holds that*

$$\mathcal{R}_{\mathrm{ID}}^{\mathrm{d}}(B) \supseteq \bigcup_{V} \{(R_{\mathrm{ID}}, R_{\mathrm{PL}}) : 0 \leq R_{\mathrm{ID}} \leq I(V;X)B \\ R_{\mathrm{PL}} \geq I(V;X|Y)B\}, \tag{99}$$

where the union is over all RVs V such that $V - X - Y$ and $I(V;X|Y)B \leq \log|\mathcal{U}|$.

Now we consider randomized secure storage for identification protocols.

Theorem 411 *It holds that*

$$\mathcal{R}_{\mathrm{ID}}(B) \supseteq \bigcup_{\epsilon>0} \bigcup_{V} \{(R_{\mathrm{ID}}, R_{\mathrm{PL}}) : 0 \leq R_{\mathrm{ID}} \leq \log|\mathcal{U}| + I(V;Y)B \\ R_{\mathrm{PL}} \geq I(V;X|Y)B\}, \tag{100}$$

where the union is over all RVs V such that $V - X - Y$ and $I(V;X|Y)B \leq \log|\mathcal{U}| - \epsilon B$.

10.4 Storage for Identification Model with Two Sources

Consider the secure storage for identification process depicted in Fig. 4. The process consists of two phases. In the first phase the system gets the message d that should be stored on the database consisting of k storage cells which can each store a value

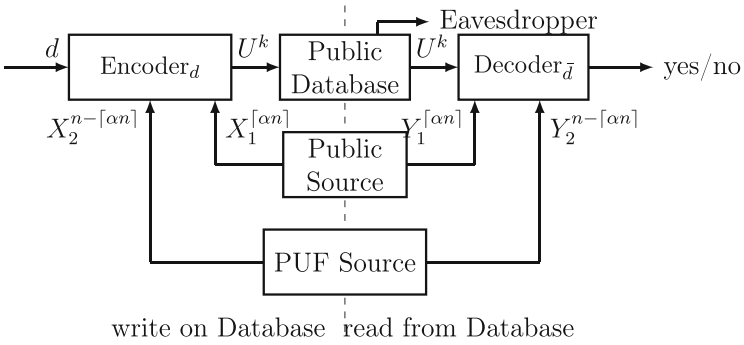


Fig. 5 Secure storage for identification process

from the alphabet $\mathcal{U}$. α is used for timesharing between the PUF source and the public source. The system reads $X_1^{\lceil \alpha n \rceil}$ from the public source and $X_2^{n-\lceil \alpha n \rceil}$ from the PUF source. The system then generates U^k from $(X_1^{\lceil \alpha n \rceil}, X_2^{n-\lceil \alpha n \rceil})$ using an encoder depending on d and stores U^k on the public database. In the second phase the system reads U^k from the database, $Y_1^{\lceil \alpha n \rceil}$ from the public source and $Y_2^{n-\lceil \alpha n \rceil}$ from the PUF source. The system then uses a decoder, depending on the message of interest $\bar{d}$, to decide whether $\bar{d}$ is the message stored on the database, making use of $(Y_1^{\lceil \alpha n \rceil}, Y_2^{n-\lceil \alpha n \rceil})$ and U^k (Fig. 5).

We now define an information theoretic model of the storage process for identification.

Definition 412 Let $k, n \in \mathbb{N}$ and $1 \geq \alpha \geq 0$. The *storage for identification model* consists of the alphabet $\mathcal{U}$, two discrete memoryless multiple sources (DMMSs) $P_{X_1Y_1}$ and $P_{X_2Y_2}$ on the alphabets $\mathcal{X}_1 \times \mathcal{Y}_1$ and $\mathcal{X}_2 \times \mathcal{Y}_2$ respectively, a set of (possibly randomized) encoders $\{\Phi_d\}_{d\in\mathcal{D}}$, $\Phi_d \colon \mathcal{X}_1^{\lceil \alpha n \rceil} \times \mathcal{X}_2^{n-\lceil \alpha n \rceil} \to \mathcal{U}^k$ for all $d \in \mathcal{D}$ and a set of (possibly randomized) decoders $\{\Psi_d\}_{d\in\mathcal{D}}$, $\Psi_d \colon \mathcal{U}^k \times \mathcal{Y}_1^{\lceil \alpha n \rceil} \times \mathcal{Y}_2^{n-\lceil \alpha n \rceil} \to \{0, 1\}$ for all $d \in \mathcal{D}$. Let $X_1^{\lceil \alpha n \rceil}$ and $Y_1^{\lceil \alpha n \rceil}$ be the random variables (RVs) generated from $P_{X_1Y_1}$ and $X_2^{n-\lceil \alpha n \rceil}$ and $Y_2^{n-\lceil \alpha n \rceil}$ be the RVs generated from $P_{X_2Y_2}$. We define $X^n = (X_1^{\lceil \alpha n \rceil}, X_2^{n-\lceil \alpha n \rceil})$ and $Y^n = (Y_1^{\lceil \alpha n \rceil}, Y_2^{n-\lceil \alpha n \rceil})$.We call $(\{\Phi_d\}_{d\in\mathcal{D}}, \{\Psi_d\}_{d\in\mathcal{D}}, \alpha)$ a *storage for identification protocol*.

10.5 Achievability Definition Two Sources

Now the properties that storage for identification protocols should have so that they are considered good storage for identification protocols intuitively are considered. It is reasonable to require a small probability that an error of the first kind occurs when using the decoder d to find out whether d is stored on the database or not. The error of the second kind should also occur with a small probability.

An eavesdropper who reads from the public database and who wants to find out whether d is stored on the database or not is considered here. The eavesdropper also has access to the public source. The sum of the probabiltiy that the eavesdropper makes an error of the first kind and the probability that the eavesdropper makes an error of the second kind should be close to one.

The largest possible identification rate is of most interest, where the number of storage cells is considered as a resource. A fixed ratio B of the number of symbols read from the two sources and the number of storage cells in the database is considered.

The output of the PUF source uniquely characterizes a device, so the authors in [13] possibly want to reuse parts of it. That is why the attacker should not have a lot of information about the PUF source output $X_2^{n-\lceil \alpha n \rceil}$.

This motivates the following definition of achievability for the storage for identification model.

Definition 413 Let $B > 0$. We call the tuple $(R_{\mathrm{ID}}, R_{\mathrm{PL}})$ $R_{\mathrm{ID}}, R_{\mathrm{PL}} \geq 0$ an *achievable rate pair* for the storage for identification model if for every $\delta > 0$ there is a $k_0 = k_0(\delta)$ such that for all $k \geq k_0$ and $n = \lceil B \cdot k \rceil$ there exists a storage for identification protocol such that for all $d, \bar{d} \in \mathcal{D}$, $d \neq \bar{d}$

$$\Pr(\Psi_d(\Phi_d(X^n), Y^n) = 0) \leq \delta \tag{101}$$

$$\Pr(\Psi_d(\Phi_{\bar{d}}(X^n), Y^n) = 1) \leq \delta$$

$$\begin{aligned} &\Pr(\Psi_d^E(\Phi_d(X^n), X_1^{\lceil \alpha n \rceil}, Y_1^{\lceil \alpha n \rceil}) = 0) \\ &+ \Pr(\Psi_d^E(\Phi_{\bar{d}}(X^n), X_1^{\lceil \alpha n \rceil}, Y_1^{\lceil \alpha n \rceil}) = 1) \geq 1 - \delta \end{aligned} \tag{102}$$

$$\frac{1}{k} \log \log |\mathcal{D}| \geq R_{\mathrm{ID}} - \delta$$

$$\frac{1}{k} I(\Phi_d(X^n), X_1^{\lceil \alpha n \rceil}, Y_1^{\lceil \alpha n \rceil}; X_2^{n-\lceil \alpha n \rceil}) \leq R_{\mathrm{PL}} + \delta$$

for all strategies $\{\Psi_d^E\}_{d \in \mathcal{D}}$ of the eavesdropper. We denote the set of all rate pairs that are achievable using such storage for identification protocols by *capacity region* $\mathcal{R}_{\mathrm{ID},2}(B)$.

11 Secure Communication and Identification Systems: Effective Performance Evaluation on Turing Machines

In this section, we present a summary of the results in [32]. In this work, a framework based on Turing machines is developed which provides a theoretical basis for effectively deciding whether or not a communication system meets pre-specified requirements on spectral efficiency and security. A particular key issue

for this is to decide whether or not the performance functions, i.e., capacities, of communication scenarios are Turing computable. Furthermore, a general model of a communication system of interest is introduced and the different communication and identification tasks are discussed. Certain communication systems including the wiretap channel are discussed. In addition, the question of computability is addressed from a more general point of view and a general necessary condition is derived under which the capacity function is not computable.

11.1 Verification Framework

11.1.1 Turing Machine

The task of a Turing machine $\mathfrak{T}$ is to verify the spectral efficiency and security of a given communication scenario $\mathrm{CS} \in \mathcal{CS}$ which is specified by its communication requirements, the underlying communication channel $\mathrm{CH} \in \mathcal{CH}$, and a communication protocol $\mathrm{CP} \in \mathcal{CP}$. This specifies for every $n \in \mathcal{N}$ an encoder-decoder pair $(\mathrm{Enc}_n, \mathrm{Dec}_n)$ for the legitimate users of the channel where n denotes the number of available resources such as block length for coding or number of used frequency bands. The encoder-decoder pair $(\mathrm{Enc}_n, \mathrm{Dec}_n)$ operates at a given rate R. Finally, the parameter k specifies the efficiency of the communication protocol where $1/k$ is the maximum gap of R to the information theoretic optimal performance C. This leads to the following definition.

Definition 414 A Turing machine $\mathfrak{T}$ given by

$$\mathfrak{T} : \mathcal{CS} \times \mathcal{CH} \times \mathcal{CP} \times \mathcal{N} \to \{\text{yes / no}\}$$

is a mapping with

$$\mathfrak{T}(\mathrm{CS}, \mathrm{CH}, \mathrm{CP}, k) = \text{yes}$$

if and only if the communication protocol CP satisfies the performance requirements and

$$C - R < \frac{1}{k} \tag{103}$$

where C denotes the supremum of the information theoretic achievable rates, i.e., the capacity.

11.1.2 Computability

Definition 415 A *rapidly converging Cauchy representation* of a computable real x is a sequence $\{x_i\}_{i=1}^{\infty}$ of rationals that converges to x rapidly, i.e., for every i and $j \geq i$ it holds that $|x_j - x_i| < 2^{-i}$.

Definition 416 A function $f : \mathcal{R}_c \to \mathcal{R}_c$ is called *Borel computable* if there is an algorithm that transforms each given rapidly converging Cauchy representation of a computable real x into a corresponding representation for $f(x)$.

It is to that Turing's definition of computability conforms to the definition of Borel computability. There are weaker forms of computability known as *Markov computability* and *Banach-Mazur computability*, of which the latter one is the weakest form of computability.

Definition 417 A function $f : \mathcal{R}_c \to \mathcal{R}_c$ is called *Markov computable* if there is an algorithm that converts an algorithm for a computable real x into an algorithm for $f(x)$.

Definition 418 A function $f : \mathcal{R}_c \to \mathcal{R}_c$ is called *Banach-Mazur computable* if f maps any given computable sequence $\{x_n\}_{n=1}^{\infty}$ of real numbers into a computable sequence $\{f(x_n)\}_{n=1}^{\infty}$ of real numbers.

In particular, Borel and Markov computability imply Banach-Mazur computability, but not vice versa. Fig. 6 illustrates the relations between these computability notions. Of particular importance for the following analysis is the observation that every Banach-Mazur computable function is necessarily continuous and, as a consequence, a discontinuous function cannot be Banach-Mazur computable. For a detailed treatment and overview of the logical relations between different notions of computability, Boche Et al., refer to [12].

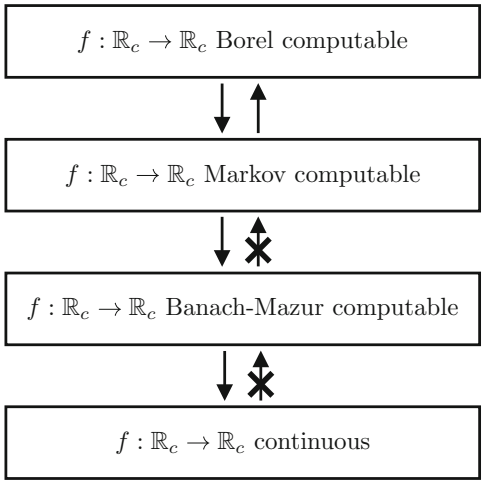


Fig. 6 Logical relation between different notions of computability. Of particular interest is the relation between Banach-Mazur computability and continuity: For a function $f : \mathcal{R}_c \to \mathcal{R}_c$ to be Banach-Mazur computable, it necessarily must be continuous

11.2 Communication Scenarios

Depending on the presence (or absence) of certain users, the general communication model specializes to certain communication scenarios. In particular, the Jammer as an active adversary will account for induced channel uncertainty at the legitimate users (Fig. 7).

11.2.1 Point-to-Point Channel

If only Alice and Bob are present (without Eve and the Jammer), the model reduces to the classical point-to-point channel, i.e., CS $=$ {point-to-point channel} and CH $= \{W\}$ with $W : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ the channel between Alice and Bob.

11.2.2 Wiretap Channel

If there is no Jammer, we have the classical wiretap channel [82], i.e., CS $=$ {wiretap channel}. The underlying channel is given by the pair of channels CH $= \{(W, V)\}$ where $W : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ denotes the channel from Alice to Bob and $V : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$ from Alice to Eve.

11.2.3 Compound Channel

If there is no Eve, but a Jammer, the transmission from Alice to Bob is affected by the channel input from the Jammer. If the Jammer's input $s \in \mathcal{S}$ is fixed throughout the whole duration of transmission, this corresponds to the compound channel [19, 81], i.e., CS $=$ {compound channel}, and the underlying channel is given by a whole family of channels CH $= \{\mathcal{W}\}$ with $\mathcal{W} = \{W_s\}_{s\in\mathcal{S}}$ and $W_s : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ denotes the channel for $s \in \mathcal{S}$.

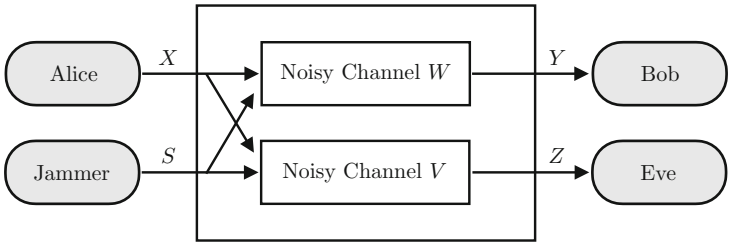


Fig. 7 General model of the communication system including a passive eavesdropper and an active jammer

11.2.4 Compound Wiretap Channel

This model accounts for secure communication under channel uncertainty by combining the wiretap channel and compound channel as CS = {compound wiretap channel} with CH = $\{(\mathcal{W}, \mathcal{V})\}$ where $(\mathcal{W}, \mathcal{V}) = \{(W_s, V_s)\}_{s\in\mathcal{S}}$ and where $W_s : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ and $V_s : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$ are the channels to Bob and Eve for $s \in \mathcal{S}$ respectively [15, 59].

11.2.5 Channel with an Active Jammer

This is a generalization of the compound channel in the sense that the Jammer's input is now allowed to vary in an arbitrary and unknown manner from channel input to channel input which is also known as the *arbitrarily-varying channel (AVC)* [1, 20, 43]. Accordingly, for transmission of block length n, state sequences $s^n \in \mathcal{S}^n$ of length n are taken into account. Therefore, we have CS = {AVC} and CH = $\{\mathfrak{W}\}$ where $\mathfrak{W} = \{W_{s^n}\}_{s^n\in\mathcal{S}^n}$.

11.2.6 Wiretap Channel with an Active Jammer

Finally, this setup combines the wiretap channel with the AVC and is accordingly known as the *arbitrarily-varying wiretap channel (AVWC)* [18, 60, 62, 77]. We have CS = {AVWC} and CH = $\{(\mathfrak{W}, \mathfrak{V})\}$ where $(\mathfrak{W}, \mathfrak{V}) = \{(W_{s^n}, V_{s^n})\}_{s^n\in\mathcal{S}^n}$.

11.3 Computability of Communication Scenarios

Boche Et.al, show that with the restriction to computable channels, the secrecy capacity is Borel computable. As a consequence, it is also computable by Turing machines so that the following equation is satisfied for any k implying that the information theoretic performance requirements and the efficiency of the communication protocol can be effectively verified by the Turing machine.

$$C_S(W, V) - R < \frac{1}{k}, \tag{104}$$

Theorem 419 *The secrecy capacity* $C_S(W, V) = \max_{P_{UX}}[I(U; Y) - I(U; Z)]$ *is Borel computable.*

11.4 General Computability Analysis

11.4.1 Capacity Function on the Set of Channels

Definition 420 The continuity of a function $F(\cdot)$ is defined as follows:

1. The channel $U \in \mathcal{CH}$ is a *continuity point* of $F(\cdot)$ if for all sequences $\{U_n\}_{n=1}^{\infty}$ with

$$\lim_{n\to\infty} d(U_n, U) = 0 \tag{105}$$

   we have

$$\lim_{n\to\infty} F(U_n) = F(U).$$

2. The channel $U \in \mathcal{CH}$ is a *discontinuity point* of $F(\cdot)$ if 1) does not hold, i.e., if there is a sequence $\{U_n\}_{n=1}^{\infty}$ that satisfies (105) but

$$\limsup_{n\to\infty} F(U_n) > \liminf_{n\to\infty} F(U_n) \tag{106}$$

   is satisfied.
3. The function $F(\cdot)$ is a *continuous function* if all DMCs $U \in \mathcal{CH}$ are continuity points according to 1).

Further, let $\mathcal{D}(F)$ be the set of those channels that are discontinuity points of F.

In [32], the following sets of channels are introduced.

$$\mathcal{N}(f) = \{U \in \mathcal{CH} : f(U) = 0\}$$
$$\partial\mathcal{N}(f) = \{U \in \mathcal{N}(f) : \forall\, \epsilon > 0\ \exists\, U_\epsilon \notin \mathcal{N}(f) \text{ such that } d(U, U_\epsilon) < \epsilon\}$$
$$\mathcal{N}_+(C) = \{U \in \mathcal{CH} : C(U) > 0\}.$$

For the following analysis, it is desirable to discuss the sets above and their properties in greater detail.

Lemma 421 *We have*

1. $\mathcal{N}_+(C)$ *is an open set*
2. $\mathcal{N}(f)$ *is a closed set*
3. $\partial\mathcal{N}(f)$ *is a closed set.*

We obtain the following result.

Theorem 422 *It holds that*

$$\mathcal{D}(F) = \mathcal{N}_+(C) \cap \partial\mathcal{N}(f) \tag{107}$$

and, in particular, that $\mathcal{D}(F) \neq \emptyset$ *if and only if* $\mathcal{N}_+(C) \cap \partial\mathcal{N}(f) \neq \emptyset$.

11.4.2 Computable Channels

Theorem 423 *It holds that*

$$\mathcal{D}_c(F) = \mathcal{D}(F) \cap \mathrm{CH}_c.$$

Theorem 424 *If we have for a communication scenario*

$$\mathcal{D}_c(F) = \mathcal{D}(F) \cap \mathrm{CH}_c \neq \emptyset, \tag{108}$$

then F is not Banach-Mazur computable.

11.5 Channel with an Active Jammer

11.5.1 Deterministic Codes

To present and discuss the deterministic code capacity of an AVC, the following definition of symmetrizability is needed.

Definition 425 An AVC $\mathfrak{W}$ is called *symmetrizable* if there exists a stochastic matrix $\sigma : \mathcal{X} \to \mathcal{P}(\mathcal{S})$ such that

$$\sum_{s^n \in \mathcal{S}^n} W(y|x, s)\sigma(s|x') = \sum_{s^n \in \mathcal{S}^n} W(y|x', s)\sigma(s|x) \tag{109}$$

holds for all $x, x' \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Roughly speaking, for such a symmetrizable AVC, the Jammer is able to "simulate" a valid channel input which makes it impossible for the receiver to decide on the correct codeword.

Theorem 426 *The deterministic code capacity $C(\mathfrak{W})$ of the AVC $\mathfrak{W}$ is not Banach-Mazur computable and therewith not Turing computable.*

11.5.2 Random Codes

The deterministic code capacity $C(\mathfrak{W})$ of the AVC $\mathfrak{W}$ is

$$C(\mathfrak{W}) = \begin{cases} \max_{P_X} \min_{q \in \mathcal{P}(\mathcal{S})} I(X; \overline{Y}_q) & \text{if } \mathfrak{W} \text{ is non-symmetrizable} \\ 0 & \text{if } \mathfrak{W} \text{ is symmetrizable} \end{cases} \tag{110}$$

The random code capacity $C_{\mathrm{ran}}(\mathfrak{W})$ of the AVC $\mathfrak{W}$ is

$$C_{\mathrm{ran}}(\mathfrak{W}) = \max_{P_X} \min_{q \in \mathcal{P}(\mathcal{S})} I(X; \overline{Y}_q). \tag{111}$$

Comparing the deterministic code capacity (110) with the random code capacity (111), we observe that $C(\mathfrak{W})$ displays a dichotomy: it either equals $C_{\mathrm{ran}}(\mathfrak{W})$ or else is zero.

Theorem 427 *The random code capacity $C_{ran}(\mathfrak{W})$ of the AVC $\mathfrak{W}$ is Borel computable.*

11.6 Wiretap Channel with an Active Jammer

Theorem 428 *The deterministic code secrecy capacity $C_{\mathrm{S}}(\mathfrak{W}, \mathfrak{V})$ of the AVWC $(\mathfrak{W}, \mathfrak{V})$ is not Banach-Mazur computable and therewith not Turing computable.*

11.7 Computability of Identification Scenarios

11.7.1 Identification over Point-to-Point Channels

The identification capacity $C_{\mathrm{ID}}(W)$ of the DMC W is known [9, 50] and it is not hard to see that the identification capacity is Borel computable.

Theorem 429 *The identification capacity $C_{\mathrm{ID}}(W) = \max_{P_X} I(X; Y)$ of the DMC W is Borel computable.*

11.7.2 Secure Identification over Point-to-Point Channels

The secure identification capacity $C_{\mathrm{SID}}(W, V)$ of the wiretap channel (W, V) has been established in [10]. Boche Et. al, show that it is not Banach-Mazur computable.

Theorem 430 *The secure identification capacity*

$$C_{SID}(W, V) = \begin{cases} C_{\mathrm{ID}}(W) & \text{if } C_{\mathrm{S}}(W, V) > 0 \\ 0 & \text{if } C_{\mathrm{S}}(W, V) = 0 \end{cases}$$

of the wiretap channel (W, V) is not Banach-Mazur computable.

11.7.3 Robust Identification over Compound Channels

The robust identification capacity $C_{\mathrm{ID}}(\mathcal{W})$ of the compound channel $\mathcal{W}$ has been established in [24] and its Borel computability is immediately obtained.

Theorem 431 *The robust identification capacity*

$$C^{ID}(\mathcal{W}) = \max_{P_X} \min_{s \in \mathcal{S}} I(X; Y_s)$$

of the compound channel $\mathcal{W}$ is Borel computable.

11.7.4 Robust and Secure Identification over Compound Channels

The robust and secure identification capacity $C_{\mathrm{SID}}(\mathcal{W}, \mathcal{V})$ of the compound wiretap channel $(\mathcal{W}, \mathcal{V})$ has been established in [24]. In this work, it is shown that $C_{\mathrm{SID}}(\mathcal{W}, \mathcal{V})$ is not Banach-Mazur computable.

Theorem 432 *The robust and secure identification capacity*

$$C_{SID}(\mathcal{W}, \mathcal{V}) = \begin{cases} C_{\mathrm{ID}}(\mathcal{W}) & \text{if } C_{\mathrm{S}}(\mathcal{W}, \mathcal{V}) > 0 \\ 0 & \text{if } C_{\mathrm{S}}(\mathcal{W}, \mathcal{V}) = 0 \end{cases}$$

of the compound wiretap channel $(\mathcal{W}, \mathcal{V})$ is not Banach-Mazur computable.

11.7.5 Robust Identification over Channels with Active Jammer

The robust identification capacity $C_{\mathrm{ID}}(\mathfrak{W})$ of the AVC $\mathfrak{W}$ has been established in [7]. We have the following result.

Theorem 433 *The robust identification capacity*

$$C_{\mathrm{ID}}(\mathfrak{W}) = \begin{cases} C_{ran}(\mathfrak{W}) & \text{if } \mathfrak{W} \text{ is non-symmetrizable} \\ 0 & \text{if } \mathfrak{W} \text{ is symmetrizable} \end{cases}$$

of the AVC $\mathfrak{W}$ is not Banach-Mazur computable.

11.7.6 Robust and Secure Identification over Wiretap Channels with Active Jammer

The robust and secure identification capacity $C_{\mathrm{SID}}(\mathfrak{W}, \mathfrak{V})$ of the AVWC $(\mathfrak{W}, \mathfrak{V})$ has been established in [25]. We have the following result.

Theorem 434 *The robust and secure identification capacity*

$$C_{SID}(\mathfrak{W},\mathfrak{V}) = \begin{cases} C_{ran}(\mathfrak{W}) & \text{if } C_{S,ran}(\mathfrak{W},\mathfrak{V}) > 0 \\ & \text{and } \mathfrak{W} \text{ is non-symmetrizable} \\ 0 & \text{otherwise} \end{cases}$$

of the AVWC $(\mathfrak{W},\mathfrak{V})$ *is not Banach-Mazur computable.*

12 Code Reverse Engineering Problem for Identification Codes

In this section we review the results of [34].

In [34], Bringer and Chabanne formalize the Code Reverse Engineering (CRE) problem for identification codes and they obtain general estimations of the difficulty of this new problem either for independent received messages or for not independent ones. They show that, in fact, an adversary cannot solve this problem easily. The results cited in [34] are based on those of [40]. In [34], the authors consider different cases taking into account the noise over the channel and the capacity of the adversary to isolate or not the communications of a low-cost contactless device (CLD). Furthermore, they apply these results to the BCCK identification protocol, introduced in [8] and based on the use of identification codes. The BCCK protocol relies on a construction of identification codes by Moulin and Koetter [61] using Reed-Solomon codes.

12.1 CRE for Identification Codes

Definition 435 (Identification CRE Problem) Let $\mathcal{X}$, $\mathcal{Y}$ be two alphabets, μ, N be two integers, λ_1 and λ_2 be two values between 0 and 1, and $\mathcal{C}$ be a family of identification codes from $\mathcal{X}$ to $\mathcal{Y}$, all with parameters $(\mu, N, \lambda_1, \lambda_2)$.

- Let $C = \{(Q(\cdot|i), D_i)\}_{i\in\{1,\ldots,N\}}$ be a code chosen randomly in $\mathcal{C}$ and $\bar{i} = (i^1,\ldots,i^M)$ be M random messages chosen independently of C to be encoded over the channel.
- Given the received messages $\bar{y} = (y^1,\ldots,y^M)$, the problem is to guess which C has been used.

Lemma 436 *For independent choices of* $i^1,\ldots,i^M$*, the conditional entropy* $H(C|\bar{y})$ *of the identification code* C *given the received messages* $\bar{y} = (y^1,\ldots,y^M)$ *satisfies*

$$H(C|\bar{y}) \geq \log_2(|\mathcal{C}|) - M(I(i;y) + I(y;C|i) - I(i;y|C)) \tag{112}$$

i.e.,

$$H(C|\bar{y}) \geq \log_2(|\mathcal{C}|) - M(H(y) - H(i) + H(i|C, y) - H(y|C, i)) \tag{113}$$

where i and y are two random variables distributed respectively as the i^j's and the y^j's.

One important difference with the CRE problem for transmission codes is that the solution is not trivial even for a noiseless channel. In fact, as the encoding is non deterministic and there are non-empty intersections between different encoding sets, that makes the reverse engineering of identification codes non trivial in many cases.

Corollary 437 *Let $\mathcal{X} = \mathcal{Y}$ be of size q. Assume that all encoding sets are of the same size in $\mathcal{C}$, i.e., there exists $\tau = \tau(\mathcal{C})$ a parameter such that for every code $F \in \mathcal{C}$, the encoding sets of A defined by the probability distribution Q_C are all of the same size: for all $i \in \{1, \ldots, N\}$, $|\{x|Q_C(x|i) > 0\}| = \tau(\mathcal{C})q^\mu$. Assume also that we are in the context of a noiseless channel. We have*

$$H(C|\bar{y}) \geq \log_2(|\mathcal{C}|) - M(\log_2(1/\tau(\mathcal{C}))) \tag{114}$$

12.2 Application to BCCK Protocol

Now comes the main contribution of [34]. The authors use the CRE problem for identification codes to study the security of the BCCK identification protocol from an information theory perspective. Here is the scheme in details.

1. Setting: here is the description of how the whole identification scheme is designed. Now, the class of identification codes that will be used is considered. Let $\mathbb{F}_q$ be a finite field of size q, $k \leq n \leq q - 1$, we define $\mathcal{C}$ the set of Moulin-Koetter ($\mu = \log_2 n + \log_2 q$, $N = q^k$, $\lambda_1 = 0$, $\lambda_2 = \frac{k-1}{n}$) identification codes from $\{0, 1\}$ to $\{0, 1\}$. Let $\mathbb{F}_q[X]_{k-1} = \{P \in \mathbb{F}_q[X], \deg P < k\}$ be the set of all polynomials over $\mathbb{F}_q$ of degree at most $k - 1$. We sort the set $\mathbb{F}_q[X]_{k-1}$ following an arbitrary choice as $\mathbb{F}_q[X]_{k-1} = \{P_1, \ldots, P_N\}$ and thus index it with integers $i \in \{1, \ldots, N\}$. An identification code $C \in \mathcal{C}$ is defined according to some evaluation domain $F_C = \{\alpha_{C,1}, \ldots, \alpha_{C,n}\} \subset \mathbb{F}_q$ with:

   - encoding sets defined by $A_{F_C,P_i} = \{(j, P_i(\alpha_{C,j}))|j \in \{1, \ldots, n\}\}$ for $i \in \{1, \ldots, N\}$;
   - for all $i \in \{1, \ldots, N\}$, the encoding distribution $Q(\cdot|i)$ is taken as the uniform distribution over A_{F_C,P_i};
   - for all $i \in \{1, \ldots, N\}$, the decoding set D_i is defined as A_{F_C,P_i}.

   This doing, a random code $C \in \mathcal{C}$ is determined by the random choice of n elements in $\mathbb{F}_q$. The size of $\mathcal{C}$ is $\binom{q}{n}n!$.

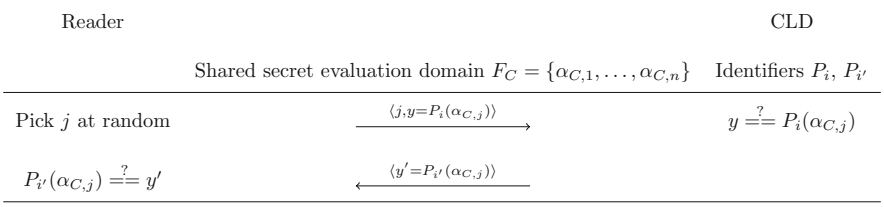


Fig. 8 Enhanced BCCK identification protocol with secret identification code

The protocol still follows [33] but with a random choice of a code C in the family $\mathcal{C}$ that becomes a secret shared among all parties (reader and devices). More precisely it is the evaluation domain F_C that is confidential. One instance of the protocol is illustrated in Fig. 8.

2. Adversary: It is assumed that the adversary knows the family $\mathcal{C}$ of identification codes of the system but he does not know the specific code that is in use. When eavesdropping on the channel the queries made by a reader and the answers produced by the devices, the adversary will see a number of messages $(j, y = P_i(\alpha_{C,j}))$ and $y' = P_{i'}(\alpha_{C,j}))$. To be able to track a particular device, he needs to determine when the same polynomial has been used with two different values $\alpha_{C,j}, \alpha_{C,j'}$. To do so, he must recover information on the evaluation domain, i.e., learn information on the code that is used. To simplify the analysis, the authors in [34] assume that, when trying to reverse engineer the identification code, the adversary eavesdrops the first message only $(\langle j, y = P_i(\alpha_{C,j})\rangle)$ when a reader communicates with a device. This is in fact easy to enforce by taking 2 different random codes, one for the reader query, the second one for the device answer.

Based on these results, security results on the protocol are derived.

- direct application: Assuming that the channel is noiseless, Corollary 437 leads to the following result.

Lemma 438 *Given M independent eavesdropped messages $y^1, \ldots, y^M$ in the enhanced BCCK identification protocol, the uncertainty on the knowledge of the adversary on the secret identification code C that is used satisfies*

$$H(C|y^1, \ldots, y^M) \geq \log_2 \left(\binom{q}{n} n! \right) - M \log_2 q$$

This underlines the difficulty of the CRE problem in this setting when n and q grow to infinity for M polynomial in $\log_2 n$.

- specific lower bound: Via an analysis specific to the Moulin-Koetter construction, the new following result is derived.

Proposition 439 *For the family $\mathcal{C}$ of Moulin-Koetter identification codes defined as above over $\mathbb{F}_q$ with $k \leq n \leq q-1$, for independent choices of M messages $i^1, \ldots, i^M$ to be encoded for a random choice of $C \in \mathcal{C}$, we have exactly*

$$H(C|y^1, \ldots, y^M) = H(C) = \log_2(|\mathcal{C}|) = \log_2 \binom{q}{n} n!$$

where $y^1, \ldots, y^M$ are the received messages, independently and randomly chosen in the encoding sets of $i^1, \ldots, i^M$, and eavesdropped by the adversary (without noise).

13 Discrete Identification

In order to achieve the double exponential growth in the identification results, the transmitter needs a local randomized source, and depending on it, it encodes the messages into the channel. There are applications where it is difficult or impossible to implement such a code, because the encoder has to process a bit sequence of exponential length. Therefore, for such applications, which include molecular communication, it is important to also consider deterministic identification (DI) codes (these are also previously referred to as non-randomized identification (NRI) codes). For DI codes, local randomization is not available to the transmitter. It was shown in [54] that the DI capacity of a binary symmetric channel is 1 bit per channel usage. The work inspired Ahlswede and Dueck to their research in [8] described above. In [5], it was stated that the DI capacity of a discrete memoryless channel (DMC) with a stochastic matrix W is given by the logarithm of the number of unique row vectors of W. For the proof, a reference was made to [2], which contains no identification and deals with a completely different model of an arbitrarily varying channel. The first rigorous proof of this statement was given in [67]. This result shows that in the deterministic setting, the number of messages scales exponentially with block length, as in the traditional transmission setting. Nevertheless, the achievable identification rates are significantly higher than those of transmission. Moreover, deterministic codes often have the advantage of simpler implementation and analysis. Moreover, in [67] and [68], the DMC and the Gaussian channel were analyzed with input constraints. Such a constraint is often associated with limited power supply or control. It is worth noting that the DI with power constraint for the Gaussian channel is infinite. More precisely, the capacitance is infinite in the exponential domain and zero in the double-exponential domain. In the follow-up work [69], the authors then considered whether there is another scaling such that the DI capacity of the Gaussian channel is finite. They found that for Gaussian channels, the number of messages scales as n^{nR}, and develop lower and upper bounds on the DI capacity at this scale. They also consider deterministic identification for Gaussian channels with fast fading and slow fading where channel side information (CSI) is available at the decoder. For slow fading, the DI capacity is infinite in the

exponential scale unless the fading gain can be zero or arbitrarily close to zero (with positive probability), in which case the DI capacity is zero. Compared to the double exponential scale in RI coding, the scale here is much smaller. Another surprising result was obtained in [57], where the authors considered DI over Gaussian channels with noise-free feedback. They showed that if the noise variance is positive, any rate can be achieved for identification over the Gaussian channel with noise-free feedback. The result means that for any chosen scaling, the corresponding DI capacity is infinite. An extended version of this text and applications of identification can be find in [39].

14 Private Interrogation of Devices via Identification Codes

Contact-less devices are generally assumed to respond automatically to any verifier scan. In [65], it was suggested that the verifier directly addresses the device with which it wants to communicate. To this aim, the verifier broadcasts the device identifier and then the corresponding device responds accordingly. However, the emission of the device identifier enables an eavesdropper to track it. In [33], the authors follow this idea and look for a solution which does not require many computations and many communications efforts, while preventing an eavesdropper to be able to track a particular device. Changing the paradigm from the situation where a device initiates the protocol to a situation where the device identifies first the interrogation request enables to envisage new solutions. The authors in [33] developed a scheme, which does not rely neither on hash functions nor on a random generator on the device side and show that this solution is very efficient in terms of channel usage.

14.1 Identification Codes

The general definition is skipped. Only the Moulin-Koetter definition is considered here.

Definition 440 Let $\mathbb{F}_q$ be a finite field of size $q, k \leq n \leq q-1$ and an evaluation domain $F = \{\alpha_1, \ldots, \alpha_n\} \in \mathbb{F}_q$. Set $A_p = \{(j, P(\alpha_j)) | j \in \{1, \ldots, n\}\}$ for P any polynomial on $\mathbb{F}_q$ of degree at most $k-1$. The Moulin-Koetter RS-Identification Codes are defined by the family of encoding and decoding sets $\{(A_P, A_P)\}_{P \in \mathbb{F}_q[X], \deg P < k}$. This leads to a $(\log_2 n + \log_2 q, q^k, 0, \frac{k-1}{n})$ identification code from $\{0,1\}$ to $\{0,1\}$. Using a Reed-Solomon code of dimension k, this gives $\lambda_2 = \frac{k-1}{n}$ since $d = n - k + 1$ (Reed-Solomon codes are Maximum Distance Separable).

A set of $M < q^k$ devices is constructed, and each of them is associated with a different random polynomial $p_l \in \mathbb{F}_q[X]$ of degree less than $k-1$. The memory of

these devices is then filled with a set of $p_l(\alpha_j)$, for $\alpha_j \in F$, with F a public subset of $\mathbb{F}_q$, i.e., the devices contain the evaluation of p_l over a subset of $\mathbb{F}_q$. The verifier is given the polynomial p_l. When the verifier wants to initiate communication with the device number l associated with the identifier p_l, it selects a random $\alpha_j \in F$ and sends $(j, p_l(\alpha_j))$ over the wireless channel. A device that receives this message checks whether the value stored in its memory at the corresponding address is equal to $p_l(\alpha_j)$, i.e. computes an equality test of two bit strings. If the test is successful, it replies and goes through the authentication protocol that will be later described. Otherwise, it remains silent.

14.2 Protocol for Interrogation

The aim in [33] is for a ContactLess Device (CLD) to recognize itself into a verifier request, but authentication of the CLD toward the verifier is handled as well. System set-up:

- Setup Authority (1^l) generates a set of parameters generates a set of parameters KA_p defining two integers μ, N, two alphabets $\mathcal{X}, \mathcal{Y}$ and two error rates λ_1, λ_2. No private parameter is defined.
- Setup Verifier KA_p constructs an $(\mu, N, \lambda_1, \lambda_2)$ identification code from $\mathcal{X}$ to $\mathcal{Y}$ following the general definition of Ahlswede and Dueck, $\mathcal{IC} = \{(Q(\cdot|i), \mathcal{D}_i)\}_{i\in\{1,\ldots,N\}}$ and sets $KV_p = \mathcal{IC}$. $\mathcal{IC}$ is based on the Moulin-Koetter construction [61].
- Setup CLD KV_p(SN) first returns randomly chosen $(i, j) \in \{1, \ldots, N\}, i \neq j$ as the parameters of the CLD identified by SN. It then initializes the CLD with the storage of a description of the decoding set D_i of the identified i and the description of $Q(\cdot|i)$, the encoding probability mass function for index j. It also stores (i, j, SN) in the verifier database.

A verifier and a set of devices are set-up as above and the following steps are then processed to interrogate and authenticate a specific CLD.

- The verifier, who wants to interrogate the CLD of identifier SN, recovers its identifier i in the database and encodes it via $Q(\cdot|i)$ into a message $x \in \mathcal{X}^n$. The verifier broadcasts the message (ACK, x), where ACK is an acknowledgment number which will help the verifier to sort the received answers when it emits simultaneously several such messages.
- Any listening CLD that receives the message (ACK, y) uses its own decoding set $D_{i_{\text{CLD}}}$ to determine whether y encodes i_{CLD}.
- f a CLD identifies y as an encoding of its identifier i_{CLD}, then it sends the message (ACK, x') to the verifier, where ACK is the incoming acknowledgment number and x' is an encoding of j_{CLD} obtained via $Q(\cdot|j_{\text{CLD}})$.

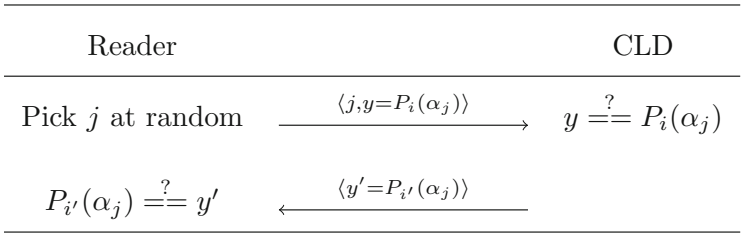


Fig. 9 CLD identification via Moulin-Koetter identification codes

- Upon receiving this message, the verifier then checks whether the received message y' is a member of the decoding set D_j of the aimed CLD. If so, then the CLD is declared as authenticated.

Now, the Moulin-Koetter setting is considered. In this setting, a set of CLDs is constructed where each of them—say CLD_l is associated with two different random polynomial identifiers $p_l, p'_l \in \mathbb{F}_q[X]$ o f degree at most $k-1$. Here p_l and p'_l are good descriptions of the associated encoding functions and the decoding sets; they are both stored on the CLD side and on the verifier database. When the verifier wants to initiate communication with CLD_l (with identifiers p_l, p'_l), it selects a random $\alpha_j \in F \subset \mathbb{F}_q[x]$ and broadcasts $(ACK, j, p_l(\alpha_j))$ over the wireless channel. A CLD with identifiers p, p' that receives this message checks whether the polynomial p stored in its memory evaluated in α_j is equal to $p_l(\alpha_j)$. If the test is successful, it responds with the value $(ACK, p'(\alpha_j))$. Otherwise, it remains silent. The verifier authenticates the CLD if the received value $p'(\alpha_j)$ is equal to $p'_l(\alpha_j)$. The description is depicted in Fig. 9.

14.3 Security Analysis

Here, we state the most important results cited in [33].

14.3.1 Effect of Passive Eavesdropping

Proposition 441 *Assume that the number M of devices simultaneously queried by the verifier is such that $\sqrt{q} \geq M \geq e\sqrt{n/k}$ (with $e = \exp(1)$). Then a passive adversary, who eavesdrops at most T requests with $T < M^2 k$, cannot reconstruct the polynomial identifiers, except with a negligible probability.*

Proposition 442 *Assume $\sqrt{q} \geq M \geq e\sqrt{n/k}$ and $T < M^2 k$. A passive adversary cannot determine whether two requests correspond to the same CLD except if there is a collision, that happens only with probability $1/\sqrt{n}$.*

14.3.2 Security Against Impersonation

In the protocol described in Sect. 14.2, a CLD replies to the verifier only if it believes that the verifier is legitimate. It is thus close to mutual authentication—although here the authentication of the verifier is only probabilistic with respect to the false-positive error rate of an identification code. It is a weaker result than general verifier authentication: a verifier cannot be impersonated in order to interrogate a pre-fixed CLD.

Proposition 443 *Assume* $\sqrt{q} \geq M \geq e\sqrt{n/k}$ *and* $T < M^2k$. *In the scheme described in [33], given a non-corrupted CLD, an adversary cannot impersonate a verifier to interrogate this specific CLD, without replaying an eavesdropped transcript, except with probability* $1/q$.

Proposition 444 *Assume* $\sqrt{q} \geq M \geq e\sqrt{n/k}$ *and* $T < M^2k$. *The scheme described in [33] is secure against impersonation of a CLD, i.e. an adversary will fail with probability* $1 - \frac{1}{q}$.

14.3.3 Privacy

Proposition 445 *If* $\sqrt{q} \geq M \geq e\sqrt{n/k}$ *and* $T < M^2k$, *then the scheme described in [33] is weak private.*

For more details about weak privacy, we refer the reader to [33].

14.3.4 Advantages for Very Low-Cost Devices

For low-cost devices, instead of storing the two polynomial identifiers p, p', we store directly the values $p(\alpha_1), \ldots, p(\alpha_n)$ and $p'(\alpha_1), \ldots, p'(\alpha_n)$ within the device. So doing, no computation is needed on the device side. Depending on the amount of memory available per device, we can also limit the number of such values by restricting ourselves to a basis of evaluation of size $L < n$, e.g., $(\alpha_1, \ldots, \alpha_L)$.

15 Applications of Identification

In this section, we present possible application examples cited in [24]. In [24], it is pointed out that the theory in particular also offers applications for the model with a transmitter and a receiver. Frequently, it is assumed that identification addresses several recipients. However, this is not absolutely necessary.

1. **Industry 4.0**:
   In production engineering, sensors are used to monitor the correct sequence of the production. The sensor data is encoded into states and transmitted to a central

unit (CU). The CU reads out and processes the states for plausibility checks. If an error occurs, the receiver is interested in whether there is a critical state in the system. The exact sensor measurement data are of minor significance for him. Therefore, an identification code can be used in this case.

In example 5, a completely different application is presented which, however, operates on the same idea.

2. **Online sales**:
   In the case of online shopping, customers have certain interests. These interests can be derived from the buying and surfing behavior of the customers in the corresponding online shop. By using an identification code, the platform operator (receiver in the identification framework) can identify whether certain items are of interest to the customer or not. This information can be used in order to optimize advertising campaigns and shop structure.

   In this example, the store owner is the receiver in the identification scheme. He wonders if a certain article in his shop is interesting for the customer.
3. **Hardware store**:
   In a hardware store, it can be interesting for the customers whether there are any offers for them at their current location. This customer requested can be handled with the help of an identification code. In a next step, potential offers can be sent to the customer using a transmission code. In contrast to 2), the customer embodies the receiver in the identification scheme in this application example.

   In this example, the customer is the receiver. Otherwise the principle is similar to example 3. The peculiarity is a two-part procedure. Information can be sent to the customers in a targeted way.
4. **Vehicle-to-X communication**:
   In next-generation driver assistance systems, the sensor data collected by the vehicle is enriched by additional information gained from other traffic participants by communication allowing, for plausibility checks concerning the desired driving maneuvers. In this case, a vehicle may ask whether a certain message, concerning the future movement of an adjacent vehicle, was transmitted or not. Since the neighboring vehicle cannot anticipate the desired movement, when no previous information exchange has happened, it is not aware in which of its transmitted messages the first vehicle is interested in. In the previously mentioned scenario, relying on low-latency identification of certain contradictory driving maneuvers, an identification code can lead to significant performance improvements compared to a classical transmission code.
5. **Healthcare**:
   In medical applications, a patient may be equipped with multiple, wirelessly connected sensors in order to monitor his health status. Besides simply reporting the status of physiological functions, there may occur critical events which can be anticipated from an unfavorable combination of different individual body signals. Thus, if a central information merging unit (CU) receives a particular measurement from one of the sensors, let's say sensor 1, the CU may calculate the potential measurements of the other sensors resulting in an overall critical

state of the patient. Thus, the CU may ask, whether a certain message (encoded measurement) was transmitted by, e.g., sensor 2. This setting is directly-related to the introduced identification framework.

16 Omnisophie

In his book "Omnisophie", Dueck compared the identification process to the "human flash-mode"; the body's response to stress. Dueck believes that our body response is very similar to an identification scheme, where we, in case of a stress alarm, decide whether we turn on our reaction system. Yes or No! Zero or one! Dueck explained:

"After a car crash, we are in state of shock. It is more terrible than we could bear. Before exams sit some of us apathetic. I have lost my house keys and sit in the cold, 40cm away from" the happiness "inside the house. Endorphin makes us focused only on our destiny and we just switch off! It is here about whether we turn on our system! yes or no! zero or one!!" Now, let us know about the story behind identification. How did Dueck come up with the identification scheme with Ahlswede?

" One day, my doctoral supervisor Rudolf Ahlswede came to my office. At that time, I was a professor in the faculty of mathematics at the university of Bielefeld. He showed me the work of Joseph Ja'Ja', university of Maryland, which was called "Identification is easier than decoding". That was the first time we realized the problem of identification. Ahlswede said: "this seems very important somehow! It is worth to take a look!" The work was easy to understand. It is about a new approach: Identification.

It is said that messages are transmitted over noisy channels. The noisier the channel is, the more difficult the transmission is. The Shannon theory is about how mathematicians and communications engineers insert the minimum amount of redundancy so that the message can be decoded with a negligible probability of error. Written languages are quiet redundant, that is why parents can usually correct their children's dictations without knowing what was dictated. The classical transmission consists of the following steps.

- We have a message to be transmitted.
- This message is converted to a Morse code and redundancy is eventually added. This process is called channel coding.
- This Morse text is transmitted over the channel.
- The message is then affected by noise.
- The channel output is received.
- The channel output is then decoded.

The theory of Shannon is concerned with the maximum rate at which information can be transmitted over a communications channel. While the mathematicians

elaborate smarter transmission methods, the engineers are "unfortunately" inventing better transmitters with less power consumption as well as with much stronger signals.

However, the approach of Ja'Ja' was concerned with identification. In the following, we will give some examples. When I traveled to Hungary in 1980, I had to submit my documents and then wait for a visa stamping. The lucked ones were called by loudspeakers to get back their entry permit papers. They are then allowed to go. In 1980, the loudspeakers were so noisy that we can barely hear our names. There were a lot of people and it was really crowded. The voice of the announcer was unclear and we could barely notice it was Hungarian. I was wondering how my name Dueck with "ue" would be pronounced. Would the announcer repeat it if I did not understand?!

Waiting and hearing indistinct names was a so difficult exercise. Actually, it did not matter to understand the name that was called but to distinguish your name! So, I did not have to decode the called names but just to decide whether Dueck was called or not! Yes, Dueck or No, it is not! Hopefully you would have flinched and thought of Ja'Ja' work. Aha! we have to be careful, it is about Yes or No. Let us take another example: When the German teacher is circling around and looking for someone who writes the hourly report then we, the students tremble and wonder who will be the luckiest one? I always sat there and waited. Would she say "Dueck!"? Yes or No?

Thus, it is not about "what happens" but whether a certain event occurs! In the theory of identification the decoder is not really interested in what the received message is, but he only wants to decide whether a message, which is of special interest to him, had been sent or not. This afternoon, I read the work of Joseph Ja'Ja' saying "Yeah, sure. This should be an interesting question and a natural good solution. What does Ahlswede mean?" Rudolf Ahlswede is famous for his insight into big things. He is a real phenomenon. He has an extraordinary sense of anticipating! Indeed, the theory of identification is becoming an increasingly important area in communications, which enlarged the basis of information theory. Rudolf said secretly: "It is double exponential". I thought about that. Next morning, we both said: "Yes, it's double exponential." I, a naive over-optimist, said: "I know what's coming out." Then, I tapped: 2 to the power of 2^{nC}! Very aesthetic, exactly matching the Shannon law!"

References

1. R. Ahlswede, Elimination of correlation in random codes for arbitrarily varying channels. Z. Wahrsch. Verw. Gebiete **44**(2), 159–175 (1978)
2. R. Ahlswede, A method of coding and its application to arbitrarily varying channels. J. Comb. Inf. Sys. Scien **5**(1), (1980)
3. R. Ahlswede, V.B. Balakirsky, Identification under random processes. Probl. Inf. Transm. **32**(1), 123–138 (1996)
4. R. Ahlswede, V. Blinovsky, Classical capacity of classical-quantum arbitrarily varying channels. IEEE Trans. Inf. Theory **53**(2), 526–533 (2007)

5. R. Ahlswede, N. Cai, Identification without randomization. IEEE Trans. Inf. Theory **45**(7), 2636–2642 (1999)
6. R. Ahlswede, I. Csiszar, Common randomness in information theory and cryptography. I. Secret sharing. IEEE Trans. Inf. Theory **39**(4), 1121–1132 (1993)
7. R. Ahlswede, I. Csiszar, Common randomness in information theory and cryptography. II. CR capacity. IEEE Trans. Inf. Theory **44**(1), 225–240 (1998)
8. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inf. Theory **35**, 15–29 (1989)
9. R. Ahlswede, G. Dueck, Identification in the presence of feedback—a discovery of new capacity formulas. IEEE Trans. Inf. Theory **35**, 30–36 (1989)
10. R. Ahlswede, Z. Zhang, New directions in the theory of identification via channels. IEEE Trans. Inf. Theory **41**, 1040–1050 (1995)
11. N. Alon, The Shannon capacity of a union. Combination **18**, 301–310 (1998)
12. J. Avigad, V. Brattka, Computability and analysis: the legacy of Alan Turing, in *Turing's Legacy: Developments from Turing's Ideas in Logic*, ed. by R. Downey (Cambridge University Press, Cambridge, 2014)
13. S. Baur, C. Deppe, H. Boche, Secure storage for identification, in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Kalamata, 2018, pp. 1–5
14. C.H. Bennett, P.W. Shor, J.A. Smolin, A.V. Thapliyal, Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem. IEEE Trans. Inf. Theory **48**(10), 2637–2655 (2002)
15. I. Bjelaković, H. Boche, J. Sommerfeld, Secrecy results for compound wiretap channels. Probl. Inf. Transm. **49**(1), 73–98 (2013)
16. I. Bjelaković, H. Boche, Classical capacities of compound and averaged quantum channels. IEEE Trans. Inf. Theory **55**(7), 3360–3374 (2009)
17. I. Bjelacovic, H. Boche, J. Sommerfeld, Secrecy results for compound wiretap channels. Probl. Inf. Transm. **49**, 73–98 (2013)
18. I. Bjelacovic, H. Boche, J. Sommerfeld, Capacity results for compound wiretap channels. Proc. ISIT 60–64 (2011)
19. D. Blackwell, L. Breiman, A.J. Thomasian, The capacity of a class of channels. Ann. Math. Stat **30**(4), 1229–1241 (1959)
20. D. Blackwell, L. Breiman, A.J. Thomasian, The capacities of certain channel classes under random coding. Ann. Math. Stat. **31**(3), 558–567 (1960)
21. H. Boche, M. Cai, C. Deppe, J. Nötzel, Classical-quantum arbitrarily varying wiretap channel: common randomness assisted code and continuity. Quantum Inf. Process **16**(1), 1–48 (2016)
22. H. Boche, M. Cai, C. Deppe, J. Nötzel, Secret message transmission over quantum channels under adversarial quantum noise: secrecy capacity and super-activation. J. Math. Phys. **60**, 062202 (2019)
23. H. Boche, C. Deppe, Secure identification under jamming attacks, in 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Kalamata, 2018, pp. 1–5
24. H. Boche, C. Deppe, Secure identification for wiretap channels: robustness, super-additivity and continuity. IEEE Trans. Inf. Forensics Secur. **13**(7), 1641–1655 (2018)
25. H. Boche, C. Deppe, Secure identification under passive eavesdroppers and active jamming attacks. IEEE Trans. Inf. Forensics Secur. **14**(2), 472–485 (2019)
26. H. Boche, C. Deppe, A. Winter, Secure and robust identification via classical-quantum channels. IEEE Trans. Inf. Theory **65**(10), 6734–6749 (2019)
27. H. Boche, J. Nötzel, Arbitrarily small amounts of correlation for arbitrarily varying quantum channel. J. Math. Phys. **54**(112), (2013). https://doi.org/10.1063/1.4825159
28. H. Boche, R.F. Schäfer, H.V. Poor, On the continuity of the secrecy capacity of compound and arbitrarily varying wiretap channels. IEEE Trans. Inf. Forensics Secur. **10**(12), 2531–2546 (2015)

29. H. Boche, R.F. Schaefer, H.V. Poor, Characterization of super-additivity and discontinuity behavior on the capacity of arbitrarily varying channels under list decoding, in *Proceedings of the IEEE International Symposium on Information Theory (ISIT '17)*, Aachen, Germany, June 2017, pp. 2820–2824
30. H. Boche, R. F. Schaefer, H. V. Poor, Identification capacity of correlation-assisted discrete memoryless channels: analytical properties and representations, in *Proceedings of the IEEE International Symposium on Information Theory (ISIT '19)*, Paris, France, July 2019, pp. 470–474
31. H. Boche, R.F. Schaefer, H.V. Poor, Identification capacity of channels with feedback: discontinuity behavior, super-activation, and Turing computability. IEEE Trans. Inf. Theory **66**(10), 6184–6199 (2020)
32. H. Boche, R.F. Schaefer, H. V. Poor, Secure communication and identification systems—effective performance evaluation on turing machines. IEEE Trans. Inf. Forensics Secur. **15**, 1013–1025 (2020)
33. J. Bringer, H. Chabanne, G. Cohen, B. Kindarji, Private interrogation of devices via identification codes, in *Progress in Cryptology - INDOCRYPT 2009. INDOCRYPT 2009*, ed. by B. Roy, N. Sendrier. Lecture Notes in Computer Science, vol. 5922 (Springer, Berlin, 2009)
34. J. Bringer, H. Chabanne, Code reverse engineering problem for identification codes. IEEE Trans. Inf. Theory **58**(4), 2406–2412 (2012)
35. M.V. Burnashev, On method of types, approximation of output measures and ID-capacity for channels with continuous alphabets, in *Proceedings of the 1999 IEEE Information Theory and Communications Workshop (Cat. No.99EX253)*, June 1999, pp. 80–81
36. M.V. Burnashev, On method of types, approximation of output measures and ID-capacity for channels with finite alphabets. Probl. Inf. Transm. **36**(3), 195–212 (2000)
37. H. Buhrman, R. Cleve, J. Watrous, R. de Wolf, Quantum fingerprinting. Phys. Rev. Lett. **87**, 167902 (2001)
38. N. Cai, A. Winter, R.W. Yeung, Quantum privacy and quantum wiretap channels. Probl. Inf. Transm. **40**(4), 318–336 (2004)
39. J. Cabrera, H. Boche, C. Deppe, R.F. Schaefer, C. Scheunert, F.H.P. Fitzek, 6G and the post-Shannon-theory, in *Shaping Future 6g Networks: Needs, Impacts and Technologies*, ed. by E. Bertin, N. Crespi, T. Magedanz. (Wiley-Blackwell, 2021)
40. M. Cluzeau, J. Tillich, On the code reverse engineering problem, in *2008 IEEE International Symposium on Information Theory*, Toronto, 2008, pp. 634–638
41. I. Csiszár, J. Körner, Broadcast channels with confidential messages. IEEE Trans. Inf. Theory **24**(3), 339–348 (1978)
42. I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems* (Academic Press, New York, 1981)
43. I. Csiszár, P. Narayan, The capacity of the arbitrarily varying channel revisited: positivity, constraints. IEEE Trans. Inf. Theory **34**(2), 181–193 (1988)
44. S. Derebeyoğlu, C. Deppe, R. Ferrara, Performance analysis of identification codes. Entropy **22**, 1067 (2020)
45. I. Devetak, The private classical capacity and quantum capacity of a quantum channel. IEEE Trans. Inf. Theory **51**(1), 44–55 (2005)
46. F. Dupuis, The Decoupling Approach to Quantum Information Theory, Ph.D. thesis, Université de Montréal, 2009
47. R. Ericson, Exponential error bounds for random codes in the arbitrarily varying channel. IEEE Trans. Inf. Theory **31**, 4248 (1985)
48. G. Fettweis, H. Boche, T. Wiegand, E. Zielinski, H. Schotten, P. Merz, S. Hirche, A. Festag, W. Häffner, M. Meyer, E. Steinbach, R. Kraemer, R. Steinmetz, F. Hofmann, P. Eisert, R. Scholl, F. Ellinger, E. Weiß, I. Riedel, The Tactile Internet, ITU-T Technology Watch Report, 2014
49. C.A. Fuchs, J. van de Graaf, Cryptographic distinguishability measures for quantum-mechanical states. IEEE Trans. Inf. Theory **45**(4), 1216–1227 (1999)
50. T.S. Han, S. Verdú, New results in the theory of identification via channels. IEEE Trans. Inf. Theory **38**, 14–25 (1992)

51. T.S. Han, S. Verdú, Approximation theory of output statistics. IEEE Trans. Inf. Theory **39**, 752–772 (1993)
52. P. Hayden, A. Winter, Weak decoupling duality and quantum identification. IEEE Trans. Inf. Theory **58**(7), 4914–4929 (2012)
53. T. Ignatenko, F.M.J. Willems, Biometric security from an information-theoretical perspective, in *Biometric Security from an Information—Theoretical Perspective* (2012)
54. J. JaJa, Identification is easier than decoding, in *Annual Symposium on Foundations of Computer Science (SFCS)* (1985), pp. 43–50
55. K. Kurosawa, T. Yoshida, Strongly universal hashing and identification codes via channels. Trans. Inf. Theory **45**, 2091–2095 (1999)
56. W. Labidi, C. Deppe, H. Boche, Secure identification for gaussian channels, in *ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 2872–2876
57. W. Labidi, H. Boche, C. Deppe, M. Wiese, Identification over the Gaussian channel in the presence of feedback (2021). arXiv:2102.01198
58. L. Lai, S. Ho, H.V. Poor, Privacy–security trade-offs in biometric security systems—part I: single use case. IEEE Trans. Inf. Forensics Secur. **6**(1), 122–139 (2011)
59. Y. Liang, G. Kramer, H.V. Poor, S. Shamai, Compound wiretap channels. EURASIP J. Wirel. Commun. Netw. **2009**, 142374 (2009)
60. E. MolavianJazi, M. Bloch, J.N. Laneman, Arbitrary jamming can preclude secure communication, in *Proceedings of the 47th Annual Allerton Conf. Commun., Control, Computing*, Monticello, 2009, pp. 1069–1075
61. P. Moulin, R. Koetter, A framework for the design of good watermark identification codes, in *Proc. SPIE, Secur., Steganogr., Watermarking Multimedia Contents VIII*, vol. 6072, 60721H-1—60721H-10, 2006
62. J. Nötzel, M. Wiese, H. Boche, The arbitrarily varying wiretap channel—Secret randomness, stability, and super-activation. IEEE Trans. Inf. Theory **62**(6), 3504–3531 (2016)
63. Y. Oohama, Converse coding theorem for the identification via multiple access channels, in *Proceedings of the IEEE Information Theory Workshop*, Bangalore, India, 2002, pp. 155–158
64. Y. Oohama, Converse coding theorems for identification via channels. IEEE Trans. Inf. Theory **59**(2), 744–759 (2013)
65. PEARS, Privacy Ensuring Affordable RFID System. European Project
66. J.M. Renes, Approximate quantum error correction via complementary observables (2010). arXiv[quant-ph]:1003.1150
67. M.J. Salariseddigh, U. Pereg, H. Boche, C. Deppe, Deterministic identification over channels with power constraints, arXiv preprint arXiv:2010.04239
68. M.J. Salariseddigh, U. Pereg, H. Boche, C. Deppe, Deterministic identification over channels with power constraints, in *IEEE International Conference on Communications* (IEEE, 2021)
69. M.J. Salariseddigh, U. Pereg, H. Boche, C. Deppe, Deterministic identification over fading channels, in IEEE Information Theory Workshop, Riva del Garda, Italy, April 2020, extended version available at https://arxiv.org/abs/2010.10010
70. C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948)
71. C.E. Shannon, The zero error capacity of a noisy channel. IRE Trans. Inf. Theory **2**, 8–19 (1956)
72. R.I. Soare, *Recursively Enumerable Sets and Degrees* (Springer, Berlin, 1987)
73. H. Tyagi, A. Vardy, Universal hashing for information-theoretic security. Proc. IEEE **103**(10), 1781–1795 (2015)
74. A.M. Turing, On computable numbers with an application to the entscheidungsproblem. Proc. Lond. Math. Soc. **2**(42), 230–265 (1936)
75. A.M. Turing, On computable numbers with an application to the entscheidungsproblem: correction. Proc. Lond. Math. Soc. **2**(43), 544–546 (1937)
76. S. Verdu, V.K. Wei, Explicit construction of optimal constant-weight codes for identification via channels. IEEE Trans. Inf. Theory **39**(1), 30–36 (1993)

77. M. Wiese, J. Nötzel, H. Boche, A channel under simultaneous jamming and eavesdropping attack—Correlated random coding capacities under strong secrecy criteria. IEEE Trans. Inf. Theory **62**(7), 3844–3862 (2016)
78. M. Wiese, H. Boche, Semantic security via seeded modular coding schemes and Ramanujan graphs. IEEE Trans. Inf. Theory **67**(1), 52–80 (2021)
79. A. Winter, Quantum and classical message identification via quantum channels, in *Festschrift "A. S. Holevo 60"*, ed. by O. Hirota (Rinton Press, 2004), pp. 171–188. Reprinted: Quantum Inf. Comput. 4(6&7), 563–578 (2004)
80. A. Winter, Identification via quantum channels, in *Information Theory, Combinatorics, and Search Theory*. Lecture Notes in Computer Science, vol. 7777 (Springer, Berlin, 2013), pp. 217–233
81. J. Wolfowitz, Simultaneous channels. Arch. Ration. Mech. Anal. **4**(4), 371–386 (1960).
82. A.D. Wyner, The wire-tap channel. Bell Syst. Tech. J. **54**(8), 1355–1387 (1975)

Correction to: Identification and Other Probabilistic Models

Correction to:
R. Ahlswede, *Identification and Other Probabilistic Models*, Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8

Due to an unfortunate oversight incorrect information was captured online regarding the authorship of chapters "Testing of Hypotheses and Identification" and "New Results in Identification Theory".

This has been updated online.

The updated online version of this chapters can be found at
https://doi.org/10.1007/978-3-030-65072-8_23
https://doi.org/10.1007/978-3-030-65072-8_27

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8_28

Supplement

1 Abschied–Ein Gedicht von Alexander Ahlswede

Es füllten mich die Träume
wie immer mit altem Schwank,
im Schatten hoher Bäume
nahm ich den Zaubertrank.

Und nun trat plötzlich jenes
Los in mein stöberndes Licht:
zuviel erbittert Geschehenes,
und Hoffnung gibt es nicht.

Doch tragen die Schatten noch Bilder
vom geliebten Vater dahin,
und meine Seele wird milder,
gab er meiner Liebe doch Sinn.

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8

Rudolf Ahlswede and his son Alexander Ahlswede

2 Gunter Dueck: Memories of Rudolf Ahlswede

Around 1974, when I met Rudi Ahlswede for the first time, he was a guest professor at the "Institute of Statistics" of the University of Göttingen, which was led by Ulrich Krengel. The entrance room of this institute was a kind of a lounge—open for all scientists and students. We could browse there through the most recent journals—and more important—play blitz chess and drink a lot of coffee. Being a student of statistics, I joined the institute nearly daily. Ulrich Krengel was like a father of a big math-stat family. In this entrance lounge, the students could listen "live" to real scientific work: Drinking coffee and loud passionated discussions. Legendary: the heated debates of Lee K. Jones, Gyula O.H. Katona, and Ahlswede about topics of combinatorics. Now, writing this more than 40 years later, I can still hear their voices in my ears…

Ahlswede began lecturing information theory for the students—promising newest research results. I asked Ulrich Krengel for the job to write a script of the upcoming lectures. So I began working with RA as a student. In these times, Paul Erdös was famous to offer smaller—*Prizes for solutions of hard math-ematical problems. "I solved an Erdös-Problem of 200"*—this could be said with all honors. In one of his first lectures, Ahlswede mentioned the prize tactics of Paul Erdös, explained an unsolved mathematical problem in information theory to us and offered 100 Germans Marks and a doctor title for a solution. Afterwards, writing

the script in my flat, I found a solution! I ran back to the institute and proudly presented my idea. However, RA found an error, what a pity! We would need another deep argument to fix my idea. One hour later, RA came with a solution: Gyula Katona was just finishing a paper on an important combinatorial problem, and just this theorem filled easily the gap in "my proof". This lucky punch paid off: I got 50 German Marks, wrote my first joint scientific paper, got no doctor title, and RA and I started a close research relationship which lasts until 1987, when I left for a position at the IBM Scientific Center in Heidelberg.

In 1976, Ahlswede was appointed full professor at the new University of Bielefeld—as the first professor of applied mathematics in the math faculty. I joined him as an assistant professor in the brandnew university building. I remember a short talk in a crowded space in the university hall: "Are you really willing to research in a new field of research? Do you really know that applied math is in the beginning? Are you really aware that there might be no professorship in information theory at all when you will be qualified to earn one? Think over it!"—I did not think, I said: "Yes, yes, yes." A decade later, Ahlswede would be right: I did not find any position which I could apply for—and moved to IBM to work on statistics and optimization.

RA put his heart and soul in a life for math and research. He expected the same for everyone. We got just a 1-year contract as an assistant. "Your dissertation should be ready after 1 year. If it is nearly ready at that time—okay, I could consider a second year."—"All the other people here get longer contracts…—"Hey, you said, you want to become a professor. Right?"—"Yes."—"Then, finally, you have to come up with famous results. Right?"—"Yes."—"For becoming a full professor, you have to present at least ten quality papers, some of them famous ones. If you need more than 2 years for the first paper, how long do you think it will take to achieve your goal? I advise you: Solve your problems fast and begin with the famous papers. This is the easiest path." On another occasion: "Think about serious problems only. Try those whose solution could win the IEEE Information Theory Society Prize Paper Award. Longstanding conjectures might be easy to prove, because people have given up and new technology and recent research might have laid a new ground."—Or: "Please do not waste too much time on lecturing. It's a nice time with students but try to stick to research only."

RA was the most passionate scientist I ever met. He was solving problems around the clock. In the theater with his wife, he has been reportedly watched writing ideas on a sheet of paper. When my wife and friends visited the famous Mont Saint Michel in France (please google for a picture to understand), we were ready to walk through the mud flat to reach the monastery. In the distance, we saw a thinking man walking from the monastery to the landside. "I recognize this body language—it looks like RA!" After some minutes: Yes, it was RA, with some notes.

Oh, I used to work in a very different time frame. This caused frequent (mild) conflicts. All the time, RA looked doubtful when I left my office exactly at 4.30 p.m.—every single day. Explanation: My wife worked as a librarian at the university, she had fixed working hours: 8 a.m. to 4.30 p.m. We owned only one car to go to work, and RA could not imagine researchers thinking at daytime only. "Prussian! You're a Prussian! Prussian steadiness!", he grumbled very often. For

me, it worked—my doctor thesis was finished after 15 months. No applause, please: I was not the fastest finisher in our floor, the 1-year contract strategy just turned out to be very successful.

In 1976/1977, RA gave short lecture on some forty most important problems of information theory. "Please solve some of them." We tried our very best. Somewhat later, Edward van der Meulen from Leuven (Belgium) published an IEEE survey paper on these most important problems, and we felt really troubled to learn there that we were in fact working on the hardest problems of all...

Excellence and Excellence! Only excellence counted on our floor. Professors who served in organizations or faculty committees or who liked to be a dean or to organize large conferences, were called "clerks" by RA. "Clerks, pure clerks, only clerks!", he shouted angrily if someone demanded administrative things.

RA was very proud that no one ever solved a problem after he had worked some time on it. It was a question of honor. Especially, he had thought a long time about the proof of the strong converse for the multiple access channel in order to have a complete solution set for his pioneering work on this channel. Within 10 years, he had many ideas—all of them failed. In 1980, he failed with a further bright idea. He wrote a complete paper and noticed the error while typing the last line. He was really angry for many days and brought the problem to my attention. At the next Saturday morning, under the shower, I had a simple idea to prove everything. On Monday morning, I presented my idea: ten lines on a table. He shook his head in aroused disbelief. I returned to my office next door to him and waited for his final comments. Some minutes later, I heard him slamming his office door, he left. I waited nervously for his return. And waited and waited for some days. At Friday morning RA returned to my office room: "Please start your Habilitation ("professor exam") process with your proof!", still upset. This made my day, of course, and, on the other side, I felt very sorry or guilty...

After my Habilitation, I got a 5-year contract as a professor of Math in Bielefeld. Our daughter Anne was born, Johannes followed in 1986. In this year, RA came with a preprint of a rather "easy to see solution" of a new problem of identification. "However", he thought over it, "it is a new and exciting question." We thought about it over the night. Next day: "It could be double exponential..." It was a strange feeling. It seemed that identification (extracting a yes-no answer from a vast amount of information) would be exponentially faster than to read/decode this information. After two or three days, we knew the solution and we were certain how to prove everything. However, it took months of hard work to complete the proof. Johannes, our baby, tried not to sleep at all. I used the long hours of night service for the proof... My contract as a professor expired in this time, I felt very sad. In 5 years, there was not a single professor position to fill with something in the near of information theory. Now, I could remember: "Are you really sure to research in a new field? Are you really aware...?"

1987, I switched to the IBM Scientific Center in Heidelberg. There, I invented some new optimization algorithms which we applied to industry problems. I was appointed a manager in 1990 and formed a new optimization business group with a few Million revenue. Some days before Christmas 1990, I got a letter from the IEEE:

"Your joint paper on identification won the IEEE Information Theory Society Prize Paper Award." I had tears in my eyes... I was named IEEE fellow in the sequel which was rewarded by IBM with a "Senior Technical Staff" title and a director's salary. At that time, a new university was looking for a professor in information theory. They called me, and I declined to apply—I had found a new destination within IBM. A decade later, I learned that RA had pushed hard the request of such a professorship to "bring me back to research", and that he was very disappointed that I preferred to stay with IBM. (He never called/contacted me in this matter!)

Anyway, I have to be deeply thankful. My IBM career was boosted strongly by the IEEE Fellowship. I benefited from my RA experiences to "originate a culture of excellence". I managed successfully people by setting high bars for the IBM researchers and to celebrate all their successes. "Gunter, you have been extremely demanding, but after some years we recognized that we really enjoyed this time." I just tried to bring RA-spirit to my department. This RA-Spirit might have been the basement of all my achievements. He was my real mentor. A million thanks.

Gunder Dueck

From left to right: Marat Burnashev, Gunter Dueck and Rudolf Ahlswede

Author Index

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8

Subject Index

R. Ahlswede, *Identification and Other Probabilistic Models*,
Foundations in Signal Processing, Communications and Networking 16,
https://doi.org/10.1007/978-3-030-65072-8