CODING THEOREM AND MEMORY CONDITIONS FOR ABSTRACT CHANNELS WITH TIME STRUCTURE

von

MARTIN MITTELBACH

MARTIN MITTELBACH

CODING THEOREM AND MEMORY CONDITIONS FOR ABSTRACT CHANNELS WITH TIME STRUCTURE

TECHNISCHE UNIVERSITÄT DRESDEN

Coding Theorem and Memory Conditions for Abstract Channels with Time Structure

Martin Mittelbach

von der Fakultät Elektrotechnik und Informationstechnik
der Technischen Universität Dresden

zur Erlangung des akademischen Grades

D O K T O R I N G E N I E U R
(Dr.-Ing.)

genehmigte Dissertation

Vorsitzende: Prof. Dr.-Ing. habil. Renate Merker
Gutachter: Prof. Dr.-Ing. Eduard A. Jorswieck, Prof. Dr.-Ing. Dr. rer. nat. Holger Boche
Tag der Einreichung: 07. 07. 2014, Tag der Verteidigung: 04. 12. 2014

Je abstrakter, desto anwendbarer.

Danksagung

An dieser Stelle möchte ich mich aufrichtig für all die Unterstützung bedanken, die ganz wesentlich zum Gelingen der vorliegenden Dissertation beigetragen hat.

Zuerst möchte ich Prof. Eduard Jorswieck für die geduldige Betreuung meiner Doktorarbeit und die Tätigkeit als Erstgutachter danken. Vor allem danke ich ihm für die ausgezeichneten Arbeitsbedingungen am Lehrstuhl Theoretische Nachrichtentechnik, das mir stets entgegengebrachte große Vertrauen, die beständige Unterstützung und den gewährten Freiraum, durch den ich die Chance hatte, mich entsprechend meiner Interessen fachlich weiterzuentwickeln. Gleichermaßen bin ich dankbar für die Möglichkeit einer eigenverantwortlichen Arbeit, bei der ich den Dingen auf den Grund gehen konnte. Besonders dankbar bin ich auch dafür, dass ich das bereits als Mitarbeiter von Prof. Adolf Finger begonnene Mathematikstudium mit der Arbeit vereinbaren und parallel zu dieser erfolgreich abschließen konnte. Die erworbene mathematische Qualifikation hatte insbesondere auf die vorliegende Dissertation einen maßgeblichen Einfluss. Mein Dank gilt ebenso Prof. Finger, der diesen Weg als ehemaliger Lehrstuhlleiter mit ermöglicht und gefördert hat.

Prof. Holger Boche danke ich für die Tätigkeit als Zweitgutachter und für die wertvollen Anregungen zur Fortführung der in dieser Doktorarbeit erzielten Ergebnisse.

All meinen Kollegen danke ich für die sehr gute Arbeitsatmosphäre und Zusammenarbeit. Ausdrücklicher Dank gilt meiner Kollegin Anne Wolf. Da ich seit Langem mit Anne das Büro teile, hatte ich nicht nur eine geduldige Zuhörerin für die stets dringlichen Berichte über meine neuesten mathematischen Fort- und Rückschritte, sondern durfte in ganz besonderer Weise von ihrer großen Hilfsbereitschaft und Kollegialität profitieren. Herzlich danken möchte ich auch Sybille Siegel und Hrjehor Mark, durch deren Engagement es trotz zunehmender Herausforderungen immer wieder gelungen ist, eine rechtzeitige Verlängerung meines Arbeitsvertrages zu erreichen. Darüber hinaus konnte ich mich in all der Zeit auf die Unterstützung meiner langjährigen Kollegen Christian Scheunert, Axel Schmidt und Holger Hösel verlassen. Axel war immer zur Stelle, wenn es brannte. Über Christians kritisches Hinterfragen nachzudenken war stets lohnenswert und hat häufig zu fundierterem Verständnis geführt.

Bedanken möchte ich mich für die Durchsicht von jeweils Teilen des Manuskriptes bei meinen Kollegen Carsten Janda, Pin-Hsun Lin, Bho Matthiesen, Johannes Richter, Christian Scheunert, Axel Schmidt und Anne Wolf sowie bei meinem Bruder Johannes.

Schließlich danke ich von Herzen meiner ganzen Familie und meinen Freunden für die fortwährende Unterstützung und den verlässlichen Rückhalt. Vor allem danke ich meinen Eltern, die mich und meine Geschwister stets nach Kräften gefördert haben. Ganz besonderer Dank gilt nicht zuletzt meiner langjährigen Freundin Anja für ihr Verständnis und die aufgebrachte Geduld. Ich verdanke ihr viel von der Zeit, die ich in Ruhe am Schreibtisch verbringen konnte. Ganz sicher eine entscheidende Voraussetzung für den erfolgreichen Abschluss dieser Arbeit.

Martin Mittelbach
Dresden, April 2015

Kurzfassung

Im ersten Teil der Arbeit wird ein Kodierungstheorem und ein dazugehöriges Umkehrtheorem von Kadota und Wyner (1972) für abstrakte Kanäle mit Zeitstruktur verallgemeinert. Als wesentlichster Beitrag wird das Kodierungstheorem für eine signifikant schwächere Bedingung an das Kanalausgangsgedächtnis bewiesen, die sogenannte totale Ergodizität für block-i.i.d. Eingaben. Dieses Ergebnis wird hauptsächlich durch eine alternative Charakterisierung der Informationsratenkapazität erreicht. Es wird gezeigt, dass die von Kadota und Wyner verwendete ψ-Mischungsbedingung (asymptotische Gedächtnislosigkeit am Kanalausgang) recht einschränkend ist, insbesondere für die wichtige Klasse der Gaußkanäle. In der Tat, für Gaußkanäle wird bewiesen, dass die ψ-Mischungsbedingung äquivalent zu endlichem Gedächtnis am Kanalausgang ist. Darüber hinaus wird eine schwache Umkehrung für alle stationären Kanäle mit Zeitstruktur bewiesen. Sowohl Intersymbolinterferenz als auch Eingabebeschränkungen werden in allgemeiner und flexibler Form berücksichtigt. Aufgrund der direkten Verwendung von äußeren Maßen und der Herleitung einer angepassten Version von Feinsteins Lemma ist es möglich, auf die Standarderweiterung der σ-Algebra am Kanaleingang zu verzichten, wodurch die Darstellungen transparenter und einfacher werden. Angestrebt wird eine operationelle Perspektive. Die Verwendung eines abstrakten Modells erlaubt dabei die einheitliche Betrachtung von zeitdiskreten und zeitstetigen Kanälen.

Für abstrakte Kanäle mit Zeitstruktur werden im zweiten Teil der Arbeit Bedingungen für ein unendliches Gedächtnis am Kanalausgang systematisch analysiert. Unter Ausnutzung der Zusammenhänge zu dem umfassenden Gebiet der stark mischenden zufälligen Prozesse wird eine Hierarchie in Form einer Folge von Implikationen zwischen den verschiedenen Gedächtnisvarianten hergeleitet. Die im Beweis des Kodierungstheorems verwendete ergodentheoretische Gedächtniseigenschaft und die ψ-Mischungsbedingung von Kadota und Wyner (1972) sind dabei Bestandteil der hergeleiteten Systematik. Weiterhin werden Bedingungen für den Kanal spezifiziert, unter denen Eigenschaften von zufälligen Prozessen am Kanaleingang bei einer Transformation durch den Kanal erhalten bleiben.

Im letzten Teil der Arbeit werden sowohl Integrationskanäle als auch Hintereinanderschaltungen von Kanälen in Bezug auf Mischungsbedingungen sowie weitere für das Kodierungstheorem relevante Kanaleigenschaften analysiert. Die erzielten Ergebnisse sind nützlich bei der Untersuchung vieler physikalisch relevanter Kanalmodelle und erlauben eine komponentenbasierte Betrachtung zusammengesetzter Kanäle. Es wird eine Reihe von Beispielen untersucht, einschließlich deterministischer Kanäle, zufälliger Filter und daraus zusammengesetzter Modelle. Abschließend werden Anwendungen aus weiteren Gebieten, beispielsweise der statistischen Signalverarbeitung, diskutiert. Insbesondere die Fourier-Transformation stationärer zufälliger Prozesse wird im Zusammenhang mit starken Mischungsbedingungen betrachtet.

Abstract

In the first part of this thesis, we generalize a coding theorem and a converse of Kadota and Wyner (1972) to abstract channels with time structure. As a main contribution we prove the coding theorem for a significantly weaker condition on the channel output memory, called total ergodicity for block-i.i.d. inputs. We achieve this result mainly by introducing an alternative characterization of information rate capacity. We show that the ψ-mixing condition (asymptotic output-memorylessness), used by Kadota and Wyner, is quite restrictive, in particular for the important class of Gaussian channels. In fact, we prove that for Gaussian channels the ψ-mixing condition is equivalent to finite output memory. Moreover, we derive a weak converse for all stationary channels with time structure. Intersymbol interference as well as input constraints are taken into account in a flexible way. Due to the direct use of outer measures and a derivation of an adequate version of Feinstein's lemma we are able to avoid the standard extension of the channel input σ-algebra and obtain a more transparent derivation. We aim at a presentation from an operational perspective and consider an abstract framework, which enables us to treat discrete- and continuous-time channels in a unified way.

In the second part, we systematically analyze infinite output memory conditions for abstract channels with time structure. We exploit the connections to the rich field of strongly mixing random processes to derive a hierarchy for the nonequivalent infinite channel output memory conditions in terms of a sequence of implications. The ergodic-theoretic memory condition used in the proof of the coding theorem and the ψ-mixing condition employed by Kadota and Wyner (1972) are shown to be part of this taxonomy. In addition, we specify conditions for the channel under which memory properties of a random process are invariant when the process is passed through the channel.

In the last part, we investigate cascade and integration channels with regard to mixing conditions as well as properties required in the context of the coding theorem. The results are useful to study many physically relevant channel models and allow a component-based analysis of the overall channel. We consider a number of examples including composed models and deterministic as well as random filter channels. Finally, an application of strong mixing conditions from statistical signal processing involving the Fourier transform of stationary random sequences is discussed and a list of further applications is given.

Contents

Notation

Frequently Used Symbols

$\mathbb{N}$	positive integers
$\mathbb{N}_0$	nonnegative integers
$\mathbb{Z}$	integers
$\mathbb{R}$	real numbers
$\bar{\mathbb{R}}$	$= \mathbb{R} \cup \{-\infty, \infty\}$
$\mathbb{C}$	complex numbers
T	time indices, either $\mathbb{R}$ or $\mathbb{Z}$
$\overline{T}$	$= T \cup \{-\infty, \infty\}$
T_+, T_0	positive and nonnegative time indices
$\varnothing$	empty set
2^A	power set of set A
$\lvert A \rvert$	cardinality of set A
A^{c}	complement of set A
$\mathbb{1}_A$	indicator function of set A
$A \triangle B$	symmetric difference of sets A and B
$A \times B$	Cartesian product of sets A and B
A_x	x-section of set $A \in X \times Y$
X_u^v, X_u^+, X_-^v	sub-product space when X is a product space generated by the family $\{X_t, t \in T\}$, $= X_u^\infty$, $= X_{-\infty}^v$
$[A]$	inverse image of set A w. r. t. projection on sub-product space
θ_w	w-shift of signals with infinite duration
$\langle \cdot \rangle_w$	w-shift of objects with finite duration
$\jmath$	imaginary unit
$\mathrm{Re}(z)$	real part of $z \in \mathbb{C}$
$\mathrm{Im}(z)$	imaginary part of $z \in \mathbb{C}$
$\overline{z}$	complex conjugate of $z \in \mathbb{C}$
$\det(A)$	determinant of matrix A
$\mathrm{tr}(A)$	trace of matrix A
a'	transpose of vector a
$\mathrm{diag}(a_1, \ldots, a_n)$	diagonal matrix with entries a_i on the main diagonal
$\log$	logarithm w. r. t. base e
$(\Omega, \mathcal{F})$	measurable space

$\sigma(\mathcal{G})^\dagger$	σ-algebra generated by family of sets $\mathcal{G}$
$\mathcal{A} \vee \mathcal{B}$	$= \sigma(\mathcal{A} \cup \mathcal{B})$
$\mathcal{B}(X)$	Borel-σ-algebra on topological space X
$\mathcal{X} \otimes \mathcal{Y}$	product of σ-algebras $\mathcal{X}$ and $\mathcal{Y}$
$\mathcal{X}_u^v$, $\mathcal{X}_u^+$, $\mathcal{X}_-^v$	sub-product σ-algebra when $\mathcal{X}$ is a product σ-algebra generated by the family $\{\mathcal{X}_t, t \in T\}$, $= \mathcal{X}_u^\infty$, $= \mathcal{X}_{-\infty}^v$
$[\mathcal{A}]$	inverse image of σ-algebra $\mathcal{A}$ w. r. t. projection on sub-product space
$(\Omega, \mathcal{F}, \mu)$	(probability) measure space
$\mu \otimes \nu$	product of measures μ and ν
$\mu \ll \nu$	absolute continuity of measure μ w. r. t. to measure ν
μ_ξ	distribution of random variable ξ defined on $(\Omega, \mathcal{F}, \mu)$
δ_x	Dirac measure concentrated at x
$\mathrm{E}(\xi)$	expectation of random variable ξ
$\mathrm{var}(\xi)$	variance of random variable ξ
$\mathrm{cov}(\xi, \eta)$, $\mathrm{cor}(\xi, \eta)$	covariance, correlation of random variables ξ and η
$\mathrm{P}(F \,\vert\, \mathcal{A})$	conditional probability of set F given σ-algebra $\mathcal{A}$
$(\mathcal{A} - \mathcal{B} - \mathcal{C})^\dagger$	Markov chain in this order of σ-algebras $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$
ξ_u^v, ξ_u^+, ξ_-^v	segment of random process $\xi = \{\xi_t, t \in T\}$, $= \xi_u^\infty$, $= \xi_{-\infty}^v$
κ, $\kappa(x, B)$	channel, probability that received signal lies in B given x was transmitted
$\mathcal{C}(b, E_b)$	block code with block length b satisfying input constraint E_b
$\varrho(u_i, V)$	decoding error probability of codeword u_i w. r. t. input signal set V
$\varrho_{\max}(V)$	(maximal) decoding error probability w. r. t. input signal set V
(b, E_b, V, M, ϵ)-code	block code $\mathcal{C}(b, E_b)$ with $\lvert\mathcal{C}(b, E_b)\rvert \geq M$ and $\varrho_{\max}(V) \leq \epsilon$
$H(\mathcal{A})^\dagger$	entropy of σ-algebra $\mathcal{A}$
$H(\mathcal{A} \,\vert\, \mathcal{B})^\dagger$	conditional entropy of σ-algebra $\mathcal{A}$ given σ-algebra $\mathcal{B}$
$I(\mathcal{A}; \mathcal{B})^\dagger$	mutual information between σ-algebras $\mathcal{A}$ and $\mathcal{B}$
$I(\mathcal{A}; \mathcal{B} \,\vert\, \mathcal{C})^\dagger$	conditional mutual information between σ-algebras $\mathcal{A}$ and $\mathcal{B}$ given σ-algebra $\mathcal{C}$
$\bar{I}(\mathfrak{A}; \mathfrak{B})^\ddagger$	mutual information rate between families of σ-algebras $\mathfrak{A}$ and $\mathfrak{B}$
$D_f(\mathrm{P} \Vert \mathrm{Q})^\dagger$	f-divergence of measures P and Q
$D(\mathrm{P} \Vert \mathrm{Q})^\dagger$	relative entropy of measures P and Q
$\lVert \mathrm{P} - \mathrm{Q} \rVert_{\mathrm{tv}}$	total variation distance of measures P and Q
$\psi(\mathrm{P} \Vert \mathrm{Q})^\dagger$	ψ-variation of measures P and Q
$\alpha(\mathcal{A}; \mathcal{B})^\dagger$	α-dependence coefficient of σ-algebras $\mathcal{A}$ and $\mathcal{B}$
$\beta(\mathcal{A}; \mathcal{B})^\dagger$	β-dependence coefficient of σ-algebras $\mathcal{A}$ and $\mathcal{B}$
$\psi(\mathcal{A}; \mathcal{B})^\dagger$	ψ-dependence coefficient of σ-algebras $\mathcal{A}$ and $\mathcal{B}$

†The arguments can be random variables as well.
‡The arguments can be random processes as well.

Abbreviations

AR	autoregressive
MA	moving average
ARMA	autoregressive moving average
IIR	infinite impulse response
a. s.	almost sure(ly)
i. i. d.	independent and identically distributed
w. r. t.	with respect to

Introduction

Motivation and background. Claude Elwood Shannon established with his seminal paper *A mathematical theory of communication* (Shannon, 1948) a unifying theory of data compression and data transmission. The importance of his work was immediately recognized by engineers and mathematicians, which is apparently the reason why the republication of the original paper as monograph (Shannon and Weaver, 1949) has the modified title *The mathematical theory of communication*. With a single contribution Shannon actually created the field of modern information theory. An important idea of his approach was to describe the communication between a sender and a receiver using a stochastic model. He introduced the concepts of entropy, mutual information, source, channel, coding, and capacity to derive fundamental limits of reliable information storage and transmission. The main results of the theory are formulated in terms of source and channel coding theorems, which relate theoretical bounds of operation to optimization problems involving quantitative measures of information.

Shannon's classical problem of channel coding describes a block-based coding-decoding procedure to reliably transmit messages from a sender to a receiver in the presence of random transmission errors. The noisy part of the communication process is modeled by an information channel connecting sender and receiver in a probabilistic manner. A major information-theoretic concern of coded information transmission is the analysis of the relation between the rate of transmission and the probability of a decoding error. With a channel coding theorem this relation is characterized in terms of a quantity, called channel capacity, in the following sense: For any transmission rate below channel capacity there exists a coding-decoding procedure such that the transmitted messages, even though randomly corrupted by noise, can be inferred from the received messages with arbitrarily low error probability. The second half of such a theorem, commonly referred to as converse, states that for any transmission rate above channel capacity this is not possible.

Since Shannon started information theory a large number of publications has been devoted to the mathematically rigorous derivation of channel coding theorems. The goal was to develop the theory for models of increasing complexity and generality capturing more and more practically relevant situations and aspects. On the one hand, the complexity is determined by the space of symbols allowed for communication and the time structure of the model, i. e., finite vs. infinite alphabets and discrete vs. continuous time. One the other hand, complexity is increased by incorporating technical constraints or effects, such as intersymbol interference or memory properties of the noise process.

A main motivation for this thesis was to establish an abstract framework, that allows us to formulate a general coding theorem for a point-to-point communication link in a mathematically rigorous way under practically useful assumptions. A central objective was to include continuous-time continuous-valued transmission models because in the literature much less attention is payed to this case compared to discrete models. Moreover, the goal was a reduction to the essential channel properties required to prove the coding statements with a main focus on infinite memory conditions.

In contrast to characterizing finite memory there is a great variety of nonequivalent alternatives to model infinite memory. Infinite memory at the channel output is, roughly speaking, a form of asymptotic independence. It can be characterized either in an ergodic-theoretic sense or by strong mixing conditions, which are based on dependence measures. Because conditions of this type have various applications in different fields of engineering and mathematics, such as statistical signal processing, measure concentration, or central limit theorems, they are of interest in their own right. Therefore, we aimed at analyzing these conditions for abstract channels with time structure beyond the scope of coding theorems.

Contribution and outline. The main contribution of the first part of the thesis is a generalization of a coding theorem and a converse of Kadota and Wyner (1972) with regard to channel model, input constraints, required channel properties, and definition of information rate capacity. Kadota and Wyner considered a continuous-time channel with real-valued input and output signals, whereas we consider channels with time structure in general, i. e., discrete- as well as continuous-time channels with completely arbitrary alphabets. Regarding stationarity, causality, and channel input memory the assumed conditions are identical. However, with respect to output memory we achieve a significant generalization, which is the main contribution concerning the coding theorem. Kadota and Wyner used a property they called asymptotic output-memorylessness, which is introduced later as ψ-mixing condition. We show that this condition is quite restrictive, in particular for the important class of Gaussian channels. Actually, we prove that for Gaussian channels the ψ-mixing condition is equivalent to finite output memory. As a result, Kadota and Wyner's formulation of the coding theorem is for example not applicable to the simple stationary additive Gaussian noise channel with proper rational noise spectral density. By introducing an alternative characterization of information rate capacity we are able to prove the coding theorem under the significantly weaker condition of total ergodicity for block-i.i.d. inputs. It is a classical result that a property of this type is sufficient in the special case of a discrete-time finite-alphabet channel. The modified definition even allows us to handle the proof in the case of finite and infinite information rate capacity in exactly the same way. The relations between the different versions of information rate capacity are studied in detail.

We prove a weak converse for all stationary channels with time structure. No further restrictions on the channel properties or alphabets are required. Due to a generic characterization we have a convenient flexibility in taking intersymbol interference into account. Furthermore, input constraints are incorporated in an abstract form. In contrast to (Kadota and Wyner, 1972) we are able to accomplish this without using a standard extension of the channel input σ-algebra. This is possible as we utilize outer measures directly and derive an adequate version of Feinstein's lemma, which we believe is more transparent. A common method proposed by Holsinger (1964) and Gallager (1968, Ch. 8) is to represent a continuous-time channel by an infinite series of parallel discrete-time channels. Using a consequent measure-theoretic description and following the approach of Kadota and Wyner (1972) we are able to avoid this transformation completely. Therefore, we can treat discrete- and continuous-time channels with abstract alphabets in a unified way, which we believe is an important argument in favor of the path taken in this thesis. We achieved results in a general framework by a suitable and consistent combination of several existing approaches. With regard to the generality of the information-theoretic models and tools our formulation is mainly influenced by the work of the Russian school of information theory, in particular by Kolmogorov (1956a), Dobrushin (1963), and Pinsker (1964). The statement of the coding theorem, however, follows the style of Ahlswede (2006) and Wolfowitz (1978), since this emphasizes the operational meaning more clearly when coding theorems for transmission

are considered. For historical notes on related results and a detailed discussion see Section § 11, where also the relation to (Mittelbach, 2012) is discussed.

In the second part of the thesis we systematically analyze infinite output memory conditions for abstract channels with time structure, which are extensions of memory conditions for random processes, known as strong mixing conditions (Bradley, 2007). These mixing conditions lie between the finite output memory condition and the ergodic-theoretic mixing conditions of Adler (1961). We formulate the memory conditions for channels in the same way as for random processes. This allows us to exploit the connections to the rich field of strong mixing conditions efficiently. There are two main contributions. On the one hand, we derive a hierarchy for the nonequivalent infinite channel output memory conditions in terms of a sequence of implications, which corresponds to that known for random processes. The ergodic-theoretic memory condition used in the proof of the coding theorem and the strong mixing condition employed by Kadota and Wyner (1972) are shown to be part of this taxonomy. On the other hand, we study the interplay between memory conditions of channels and random processes. We derive sufficient conditions under which a channel transforms an input probability measure with a certain memory property into an input-output probability measure sharing the same property.

In the last part of this thesis we consider aspects that are useful to analyze concrete channel models. First, we derive results for cascade channels that allow to conclude properties of a complex channel from properties of basic building blocks. Then we study integration channels, for which the channel model can be decomposed into a deterministic function and a random noise source. This is possible for many physically relevant models. We prove results that allow to verify properties of the overall integration channel by verifying properties of the channel function and the noise source separately. The analysis of cascade and integration channels includes mixing conditions as well as properties required in the context of the coding theorem from the first part of the thesis. Finally, we specify a number of examples and discuss applications of mixing conditions. Starting with basic channel models, such as abstract deterministic channels, additive noise channels, or state-dependent channels we continue with deterministic as well as random filter channels and composed models. In several illustrative examples we use previously analyzed random processes with special mixing properties as building blocks. We apply the tools obtained throughout the thesis to analyze relevant channel properties. With respect to memory conditions we also discuss specific aspects of deterministic and composed channels.

In Chapter I we introduce a general stochastic transmission and communication model as well as information-theoretic measures and tools, which allow to formulate and prove in Chapter II a coding theorem and a weak converse for abstract channels with time structure. Chapter I provides further fundamental material on divergence and dependences measures, based on which we define and investigate in Chapter III memory conditions for random processes and channels with time structure. Cascade and integration channels are studied in Chapter IV and a variety of examples and applications is analyzed in Chapter V. The Appendix contains mathematical background material and a number of outsourced proofs.

Chapter I

Fundamentals

In this chapter we provide the material required to formulate and prove a coding theorem and a weak converse for abstract channels with time structure. We introduce a general stochastic transmission and communication model as well as information-theoretic measures and tools which allow the analysis of theoretical limits of coded information transmission. Additionally, we collect material on divergence and dependence measures based on which we define and investigate memory conditions for random processes and channels with time structure. We begin the chapter by establishing some general notation used throughout the thesis. Basic notions of probability and measure theory are used freely. As a compact reference in this regard please refer to (Kallenberg, 2002, Chs. 1–4, 6), where the notation is close to that introduced below. Further recommendable references are (Bauer, 1995, 2001; Billingsley, 1995). More advanced and frequently used background material from probability and measure theory is collected in Appendices A to C in a form suitable for the purposes of the thesis.

§1 General Notation

(1.1) Sets, σ-algebras, measurable spaces and functions. As usual $\mathbb{Z}$, $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{R}$, and $\mathbb{C}$ denote the set of integers, positive integers, nonnegative integers, real numbers, and complex numbers, respectively. By $\bar{\mathbb{R}}$ we denote the extended real line defined as $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$.

Throughout the thesis we denote by T the set of time indices. Whenever the index set T is used it can be replaced either by $\mathbb{Z}$ to model discrete-time or by $\mathbb{R}$ to model continuous-time. Occasionally, it is convenient to extend T and consider $\overline{T} = T \cup \{-\infty, \infty\}$. The sets of positive and nonnegative time indices are denoted by T_+ and T_0, respectively. In case of $T = \mathbb{Z}$ we use the interval notation to denote the set of consecutive integers contained in the interval, e. g., if $v \in T_+$, then $(0, v]$ is the short hand version of $\{1, 2, \ldots, v\}$.

We write $\varnothing$ for the empty set. For a set A we denote by A^{c} its complement and by 2^A its power set, i. e., the set of all subsets of A. The number of elements in A is denoted by $|A|$. The symmetric difference of the sets A and B is defined by $A \triangle B = (A \cap B^{\mathrm{c}}) \cup (B \cap A^{\mathrm{c}})$. We write $\mathbb{1}_A$ for the indicator function of the set A, which is one for all elements of A and zero otherwise.

A pair $(\Omega, \mathcal{F})$ consisting of a space Ω and a σ-algebra $\mathcal{F}$ is called measurable space. A partition of $(\Omega, \mathcal{F})$ is any countable family $\{A_1, A_2, \ldots\}$ of disjoint sets $A_i \in \mathcal{F}$ whose union is equal to Ω. If $\mathcal{G}$ is a family of subsets of Ω, then $\sigma(\mathcal{G})$ denotes the smallest σ-algebra containing $\mathcal{G}$. If $\mathcal{A}$ and $\mathcal{B}$ are two σ-algebras of subsets of Ω, then we also write $\mathcal{A} \vee \mathcal{B}$ instead of $\sigma(\mathcal{A} \cup \mathcal{B})$. Countable spaces are usually equipped with the corresponding power set as σ-algebra. If X is a topological space, then $\mathcal{B}(X)$ denotes the Borel-σ-algebra on X. As standard σ-algebra on $\mathbb{R}$, $\bar{\mathbb{R}}$, and $\mathbb{C}$ we consider the corresponding Borel-σ-algebra.

Suppose $(\Omega, \mathcal{F})$ and $(X, \mathcal{X})$ are measurable spaces and f is a function on Ω with values in X. If f is measurable w. r. t. the σ-algebras $\mathcal{F}$ and $\mathcal{X}$, then we say f is $\mathcal{F}/\mathcal{X}$-measurable. If f

is a real-valued or numerical function, i. e., has values in $\mathbb{R}$ or $\bar{\mathbb{R}}$ and is $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable or $\mathcal{F}/\mathcal{B}(\bar{\mathbb{R}})$-measurable, then we simply say f is $\mathcal{F}$-measurable.

(1.2) Product measurable spaces. Assume that $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ are measurable spaces. Then $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$ denotes the corresponding product measurable space consisting of the product space $X \times Y$ and the product σ-algebra $\mathcal{X} \otimes \mathcal{Y}$. The set $A_x = \{y \in Y : (x, y) \in A\}$ is called the x-section of the set $A \subset X \times Y$ for all $x \in X$.

Suppose $\{(X_t, \mathcal{X}_t), t \in T\}$ is a family of arbitrary measurable spaces and for $u \le v \in \overline{T}$ the set J is given by

$$J = \begin{cases} (u, v] & \text{if } u < v < \infty \\ (u, \infty) & \text{if } u < v = \infty \\ \{v\} & \text{if } u = v \neq \pm\infty \\ \varnothing & \text{otherwise} \end{cases}.$$

Then we denote the product space and the product σ-algebra of the subfamily related to J by

$$X_u^v = \bigtimes_{t \in J} X_t, \qquad \mathcal{X}_u^v = \bigotimes_{t \in J} \mathcal{X}_t.$$

As short hand notation we use X, X_-^v, and X_u^+ for $X_{-\infty}^{\infty}$, $X_{-\infty}^v$, and X_u^{∞}, respectively. A corresponding convention applies to product σ-algebras.

Let π_t denote the coordinate projection from X to X_t and π_u^v the projection from X to X_u^v,

$$\pi_t(x) = x_t \qquad \text{and} \qquad \pi_u^v(x) = \{\pi_t(x), t \in J\},$$

where $x = \{x_s, s \in T\}$ with $x_s \in X_s$ denotes an element of X. For the inverse image of a set $A \subset X_u^v$ w. r. t. the projection π_u^v we write

$$[A] = (\pi_u^v)^{-1}(A).$$

If A contains only one element $x_u^v \in X_u^v$, then we write $[x_u^v]$ instead of $[\{x_u^v\}]$. We extend this notation in a natural way to a family $\mathcal{A}$ of subsets of X_u^v on an element-by-element basis, i. e.,

$$[\mathcal{A}] = \{[A] : A \in \mathcal{A}\}.$$

Assume that $w \in T$ and $(X_t, \mathcal{X}_t) = (X_0, \mathcal{X}_0)$ for all $t \in T$. Then the shift operator θ_w on X is defined for any $x = \{x_s, s \in T\} \in X$ by

$$\theta_w(x) = \{\tilde{x}_t, t \in T\}, \qquad \tilde{x}_t = x_{t-w}.$$

A set $A \subset X$ is called w-invariant if

$$A = \theta_w(A) = \{\theta_w(x) : x \in X\}$$

holds. It is called (shift-) invariant, if it is w-invariant for all $w \in T$. The w-shifted version of an element $x_u^v \in X_u^v$, denoted by $\langle x_u^v \rangle_w$, is an element of X_{u+w}^{v+w} given by

$$\langle x_u^v \rangle_w = \pi_{u+w}^{v+w}\big(\theta_w(x)\big),$$

where x is any element of $[x_u^v]$. We naturally extend this notation on an element-by-element basis to a set $A \subset X_u^v$ and a family $\mathcal{A}$ of subsets of X_u^v to obtain the w-shifted versions $\langle A\rangle_w$ and $\langle\mathcal{A}\rangle_w$. With the operator $\langle\cdot\rangle_w$ we are able to emphasize the position on the time axis.

Example. As an illustrative example of the introduced notation suppose $(X_t, \mathcal{X}_t) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for all $t \in T$. The elements x_t of $X_t = X_t^t$ are thus real scalars. Assume that $v \in T_+$. If $T = \mathbb{R}$, then X_0^v is the set of all real functions x_0^v on the interval $(0, v]$. For $x \in X$, i. e., a real function on the real line, $\pi_0^v(x)$ is the part of x on the interval $(0, v]$. In contrast, for a real function x_0^v on $(0, v]$, the set $[x_0^v]$ consists of all real functions on the real line coinciding on $(0, v]$ with x_0^v. The function $\langle x_0^v\rangle_w$ is a copy of x_0^v shifted to the interval $(w, v + w]$. If $T = \mathbb{Z}$, then X_0^v is the set of all v-dimensional real vectors $x_0^v = (x_1, x_2, \ldots, x_v)$. Furthermore, X_0^+ denotes the set of real one-sided sequences $x_0^+ = (x_1, x_2, \ldots)$. For $x \in X$, i. e., a two-sided real sequence, $\pi_0^v(x)$ is the part of x with indices $\{1, 2, \ldots, v\}$. For a real vector x_0^v, in turn, $[x_0^v]$ is the set of all two-sided real sequences coinciding with x_0^v for the indices $\{1, 2, \ldots, v\}$. The vector $\langle x_0^v\rangle_w$ is a copy of x_0^v with shifted indices $\{w + 1, w + 2, \ldots, w + v\}$.

Note that the space X_t can also be a finite or countable set or it can itself consist of vectors, matrices, functions etc.

(1.3) Probability and measure spaces, random variables. The triple $(\Omega, \mathcal{F}, \mu)$ consisting of the space Ω, the σ-algebra $\mathcal{F}$, and the measure μ is called measure space. If μ is a probability measure, then $(\Omega, \mathcal{F}, \mu)$ is called probability space. Suppose ν is another measure on $\mathcal{F}$. Then we write $\mu \ll \nu$ if μ is absolutely continuous w. r. t. ν (see Paragraph A.7). By δ_ω we denote the Dirac measure (see Paragraph A.4) on $\mathcal{F}$ concentrated at some $\omega \in \Omega$. If $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ are two measure spaces, then $\mu_1 \otimes \mu_2$ denotes the product measure obtained from μ_1 and μ_2, which is defined on the product σ-algebra $\mathcal{F}_1 \otimes \mathcal{F}_2$.

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. For a random variable ξ on $(\Omega, \mathcal{F}, \mathrm{P})$ we denote by $\sigma(\xi)$ the smallest σ-algebra w. r. t. which ξ is measurable. The distribution of ξ is denoted by P_ξ. If ξ is a real, numerical, or complex random variable, then $\mathrm{E}(\xi)$ denotes the expectation of ξ, given it exists. The variance of ξ is $\mathrm{var}(\xi)$. If η is another real, numerical, or complex random variable on $(\Omega, \mathcal{F}, \mathrm{P})$, then $\mathrm{cov}(\xi, \eta)$ and $\mathrm{cor}(\xi, \eta)$ are the covariance and correlation of ξ and η. For any set $F \in \mathcal{F}$ and σ-algebra $\mathcal{A} \subset \mathcal{F}$ we denote by $\mathrm{P}(F \,|\, \mathcal{A})$ the conditional probability of F given $\mathcal{A}$. If the σ-algebras $\mathcal{A}, \mathcal{B}, \mathcal{C} \subset \mathcal{F}$ form a Markov chain in this order (see Paragraph A.2), then we write $(\mathcal{A} - \mathcal{B} - \mathcal{C})$. If the random variables ξ, η, and ζ on $(\Omega, \mathcal{F}, \mathrm{P})$ form a Markov chain in this order, then we write $(\xi - \eta - \zeta)$.

Let us adopt the notation of Paragraph 1.2. The w-shifted copy $\langle\mu\rangle_w$ of a measure μ on $\mathcal{X}_u^v$ is a measure on $\mathcal{X}_{u+w}^{v+w}$ defined by

$$\langle\mu\rangle_w(A) = \mu\big(\langle A\rangle_{-w}\big)$$

for any $A \in \mathcal{X}_{u+w}^{v+w}$. For any $t \in T$ let ξ_t be a random variable on $(\Omega, \mathcal{F}, \mathrm{P})$ with values in $(X_t, \mathcal{X}_t)$. Then ξ_u^v denotes the random variable on $(\Omega, \mathcal{F}, \mathrm{P})$ with values in $(X_u^v, \mathcal{X}_u^v)$, which is defined for any $\omega \in \Omega$ by

$$\xi_u^v(\omega) = \{\xi_t(\omega), t \in J\}.$$

We identify ξ_u^v with the family $\{\xi_t, t \in J\}$ of random variables, which is (a segment of) a random process either with discrete time (random sequence) or continuous time. As short hand notation we also use ξ, ξ_-^v, and ξ_u^+ instead of $\xi_{-\infty}^\infty$, $\xi_{-\infty}^v$, and ξ_u^∞.

Example. If we assume the situation of the example at the end of Paragraph 1.2, then ξ_0^v is a real continuous-time random process on the interval $(0, v]$ for $T = \mathbb{R}$. In contrast, if $T = \mathbb{Z}$, then $\xi_0^v = (\xi_1, \xi_2, \ldots, \xi_v)$ is a v-dimensional real random vector, $\xi_0^\infty = (\xi_1, \xi_2, \ldots)$ is one-sided and $\xi = (\ldots, \xi_{-1}, \xi_0, \xi_1, \ldots)$ is a two-sided sequences of real random variables.

(1.4) Matrices, logarithms, complex numbers. The determinant and trace of a matrix A are denoted by $\det(A)$ and $\operatorname{tr}(A)$. The transpose of a vector a is denoted by a'. By $\operatorname{diag}(a_1, \ldots, a_n)$ we represent the diagonal matrix with entries a_i on the main diagonal.

If we write $\log$, then we always assume logarithms w. r. t. base e. In addition, we always suppose $0 \log \frac{0}{x} = 0$ for $x \geq 0$ and $x \log \frac{x}{0} = +\infty$ for $x > 0$.

The real part, the imaginary part, and the complex conjugate of a number $z \in \mathbb{C}$ are denoted by $\operatorname{Re}(z)$, $\operatorname{Im}(z)$, and $\overline{z}$, respectively. The imaginary unit is $\jmath$.

§2 Abstract Channel Model

An (information) channel is an abstract stochastic model that characterizes the random corruption of data or signals transmitted from a sender to a receiver. The most general mathematical description of a channel was proposed by Kolmogorov (1956b)[1]. The basic definition given below is formulated without reference to a transmission over time and is based on (Dobrushin, 1959, Sec. 1.5)[2]. Time-structure, a term adopted from Ahlswede (2006, Sec. 1), is later added by considering product spaces as channel input and output such that input and output signals have (two-sided) infinite duration. Models of this type are advantageous in terms of defining concepts such as stationarity, ergodicity, information rate capacity etc. and are considered, e. g., by Khinchin (1956, Ch. III)[3], Feinstein (1958, Sec. 6.2), Kadota and Wyner (1972), Ihara (1993, Sec. 4.1), Kakihara (1999, Sec. 3.1), or Gray (2011, Sec. 2.2). A different approach, taken by Dobrushin (1963), Wolfowitz (1978, Sec. 5.1), Ahlswede (2006, Sec. 1), Verdú and Han (1994), or Csiszár and Körner (2011, Ch. 6) to represent transmission over time is to consider not a single but a whole collection of channels, each modeling the transmission of fixed finite duration. See (Dobrushin, 1963, Sec. 1.8) for a discussion on the two types of models.

After introducing channels with time structure we define properties such as stationarity, ergodicity, causality, and asymptotic input-memorylessness. This section contains the basic material on channels, which is directly relevant in connection with formulating an abstract coding theorem (and converse). More on channels is detailed in the later Sections §13, §14 and §15.

(2.1) Definition (Channel, input/output space/measure). Let $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ be arbitrary measurable spaces. A Markov-kernel κ from $(X, \mathcal{X})$ to $(Y, \mathcal{Y})$ is called a channel with input space $(X, \mathcal{X})$, output space $(Y, \mathcal{Y})$, and input-output space $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$. The channel κ and a probability measure μ on $\mathcal{X}$, called input probability measure, induce a probability measure $\mu\kappa$ on $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$, given by

$$\mu\kappa(C) = \int_X \kappa(x, C_x) \, \mathrm{d}\mu(x), \qquad C \in \mathcal{X} \otimes \mathcal{Y}, \tag{1}$$

[1] Russian original, see (Kolmogorov, 1963) for English and (Chintschin et al., 1967, Part IV) for German translation. See also (Kolmogorov, 1956a) for a shortened English version.

[2] Russian original, see (Dobrushin, 1963, Sec. 1.5) for English and (Dobruschin, 1963, Sec. 1.5) for German translation.

[3] Russian original, see (Khinchin, 1957, Part II, Ch. III) for English and (Chintschin et al., 1967, Part II, Ch. III) for German translation.

where C_x denotes the x-section of the set C. The joint measure $\mu\kappa$ is called input-output probability measure and the marginal measure ν on $\mathcal{Y}$ given by

$$\nu(B) = \mu\kappa(X \times B) = \int_X \kappa(x, B)\,\mathrm{d}\mu(x), \qquad B \in \mathcal{Y},$$

is called output probability measure.

(2.2) Remark. At first, due to Paragraph A.3, the definition of a channel means $\kappa(x, \cdot)$ is a probability measure on $\mathcal{Y}$ for any $x \in X$, i. e., for any channel input we have a probability distribution on the channel output space. In particular, $\kappa(x, B)$ specifies the probability that the received symbol lies in the set $B \in \mathcal{Y}$ given the transmitted symbol was $x \in X$. Secondly, the definition includes that $\kappa(\cdot, B)$ is an $\mathcal{X}/\mathcal{B}([0,1])$-measurable function on X for any $B \in \mathcal{Y}$. Whether or not this technical condition is needed depends on the problem to be analyzed. As demonstrated by Augustin (1966) the measurability condition can possibly be omitted, even for abstract channels, if only coding problems are studied, which allow the restriction to finitely supported channel input measures. However, for several questions studied in this thesis we need to consider channel input-output probability measures induced by general input measures. Furthermore, in derivations in connection with channels we often have to apply measure-theoretic tools, which require the introduced measurability of the channel. To achieve a unified presentation throughout the thesis we will therefore assume this additional measurability condition. Note, that it can be a difficult problem to verify the measurability in concrete examples. However, models of physical channels are often from the class considered in Section § 15, for which the verification of the measurability is much easier.

(2.3) Definition (Channel with time structure). Let $\{(X_t, \mathcal{X}_t), t \in T\}$ and $\{(Y_t, \mathcal{Y}_t), t \in T\}$ be two families of measurable spaces with $(X_t, \mathcal{X}_t) = (X_0, \mathcal{X}_0)$ and $(Y_t, \mathcal{Y}_t) = (Y_0, \mathcal{Y}_0)$ for all $t \in T$. We adopt all the notation for product measurable spaces related to this families from Paragraph 1.2 and call a channel κ with input product space $(X, \mathcal{X})$ and output product space $(Y, \mathcal{Y})$ a channel with time structure. If we have $T = \mathbb{Z}$ for the time index set[4] introduced at the beginning of Section § 1, then κ is called discrete-time channel and if $T = \mathbb{R}$, then κ is called continuous-time channel. Elements of X and Y or of corresponding sub-product spaces are called input and output (time) signals, respectively. The measurable spaces $(X_0, \mathcal{X}_0)$ and $(Y_0, \mathcal{Y}_0)$ are called input and output alphabet.

(2.4) Example (Alphabets). The alphabets of the channel with time structure are arbitrary in the general definition. In a typical example the channel input and output signals are real- or complex- (vector-) valued, i. e., $X_0, Y_0 \in \{\mathbb{R}, \mathbb{R}^n, \mathbb{C}, \mathbb{C}^n\}$ with $\mathcal{X}_0$ and $\mathcal{Y}_0$ taken as the corresponding Borel-σ-algebras. As a representative illustration of a channel with time structure Figure 1 shows a continuous-time channel with real-valued input and output signals, i. e. inputs and outputs are real functions on the real line. The depicted example of an output set consists of all signals having at two instances of time values in a certain interval.
In the discrete-time case the binary and discrete-valued channel are further prominent examples, where the alphabets are given by $X_0, Y_0 \in \{\{0,1\}, \{0,1,\ldots,m\}\}$ together with the corresponding power sets as σ-algebras. Of course, the input and output alphabet can also be different,

[4] It can also be meaningful to speak of a channel with time structure if T is replaced by some other totally ordered set. However, this will not be considered here.

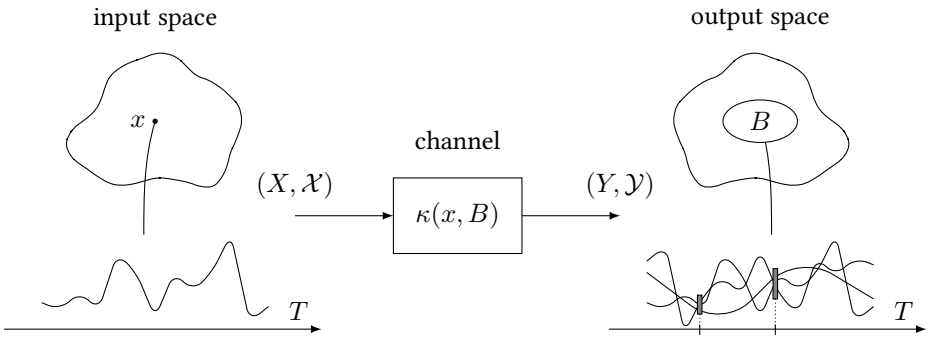


Figure 1: Channel with time structure.

which is the case, e. g., if we consider the quantization of real-valued input signals, a thresholding operation, or some modulation scheme in combination with a soft-decision-decoder.

In more advanced examples the alphabets themselves can be product or function spaces. This allows us, e. g., the following representation of a channel, which is useful in connection with coded information transmission. Assume that we partition the time index set T into segments of size $s \in T_+$, then we can consider the original channel also as a discrete-time channel with input alphabet $(X_0^s, \mathcal{X}_0^s)$ and output alphabet $(Y_0^s, \mathcal{Y}_0^s)$.

(2.5) Example (Gaussian channel). For later reference let us consider the following channel with time structure. Suppose we have real (vector-) valued output signals, i. e., $(Y_0, \mathcal{Y}_0) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ for some positive integer n. Let $\eta = \{\eta_t, t \in T\}$ denote the family of coordinate projections on the channel output space, where η_t denotes the projection from Y to Y_t. For any $x \in X$ the projection η_t is a random variable on the probability space $(Y, \mathcal{Y}, \kappa(x, \cdot))$ and η is the corresponding random process. If η is a Gaussian (vector) process (see Paragraph A.6) for all $x \in X$, then κ is called Gaussian channel. We make the assumption that all second moments of the involved random processes are finite, which is convenient for analysis and does not pose restrictions to models of practical situations. We restrict ourselves to the real case, however, the extension to the complex case is canonical. Note that we have not specified the input alphabet. The additive noise channel, where the noise is Gaussian is the most simple and prominent example of a Gaussian channel (see Paragraph 16.3). In that case we have $(X_0, \mathcal{X}_0) = (Y_0, \mathcal{Y}_0) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

(2.6) Remark (Transmission of finite duration). For the general channel with time structure the channel input signal needs to be specified for the whole set of time indices to determine the (conditional) probability of an event at the channel output. The canonical way for this channel model to represent the transmission of a signal with finite duration is to assume that the null-signal of infinite duration is sent before and after the actual transmission. This is only possible under the condition that there is a null-element in the input alphabet. The cases of practical relevance we have in mind satisfy this constraint, in particular those in Example 2.4.

Similarly, the channel with time structure characterizes the (conditional) probability of output events of infinite duration. The probability of an event $B \in \mathcal{Y}_0^s$ for finite observation time $s \in T_+$

is then given by the corresponding probability of the inverse image[5] (cylinder set) $[B]$, which means it does not matter what happened outside the observation interval $(0, s]$.

Next, we define some properties of channels with time structure, which are directly relevant in connection with proving a coding theorem (and converse) for such channels. In the rest of this section κ is a channel with time structure as introduced in Definition 2.3.

(2.7) Definition (Stationarity, ergodicity, causality, asymptotic input-memorylessness).

(i) *Stationarity.* Given $s \in T$, the channel κ is called s-stationary, if for any $x \in X$ and $B \in \mathcal{Y}$ we have

$$\kappa(x, B) = \kappa(\theta_s(x), \theta_s(B)),$$

where θ_s denotes the shift operator[6] defined in Paragraph 1.2. It is called stationary, if it is s-stationary for all $s \in T$.

(ii) *Ergodicity.* For $s \in T_+$, an s-stationary channel κ is called s-ergodic (s-ergodic for s-i.i.d. inputs), if for any s-stationary s-ergodic (s-i.i.d., see Definition B.1) channel input probability measure μ the induced input-output probability measure $\mu\kappa$ is s-stationary and s-ergodic. A stationary channel κ is called totally ergodic (totally ergodic for block-i.i.d. inputs), if for all $s \in T_+$ it is s-ergodic (s-ergodic for s-i.i.d. inputs). It is called ergodic, if for any stationary ergodic channel input probability measure μ the induced input-output probability measure $\mu\kappa$ is stationary and ergodic.

(iii) *Causality.* The channel κ is called causal if for any $t \in T$, $B \in \mathcal{Y}_-^t$, and $x, \tilde{x} \in X$ coinciding on $(-\infty, t]$ we have

$$\kappa(x, [B]) = \kappa(\tilde{x}, [B]).$$

(iv) *Asymptotic input-memorylessness.* The channel κ is called asymptotically input-memoryless for the input signal set $X' \subset X$ if for any $\epsilon > 0$ and $s \in T$ there exists a $t_I(\epsilon, s) \in T_0$ such that for any $B \in \mathcal{Y}_s^+$ and $x, \tilde{x} \in X'$ coinciding on $(s - t_I(\epsilon, s), \infty)$ we have

$$\big|\kappa(x, [B]) - \kappa(\tilde{x}, [B])\big| < \epsilon.$$

(2.8) Remark. The introduced concept of stationarity is basically considered by all authors using the same type of channel with time structure. See for example (Gray, 2011, Sec. 2.3) from the list of papers given at the beginning of this section. Stationarity of a channel actually means shift invariance, that is, the conditional probabilities specified by the channel do not change if the input signal and the output event are jointly shifted. The version given here is suitable for the discrete- and the continuous-time case. However, in the former case the definition simplifies due to the following observation: An s-stationary channel is also u-stationary for all $u = ks$ with $k \in \mathbb{Z}$. Therefore, stationarity and 1-stationarity are equivalent if $T = \mathbb{Z}$. Note that for a stationary channel κ the probability measure $\kappa(x, \cdot)$ on $\mathcal{Y}$ for fixed input $x \in X$ is usually not stationary. However, stationary input measures are transformed into stationary input-output

[5] When it is clear from the context we will not explicitly mention the space on which projections are defined to build inverse images. For example, here we have a channel κ, whose second argument is a set form $\mathcal{Y}$. Thus, the projection w. r. t. which the inverse image $[B]$ of the set B is taken, must be defined on the space Y. In connection, e. g., with mutual information the projections are usually defined on the product space $X \times Y$, as in Remarks 4.2 and 4.4.

[6] To keep notation simple θ_s denotes the shift operator on X as well as on Y.

and output measures, as stated in the lemma below. This property can also be used to define a more relaxed version of stationarity of channels as given in (Gray, 2011, p. 26).

In the literature, ergodic channels (see (Kakihara, 1999, p. 137) or (Gray, 2011, p. 28)) and totally ergodic channels (see Vajda (1967, Def. 1) or Gray (2011, p. 360)) are considered. In addition, we introduce a less restrictive version of totally ergodic channels because this is exactly the form we need to derive a block-coding theorem. For this variant we require the s-ergodicity of the input-output probability measure only for s-i.i.d. input probability measures and not, as common, for all s-ergodic input probability measures. We indicate this modification by adding "for (s-/block-) i.i.d. inputs" to the usual name. According to the comment at the beginning of Remark B.2 a natural alternative of this supplementary expression is "for (s-/block-) memoryless inputs". A nontrivial example showing that this condition is strictly weaker than total ergodicity is given in Paragraph 16.4. Further note, that ergodicity is defined only in connection with stationarity, since we will only use this restricted form. Ergodicity of a channel is an indirect condition on the so-called output memory (see Section § 13 for details, especially Theorem 13.9 and (15.4.iv) for explicit conditions) and can be interpreted as a weak form of asymptotic independence of remote output events.

The definition of causality is canonical and the given formulation is taken from (Kadota, 1972). For a causal channel the probability of an output event up to time t is determined by the input signal up to time t, i. e., no future inputs must be known to determine the probability of an output event involving current and past time indices. Since we want to interpret the index set T as time axis, it is always physically meaningful to assume the channel to be causal. In this case we do not need to specify future inputs in the situation of Remark 2.6.

Causality has the following alternative characterization: The channel κ is causal if for any $t \in T$ and $B \in \mathcal{Y}_-^t$ the function $\kappa(\cdot, [B])$ is $[\mathcal{X}_-^t]$-measurable. To obtain this equivalent definition of causality suppose κ is causal and let $t \in T$ and $B \in \mathcal{Y}_-^t$. According to (2.7.iii) there exists an $\mathcal{X}_-^t$-measurable function $g_{t,B}$ on X_-^t such that $\kappa(\cdot, [B]) = g_{t,B}\big(\pi_-^t(\cdot)\big)$, where π_-^t denotes the projection from X to X_-^t. Due to the factorization lemma (see Lemma A.10) this is equivalent to the $[\mathcal{X}_-^t]$-measurability of $\kappa(\cdot, [B])$. Causality is defined for stationary and non-stationary channels. If κ is stationary and the defining relation of causality is true for $t = 0$, then it is automatically true for all $t \in T$, which allows to simplify the definition in this case.

The condition of asymptotic input-memorylessness is adopted from (Kadota and Wyner, 1972) and means, that the probability of an output event after time s is determined within a tolerance of ϵ by the input signal after time s and $t_I(\epsilon, s)$ time indices in the past (at most). Defining asymptotic input-memorylessness for all input signals is often too restrictive. Therefore, we consider the subspace X' in the definition, which is sufficient if X' is chosen properly. As causality, the asymptotic input-memorylessness is defined for stationary and non-stationary channels. If κ is stationary, X' is a shift-invariant set, and the defining relation of asymptotic input-memorylessness is true for $s = 0$, then it is true for all $s \in T$. In this case $t_I(\epsilon, s)$ does not depend on s.

The proof of the next lemma is given, e. g., in (Gray, 2011, p. 25) or (Kakihara, 1999, p. 124) for the case $T = \mathbb{Z}$. There is actually no difference in the derivations for $T = \mathbb{R}$.

(2.9) Lemma (*Stationarity of induced input-output probability measure). If κ is an s-stationary (stationary) channel and μ is an s-stationary (stationary) channel input probability measure, then the induced input-output probability measure $\mu\kappa$ and the corresponding output probability measure ν are s-stationary (stationary).*

§3 Block Codes and Information Transmission

In this thesis, we consider Shannon's classical problem of channel coding in a general context. Channel coding means a block-based coding-decoding procedure is employed to reliably transmit messages from a sender to a receiver in the presence of random transmission errors. The noisy part of the communication process is modeled by an information channel connecting sender and receiver. A main information-theoretic concern of coded information transmission is the analysis of the relation between the speed of transmission and the probability of a decoding error. More precisely, we are interested in conditions under which transmitted messages — even though they are disturbed — can be inferred from the received messages with arbitrarily low error probability, while maintaining a certain transmission rate. So called coding theorems (and converses, see Section §9) are central results stating theoretical limits in this regard. This section introduces basic notions required to precisely formulate those results.

The subsequent definitions are formulated such that they are suitable for any channel with time-structure, including discrete- and continuous-time models with arbitrary alphabets. The effect of channel memory (also infinite) representing the potential impact of previously transmitted signals (codewords) on the current transmission is taken into account as well as possible restrictions on allowed channel inputs. Later, we want to formulate the main coding results in a way that emphasizes their operational meaning. Therefore, the definitions below are based on the style of Wolfowitz (1978) and Ahlswede (2006, Sec. 2).

The setting in this section is a channel with time structure as introduced in Definition 2.3. We adopt the notation from there and Paragraph 1.2. First, we specify constraints on the channel input signals in an abstract manner. As in (Thomasian, 1961, Sec. 4) or (Ash, 1965, Sec. 8.1) this is done based on sets. We introduce relevant quantities in connection with channel coding, such as block code, code rate, decoding error etc. The definitions are extensions of those in (Jelinek, 1968, Sec. 6.2), (Wolfowitz, 1978, Sec. 5.1), or (Ahlswede, 2006, Sec. 2). Then we explain how a code is used in a communication system to transmit information, in particular to point out, that we consider codes that allow the repeated transmission of messages with the same reliability.

(3.1) Input constraints. Usually signals used for information transmission have to meet certain constraints that result for example from technical specifications. For a transmission of duration $s \in T_+$ starting at time 0, we model such constraints in an abstract way by a set $E_s \subset X_0^s$, i. e., a subset of all possible input signals in the time period $(0, s]$. Please note that we do not require $E_s \in \mathcal{X}_0^s$, which is particularly important for continuous-time channels. From E_s we build the set

$$E_s^* = \bigtimes_{k\in\mathbb{Z}} \langle E_s \rangle_{ks} \subset X \tag{1}$$

of all signals in X that satisfy the input constraint in any time period $(ks, (k+1)s]$, $k \in \mathbb{Z}$. A signal from E_s^* represents a sequence of transmissions, each of duration s and satisfying the input constraint. In cases of practical importance the set E_s contains the null-signal such that we can represent a transmission of finite duration also by a signal from E_s^* as explained in Remark 2.6. As an example consider the amplitude or power constraint in Example 3.2. The union of all time-shifted versions of E_s^* is the shift-invariant set

$$E_s' = \bigcup_{t\in(0,s]} \theta_t(E_s^*). \tag{2}$$

Assume now that we have a constraint set for any time duration, i. e., suppose we have the family $\mathcal{E} = \{E_s \subset X_0^s, s \in T_+\}$ of sets. Then

$$E'' = \bigcup_{s \in T_+} E'_s \tag{3}$$

with E'_s as in (2), is a suitable candidate for the set of input signals w. r. t. which asymptotic input-memorylessness is defined, in particular for stationary channels (see (2.7.iv) and Remark 2.8). We are especially interested in families $\mathcal{E}$ of constraint sets, that satisfy the following regularity condition

$$\bigtimes_{k=0}^{n-1} \big\langle E_s \big\rangle_{ks} \subset E_{ns} \tag{4}$$

for any $s \in T_+$ and $n \in \mathbb{N}$. This condition is important, for example, to obtain suitable representations of the information rate capacity and also directly in the coding theorem (and converse) for channels with time structure. A useful sufficient condition for (4) is that

$$E_u \times \big\langle E_v \big\rangle_u \subset E_{u+v} \tag{5}$$

holds for all $u, v \in T_+$.

(3.2) Example (Input constraints). Assume that $s \in T_+$. Often a constraint set $E_s \subset X_0^s$ has the form[7]

$$E_s = \big\{\Phi_s \le 1\big\} = \big\{x \in X_0^s : \Phi_s(x) \le 1\big\}, \tag{1}$$

where Φ_s is a nonnegative functional on X_0^s, also called cost function. Clearly, using an indicator function we can indirectly characterize any set $E_s \subset X_0^s$ by such a functional. However, reasonable examples are of the following type. Let ϕ be a nonnegative function on X_0 and assume that ϕ also denotes the identically defined function on X_t for all $t \in T$. In our first example the functional Φ_s is given by

$$\Phi_s(x) = \sup_{t \in (0,s]} \phi(x_t), \tag{2}$$

for all $x = \big\{x_t, t \in (0, s]\big\} \in X_0^s$ and in the second example Φ_s has the form

$$\Phi_s(x) = \frac{1}{s} \int_{(0,s]} \phi(x_t)\, \mathrm{d}\lambda(t). \tag{3}$$

Here we denote by λ the counting measure on the integers if $T = \mathbb{Z}$ and the one-dimensional Lebesgue measure if $T = \mathbb{R}$ (see Paragraph A.4). This allows us a common notation for the discrete- and continuous-time case. In the continuous-time case the integral is defined only for signals $x \in X_0^s$ for which the nonnegative function $\{\phi(x_t), t \in (0, s]\}$ is Lebesgue-measurable. For the remaining signals in X_0^s we set $\Phi_s(x)$ equal to an arbitrary constant larger than 1. A constraint specified by (3) is considered, e. g., in (Csiszár and Körner, 2011, p. 91) for the discrete-time case. If we have real-valued input signals as in Example 2.4 and $\phi(x_t) = |x_t|$, then (2)

[7] Without loss of generality we can choose 1 as upper bound in (1) because we can normalize the inequality.

represents the so-called amplitude constraint. For real-valued input signals and $\phi(x_t) = x_t^2$, the functional in (3) represents the important average energy or power constraint. Note that in the continuous-time case, we indeed have $E_s \notin \mathcal{X}_0^s$ for the amplitude and the average energy constraint. See (Bharucha, 1969, Sec. 5.1) and the references at the end of Paragraph A.12 for details on this fact.

Consider now the family $\mathcal{E} = \{E_s \subset X_0^s, s \in T_+\}$ of constraint sets, where E_s is given in (1). In the first example with Φ_s as in (2), we have for any $s \in T_+$

$$E_s = \bigtimes_{t\in(0,s]} J_t,$$

which implies for the sets introduced in Paragraph 3.1

$$\begin{aligned} E_s^* = E_s' &= E'' \\ &= \bigtimes_{t\in T} J_t, \end{aligned}$$

where J_t is for all $t \in T$ the inverse image of the real interval $[0, 1]$ w. r. t. ϕ. We further obtain for any $u, v \in T_+$

$$E_u \times \left\langle E_v \right\rangle_u = E_{u+v},$$

such that the regularity condition (3.1.4) holds even with equality. The second example with Φ_s as in (3) also satisfies the regularity condition. Given the signals $\hat{x} = \{\hat{x}_t, t \in (0, u]\} \in E_u$ and $\check{x} = \{\check{x}_t, t \in (0, v]\} \in E_v$, we have for the composed signal $x = (\hat{x}, \langle\check{x}\rangle_u) = \{x_t, t \in (0, u+v]\}$

$$\begin{aligned} \int_{(0,u+v]} \phi(x_t)\,\mathrm{d}\lambda(t) &= \int_{(0,u]} \phi(x_t)\,\mathrm{d}\lambda(t) + \int_{(u,u+v]} \phi(x_t)\,\mathrm{d}\lambda(t) \\ &= \int_{(0,u]} \phi(\hat{x}_t)\,\mathrm{d}\lambda(t) + \int_{(0,v]} \phi(\check{x}_t)\,\mathrm{d}\lambda(t) \\ &\le u + v, \end{aligned}$$

which implies (3.1.5) and therefore (3.1.4). Furthermore, we can show that the set

$$F = \left\{ x = \{x_t, t \in T\} \in X : \limsup_{s\to\infty} \frac{1}{2s} \int_{(-s,s]} \phi(x_t)\,\mathrm{d}\lambda(t) \le 1 \right\}$$

is shift-invariant and that

$$E'' \subset F$$

holds, with E'' as given in (3.1.3). For example, if the input signals are real-valued and $\phi(x_t) = x_t^2$, then F represents the set of all asymptotically power limited signals with infinite duration. The comment directly below (3.1.3) does also apply to the set F in connection with the constraints defined by (3), which is the reason why we are interested in this set.

(3.3) Definition (Block code with input constraint, code parameters). Suppose the quantities $b \in T_+$, $E_b \subset X_0^b$, and $W = \{1, 2, \dots, m\}$ for some $m \in \mathbb{N}$ are given. Let $u_i \in E_b$, $i \in W$, be pairwise distinct and let $B_i \in \mathcal{Y}_0^b$, $i \in W$, form a partition[8] of Y_0^b. Then

$$\mathcal{C}(b, E_b) = \big\{(u_i, B_i), i \in W\big\}$$

is called a (channel) block code for the message set W satisfying the input constraint E_b, where u_i is called codeword and B_i decoding set for message $i \in W$. The code rate of $\mathcal{C}(b, E_b)$ is defined by

$$R_{\mathcal{C}} = \frac{1}{b} \log m,$$

where $m = |\mathcal{C}(b, E_b)|$ denotes the code size and b the block length of $\mathcal{C}(b, E_b)$.

(3.4) Definition (Decoding error probability, (b, E_b, V, M, ϵ)-code). Let $\mathcal{C}(b, E_b)$ be a block code as in Definition 3.3 and let

$$U_b^* = \bigtimes_{k \in \mathbb{Z}} \big\langle \{u_i, i \in W\} \big\rangle_{kb} \tag{1}$$

be the set of all two-sided sequences of codewords. Suppose V is a b-invariant set satisfying $U_b^* \subset V \subset X$ and assume that the channel κ is b-stationary. The decoding error probability $\varrho(u_i, V)$ for codeword u_i w. r. t. the set V is defined by

$$\varrho(u_i, V) = \sup_{x \in [u_i] \cap V} \kappa(x, [B_i^{\mathrm{c}}]),$$

where B_i is the decoding set corresponding to u_i. The (maximal) decoding error probability $\varrho_{\max}(V)$ of $\mathcal{C}(b, E_b)$ is defined by

$$\varrho_{\max}(V) = \max_{i \in W} \varrho(u_i, V), \tag{2}$$

where W denotes the finite message set. The code $\mathcal{C}(b, E_b)$ is called a (b, E_b, V, M, ϵ)-code if

$$|\mathcal{C}(b, E_b)| \geq M \qquad \text{and} \qquad \varrho_{\max}(V) \leq \epsilon.$$

(3.5) Coding and decoding. A block code $\mathcal{C}(b, E_b)$ is used in the following way to communicate messages from a set W over a channel κ with time structure. Let the operation start at time 0. To transmit message $i \in W$, codeword u_i is sent within the time period $(0, b]$. If the received signal v within the same time period lies in the decoding set B_j, then v is decoded as message j. To transmit a sequence $(i_1, i_2, \dots, i_n) \in W^n$ of n messages, the described coding-decoding rule is applied n times in succession. To transmit the kth message in the sequence, the time-shifted version $\langle u_{i_k} \rangle_{(k-1)b}$ of the codeword u_{i_k} is sent within the time period $((k-1)b, kb]$. If the received signal v within $((k-1)b, kb]$ lies in the time-shifted version $\langle B_j \rangle_{(k-1)b}$ of the decoding set B_j, then v is decoded as message j. This message is then an estimate of i_k. The

[8] Assuming disjointness is essential here but the decoding sets do not have to fill up the whole space Y_0^b. However, the assumption is convenient and for our purposes it can always be established by merging the remaining part of Y_0^b with one of the sets B_i.

process of transmitting n messages ends at time nb. Please refer to Remark 2.6 for a discussion on input signals of finite duration in connection with the considered channel model.

Figure 2 illustrates the coding-decoding process for a continuous-time channel with real input and output signals. The messages $2, 1, m, \ldots$ are encoded into the real functions $u_2, u_1, u_m, \ldots$ of duration b taken from the set of codewords. The decoder identifies in each time interval of length b the decoding set containing the received noisy signal, which is represented as dot in the output space. In the example, the second message in the sequence is incorrectly decoded.

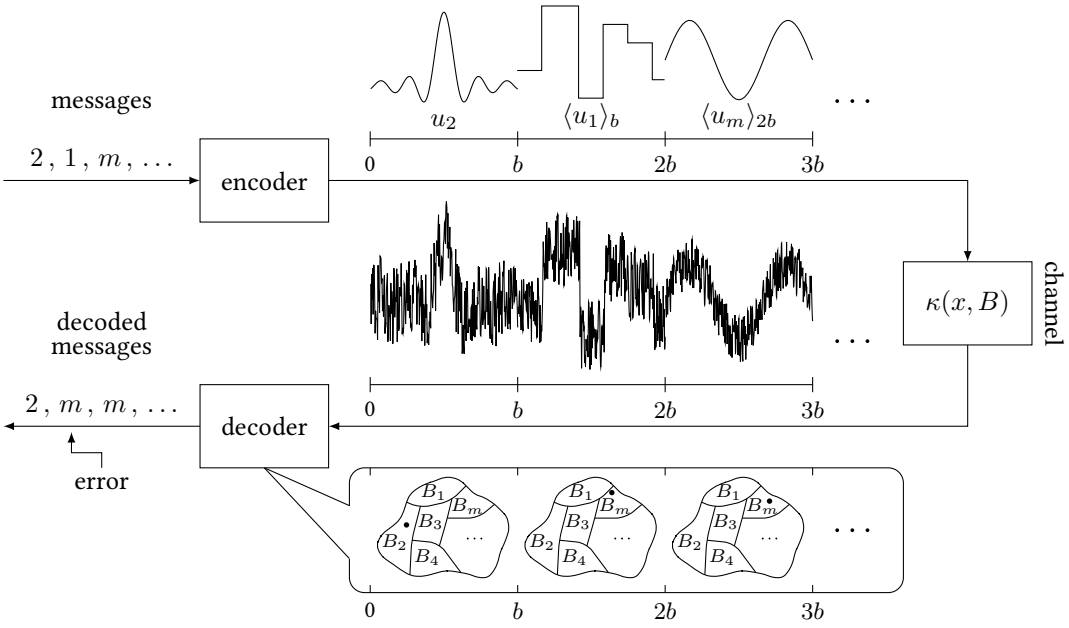


Figure 2: Illustration of coded information transmission.

The code size $|\mathcal{C}(b, E_b)|$ is the number of different messages that can be communicated at most with the code. The block length b specifies the time duration of transmitting a single message. The set E_b represents certain constraints imposed on the codewords as described in Paragraph 3.1. Examples of constraints are given in Example 3.2. The code rate $R_{\mathcal{C}}$ characterizes the speed of transmission in [nat/channel use] if $T = \mathbb{Z}$ or in [nat/second] if $T = \mathbb{R}$.

For any $t \in T$ let V_-^t and V_t^+ denote the images of the set V w. r. t. the projections from X to X_-^t and to X_t^+, respectively. Then $\varrho(u_i, V)$ is the probability that codeword u_i — a signal in the time period $(0, b]$ — is not decoded as message i, no matter what signal from the set V_-^0 has been sent before and from the set V_b^+ has been sent afterwards. The assumption of a b-stationary channel κ and a b-invariant set V implies that $\varrho(u_i, V)$ is also equal to the probability that the shifted codeword $\langle u_i \rangle_{kb}$, $k \in \mathbb{N}$, is incorrectly decoded, regardless of past signals from the set V_-^{kb} or future signals from the set $V_{(k+1)b}^+$. That means $\varrho(u_i, V)$ is the decoding error probability w. r. t. V for codeword u_i and for any version of u_i shifted by a multiple of the block length b. The maximum of this decoding error probability over all codewords of $\mathcal{C}(b, E_b)$ is represented by $\varrho_{\max}(V)$. Subsequently, we restrict ourselves to situations, where we have the described

shift-invariance of the decoding error. The introduced decoding error probability is based on (Kadota and Wyner, 1972, eq. (3)) and (Gray and Ornstein, 1979, eq. (2)). It is defined in a worst case sense over the possible past and future input history.

We require the set V to contain the set U_b^* of all sequences of codewords to make the decoding error at least robust w. r. t. past and future transmissions of codewords. Since it might be of interest to have robustness not only w. r. t. codewords, the definition allows to consider also larger sets of signals. As a typical example consider the set $E_b^* = \times_{k\in\mathbb{Z}} \langle E_b \rangle_{kb} \supset U_b^*$ already introduced in (3.1.1). For a transmission over time it is physically meaningful to assume the channel to be causal. Then the decoding error does only depend on past but not future inputs.

§4 Information Measures

We introduce mutual information and entropy in conditional and unconditional form. These basic information measures go back to the work of Shannon (1948) and were generalized by the Russian school of information theory starting with the work of Gelfand et al. (1956)[9] and continued mainly by Dobrushin (1959)[10] and Pinsker (1960)[11]. A standard reference for information measures in general form is (Pinsker, 1964), where the definitions are given for random variables with abstract alphabets. Even though this reference is the basis for this section, the difference in the presentation below is that the material is given for σ-algebras. Random variables are treated as special case. Having both versions gives us a nice flexibility in formulating results and proofs. When we work directly with measures the σ-algebra based version is usually more natural and convenient. However, some results are expressed more clearly in terms of random variables. We adopt the σ-algebra based version for mutual information and entropy from (Bradley, 2007, Ch. 5). For a general form of conditional entropy see also (Billingsley, 1965, Secs. 6 and 12). The definition of conditional mutual information is a modified version of the random variable based formulation of Wyner (1978). The advantage of Wyner's direct definition is that it generalizes Dobrushin's indirect definition considered in (Pinsker, 1964) to cases, for which certain regular conditional probabilities do not exist (Pinsker, 1964, see translator's remarks to Ch. 3). As a reference of information measures for abstract alphabets see also (Gray, 2011, Sec. 7.4). The general definitions introduced below are required in particular to handle continuous-time models.

In this section we further give an integral representation of mutual information and a list of relevant properties of the information measures. Finally, we introduce the (mutual) information rate and consider important special cases, for which the information rate exists. We omit physical or engineering interpretations of the defined quantities since we are only interested in their application as mathematical tools to prove coding results. Throughout this section, the setting will be a given abstract probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Unless stated otherwise all random variables are defined on this space.

(4.1) Definition (Mutual information, entropy). Suppose $\mathcal{A}$ and $\mathcal{B}$ are sub-σ-algebras of $\mathcal{F}$. Then the mutual information $I(\mathcal{A};\mathcal{B})$ between $\mathcal{A}$ and $\mathcal{B}$ is defined by

$$I(\mathcal{A};\mathcal{B}) = \sup \sum_{i=1}^{m} \sum_{j=1}^{n} \mathrm{P}(A_i \cap B_j) \log \frac{\mathrm{P}(A_i \cap B_j)}{\mathrm{P}(A_i)\mathrm{P}(B_j)},$$

[9] Russian original, see (Shiryaev, 1992, No. 2) for English and (Gelfand et al., 1958, Ch. II) for German translation.
[10] see footnote 2 on page 8
[11] Russian original, see (Pinsker, 1964) for English and (Pinsker, 1963) for German translation.

where the supremum is taken w. r. t. all partitions $\{A_1, A_2, \dots, A_m\}$ and $\{B_1, B_2, \dots, B_n\}$ of Ω with $A_i \in \mathcal{A}$ and $B_j \in \mathcal{B}$. The entropy $H(\mathcal{A})$ of $\mathcal{A}$ is defined by

$$H(\mathcal{A}) = I(\mathcal{A}; \mathcal{A}).$$

Suppose ξ and η are random variables with values in arbitrary measurable spaces. Then the mutual information $I(\xi; \eta)$ between ξ and η is defined by

$$I(\xi; \eta) = I(\sigma(\xi); \sigma(\eta))$$

and the entropy $H(\xi)$ of ξ by

$$H(\xi) = H(\sigma(\xi)).$$

(4.2) Remark. Assume that the random variables ξ and η have values in the measurable spaces $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$. Then we have the identity

$$I(\sigma(\xi); \sigma(\eta)) = I\big([\mathcal{X}]; [\mathcal{Y}]\big), \tag{1}$$

where the left-hand side refers to the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ on which the random variables ξ and η are defined as in Definition 4.1. The right-hand side refers to the probability space $(X \times Y, \mathcal{X} \otimes \mathcal{Y}, \mathrm{P}_{\xi,\eta})$ and has the form

$$I\big([\mathcal{X}]; [\mathcal{Y}]\big) = \sup \sum_{i=1}^{m} \sum_{j=1}^{n} \mathrm{P}_{\xi,\eta}(A_i \times B_j) \log \frac{\mathrm{P}_{\xi,\eta}(A_i \times B_j)}{\mathrm{P}_{\xi}(A_i)\mathrm{P}_{\eta}(B_j)}, \tag{2}$$

where P_{ξ}, P_{η}, and $\mathrm{P}_{\xi,\eta}$ denote the distribution of ξ, η, and (ξ, η), respectively. The supremum is taken w. r. t. all partitions $\{A_1, A_2, \dots, A_m\}$ of X and $\{B_1, B_2, \dots, B_n\}$ of Y with $A_i \in \mathcal{X}$ and $B_j \in \mathcal{Y}$. The equality in (1) follows from the correspondence between the involved partitions of Ω and $X \times Y$ as shown, e. g., in (Mittelbach, 2012, Par. 2.2). The representation in (2) is used in (Dobrushin, 1963) or (Pinsker, 1964) as definition of the mutual information between ξ and η.

There exists a useful integral representation of the mutual information, which was independently obtained by Gelfand and Yaglom (1957)[12] and Pérez (1957). The version given below is based on (Bradley, 2007, Th. 5.6).

(4.3) Theorem (*Integral form of mutual information). Let $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ be the product of the measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, assume that $\mathcal{A}$ is a sub-σ-algebra of $\mathcal{F}_1$ and $\mathcal{B}$ is a sub-σ-algebra of $\mathcal{F}_2$. Let P be a probability measure on $\mathcal{F}_1 \otimes \mathcal{F}_2$ and let P_1 and P_2 denote the marginal measures of P on $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. Further, let P' denote the restriction of P to $\mathcal{A} \otimes \mathcal{B}$ and Q the restriction of the product measure $\mathrm{P}_1 \otimes \mathrm{P}_2$ to $\mathcal{A} \otimes \mathcal{B}$. Then we have*

$$I\big([\mathcal{A}]; [\mathcal{B}]\big) = \begin{cases} \displaystyle\int_{\Omega_1 \times \Omega_2} \log f \, \mathrm{dP}' & \text{if} \quad \mathrm{P}' \ll \mathrm{Q} \\ \infty & \text{otherwise} \end{cases}, \tag{1}$$

where f denotes the Q-density of P', if P' is absolutely continuous w. r. t. Q (see Paragraph A.7).

[12] Russian original, see (Gelfand and Yaglom, 1959) for English and (Gelfand et al., 1958, Ch. I) for German translation.

(4.4) Remark. The integral in (4.3.1) can be rewritten as

$$\int_{\Omega_1\times\Omega_2} \log f \,\mathrm{dP}' = \int_{\Omega_1\times\Omega_2} f \log f \,\mathrm{dQ}$$

since f is the Q-density of P'.

Consider now the setting in Remark 4.2. Applying the result from Theorem 4.3 in the probability space $(X\times Y, \mathcal{X}\otimes\mathcal{Y}, \mathrm{P}_{\xi,\eta})$ yields

$$I(\xi;\eta) = I\big([\mathcal{X}];[\mathcal{Y}]\big) = \int_{X\times Y} \log f_{\xi,\eta} \,\mathrm{dP}_{\xi,\eta} \tag{1}$$

if $\mathrm{P}_{\xi,\eta}$ is absolutely continuous w. r. t. $\mathrm{P}_\xi\otimes\mathrm{P}_\eta$, where $f_{\xi,\eta}$ denotes the $\mathrm{P}_\xi\otimes\mathrm{P}_\eta$-density of $\mathrm{P}_{\xi,\eta}$. Commonly, $\log f_{\xi,\eta}$ is called information density of ξ and η. Using the integral transformation formula we have for (1)

$$\int_{X\times Y} \log f_{\xi,\eta}(x,y) \,\mathrm{dP}_{\xi,\eta}(x,y) = \int_\Omega \log f_{\xi,\eta}\big(\xi(\omega),\eta(\omega)\big) \,\mathrm{dP}'(\omega), \tag{2}$$

where P' denotes the restriction of P to the σ-algebra $\sigma(\xi)\vee\sigma(\eta)$.

Assume now that the underlying probability space $(\Omega,\mathcal{F},\mathrm{P})$ on which ξ and η are defined has product structure, i. e., $\Omega=\Omega_1\times\Omega_2$ and $\mathcal{F}=\mathcal{F}_1\otimes\mathcal{F}_2$. Let Q denote the restriction of $\mathrm{P}_1\otimes\mathrm{P}_2$ to $\sigma(\xi)\vee\sigma(\eta)$, where P_1 and P_2 denote the marginal measures of P on $\mathcal{F}_1$ and $\mathcal{F}_2$. Suppose

$$\mathrm{P}_\xi\otimes\mathrm{P}_\eta = (\mathrm{P}_1\otimes\mathrm{P}_2)_{\xi,\eta} \tag{3}$$

holds, where $(\mathrm{P}_1\otimes\mathrm{P}_2)_{\xi,\eta}$ denotes the distribution of (ξ,η) if the probability measure P on the underlying probability space is replaced by $\mathrm{P}_1\otimes\mathrm{P}_2$. Then the integral transformation formula yields that $f_{\xi,\eta}(\xi,\eta)$ is the Q-density of P'. The condition in (3) is satisfied, e. g., if the random variables ξ and η can be represented as the compositions

$$\xi = \xi'\circ\pi_1 \qquad \text{and} \qquad \eta = \eta'\circ\pi_2.$$

Here ξ' is a random variable on $(\Omega_1,\mathcal{F}_1,\mathrm{P}_1)$ with values in $(X,\mathcal{X})$ and η' is a random variable on $(\Omega_2,\mathcal{F}_2,\mathrm{P}_2)$ with values in $(Y,\mathcal{Y})$. Further, π_1 and π_2 denote the projections from $\Omega_1\times\Omega_2$ to Ω_1 and to Ω_2, respectively. In this situation we have $\sigma(\xi)\subset[\mathcal{F}_1]$, $\sigma(\eta)\subset[\mathcal{F}_2]$, and $\sigma(\xi)\vee\sigma(\eta) = \sigma(\xi')\otimes\sigma(\eta')$. Therefore, an integral representation of the mutual information $I(\xi;\eta)$ on the underlying probability space, as in (2), can be obtained directly by applying Theorem 4.3 in the probability space $(\Omega_1\times\Omega_2, \mathcal{F}_1\otimes\mathcal{F}_2, \mathrm{P})$. See (Mittelbach, 2012, Par. 1.7, 2.12) for details and (Gray, 2011, Lem. 7.5) for a related result. If it is advantageous to consider the mutual information in the underlying probability space or in the space of values depends on the specific application.

(4.5) Definition (Conditional mutual information, conditional entropy). Let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be sub-σ-algebras of $\mathcal{F}$. Then the conditional mutual information $I(\mathcal{A};\mathcal{B}|\mathcal{C})$ between $\mathcal{A}$ and $\mathcal{B}$ given $\mathcal{C}$ is defined by

$$I(\mathcal{A};\mathcal{B}|\mathcal{C}) = \sup \sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{E}\left[\mathrm{P}(A_i\cap B_j|\mathcal{C}) \log \frac{\mathrm{P}(A_i\cap B_j|\mathcal{C})}{\mathrm{P}(A_i|\mathcal{C})\mathrm{P}(B_j|\mathcal{C})}\right],$$

where the supremum is taken w. r. t. all partitions $\{A_1, A_2, \dots, A_m\}$ and $\{B_1, B_2, \dots, B_n\}$ of Ω with $A_i \in \mathcal{A}$ and $B_j \in \mathcal{B}$. The conditional entropy $H(\mathcal{A}|\mathcal{C})$ of $\mathcal{A}$ given $\mathcal{C}$ is defined by

$$H(\mathcal{A}|\mathcal{C}) = I(\mathcal{A}; \mathcal{A}|\mathcal{C}).$$

Suppose ξ, η, and ζ are random variables with values in arbitrary measurable spaces. Then the conditional mutual information $I(\xi; \eta|\zeta)$ between ξ and η given ζ is defined by

$$I(\xi; \eta|\zeta) = I(\sigma(\xi); \sigma(\eta)|\sigma(\zeta))$$

and the conditional entropy $H(\xi|\zeta)$ of ξ given ζ by

$$H(\xi|\zeta) = H(\sigma(\xi)|\sigma(\zeta)).$$

(4.6) Remark. The conditional probability for the trivial σ-algebra $\{\Omega, \varnothing\}$ satisfies $\mathrm{P}(A) = \mathrm{P}(A|\{\Omega, \varnothing\})$ for all $A \in \mathcal{F}$. Therefore, we have

$$I(\mathcal{A}; \mathcal{B}) = I(\mathcal{A}; \mathcal{B}|\{\Omega, \varnothing\}),$$

i. e., Definition 4.1 is a special case of Definition 4.5.

(4.7) Fundamental properties of information measures. Let $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be sub-σ-algebras of $\mathcal{F}$. Then the following properties hold.

(i) *Nonnegativity:*

$$\begin{aligned} 0 \le I(\mathcal{A}; \mathcal{B}), \qquad & I(\mathcal{A}; \mathcal{B}) = 0 \iff \mathcal{A}, \mathcal{B} \text{ independent} \\ 0 \le I(\mathcal{A}; \mathcal{B}|\mathcal{C}), \qquad & I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = 0 \iff (\mathcal{A} - \mathcal{C} - \mathcal{B}) \end{aligned}$$

(ii) *Monotonicity:* If $\mathcal{A}_1 \subset \mathcal{A}_2$, then

$$\begin{aligned} I(\mathcal{A}_1; \mathcal{B}) &\le I(\mathcal{A}_2; \mathcal{B}), \\ I(\mathcal{A}_1; \mathcal{B}|\mathcal{C}) &\le I(\mathcal{A}_2; \mathcal{B}|\mathcal{C}). \end{aligned}$$

(iii) *Symmetry:*

$$\begin{aligned} I(\mathcal{A}; \mathcal{B}) &= I(\mathcal{B}; \mathcal{A}), \\ I(\mathcal{A}; \mathcal{B}|\mathcal{C}) &= I(\mathcal{B}; \mathcal{A}|\mathcal{C}). \end{aligned}$$

(iv) *Chain rule:*

$$\begin{aligned} I(\mathcal{A}_1 \vee \mathcal{A}_2; \mathcal{B}) &= I(\mathcal{A}_1; \mathcal{B}) + I(\mathcal{A}_2; \mathcal{B}|\mathcal{A}_1) \\ I(\mathcal{A}_1 \vee \mathcal{A}_2; \mathcal{B}|\mathcal{C}) &= I(\mathcal{A}_1; \mathcal{B}|\mathcal{C}) + I(\mathcal{A}_2; \mathcal{B}|\mathcal{A}_1 \vee \mathcal{C}) \end{aligned}$$

(v) *(Conditional) entropy and (conditional) mutual information:*

$$\begin{aligned} H(\mathcal{A}) &= I(\mathcal{A}; \mathcal{B}) + H(\mathcal{A}|\mathcal{B}) \\ H(\mathcal{A}|\mathcal{C}) &= I(\mathcal{A}; \mathcal{B}|\mathcal{C}) + H(\mathcal{A}|\mathcal{B} \vee \mathcal{C}) \end{aligned}$$

In particular,

$$\begin{aligned} H(\mathcal{A}|\mathcal{B}) &\le H(\mathcal{A}), \qquad & I(\mathcal{A}; \mathcal{B}) &\le H(\mathcal{A}), \\ H(\mathcal{A}|\mathcal{B} \vee \mathcal{C}) &\le H(\mathcal{A}|\mathcal{C}), \qquad & I(\mathcal{A}; \mathcal{B}|\mathcal{C}) &\le H(\mathcal{A}|\mathcal{C}). \end{aligned}$$

(vi) *Independence:* If $\mathcal{A}_1$ and $\mathcal{A}_2$ are independent, then the inequality

$$I(\mathcal{A}_1 \vee \mathcal{A}_2; \mathcal{B}_1 \vee \mathcal{B}_2) \geq I(\mathcal{A}_1; \mathcal{B}_1) + I(\mathcal{A}_2; \mathcal{B}_2).$$

holds. Equality holds if $\mathcal{A}_1 \vee \mathcal{B}_1$ and $\mathcal{A}_2 \vee \mathcal{B}_2$ are independent.

(vii) *Atomic σ-algebras:* Suppose $\mathcal{A}$ and $\mathcal{B}$ are completely atomic σ-algebras (see Paragraph A.5) with finitely many or countably many atoms. Then

$$I(\mathcal{A}; \mathcal{B}) = \sum_{i,j} \mathrm{P}(A_i \cap B_j) \log \frac{\mathrm{P}(A_i \cap B_j)}{\mathrm{P}(A_i)\mathrm{P}(B_j)},$$

$$I(\mathcal{A}; \mathcal{B}\,|\,\mathcal{C}) = \sum_{i,j} \mathrm{E}\left[\mathrm{P}(A_i \cap B_j\,|\,\mathcal{C}) \log \frac{\mathrm{P}(A_i \cap B_j\,|\,\mathcal{C})}{\mathrm{P}(A_i\,|\,\mathcal{C})\mathrm{P}(B_j\,|\,\mathcal{C})}\right],$$

where $A_1, A_2, \ldots$ and $B_1, B_2, \ldots$ are the atoms of $\mathcal{A}$ and $\mathcal{B}$, respectively. In particular,

$$H(\mathcal{A}) = -\sum_i \mathrm{P}(A_i) \log \mathrm{P}(A_i),$$

$$H(\mathcal{A}\,|\,\mathcal{C}) = -\sum_i \mathrm{E}\big[\mathrm{P}(A_i\,|\,\mathcal{C}) \log \mathrm{P}(A_i\,|\,\mathcal{C})\big].$$

If $\mathcal{A}$ has finitely many, say m, atoms, then

$$H(\mathcal{A}) \leq \log m$$

with equality if and only if all atoms have the same probability $1/m$.

(viii) If $\mathcal{A}$ fails to be completely atomic, then

$$H(\mathcal{A}) = \infty.$$

(4.8) Remark. The list in Paragraph 4.7 is a selection of well-known properties required later. From the given σ-algebra based formulations the versions for random variables are immediately obtained, e. g., if ξ_1, ξ_2, η are random variables, then the first equation in (4.7.iv) has the form

$$I(\xi_1, \xi_2; \eta) = I(\xi_1; \eta) + I(\xi_2; \eta\,|\,\xi_1).$$

Clearly, nonnegativity and monotonicity also hold for entropy and conditional entropy as a special case. The monotonicity and the symmetry directly follow from the definitions. For a proof of the relations in the first line of (4.7.i) see (Bradley, 2007, Th. 5.3). The relations in the second line and the chain rules in (4.7.iv) follow from the proofs in Wyner (1978) basically by replacing there the σ-algebras generated by random variables with general σ-algebras. For a definition of a Markov chain see Paragraph A.2. Applying the chain rules to $I(\mathcal{A}; \mathcal{A} \vee \mathcal{B})$ and $I(\mathcal{A}; \mathcal{A} \vee \mathcal{B}\,|\,\mathcal{C})$ and using the fact that $\mathcal{A}$, $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{A}$, $\mathcal{A} \vee \mathcal{C}$, $\mathcal{B}$ both form a Markov chain yields the equalities in (4.7.v). These equalities and the nonnegativity of the information measures imply the inequalities in (4.7.v). The results in (4.7.vi) are also based on the chain rule and on (4.7.i) as shown, e. g., in (Ihara, 1993, Lem. 4.2.2).

The representation of mutual information for completely atomic σ-algebras in (4.7.vii) is derived in (Bradley, 2007, Rmk. 5.5.f). The corresponding result for the conditional mutual information is obtain similarly. See (Billingsley, 1965, Sec. 12) for the special case of conditional entropy.

The principle argument is, that the sums in the definitions of the information measures increase for refinements of partitions. A σ-algebra is completely atomic, e. g., if the probability measure defined on it is the (countable) convex combination of Dirac measures (see Paragraph A.4). Further special cases of completely atomic σ-algebras are finite σ-algebras and those generated by discrete random variables or finite partitions (see Paragraph A.5). For example assume that ξ and η are random variables with values in $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$, where $X = \{a_1, a_2, \ldots\}$ and $Y = \{b_1, b_2, \ldots\}$ are finite or countable sets and $\mathcal{X}$ and $\mathcal{Y}$ are the corresponding power sets. Then $\sigma(\xi)$ and $\sigma(\eta)$ are completely atomic σ-algebras and, e. g., the first identity in (4.7.vii) has the form

$$I(\xi;\eta) = \sum_{i,j} \mathrm{P}(\xi = a_i, \eta = b_j) \log \frac{\mathrm{P}(\xi = a_i, \eta = b_j)}{\mathrm{P}(\xi = a_i)\mathrm{P}(\eta = b_j)}.$$

The entropy of a completely atomic σ-algebra with m atoms is upper bounded by $\log m$ due to the concavity of the logarithm as derived, e. g., in (Gallager, 1968, Th. 2.3.1), where the case of equality is also derived. That non-completely atomic σ-algebras have infinite entropy as claimed in (4.7.viii) is shown in (Bradley, 2007, p. 170).

(4.9) Example (Conditional mutual information and Markov kernel). Let $(X, \mathcal{X})$, $(Y_1, \mathcal{Y}_1)$, and $(Y_2, \mathcal{Y}_2)$ be measurable spaces. Suppose μ is a probability measure on $\mathcal{X}$, κ is a Markov kernel from $(X, \mathcal{X})$ to $(Y_1 \times Y_2, \mathcal{Y}_1 \otimes \mathcal{Y}_2)$, and $\mu\kappa$ denotes the probability measure on $\mathcal{X} \otimes \mathcal{Y}_1 \otimes \mathcal{Y}_2$ induced by μ and κ as in Definition 2.1. Further assume that ξ, η_1, and η_2 denote the projections from $X \times Y_1 \times Y_2$ to X, Y_1, and Y_2, respectively.

For fixed $x \in X$ let $I(\eta_1(x); \eta_2(x) \,|\, x)$ denote the mutual information between $\eta_1(x)$ and $\eta_2(x)$, where $\eta_1(x) = \eta_1(x, \cdot, \cdot)$ and $\eta_2(x) = \eta_2(x, \cdot, \cdot)$ are considered as random variables on the probability space $(Y_1 \times Y_2, \mathcal{Y}_1 \otimes \mathcal{Y}_2, \kappa(x, \cdot))$. If ξ, η_1, and η_2 are considered as random variables on the probability space $(X \times Y_1 \times Y_2, \mathcal{X} \otimes \mathcal{Y}_1 \otimes \mathcal{Y}_2, \mu\kappa)$, then we have

$$I(\eta_1; \eta_2 \,|\, \xi) = \int_X I(\eta_1(x); \eta_2(x) \,|\, x) \,\mathrm{d}\mu(x)$$

for the mutual information between η_1 and η_2 given ξ. This result is derived similar to the example in (Pinsker, 1964, pp. 32-34).

Now, we introduce the commonly used and maybe most natural version of (mutual) information rate. Pinsker (1964, Sec. 5.4) gives a number of alternative definitions and derives in the case of stationarity conditions under which the various forms are equal (Pinsker, 1964, Secs. 7, 8). Some of these definitions have mathematical advantages, e. g., they allow to infer properties of information rates for general processes from those known for finite-alphabet processes. See also (Gray and Kieffer, 1980) and (Gray, 2011, Sec. 8.1, 8.2, 8.4) in this regard. However, the basic definition below is sufficient in the context of this thesis.

(4.10) Definition (Information rate). Let $\mathfrak{A} = \{\mathcal{A}_t, t \in T_+\}$ and $\mathfrak{B} = \{\mathcal{B}_t, t \in T_+\}$ be two families of sub-σ-algebras of $\mathcal{F}$. Then the (mutual) information rate $\bar{I}(\mathfrak{A}; \mathfrak{B})$ of $\mathfrak{A}$ and $\mathfrak{B}$ is defined by

$$\bar{I}(\mathfrak{A}; \mathfrak{B}) = \lim_{s \to \infty} \frac{1}{s} I\Bigg(\bigvee_{t \in (0,s]} \mathcal{A}_t; \bigvee_{t \in (0,s]} \mathcal{B}_t \Bigg)$$

if the limit exists. If $\xi = \{\xi_t, t \in T\}$ and $\eta = \{\eta_t, t \in T\}$ are two random processes, then the (mutual) information rate $\bar{I}(\xi;\eta)$ of ξ and η is defined by

$$\bar{I}(\xi;\eta) = \bar{I}\Big(\{\sigma(\xi_t), t \in T_+\}; \{\sigma(\eta_t), t \in T_+\}\Big).$$

(4.11) Remark. Clearly, we can also write

$$\bar{I}(\xi;\eta) = \lim_{s\to\infty} \frac{1}{s} I(\xi_0^s; \eta_0^s).$$

From (Pinsker, 1964, Th. 7.4.2) and the comments given there we obtain the following sufficient conditions for the existence of the information rate.

(4.12) Lemma. *Let $\xi = \{\xi_t, t \in T\}$ and $\eta = \{\eta_t, t \in T\}$ be random processes such that the pair process $\{(\xi_t, \eta_t), t \in T\}$ is stationary. In the continuous-time case assume that ξ and η are continuous in the sense of Pinsker (see Definition B.5). If*

$$I(\xi_-^0; \xi_t^+) < \infty \qquad \text{or} \qquad I(\xi_-^0; \xi_0^t) < \infty \tag{1}$$

holds for some $t \in T_+$, then the information rate $\bar{I}(\xi;\eta)$ exists.

The next lemma is an important basic result, which is used in the form of Corollary 4.14 to prove the central coding theorem (and converse) for abstract channels with time structure. The special case in Corollary 4.14 is proved in (Mittelbach, 2012, Lem. 2.25, 2.26). The generalized form in Lemma 4.13, that is based on σ-algebras and conditions for (conditional) mutual informations rather than distributions, is shown similarly. The proof is given in Paragraph E.1 of Appendix E for the sake of consistency and completeness of the presentation. Note that the existence of the information rate considered in Corollary 4.14 is already guaranteed by Lemma 4.12 because the second condition in (4.12.1) is satisfied. However, later we will make use of the monotonicity result given in the corollary. This monotonicity is used also by Kadota and Wyner (1972) to prove a coding theorem for continuous-time channels. However, their proof in (Kadota and Wyner, 1972, Appendix II) regarding the monotonicity is not correct as shown in (Mittelbach, 2012, Rmk. 2.27, Exp. 2.28). See also Paragraph 16.6 illustrating this issue.

(4.13) Lemma. *Let $\mathfrak{A} = \{\mathcal{A}_k, k \in \mathbb{N}\}$ be an independent family and $\mathfrak{B} = \{\mathcal{B}_k, k \in \mathbb{N}\}$ be an arbitrary family of sub-σ-algebras of $\mathcal{F}$.*

(i) If for all $n \in \mathbb{N}$ and $k = 2, 3, \ldots, n+1$

$$I\Big(\mathcal{A}_{k-1}; \bigvee_{l=1}^{n} \mathcal{B}_l \,\Big|\, \bigvee_{l=1}^{k-2} \mathcal{A}_l\Big) = I\Big(\mathcal{A}_k; \bigvee_{l=2}^{n+1} \mathcal{B}_l \,\Big|\, \bigvee_{l=2}^{k-1} \mathcal{A}_l\Big) \tag{1}$$

holds, then the sequence $\big\{n^{-1} I\big(\bigvee_{l=1}^{n} \mathcal{A}_l; \bigvee_{l=1}^{n} \mathcal{B}_l\big), n \in \mathbb{N}\big\}$ is monotonically increasing.

(ii) If for all $m, n \in \mathbb{N}$

$$I\Big(\bigvee_{l=1}^{n} \mathcal{A}_l; \bigvee_{l=1}^{n} \mathcal{B}_l\Big) = I\Big(\bigvee_{l=m+1}^{m+n} \mathcal{A}_l; \bigvee_{l=m+1}^{m+n} \mathcal{B}_l\Big) \tag{2}$$

holds, then the sequence $\big\{I\big(\bigvee_{l=1}^{n} \mathcal{A}_l; \bigvee_{l=1}^{n} \mathcal{B}_l\big), n \in \mathbb{N}\big\}$ is superadditive (see Paragraph D.1).

In (i) *as well as in* (ii) *the information rate* $\bar{I}(\mathfrak{A};\mathfrak{B})$ *exists and is given by*

$$\bar{I}(\mathfrak{A};\mathfrak{B}) = \sup_{n\geq 1} \frac{1}{n} I\Big(\bigvee_{l=1}^{n} \mathcal{A}_l; \bigvee_{l=1}^{n} \mathcal{B}_l \Big).$$

(4.14) Corollary (*Information rate for independent and stationary random sequences). Let* $\xi = \{\xi_k, k \in \mathbb{Z}\}$ *be a sequence of independent and* $\eta = \{\eta_k, k \in \mathbb{Z}\}$ *be a sequence of arbitrary random variables such that the pair sequence* $\{(\xi_k, \eta_k), k \in \mathbb{Z}\}$ *is stationary. Then the sequence* $\{I(\xi_0^n; \eta_0^n), n \in \mathbb{N}\}$ *is superadditive, the sequence* $\{n^{-1} I(\xi_0^n; \eta_0^n), n \in \mathbb{N}\}$ *is monotonically increasing, the information rate* $\bar{I}(\xi;\eta)$ *exists, and it is given by*

$$\bar{I}(\xi;\eta) = \sup_{n\geq 1} \frac{1}{n} I(\xi_0^n; \eta_0^n).$$

Proof. In Lemma 4.13 we put $\mathcal{A}_k = \sigma(\xi_k)$ and $\mathcal{B}_k = \sigma(\eta_k)$ for all $k \in \mathbb{N}$. Since $\{\xi_k, k \in \mathbb{N}\}$ is a sequence of independent random variables $\{\mathcal{A}_k, k \in \mathbb{N}\}$ is an independent family of sub-σ-algebras of $\mathcal{F}$. Condition (4.13.1) as well as (4.13.2) are satisfied due to the stationarity of the pair sequence $\{(\xi_k, \eta_k), k \in \mathbb{Z}\}$. □

§5 Information Rate Capacity

The information rate capacity is an important performance parameter of a channel with time structure, which represents, roughly speaking, the (asymptotic) solution of an extremum problem involving the mutual information (rate) between the channel input and output. Under suitable conditions it has an operational meaning as fundamental limit of reliable information transmission as shown with a coding theorem (and converse) in Section §9. In this section, κ is a channel with time structure as introduced in Definition 2.3 and $\mathcal{E} = \{E_s \subset X_0^s, s \in T_+\}$ is a family of input constraints, specified, e. g., by cost functions as in Example 3.2. First, we define the information rate capacity of κ for the constraints $\mathcal{E}$ and make some remarks on the benefits of the introduced version compared to what is proposed in the literature. Then we derive useful equivalent representations of the information rate capacity of stationary channels. More comments and results in this regard are given in Section §10 and in the discussion in Section §11.

(5.1) Definition (Information rate capacity). The information rate capacity of the channel κ for the constraints $\mathcal{E}$ is defined by

$$C = \limsup_{s\to\infty} \frac{1}{s} C_s \qquad \text{with} \qquad C_s = \sup_{\mu\in\mathcal{P}_s} I\big([\mathcal{X}_0^s]; [\mathcal{Y}_0^s]\big). \tag{1}$$

For any $s \in T_+$ we denote by $\mathcal{P}_s$ the set of all s-i.i.d. probability measures (see Definition B.1)

$$\mu = \bigotimes_{k\in\mathbb{Z}} \langle \mu_0 \rangle_{ks},$$

on the channel input σ-algebra $\mathcal{X}$, where the probability measure μ_0 on $\mathcal{X}_0^s$ has the form

$$\mu_0 = \sum_{i=1}^{m} p_i \delta_{a_i}. \tag{2}$$

The number m of summands in (2) is finite and δ_{a_i} denotes the Dirac measure (see Paragraph A.4)

on $\mathcal{X}_0^s$ for the signal $a_i \in E_s$. The probabilities p_i are positive[13] and satisfy $\sum_{i=1}^m p_i = 1$. The signals $a_1, a_2, \dots, a_m$ are assumed to be pairwise distinct[13].

(5.2) Remark. We have implicitly used the identities

$$X = \bigtimes_{t\in T} X_t = \bigtimes_{k\in\mathbb{Z}} X_{ks}^{(k+1)s} \qquad \text{and} \qquad \mathcal{X} = \bigotimes_{t\in T} \mathcal{X}_t = \bigotimes_{k\in\mathbb{Z}} \mathcal{X}_{ks}^{(k+1)s}.$$

Let ξ_t denote the projection from $X \times Y$ to X_t and η_t the projection from $X \times Y$ to Y_t for all $t \in T$. Then ξ_t and η_t are actually random variables on the channel input-output probability space $(X \times Y, \mathcal{X} \otimes \mathcal{Y}, \mu\kappa)$ and C_s can also be written as

$$C_s = \sup_{\mu\in\mathcal{P}_s} I\big(\xi_0^s; \eta_0^s\big).$$

To define the information rate capacity of a channel with time structure the time index set T is at first partitioned into segments of size $s \in T_+$. In the time period $(0, s]$ we have the input probability measure μ_0, for which the probability mass is concentrated on a finite number of input signals $a_1, a_2, \dots, a_m$ satisfying the constraint E_s. The probability measure $\langle\mu_0\rangle_{ks}$ is a shifted copy of μ_0 to the time period $(ks, (k+1)s]$ and the probability measure μ is the countable infinite product of all shifted versions of μ_0, i. e., an s-i.i.d. probability measure. The set $\mathcal{P}_s$ is composed of all s-i.i.d. probability measures μ generated by μ_0 of the previously specified form. The quantity C_s is the supremum of the mutual information between the channel input and output within the time period $(0, s]$, where the supremum is taken w. r. t. all input probability measures μ from $\mathcal{P}_s$. Finally, normalizing C_s by the time duration s and taking the limit superior w. r. t. s defines the information rate capacity C in [nat/channel use] if $T = \mathbb{Z}$ or in [nat/second] if $T = \mathbb{R}$. Note that optimizing w. r. t. s-i.i.d., i. e., s-stationary and s-memoryless input probability measures, is associated with the potential application as performance parameter for information transmission using block codes.

The definition of information rate capacity is based on the one of Kadota and Wyner (1972), however, with a significant difference regarding the structure of the probability measure μ_0. In Definition 5.1 only probability measures μ_0 with finite support on the constraint set E_s are considered. In the generalization of Kadota and Wyner's definition – they consider the special case of a continuous-time channel with real-valued input and output signals – to abstract channels with time structure (5.1.2) is replaced by a probability measure μ_0 for which the outer μ_0-measure of the constraint set E_s is equal to 1 (see Definition 10.1). The essential advantage of considering only finitely supported probability measures is that the proof of the central coding theorem in Section §9 is identical for the cases of finite and infinite information rate capacity. This allows a simplified proof and, more importantly, a weakening of the conditions on the channel properties. The limitations of Kadota and Wyner's version in this respect are discussed in Remark 10.4. The relations between the different versions of information rate capacity are analyzed in detail in Section §10 showing together with the theorems in Section §9 that Definition 5.1 provides the adequate form.

An additional benefit of the introduced definition is that it is more closely related to the (possible) operational meaning in connection with coded information transmission. Using this specific form is inspired by the work of Kemperman (1974). For the special case of a discrete-time memoryless channel (see Example 13.8) also Gallager (1968, p. 318/324), Wagner (1968) or

[13] This assumption is not necessarily required but convenient and means no loss of generality.

Schwarte (1996) considered a capacity definition based on finitely supported input measures. Further note that neither $E_s \in \mathcal{X}_0^s$ nor $\{a_k\} \in \mathcal{X}_0^s$ is required for the definition of μ_0 in (5.1.2). This is a convenient technical detail making the use of so-called standard extensions (see Paragraph A.12) needless, which is required following Kadota and Wyner's approach. A continuous-time channel with real-valued input signals and an amplitude or average power constraint (see Examples 2.4 and 3.2), is an example, where this is relevant because $E_s \notin \mathcal{X}_0^s$ and $\{a_k\} \notin \mathcal{X}_0^s$.

In the next lemma, representations of the information rate capacity are given, which are useful, e. g., to prove a coding theorem (or converse) for stationary channels or to show under suitable conditions the equality of the previously discussed versions of information rate capacity. The identity (5.3.1) below was shown in (Mittelbach, 2012, Lem. 5.9) for a continuous-time channel with real-valued input signals and an amplitude or average power constraint. A proof of the subsequent generalized form including the derivation of the second identity (5.3.2) is given in Paragraph E.2 of Appendix E.

(5.3) Lemma (*Information rate capacity of stationary channels). Consider the information rate capacity $C = \limsup_{s\to\infty} C_s/s$ introduced in Definition 5.1 and assume that the channel κ is stationary and the family $\mathcal{E}$ of input constraints satisfies the regularity condition (3.1.4). Then we have the identities*

$$C = \sup_{s\in T_+} \frac{1}{s} C_s, \tag{1}$$

$$C = \sup_{\mu\in\mathcal{P}} \bar{I}(\mu), \qquad \bar{I}(\mu) = \lim_{s\to\infty} \frac{1}{s} I\big([\mathcal{X}_0^s]; [\mathcal{Y}_0^s]\big), \tag{2}$$

where $\mathcal{P} = \bigcup_{s\in T_+} \mathcal{P}_s$ with $\mathcal{P}_s$ as defined in Definition 5.1.

(5.4) Remark. If $\xi = \{\xi_t, t \in T\}$ and $\eta = \{\eta_t, t \in T\}$ are the families of projections ξ_t and η_t introduced at the beginning of Remark 5.2, then with Remark 4.11 we can write

$$\bar{I}(\mu) = \lim_{s\to\infty} \frac{1}{s} I(\xi_0^s; \eta_0^s) = \bar{I}(\xi; \eta),$$

i. e., according to (5.3.2) the information rate capacity can be represented as supremum of information rates. This is actually the real justification, at least under the conditions of Lemma 5.3, of using the term information rate capacity adopted from (Gray, 2011, Sec. 14.4). Since we only consider situations, where this conditions are satisfied we use the name in general. Note that the information rate $\bar{I}(\xi; \eta)$, i. e., the limit, indeed exists for all $\mu \in \mathcal{P}$ as shown in the proof of Lemma 5.3 in Paragraph E.2. A result in the direction of (5.3.2) is given, e. g., in (Gray, 2011, Lem. 14.5). However, only for the very special case of a causal stationary discrete-time channel with no input memory and finite alphabets.

Identity (5.3.2) relates two basic approaches to define information capacities for channels with time structure. The quantity in Definition 5.1 is the limit (superior) of suprema of mutual informations for finite duration. In contrast, (5.3.2) characterizes the capacity as supremum of information rates, i. e., of limits. The former type is also considered by Gallager (1968, p. 370) or Wolfowitz (1978, p. 56). It is the more natural generalization of Shannon's original definition of information capacity, is more amenable to numerical calculations, and indicates more clearly the

relationship to its operational meaning for block-coded information transmission, which is what we are interested in. The latter form can be advantageous for mathematical reasons and is called process definition by Gray (2011, p. 369). For the information-theoretic analysis of channels with memory it has a long history from (Khinchin, 1957, p. 91) or (Feinstein, 1958, p. 87) to (Kakihara, 1999, p. 167) or (Gray, 2011, p. 367). See also (Gray and Ornstein, 1979, p. 302) for further classical references and (Wolfowitz, 1978, p. 60/61) for remarks on the different approaches.

§6 f-Divergence

As overviewed in (Gibbs and Su, 2002) there are many ways to quantify the difference between two (probability) measures. We are interested in the so-called f-divergence introduced by Csiszár (1963), which specifies a whole class of such quantities. After defining f-divergence we collect some material including alternative representations and useful properties. We are particularly interested in relative entropy and total variation distance and properties of these special f-divergences. Based on divergence measures we define in Section §7 (and already defined in Section §4 with the mutual information) dependence measures, which, in turn, are used later to define different types of memory for random processes and channels with time structure. Furthermore, we can express the property of asymptotic input-memorylessness introduced in (2.7.iv) using the total variation distance, which is helpful in later derivations.

The material on the f-divergence is taken from (Csiszár, 1963, 1967). Further references are given below, if results are taken from somewhere else. Throughout this section, $(X, \mathcal{X})$ will be a measurable space and P and Q are probability measures on $\mathcal{X}$. Furthermore, f is a real-valued convex function[14] on $(0, \infty)$.

(6.1) Definition (f-divergence). The f-divergence of P and Q is defined as

$$D_f(\mathrm{P}\|\mathrm{Q}) = \sup \sum_{i=1}^{n} \mathrm{Q}(A_i) f\left(\frac{\mathrm{P}(A_i)}{\mathrm{Q}(A_i)}\right), \tag{1}$$

where the supremum is taken w. r. t. all partitions $\{A_1, A_2, \ldots, A_n\}$ of X with $A_i \in \mathcal{X}$.

Suppose ξ and η are random variables with values in $(X, \mathcal{X})$ such that P and Q are the distributions of ξ and η, respectively. Then the f-divergence of ξ and η is defined by

$$D_f(\xi\|\eta) = D_f(\mathrm{P}\|\mathrm{Q}).$$

In the next theorem an integral form of the f-divergence is given, which is useful in calculations. Often this representation is taken as definition, e. g., in (Csiszár, 1963, § 1).

(6.2) Theorem (*Integral form of f-divergence). Assume that λ is a σ-finite[15] measure w. r. t. which* P *and* Q *are absolutely continuous (see Paragraph A.7). If p and q denote the corresponding λ-densities, then the f-divergence of* P *and* Q *is given by*

$$D_f(\mathrm{P}\|\mathrm{Q}) = \int_X q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}\lambda(x). \tag{1}$$

[14] The following conventions are used to avoid meaningless expressions: $f(0) = \lim_{u \to +0} f(u)$, $0 \cdot f\left(\frac{0}{0}\right) = 0$, and $0 \cdot f\left(\frac{a}{0}\right) = \lim_{\epsilon \to +0} \epsilon f\left(\frac{a}{\epsilon}\right) = a \lim_{u \to \infty} \frac{f(u)}{u}$ for $0 < a < \infty$.

[15] If the space X is the countable union of $\mathcal{X}$-measurable sets with finite λ-measure, then λ is called σ-finite.

(6.3) Remark. A measure λ with the required properties always exists, e. g., $\lambda = \mathrm{P} + \mathrm{Q}$ is a possible choice. The λ-densities can be chosen nonnegative and finite everywhere. Due to the convexity of f, the integral is always meaningful. Its value does not depend on the particular choice of λ.

If P is absolutely continuous w. r. t. Q, then (6.2.1) simplifies because we can choose $\lambda = \mathrm{Q}$ so that $q(x) = 1$ for all $x \in X$. If $\lim_{u\to\infty} f(u)/u = \infty$, then we have

$$D_f(\mathrm{P}\|\mathrm{Q}) = \begin{cases} \displaystyle\int_X f\big(p(x)\big)\,\mathrm{dQ}(x) & \text{if} \quad \mathrm{P} \ll \mathrm{Q} \\ \infty & \text{otherwise} \end{cases}, \tag{1}$$

where p denotes the Q-density of P, given P is absolutely continuous w. r. t. Q, which is denoted by $\mathrm{P} \ll \mathrm{Q}$.

(6.4) Properties of f-divergence.

(i) *Lower bound:*

$$D_f(\mathrm{P}\|\mathrm{Q}) \geq f(1)$$

If f is strictly convex, then equality holds if and only if $\mathrm{P} = \mathrm{Q}$.

(ii) *Data processing inequality:* Let $(Y, \mathcal{Y})$ be a measurable space and assume that K is a Markov kernel from $(X, \mathcal{X})$ to $(Y, \mathcal{Y})$. Let $\bar{\mathrm{P}}$ and $\bar{\mathrm{Q}}$ denote the probability measures on $\mathcal{Y}$ given by

$$\bar{\mathrm{P}}(B) = \int_X K(x, B)\,\mathrm{dP}(x), \qquad \bar{\mathrm{Q}}(B) = \int_X K(x, B)\,\mathrm{dQ}(x)$$

for all $B \in \mathcal{Y}$. Then we have the inequality

$$D_f(\mathrm{P}\|\mathrm{Q}) \geq D_f(\bar{\mathrm{P}}\|\bar{\mathrm{Q}}).$$

In particular, if g is an $\mathcal{X}/\mathcal{Y}$-measurable function on X with values in Y, then

$$D_f(\mathrm{P}\|\mathrm{Q}) \geq D_f(\mathrm{P}_g\|\mathrm{Q}_g)$$

holds, where P_g and Q_g denote the image measures of P and Q w. r. t. g. If f is strictly convex, then equality holds if and only if g is a sufficient statistic.

(6.5) Remark. Note that in general the f-divergence is not a metric because it is not symmetric. However, if $\hat{f}$ is the real convex function given by

$$\hat{f}(u) = uf\left(\frac{1}{u}\right)$$

for all $u \in (0, \infty)$, then we have

$$D_{\hat{f}}(\mathrm{Q}\|\mathrm{P}) = D_f(\mathrm{P}\|\mathrm{Q}).$$

Also, in general the f-divergence does not satisfy the triangle inequality.

In statistics, a number of special f-divergences are widely used. In (Csiszár, 1967, § 1), (Csiszár and Shields, 2004, p. 447), or (Gilardoni, 2010, Sec. I.C) several popular examples are given. Subsequently, we consider the total variation distance and the relative entropy.

(6.6) Definition (Total variation distance). If the function f is given by

$$f(u) = |u - 1|$$

for all $u \in (0, \infty)$, then the f-divergence is called total variation distance and we write

$$D_f(\mathrm{P}\|\mathrm{Q}) = \|\mathrm{P} - \mathrm{Q}\|_{\mathrm{tv}}.$$

(6.7) Remark. Let $\{A_1, A_2, \ldots, A_n\}$ be a partition of X with $A_i \in \mathcal{X}$. Without loss of generality we assume that for some $m \in \{0, 1, \ldots, n\}$ the inequality $\mathrm{P}(A_i) - \mathrm{Q}(A_i) \geq 0$ holds for all $1 \leq i \leq m$ and the inequality $\mathrm{P}(A_i) - \mathrm{Q}(A_i) \leq 0$ holds for all $m < i \leq n$. Since the A_i's form a partition, we have

$$\begin{aligned}\sum_{i=1}^{n} |\mathrm{P}(A_i) - \mathrm{Q}(A_i)| &= \big(\mathrm{P}(G) - \mathrm{Q}(G)\big) + \big(\mathrm{Q}(G^{\mathrm{c}}) - \mathrm{P}(G^{\mathrm{c}})\big) \\ &= 2\,|\mathrm{P}(G) - \mathrm{Q}(G)|,\end{aligned}$$

where $G = \bigcup_{i=1}^{m} A_i$. Thus, Definition 6.1 simplifies in the case of the total variation distance to

$$\|\mathrm{P} - \mathrm{Q}\|_{\mathrm{tv}} = 2 \sup_{A \in \mathcal{X}} |\mathrm{P}(A) - \mathrm{Q}(A)|, \tag{1}$$

which is the form commonly used as definition. Note that in the literature the notation is not consistent. The quantity is also called variation (Pinsker, 1964, p. 6), variation(al) distance (Csiszár, 1967, p. 301), or total variation (Kemperman, 1969, p. 2174). Furthermore, the factor 2 appearing on the right-hand side of (1) is often omitted.

It is easily verified that the total variation distance is a metric, which is bounded by

$$0 \leq \|\mathrm{P} - \mathrm{Q}\|_{\mathrm{tv}} \leq 2. \tag{2}$$

From (1) we see that $\|\mathrm{P} - \mathrm{Q}\|_{\mathrm{tv}} = 0$ holds if and only if $\mathrm{P} = \mathrm{Q}$.

(6.8) Definition (Relative entropy). If the function f is given by

$$f(u) = u \log u$$

for all $u \in (0, \infty)$, then the f-divergence is called relative entropy and we write

$$D_f(\mathrm{P}\|\mathrm{Q}) = D(\mathrm{P}\|\mathrm{Q}).$$

(6.9) Remark. Since the relative entropy was introduced by Kullback and Leibler (1951) it is also called Kullback-Leibler divergence. There are many other names used in the literature, including generalized entropy (Pinsker, 1964, p. 19) and I-divergence (Csiszár, 1967, p. 301).

Due to (6.4.i) the relative entropy is nonnegative and because f is strictly convex it is 0 if and only if the two probability measures are identical.

The properties in Paragraph 6.4 are valid for any f-divergence. In addition, we want to list some specific properties of the relative entropy and the total variation distance useful later on.

(6.10) Properties of relative entropy and total variation distance.

(i) *Pinsker's inequality:*

$$\|\mathrm{P}-\mathrm{Q}\|_{\mathrm{tv}} \le \sqrt{2D(\mathrm{P}\|\mathrm{Q})}$$

(ii) *Relative entropy for product measures:* Suppose $(X_1, \mathcal{X}_1)$ is a measurable space and P_1 and Q_1 are probability measures on $\mathcal{X}_1$. Further, suppose $(X_2, \mathcal{X}_2)$ is a measurable space and P_2 and Q_2 are probability measures on $\mathcal{X}_2$. Then we have

$$D(\mathrm{P}_1 \otimes \mathrm{P}_2 \| \mathrm{Q}_1 \otimes \mathrm{Q}_2) = D(\mathrm{P}_1\|\mathrm{Q}_1) + D(\mathrm{P}_2\|\mathrm{Q}_2).$$

(iii) *Relative entropy between random sequences:* Let $\xi = \{\xi_k, k \in \mathbb{N}\}$ and $\eta = \{\eta_k, k \in \mathbb{N}\}$ be two sequences of random variables, where ξ_k and η_k have values in the same measurable space $(X_k, \mathcal{X}_k)$. Then we have

$$D(\xi\|\eta) = \lim_{n\to\infty} D(\xi_0^n \| \eta_0^n).$$

(iv) *Singularity of Gaussian processes:* If $\xi = \{\xi_t, t \in T\}$ and $\eta = \{\eta_t, t \in T\}$ are two Gaussian processes (either discrete- or continuous-time) and

$$D(\xi\|\eta) + D(\eta\|\xi) = \infty,$$

then

$$\|\mathrm{P}-\mathrm{Q}\|_{\mathrm{tv}} = 2,$$

where P and Q denote the distribution of ξ and η.

(v) *Total variation distance and integration:* For any real-valued $\mathcal{X}$-measurable function g on X satisfying $|g| \le \mathrm{c}$ for some positive constant c we have

$$\left| \int_X g \,\mathrm{dP} - \int_X g \,\mathrm{dQ} \right| \le \mathrm{c}\|\mathrm{P}-\mathrm{Q}\|_{\mathrm{tv}}.$$

(6.11) Remark. The inequality in (6.10.i) goes back to Pinsker (1964, Sec. 2.3). A proof of the given form can be found in (Csiszár, 1967, Th. 4.1) or (Kemperman, 1969, Sec. 6.2). The inequality provides a useful upper bound of the total variation distance, in particular, because often it is difficult to calculate the total variation distance exactly.

To prove the factorization result in (6.10.ii) we observe that the product measure $\mathrm{P}_1 \otimes \mathrm{P}_2$ is absolutely continuous w. r. t. the product measure $\mathrm{Q}_1 \otimes \mathrm{Q}_2$ if and only if P_1 is absolutely continuous w. r. t. Q_1 and P_2 is absolutely continuous w. r. t. Q_2. Then $\mathrm{P}_1 \otimes \mathrm{P}_2$ has $\mathrm{Q}_1 \otimes \mathrm{Q}_2$-density $p_1 p_2$, where p_1 is the Q_1-density of P_1 and p_2 is the Q_2-density of P_2. With (6.3.1) and the properties of the logarithm we obtain the assertion.

Regarding (6.10.iii) the defining relation in (6.1.1) or the data processing inequality in (6.4.ii) (see (Csiszár, 1963, Folg. 4)) yield the monotonicity

$$D(\xi_0^n \|\eta_0^n) \le D(\xi_0^{n+1} \|\eta_0^{n+1}).$$

Therefore, the limit (possibly infinite) exists and we have

$$\lim_{n\to\infty} D(\xi_0^n \| \eta_0^n) \le D(\xi \| \eta).$$

The reversed inequality can be shown with Dobrushin's theorem (Pinsker, 1964, Th. 2.4.1) similar to the proof given in (Pinsker, 1964, p. 12).

The result in (6.10.iv) is quite interesting because Pinsker's inequality usually forces the direction of an implication involving relative entropy and total variation distance. It is a direct consequence of a theorem of Hájek (1958)[16] and Feldman (1958)[17] and the characterization of total variation distance given in (6.7.1). A proof is also given in (Ibragimov and Rozanov, 1978, Ch. III, Th. 1) or (Hida and Hitsuda, 2007, Th. 6.1).

The inequality in (6.10.v) follows from the characterization of the total variation distance given in (Gibbs and Su, 2002, p. 7).

(6.12) Example (Relative entropy of Gaussian random vectors and sequences). Suppose $\xi_0^n = (\xi_1, \xi_2, \dots, \xi_n)$ is a real, n-dimensional Gaussian random vector with invertible covariance matrix Σ_ξ and expectation vector m_ξ. Further, let $\eta_0^n = (\eta_1, \eta_2, \dots, \eta_n)$ be a another real, n-dimensional Gaussian random vector with invertible covariance matrix Σ_η and expectation vector m_η. Then the relative entropy of ξ_0^n and η_0^n is given by

$$D(\xi_0^n \| \eta_0^n) = \frac{1}{2}\Big(\log\det\Big(\Sigma_\eta \Sigma_\xi^{-1}\Big) + \operatorname{tr}\Big(\Sigma_\xi \Sigma_\eta^{-1}\Big) - n\Big) + \frac{1}{2}\operatorname{tr}\Big(\Sigma_\eta^{-1} d d'\Big),$$

where d' denotes the transpose of the vector $d = m_\xi - m_\eta$. In particular, we have

$$D(\xi_k \| \eta_k) = \frac{1}{2}\left(\log\left(\frac{\operatorname{var}(\eta_k)}{\operatorname{var}(\xi_k)}\right) + \frac{\operatorname{var}(\xi_k)}{\operatorname{var}(\eta_k)} - 1\right) + \frac{1}{2\operatorname{var}(\eta_k)}\big(\mathrm{E}(\xi_k) - \mathrm{E}(\eta_k)\big)^2. \tag{1}$$

This result is taken from (Kullback, 1968, p. 189).

To discuss some specific examples let us define the matrices

$$A = \operatorname{diag}(a_1, a_2, \dots, a_n), \qquad B = \operatorname{diag}(b_1, b_2, \dots, b_n), \qquad S(\rho) = \sigma^2\Big(\rho^{|i-j|}\Big)_{i,j=1}^n,$$

where a_k, b_k, and σ are positive constants and ρ is a real constant satisfying $|\rho| < 1$. The matrix $S(\rho)$ is a scaled Kac-Murdock-Szegö matrix (see Paragraph D.2) and therefore a symmetric Toeplitz matrix.

(i) *Equal covariance.* If $\Sigma_\xi = \Sigma_\eta$, then we have

$$D(\xi_0^n \| \eta_0^n) = \frac{1}{2}\operatorname{tr}\Big(\Sigma_\eta^{-1} d d'\Big) = \frac{1}{2} d' \Sigma_\eta^{-1} d. \tag{2}$$

For the special case $\Sigma_\xi = \Sigma_\eta = A$ the components of the Gaussian vectors ξ_0^n and η_0^n are independent and we obtain from (6.10.ii) and (1)

$$D(\xi_0^n \| \eta_0^n) = \sum_{k=1}^n D(\xi_k \| \eta_k) = \frac{1}{2}\sum_{k=1}^n \frac{d_k^2}{a_k}, \tag{3}$$

[16] Russian original, see (Hájek, 1961) for English translation.

[17] See (Feldman, 1959) for corrections.

where d_k is the kth component of the vector d. Of course, $D(\xi_0^n\|\eta_0^n)$ can also be calculated directly from (2).

For the special case $\Sigma_\xi = \Sigma_\eta = S(\rho)$ we obtain

$$D(\xi_0^n\|\eta_0^n) = \frac{1}{2(1-\rho^2)\sigma^2}\left(\sum_{k=1}^{n} d_k^2 + \rho^2\sum_{k=2}^{n-1} d_k^2 - 2\rho\sum_{k=1}^{n-1} d_k d_{k+1}\right) \tag{4}$$

by using (2) and (D.2.1). We can upper bound (4) by

$$D(\xi_0^n\|\eta_0^n) \le \frac{2}{(1-\rho^2)\sigma^2}\sum_{k=1}^{n} d_k^2 \tag{5}$$

using $|\rho| < 1$ and the Cauchy-Schwarz inequality for real vectors. For $\rho = 0$ the covariance matrices are equal to the identity matrix and in (4) only the first sum remains.

(ii) *Equal expectation.* If $m_\xi = m_\eta$, then we have

$$D(\xi_0^n\|\eta_0^n) = \frac{1}{2}\Big(\log\det\Big(\Sigma_\eta\Sigma_\xi^{-1}\Big) + \mathrm{tr}\Big(\Sigma_\xi\Sigma_\eta^{-1}\Big) - n\Big). \tag{6}$$

Suppose the covariance matrices are given by $\Sigma_\xi = A$ and $\Sigma_\eta = B$. Then either from (6) or from (6.10.ii) together with (1) we obtain

$$D(\xi_0^n\|\eta_0^n) = \sum_{k=1}^{n} D(\xi_k\|\eta_k) = \frac{1}{2}\sum_{k=1}^{n}\left(\log\left(\frac{b_k}{a_k}\right) + \frac{a_k}{b_k} - 1\right). \tag{7}$$

Alternatively, assume that $\Sigma_\xi = S(\rho_\xi)$ and $\Sigma_\eta = S(\rho_\eta)$, where $|\rho_\xi| < 1$ and $|\rho_\eta| < 1$. Applying (D.2.1) and (6) yields

$$D(\xi_0^n\|\eta_0^n) = (n-1)\left(\frac{1}{2}\log\left(\frac{1-\rho_\eta^2}{1-\rho_\xi^2}\right) + \frac{\rho_\eta^2 - \rho_\xi\rho_\eta}{1-\rho_\eta^2}\right), \tag{8}$$

where we have also used (D.2.2).

(iii) *Random sequences.* Assume that $\xi = \{\xi_k, k \in \mathbb{N}\}$ and $\eta = \{\eta_k, k \in \mathbb{N}\}$ are second order Gaussian random sequences and as before let us put $d_k = \mathrm{E}(\xi_k) - \mathrm{E}(\eta_k)$. According to (6.10.iii) the relative entropy of ξ and η is given by the limit

$$D(\xi\|\eta) = \lim_{n\to\infty} D(\xi_0^n\|\eta_0^n). \tag{9}$$

We consider some examples related to those of part (i) and (ii). First, suppose ξ and η are sequences of independent Gaussian random variables. If

$$\mathrm{var}(\xi_k) = \mathrm{var}(\eta_k) = a_k > 0,$$

then $D(\xi_0^n\|\eta_0^n)$ in (9) is given by (3). If $d_k = 0$ and

$$\mathrm{var}(\xi_k) = a_k > 0 \qquad \text{and} \qquad \mathrm{var}(\eta_k) = b_k > 0,$$

then $D(\xi_0^n \| \eta_0^n)$ in (9) is given by (7). Now suppose the covariances are given by

$$\operatorname{cov}(\xi_i, \xi_j) = \operatorname{cov}(\eta_i, \eta_j) = \sigma^2 \rho^{|i-j|}, \tag{10}$$

where $\sigma^2 > 0$ and $|\rho| < 1$. Then $D(\xi_0^n \| \eta_0^n)$ in (9) is equal to (4). If $d_k = 0$ and

$$\operatorname{cov}(\xi_i, \xi_j) = \sigma^2 \rho_\xi^{|i-j|} \qquad \text{and} \qquad \operatorname{cov}(\eta_i, \eta_j) = \sigma^2 \rho_\eta^{|i-j|},$$

where $\sigma^2 > 0$, $|\rho_\xi| < 1$ and $|\rho_\eta| < 1$, then $D(\xi_0^n \| \eta_0^n)$ in (9) is equal to (8).

Depending on the moments the limit in (9) is either some finite value or is infinite for the first three examples. However, in the last example we take the limit of (8) as $n \to \infty$, i. e., there are only two possibilities:

$$D(\xi \| \eta) = \begin{cases} 0 & \text{if} \quad \rho_\xi = \rho_\eta \\ \infty & \text{if} \quad \rho_\xi \neq \rho_\eta \end{cases}.$$

Together with (6.10.iv) we further conclude that for $\rho_\xi \neq \rho_\eta$ the total variation distance between the distribution P of ξ and the distribution Q of η takes the maximum possible value, i. e.,

$$\|\mathrm{P} - \mathrm{Q}\|_{\mathrm{tv}} = 2.$$

Consider the random sequences $\hat{\xi} = \{\hat{\xi}_k, k \in \mathbb{N}\}$ and $\hat{\eta} = \{\hat{\eta}_k, k \in \mathbb{N}\}$ with $\hat{\xi}_k = c_k \xi_k$ and $\hat{\eta}_k = c_k \eta_k$, where the c_k's are nonzero constants. From (6.4.ii) we have

$$D(\hat{\xi} \| \hat{\eta}) = D(\xi \| \eta),$$

because the random sequences $\hat{\xi}$ and $\hat{\eta}$ are obtained from the random sequences ξ and η by the same bijective transformation. For the previously discussed example it follows that the total variation distance has still the maximum value, even if the processes are multiplied by rapidly decaying (but nonzero) constants. This illustrates the sensitivity of the total variation distance as a metric for distributions of random processes.

§7 Dependence Measures

To quantify "how dependent" two random variables are, it is natural to measure somehow the difference between the joint distribution and the product of the marginal distributions. As a measure of difference we can use, for example, the f-divergence introduced in the previous section. A specific measure of this type is the mutual information considered in detail in Section §4. There are many more dependence coefficients used in the literature that are not based on the f-divergence, e. g., the α- or ψ-dependence coefficient. The book of Bradley (2007) is a comprehensive reference in this regard.

In this section we define the α-, β- and ψ-dependence coefficient and collect some relevant properties. We further introduce the ψ-variation as a generalization of the ψ-coefficient. Based on the material of this section we define and analyze memory conditions for random processes and channels with time structure in Sections §12 and §13.

(7.1) f-divergence as a measure of dependence. Suppose ξ and η are random variables on the same probability space $(\Omega, \mathcal{F}, \mu)$ with values in the measurable spaces $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$, respectively. Then, the mutual information between ξ and η is given by

$$I(\xi;\eta) = D\big(\mu_{\xi,\eta} \| \mu_\xi \otimes \mu_\eta\big), \tag{1}$$

which can be seen from the representation (4.4.1) and the integral form of relative entropy given in (6.3.1). Thus, mutual information measures the difference between the joint distribution $\mu_{\xi,\eta}$ and the product $\mu_\xi \otimes \mu_\eta$ of the marginal distributions in terms of relative entropy. It is therefore a measure[18] of dependence, which is zero if and only if the random variables are independent.

Note that there is another representation of mutual information in terms of relative entropy, which is, for example, useful to solve optimization problems related to the computation of the information capacity of a channel. Based on the identity (see (Csiszár, 1978, Sec. 4))

$$D\big(\mu_{\xi,\eta} \| \mu_\xi \otimes \nu\big) = I(\xi;\eta) + D(\mu_\eta \| \nu)$$

and Remark 6.9 we have

$$I(\xi;\eta) = \min_\nu D\big(\mu_{\xi,\eta} \| \mu_\xi \otimes \nu\big),$$

where the minimum is taken w. r. t. all probability measures ν on $\mathcal{Y}$. See (Huang and Meyn, 2005) for an application in the context of calculating information capacities.

Any other f-divergence, which is zero if and only if the probability measures are equal, is suitable to define a dependence coefficient. For example, the well-known β-dependence coefficient is based on the total variation distance. It is given by

$$\beta(\xi;\eta) = \frac{1}{2}\big\|\mu_{\xi,\eta} - \mu_\xi \otimes \mu_\eta\big\|_{\mathrm{tv}}. \tag{2}$$

See (Bradley, 2007, Def. 3.3, Cor. 3.30) for the form given here.

Before we define further non-f-divergence-based dependence measures we introduce the ψ-variation. It is a more general quantity than the ψ-dependence coefficient, just as relative entropy is a generalization of mutual information. The term ψ-variation is coined by the author. The author is not aware of any reference, where the generalized ψ-variation is considered.

(7.2) Definition (ψ-variation). Let $(X, \mathcal{X})$ be a measurable space and suppose P and Q are probability measures on $\mathcal{X}$. The ψ-variation of P and Q is defined by

$$\psi(\mathrm{P}\|\mathrm{Q}) = \sup\left|\frac{\mathrm{P}(A)}{\mathrm{Q}(A)} - 1\right|,$$

where the supremum is taken w. r. t. all $A \in \mathcal{X}$ with $\mathrm{Q}(A) > 0$.

Suppose ξ and η are random variables with values in $(X, \mathcal{X})$ such that P and Q are the distributions of ξ and η, respectively. Then we call

$$\psi(\xi\|\eta) = \psi(\mathrm{P}\|\mathrm{Q})$$

the ψ-variation of ξ and η.

[18] Here, measure is not meant in a measure-theoretic sense.

(7.3) Properties of ψ-variation. Let $(X, \mathcal{X})$ be a measurable space and suppose P and Q are probability measures on $\mathcal{X}$.

(i) *Nonnegativity:*

$$\psi(\mathrm{P}\|\mathrm{Q}) \geq 0$$

Equality holds if and only if $\mathrm{P} = \mathrm{Q}$.

(ii) *Data processing inequality:* Let $(Y, \mathcal{Y})$, K, $\bar{\mathrm{P}}$, $\bar{\mathrm{Q}}$, g, P_g, and Q_g be given as in (6.4.ii) and assume that the Markov kernel K has the form of an integration channel (see Definition 15.1). Then the inequality

$$\psi(\mathrm{P}\|\mathrm{Q}) \geq \psi(\bar{\mathrm{P}}\|\bar{\mathrm{Q}})$$

holds. In particular, we have

$$\psi(\mathrm{P}\|\mathrm{Q}) \geq \psi(\mathrm{P}_g\|\mathrm{Q}_g).$$

(iii) *Probability measures on product spaces:* Assume that $(X, \mathcal{X})$ is a product space given by $(X, \mathcal{X}) = (X_1 \times X_2, \mathcal{X}_1 \otimes \mathcal{X}_2)$. Then we have

$$\psi(\mathrm{P}\|\mathrm{Q}) = \sup \left| \frac{\mathrm{P}(A_1 \times A_2)}{\mathrm{Q}(A_1 \times A_2)} - 1 \right|,$$

where the supremum is taken w. r. t. all $A_1 \in \mathcal{X}_1$ and $A_2 \in \mathcal{X}_2$ with $\mathrm{Q}(A_1 \times A_2) > 0$.

(7.4) Remark. Assertion (7.3.i) is evident. A proof of (7.3.ii) is given in Paragraph E.3 in Appendix E. We conjecture that the assertion is true for all Markov kernels, not only for those considered here.

Assertion (7.3.iii) means that the ψ-variation for measures on product spaces is already determined by all rectangles. We do not need to consider all sets from the product-σ-algebra to calculate the supremum. A proof is given in Paragraph E.4 in Appendix E. This result is the bridge to the well-known ψ-dependence coefficient.

(7.5) Definition (α-, β-, and ψ-dependence coefficient). Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and $\mathcal{A}$ and $\mathcal{B}$ be sub-σ-algebras of $\mathcal{F}$. Then we define

$$\alpha(\mathcal{A}; \mathcal{B}) = \sup \big| \mu(A \cap B) - \mu(A)\mu(B) \big|,$$

where the supremum is taken w. r. t. all $A \in \mathcal{A}$ and $B \in \mathcal{B}$, and

$$\psi(\mathcal{A}; \mathcal{B}) = \sup \left| \frac{\mu(A \cap B)}{\mu(A)\mu(B)} - 1 \right|,$$

where the supremum is taken w. r. t. all $A \in \mathcal{A}$ and $B \in \mathcal{B}$ satisfying the condition $\mu(A)\mu(B) > 0$. Further, we define

$$\beta(\mathcal{A}; \mathcal{B}) = \sup \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} \big| \mu(A_i \cap B_j) - \mu(A_i)\mu(B_j) \big|,$$

where the supremum is taken w. r. t. all partitions $\{A_1, A_2, \dots, A_m\}$ and $\{B_1, B_2, \dots, B_n\}$ of Ω with $A_i \in \mathcal{A}$ and $B_j \in \mathcal{B}$.

Suppose ξ and η are random variables defined on $(\Omega, \mathcal{F}, \mu)$. Then we call

$$\begin{aligned}\alpha(\xi;\eta) &= \alpha(\sigma(\xi);\sigma(\eta))\\ \beta(\xi;\eta) &= \beta(\sigma(\xi);\sigma(\eta))\\ \psi(\xi;\eta) &= \psi(\sigma(\xi);\sigma(\eta))\end{aligned}$$

the α-dependence coefficient, β-dependence coefficient, and ψ-dependence coefficient of ξ and η, respectively.

(7.6) Remark. The definitions are taken from (Bradley, 2007, Def. 3.3). Please note the different nature of the coefficients. The α-dependence measure is a supremum of differences, whereas the ψ-dependence measure is a supremum of ratios. For both coefficients pairs of sets are considered, whereas the β-dependence coefficient is defined w. r. t. pairs of partions.

Assume that the random variables ξ and η in Definition 7.5 have values in the spaces $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$, respectively. Then we have

$$\begin{aligned}\alpha(\xi;\eta) &= \sup\left|\mu_{\xi,\eta}(F\times G) - \mu_\xi \otimes \mu_\eta(F\times G)\right| \\ &= \sup\left|\mu_{\xi,\eta}(A\cap B) - \mu_{\xi,\eta}(A)\mu_{\xi,\eta}(B)\right| \\ &= \alpha\big([\mathcal{X}];[\mathcal{Y}]\big)\end{aligned} \tag{1}$$

where μ_ξ, μ_η, and $\mu_{\xi,\eta}$ denote the distribution of ξ, η, and (ξ,η). According to Definition 7.5 the quantity $\alpha(\xi;\eta)$ refers to the ground probability space $(\Omega, \mathcal{F}, \mu)$. The quantity $\alpha([\mathcal{X}];[\mathcal{Y}])$, however, refers to the probability space $(X\times Y, \mathcal{X}\otimes\mathcal{Y}, \mu_{\xi,\eta})$. The supremum in (1) is taken w. r. t. all $F \in \mathcal{X}$ and $G \in \mathcal{Y}$. This representation is obtained simply by moving from the ground space to the product space of values and by using the definition of a product measure. The second supremum is taken w. r. t. all $A \in [\mathcal{X}]$ and $B \in [\mathcal{Y}]$. The identity is a simple reformulation based on the definition of marginal measures.

Correspondingly, we have

$$\beta(\xi;\eta) = \frac{1}{2}\big\|\mu_{\xi,\eta} - \mu_\xi\otimes\mu_\eta\big\|_{\mathrm{tv}} \tag{2}$$

$$= \sup\left|\mu_{\xi,\eta}(F) - \mu_\xi\otimes\mu_\eta(F)\right| \tag{3}$$

$$= \beta\big([\mathcal{X}];[\mathcal{Y}]\big). \tag{4}$$

The first identity is already given in (7.1.2) and follows from (Bradley, 2007, Cor. 3.30). The supremum in (3) is taken w. r. t. all $F \in \mathcal{X}\otimes\mathcal{Y}$. This representation of the total variation distance is given in (6.7.1). The last equality is due to (Bradley, 2007, Th. 3.29).

Similarly, we have

$$\psi(\xi;\eta) = \psi\big([\mathcal{X}];[\mathcal{Y}]\big) \tag{5}$$

$$= \sup\left|\frac{\mu_{\xi,\eta}(F\times G)}{\mu_\xi\otimes\mu_\eta(F\times G)} - 1\right| \tag{6}$$

$$= \psi(\mu_{\xi,\eta}\|\mu_\xi\otimes\mu_\eta), \tag{7}$$

where the supremum (6) is taken w. r. t. all $F \in \mathcal{X}$ and $G \in \mathcal{Y}$ with $\mu_\xi \otimes \mu_\eta(F \times G) > 0$. The last identity is due to (7.3.iii) and demonstrates the connection between the ψ-variation and the ψ-dependence coefficient. Please note the difference between $\psi(\xi\|\eta)$ and $\psi(\xi;\eta)$. In the previous situation $\psi(\xi\|\eta)$ does only make sense if $(X,\mathcal{X}) = (Y,\mathcal{Y})$.

Note that in (6) it does not matter if the supremum is taken w. r. t. all sets of $\mathcal{X} \otimes \mathcal{Y}$ or only w. r. t. rectangles. However, the suprema in (1) and (3) are not equal in general.

The properties of the dependence coefficients given next are formulated for random variables. Of course, for (7.7.i)–(7.7.vi) corresponding results also hold for the σ-algebra based versions.

(7.7) Properties of α-, β-, and ψ-dependence coefficient. Let ξ, η, and ζ be random variables on the probability space $(\Omega, \mathcal{F}, \mu)$.

(i) *Nonnegativity:*

$$0 \le \alpha(\xi;\eta), \qquad 0 \le \beta(\xi;\eta), \qquad 0 \le \psi(\xi;\eta).$$

Equality holds if and only if ξ and η are independent.

(ii) *Monotonicity:* If $\xi = (\xi_1, \xi_2)$ is a random vector, then we have

$$\alpha(\xi_1;\eta) \le \alpha(\xi;\eta), \qquad \beta(\xi_1;\eta) \le \beta(\xi;\eta), \qquad \psi(\xi_1;\eta) \le \psi(\xi;\eta).$$

(iii) *Independence:* If $\xi = (\xi_1, \xi_2)$ and $\eta = (\eta_1, \eta_2)$ are random vectors and (ξ_1, η_1) and (ξ_2, η_2) are independent, then we have

$$\begin{aligned}
\alpha(\xi;\eta) &\le \alpha(\xi_1;\eta_1) + \alpha(\xi_2;\eta_2) \\
\beta(\xi;\eta) &\le 1 - \big(1 - \beta(\xi_1;\eta_1)\big)\big(1 - \beta(\xi_2;\eta_2)\big) \\
&\le \beta(\xi_1;\eta_1) + \beta(\xi_2;\eta_2) \\
\psi(\xi;\eta) &\le \big(1 + \psi(\xi_1;\eta_1)\big)\big(1 + \psi(\xi_2;\eta_2)\big) - 1.
\end{aligned}$$

(iv) *Markov chain:* If we have the Markov chain $(\xi - \eta - \zeta)$, then the following equalities hold.

$$\alpha(\xi;\eta,\zeta) = \alpha(\xi;\eta), \qquad \beta(\xi;\eta,\zeta) = \beta(\xi;\eta), \qquad \psi(\xi;\eta,\zeta) = \psi(\xi;\eta)$$

(v) *Inequalities for α-, β-dependence coefficient and mutual information:*

$$\alpha(\xi;\eta) \le \beta(\xi;\eta) \le \sqrt{\tfrac{1}{2} \sqrt{I(\xi;\eta)}}$$

(vi) *Relation between ψ-dependence coefficient and mutual information:*

$$I(\xi;\eta) \le \big(1 + \psi(\xi;\eta)\big) \log\big(1 + \psi(\xi;\eta)\big)$$

(vii) *ψ-dependence coefficient and Gaussian distribution:* If (ξ,η) is a 2-dimensional Gaussian random vector, then we have

$$\psi(\xi;\eta) < 2 \implies \operatorname{cor}(\xi,\eta) = 0.$$

(viii) *Covariance inequalities:* Let c be a nonnegative constant and $\mathcal{A}$ and $\mathcal{B}$ be sub-σ-algebras of $\mathcal{F}$. Assume that ξ is real-valued and $\mathcal{A}$-measurable satisfying $|\xi| \leq \mathrm{c}$. Further, assume that η is real-valued and $\mathcal{B}$-measurable satisfying $|\eta| \leq \mathrm{c}$. Then we have

$$\begin{aligned}\big|\mathrm{E}(\xi\eta) - \mathrm{E}(\xi)\mathrm{E}(\eta)\big| &\leq 4\mathrm{c}^2\alpha(\mathcal{A};\mathcal{B}),\\ \big|\mathrm{E}(\xi\eta) - \mathrm{E}(\xi)\mathrm{E}(\eta)\big| &\leq \mathrm{E}|\xi|\,\mathrm{E}|\eta|\,\psi(\mathcal{A};\mathcal{B}) \leq \mathrm{c}^2\psi(\mathcal{A};\mathcal{B}).\end{aligned}$$

(ix) *Special facts for independent random sequences:* Assume that $\xi = \{\xi_k, k \in \mathbb{N}\}$ and $\eta = \{\eta_k, k \in \mathbb{N}\}$ are random sequences such that the sequence $\{(\xi_k, \eta_k), k \in \mathbb{N}\}$ of 2-dimensional vectors is independent.
If the pairs (ξ_k, η_k) are identically distributed with real-valued components and ξ_1 and η_1 are not independent, then we have

$$\beta(\xi;\eta) = 1.$$

If the random variables ξ_k and η_k are binary, i. e., they can take only two possible values, then we have

$$\alpha(\xi;\eta) \leq \frac{1}{4}\sup_{k\in\mathbb{N}}|\mathrm{cor}(\xi_k,\eta_k)|.$$

Proof. Part (i). The assertion follows from (7.6.1), (7.6.3), and (7.6.6). In case of the ψ-dependence coefficient the assertion alternatively follows from (7.3.i) and (7.6.7).

Part (ii). The monotonicity follows immediately from the supremum representations of the dependence coefficients given in Definition 7.5.

Part (iii). For a derivation of this inequalities see (Bradley, 2007, Lem. 6.4, Th. 6.2).

Part (iv). For a proof see (Bradley, 2007, Th. 7.2).

Part (v). Comparing (7.6.1) and (7.6.3) yields the first inequality. To obtain the second inequality we combine (6.10.i), (7.1.1), and (7.6.2). Note, that the inequalities are given in (Bradley, 2007, Prop. 3.11 (a), Th. 5.3 (III)) with different constants.

Part (vi). The inequality is easily derived based on Remark 4.2 and (7.6.6). See also (Bradley, 2007, Prop. 5.2 (I.c), Th. 5.3 (II)) in this regard.

Part (vii). The implication is obtained by combining Prop. 3.4 (d), Prop. 3.11 (a), and Th. 9.7 (I) of (Bradley, 2007). The underlying result is due to Ibragimov and Linnik (1971, Th. 17.3.2).

Part (viii). The inequalities follow from (Bradley, 2007, Th. 4.4 (a1), (d1)).

Part (ix). The first assertion is taken from (Bradley, 2007, Th. 3.34). The second assertion is a combination of Prop. 3.11 (b), Prop. 3.20, and Th. 6.1 of (Bradley, 2007). □

§8 Tools to Prove Achievability and Converse

We collect some classic information-theoretic results in a form suitable to serve as building blocks to prove a coding theorem (and converse) for abstract channels with time structure. Namely, we state a version of Feinstein's fundamental lemma, Pinsker's version of the ergodic theorem of information theory, and Fano's inequality. The original forms are published in (Feinstein, 1954), (Pinsker, 1964), and (Fano, 1952).

(8.1) Lemma (*Feinstein's lemma with input constraint). Let* $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ *be the product of the measurable spaces* $(\Omega_1, \mathcal{F}_1)$ *and* $(\Omega_2, \mathcal{F}_2)$*, assume that* P_1 *is a probability measure on* $\mathcal{F}_1$*, and* K *is a Markov kernel from* $(\Omega_1, \mathcal{F}_1)$ *to* $(\Omega_2, \mathcal{F}_2)$*. Let* P *with*

$$\mathrm{P}(F) = \int_{\Omega_1} K(\omega_1, F_{\omega_1}) \,\mathrm{d}\mathrm{P}_1(\omega_1), \quad F \in \mathcal{F}_1 \otimes \mathcal{F}_2,$$

denote the probability measure on $\mathcal{F}_1 \otimes \mathcal{F}_2$ *induced by* P_1 *and* K *and let* P_2 *denote the marginal measure of* P *on* $\mathcal{F}_2$*. Assume that* P *has the Lebesgue decomposition (see (A.7.iii))*

$$\mathrm{P}(F) = \int_F f \,\mathrm{d}\mathrm{P}_1 \otimes \mathrm{P}_2 + \mathrm{P}(F \cap N), \quad F \in \mathcal{F}_1 \otimes \mathcal{F}_2, \tag{1}$$

where N *is a suitable* $\mathrm{P}_1 \otimes \mathrm{P}_2$*-nullset and* f *is the* $\mathrm{P}_1 \otimes \mathrm{P}_2$*-density of* P_a *given by*

$$\mathrm{P}_a(F) = \mathrm{P}(F \cap N^{\mathrm{c}}), \quad F \in \mathcal{F}_1 \otimes \mathcal{F}_2.$$

Let $\gamma \in \mathbb{R}$*,* $m \in \mathbb{N}$*, and* $A \in \mathcal{F}_1$ *be arbitrary and consider the sets*

$$G = (G_\gamma \cap (A \times \Omega_2)) \cup N,$$
$$G_\gamma = \{f > \mathrm{e}^\gamma\} = \{(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 : f(\omega_1, \omega_2) > \mathrm{e}^\gamma\}.$$

Then there exist elements $a_1, a_2, \ldots, a_m \in A$ *and a partition* $\{B_1, B_2, \ldots, B_m\}$ *of* Ω_2 *with* $B_i \in \mathcal{F}_2$ *such that*

$$K(a_i, B_i^{\mathrm{c}}) < \epsilon \tag{2}$$

holds for all $i = 1, 2, \ldots, m$*, where* ϵ *is given by*

$$\epsilon = m\mathrm{e}^{-\gamma} + \mathrm{P}(G^{\mathrm{c}}).$$

In particular, the assertion holds for

$$\epsilon = m\mathrm{e}^{-\gamma} + \mathrm{P}(G_\gamma^{\mathrm{c}}) + \mathrm{P}_1(A^{\mathrm{c}}). \tag{3}$$

(8.2) Remark. Feinstein's lemma is a tool to prove coding theorems for channels with memory in a transparent way. Sometimes this is called the method of maximal codes, which is different from Shannon's original random coding approach. There is a variety of versions of the lemma. The original form of Feinstein (1954) is formulated for discrete spaces with a finite number of elements. Feinstein also considered the application to discrete-time stationary channels and the extension to arbitrary measurable output spaces. A formulation of the lemma for discrete-time causal stationary channels with finite memory is also given by Khinchin (1957, Part II, Ch. IV). This work is generalized to abstract spaces and information-stable channels by Dobrushin (1963, Sec. 3). We state the lemma in purely stochastic terms without (direct) reference to a channel setting. It is essentially the version of Kadota (1970), however, extended in a way to incorporate input constraints represented by the set A. It is stated for abstract measurable spaces without any further constraints. Input constraints are also included in the formulations of Thomasian (1961, Th. 2) or Ash (1965, Lemma 8.2.1). A version for abstract spaces without considering input constraints is also given in (Gray, 2011, Sec. 14.2) but the existence of densities is assumed there, which corresponds to the case when the set N in Lemma 8.1 is equal to the empty set. For a detailed proof of Lemma 8.1 please refer to (Mittelbach, 2012, Lem. 3.1, 3.3).

In Lemma 8.1 we have the condition $A \in \mathcal{F}_1$ for the set A representing an input constraint. This does not hold in general in the situation, where we want to apply the lemma. The corollary below has the right form for which the constraint set A does not have to lie in the σ-algebra $\mathcal{F}_1$. A comment on the possibility of using outer measures (see Paragraph A.11) in this case is also made by Kemperman (1969, Rmk. 3.1). Corollary 8.3 is proved in Paragraph E.5 of Appendix E.

(8.3) Corollary. *Consider the situation of Lemma 8.1 but assume that $A \subset \Omega_1$ is an arbitrary set for which the outer P_1-measure is equal to 1. Then the assertion of Lemma 8.1 holds for*

$$\epsilon = m\mathrm{e}^{-\gamma} + \mathrm{P}(G_\gamma^{\mathrm{c}}). \tag{1}$$

To prove a coding theorem using Feinstein's lemma it is necessary to control the upper bound ϵ of the (conditional) probability in (8.1.2), where the probability will correspond to the decoding error. This is possible with the help of the ergodic theorem of information theory. We state the theorem in a form given by Pinsker (1964, Th. 8.2.1). For a proof please refer to this reference. We adopt the random variable based formulation because it is most clearly given this way.

(8.4) Theorem (*Ergodic theorem of information theory). Assume that $\xi = \{\xi_t, t \in T\}$ and $\eta = \{\eta_t, t \in T\}$ are random processes on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Suppose the pair process $\{(\xi_t, \eta_t), t \in T\}$ is stationary and ergodic and in the continuous-time case suppose ξ and η are continuous in the sense of Pinsker (see Definition B.5). If one of the conditions in (4.12.1) is satisfied and the information rate $\bar{I}(\xi;\eta)$ is finite, then we have for any $\epsilon > 0$*

$$\lim_{s\to\infty} \mathrm{P}\left(\left|\frac{1}{s}\log f^{(s)}\big(\xi_0^s, \eta_0^s\big) - \bar{I}(\xi;\eta)\right| \geq \epsilon\right) = 0, \tag{1}$$

where $f^{(s)}$ denotes the $\mathrm{P}_{\xi_0^s} \otimes \mathrm{P}_{\eta_0^s}$-density of $\mathrm{P}_{\xi_0^s,\eta_0^s}$. By $\mathrm{P}_{\xi_0^s,\eta_0^s}$, $\mathrm{P}_{\xi_0^s}$, and $\mathrm{P}_{\eta_0^s}$ we denote the distribution of (ξ_0^s, η_0^s), ξ_0^s, and η_0^s, respectively.

(8.5) Remark. An equivalent way of expressing (8.4.1) is

$$\frac{1}{s}\log f^{(s)}(\xi_0^s, \eta_0^s) \xrightarrow[s\to\infty]{} \bar{I}(\xi;\eta) \qquad \text{(in probability)}.$$

Due to the first inequality in (4.7.ii) a finite information rate $\bar{I}(\xi;\eta)$ implies a finite mutual information $I(\xi_0^s;\eta_0^s)$ for all $s \in T_+$. Theorem 4.3 then implies that the density $f^{(s)}$ exists and therefore (8.4.1) is well-defined.

On the one hand the conditions in (4.12.1) guarantee the existence of the information rate $\bar{I}(\xi;\eta)$. On the other hand they imply the equality of the information rate with another type of information rate defined in (Pinsker, 1964, (5.4.4)), which is required in the ergodic theorem. Note that a more general condition than (4.12.1) is used in the original formulation of the theorem in (Pinsker, 1964, Th. 8.2.1). However, the employed conditions are easier to verify and general enough in the context we want to apply the theorem. From (Pinsker, 1964, Th. 7.4.2) and the comment thereafter it follows that (4.12.1) implies Pinsker's more general form.

Consider the special case when $\{\xi_t, t \in T\}$ and $\{\eta_t, t \in T\}$ are random sequences, i. e. $T = \mathbb{Z}$, and the ξ_t's are discrete random variables taking values in a finite set. Then it is easily verified with the second inequality in (4.7.v) and the last in (4.7.vii), that all conditions of the ergodic theorem are satisfied. If the η_t's are also discrete random variables taking values in a finite

set, then we have the special case formulated in (Pinsker, 1964, Th. 6.4.1), which is actually McMillan's version of the theorem.

Let us restate some comments and historical remarks from (Mittelbach, 2012, p. 69). The ergodic theorem of information theory, often called asymptotic equipartition property, is formulated in various degrees of generality. The stated form has the key advantage of being applicable to random processes with values in arbitrary space, which is exactly what we need to prove a coding theorem in the general context we are interested in. But note that the condition of finite information rate is required. In the version of Kadota (1974) this condition is replaced by one on the asymptotic behavior of a certain conditional mutual information and the theorem is formulated for real-valued processes. In the form of Barron (1985, Th. 4) there is also no condition on the finiteness of the information rate, but it only holds for random processes with values in a standard Borel space.

Shannon (1948) stated the first theorem in this direction. It was strictly proved by McMillan (1953) under more general conditions and was further generalized by Breiman (1957, 1960). These classical results all consider the discrete-time case and finite alphabets and are usually referred to as Shannon-McMillan-Breiman theorem. The Russian school of information theory derived related results based on the concept of information stability, which applies to very general situations. Girardin (2005) gives an overview over relevant work in this area with a comparison on the modes of convergence and the allowed value spaces.

Fano's inequality is a standard tool to prove a weak converse of a coding theorem. The version below is given in terms of σ-algebras and partitions since it is used later in this form. See (Billingsley, 1965, p. 81) or (Gallager, 1968, p. 78–79) for a proof of this result.

(8.6) Lemma (*Fano's inequality). Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space, assume that $\mathcal{A}$ is the σ-algebra generated by the partition $\{A_1, A_2, \dots, A_m\}$ of Ω with $A_i \in \mathcal{F}$, and $\mathcal{B}$ is the σ-algebra generated by the partition $\{B_1, B_2, \dots, B_m\}$ of Ω with $B_i \in \mathcal{F}$. Then the conditional entropy $H(\mathcal{A}|\mathcal{B})$ of $\mathcal{A}$ given $\mathcal{B}$ satisfies the inequality*

$$h_m(p_e) \geq H(\mathcal{A}|\mathcal{B}),$$

where p_e is given by

$$p_e = \sum_{i=1}^{m} \mathrm{P}(A_i \cap B_i^c)$$

and the function h_m is defined for integers $m > 1$ by

$$h_m(x) = x \log(m-1) - x \log x - (1-x) \log(1-x), \qquad x \in [0,1]. \tag{1}$$

Chapter II

Coding Theorem and Converse for Abstract Channels with Time Structure

Under quite general conditions we prove a block coding theorem and a converse for abstract channels with time structure. Thereby, we establish the operational significance of the information rate capacity C of Definition 5.1 for coded information transmission in the following sense. For any code rate below C there exists a coding-decoding procedure such that the transmitted messages, even though they are randomly corrupted by noise, can be inferred from the received messages with arbitrarily low error probability. For any code rate above C, this is not possible. In Section §9 we formulate and prove the theorems. In Section §10 we study the information rate capacity introduced by Kadota and Wyner (1972) and show that it is equal to C if it has an operational meaning in the sense of the coding theorem. The discussion of the results including the discussion of contributions and related work is postponed to Section §11.

§9 Coding Theorem and Weak Converse

In this section, κ is a channel with time structure as introduced in Definition 2.3 and C is the information rate capacity of κ for the family $\mathcal{E} = \{E_s \subset X_0^s, s \in T_+\}$ of input constraints as defined in Definition 5.1. We first state and prove the achievability result.

(9.1) Theorem (*Coding theorem). Suppose the channel κ is stationary, causal, asymptotically input-memoryless for the set E'', and totally ergodic for block-i.i.d. inputs. The set E'' is defined in (3.1.3) based on the family $\mathcal{E}$ of input constraints, which is assumed to satisfy the regularity condition (3.1.4).*

(i) If the information rate capacity C for the constraints $\mathcal{E}$ is finite, then for any $\rho \in (0, C)$, $\epsilon \in (0,1)$, and $b_0 \in T_+$ there exists a $(b, E_b, E'', \mathrm{e}^{(C-\rho)b}, \epsilon)$-code for some block length $b \geq b_0$.

(ii) If C is infinite, then for any $R > 0$, $\epsilon \in (0,1)$, and $b_0 \in T_+$ there exists a $(b, E_b, E'', \mathrm{e}^{Rb}, \epsilon)$-code for some block length $b \geq b_0$.

(9.2) Remark. To require the asymptotic input-memorylessness on the set E'' of input signals is not the only possibility. However, it is convenient for the proof to consider a shift-invariant set in connection with the assumption of a stationary channel.

For the set E_b^* defined in (3.1.1) we have $E_b^* \subset E''$ so that the assertion of the coding theorem is still valid if E'' in (9.1.i) and (9.1.ii) is replaced by E_b^*. The robustness of the code w. r. t. this set is typically sufficient (see comments in Paragraph 3.5).

Proof. (i) *Setup.* For the channel κ we adopt the notation of Definition 2.3. Let us fix some $\epsilon \in (0,1)$. If $C < \infty$ we fix some $\rho \in (0, C)$ and put $R = C - \rho$. If $C = \infty$ then let $R > 0$ be arbitrary. Due to the definition of the information rate capacity given in Definition 5.1 there

exists in both cases an $s_0 \in T_+$ and a $\mu \in \mathcal{P}_{s_0}$ such that

$$R < \frac{1}{s_0} I\big(\xi_0^{s_0}; \eta_0^{s_0}\big). \tag{1}$$

The notation is based on the projections ξ_t and η_t introduced in Remark 5.2. These are random variables on the input-output probability space $(X \times Y, \mathcal{X} \otimes \mathcal{Y}, \mu\kappa)$ with the measure $\mu\kappa$ induced by the input measure μ and the channel κ. We define the random sequences $\alpha = \{\alpha_k, k \in \mathbb{Z}\}$ and $\beta = \{\beta_k, k \in \mathbb{Z}\}$ with

$$\alpha_{k+1} = \xi_{ks_0}^{(k+1)s_0} \qquad \text{and} \qquad \beta_{k+1} = \eta_{ks_0}^{(k+1)s_0}.$$

By μ_0 we denote the probability measure on $\mathcal{X}_0^{s_0}$ from which the product measure μ is constructed as specified in Definition 5.1. Suppose μ_0 is given by

$$\mu_0 = \sum_{i=1}^{M} p_i \delta_{a_i}, \tag{2}$$

where M is a natural number, p_i is a positive probability, and δ_{a_i} denotes the Dirac measure for the element $a_i \in E_{s_0}$, where $a_i \neq a_j$ if $i \neq j$. We further define the sets

$$A_0 = \big\{a_i : i \in \{1, 2, \ldots, M\}\big\}, \qquad A_n = \bigtimes_{k=0}^{n-1} \langle A_0 \rangle_{ks_0}, \qquad A = \bigtimes_{k \in \mathbb{Z}} \langle A_0 \rangle_{ks_0}. \tag{3}$$

The channel κ is assumed to be asymptotically input-memoryless for the set E''. Due to (2.7.iv) this implies there exists a minimal $l_0 \in \mathbb{N}$, such that for any $B \in \mathcal{Y}_0^+$ and $x, \tilde{x} \in E''$ coinciding on $(-l_0 s_0, \infty)$ we have

$$\big|\kappa(x, [B]) - \kappa(\tilde{x}, [B])\big| < \frac{\epsilon}{2}. \tag{4}$$

It follows together with the stationarity of κ, the equality $\mathcal{Y}_{l_0 s_0}^+ = \langle \mathcal{Y}_0^+ \rangle_{l_0 s_0}$, and the shift-invariance of the set E'' that (4) holds for any $B \in \mathcal{Y}_{l_0 s_0}^+$ and $x, \tilde{x} \in E''$ coinciding on $(0, \infty)$.

(ii) *Ergodic theorem.* Since μ is an s_0-i.i.d. probability measure and κ is a stationary channel the properties of the random sequences α and β are such that we can apply Corollary 4.14 as derived in part (E.2.i) of the proof of Lemma 5.3. Thus, the information rate $\bar{I}(\alpha; \beta)$ exists and we have the following inequalities

$$\frac{1}{s_0} I(\alpha_1; \beta_1) \leq \frac{1}{n s_0} I(\alpha_0^n; \beta_0^n) \leq \frac{1}{s_0} \bar{I}(\alpha; \beta) \leq \frac{1}{s_0} \log M \tag{5}$$

for any $n \in \mathbb{N}$. For the inequality on the right-hand side we have also used

$$I(\alpha_0^n; \beta_0^n) \leq \log M^n. \tag{6}$$

Since μ_0 is given by (2) and the distribution of α_0^n is $\bigotimes_{k=0}^{n-1} \langle \mu_0 \rangle_{ks_0}$ the σ-algebra $\mathcal{X}_0^{ns_0}$ is completely atomic with M^n atoms (see (E.2.3) and Paragraph A.5). Taking Remark 4.2 into account, using the second inequality in (4.7.v), and the last in (4.7.vii) yields the upper bound of $I(\alpha_0^n; \beta_0^n)$.

The channel κ is assumed to be totally ergodic for block-i.i.d. inputs, thus it is s_0-ergodic for s_0-i.i.d. inputs. Because μ is an s_0-i.i.d. probability measure we obtain the s_0-stationarity and s_0-ergodicity of the input-output probability measure $\mu\kappa$ according to (2.7.ii). Therefore, the pair sequence $\{(\alpha_k, \beta_k), k \in \mathbb{Z}\}$ is stationary and ergodic. Furthermore, the information rate $\bar{I}(\alpha;\beta)$ is finite due to (5) and $I(\alpha_-^0;\alpha_1) = 0$ holds because $\{\alpha_k, k \in \mathbb{Z}\}$ is an i.i.d.-sequence. As a consequence we can apply Theorem 8.4 and obtain for any $\epsilon^* > 0$

$$\lim_{n\to\infty} \mu\kappa\Big(f^{(n)}(\alpha_0^n, \beta_0^n) \le \mathrm{e}^{n(\bar{I}(\alpha;\beta)-\epsilon^*)}\Big) = 0, \tag{7}$$

where $f^{(n)}$ denotes the density of the distribution of (α_0^n, β_0^n) w. r. t. the product of the distributions of α_0^n and β_0^n. This density exists for all $n \in \mathbb{N}$ since $\bar{I}(\alpha;\beta)$ is finite.

(iii) *Code size.* Let us fix

$$\epsilon^* = \frac{1}{3}\big(\bar{I}(\alpha;\beta) - Rs_0\big), \tag{8}$$

which is positive because of (1) and (5). Further, due to (7) we can choose a sufficiently large $n_0 \in \mathbb{N}$, such that

$$\mathrm{e}^{-n_0\epsilon^*} < \frac{\epsilon}{4} \tag{9}$$

$$Rl_0s_0 + 1 < n_0\epsilon^* \tag{10}$$

$$\mu\kappa\big((\alpha_0^{n_0}, \beta_0^{n_0}) \in G_\gamma^{\mathrm{c}}\big) < \frac{\epsilon}{4} \tag{11}$$

hold simultaneously, where

$$G_\gamma = \big\{f^{(n_0)} > \mathrm{e}^\gamma\big\} \qquad \text{and} \qquad \gamma = n_0(\bar{I}(\alpha;\beta) - \epsilon^*). \tag{12}$$

We continue by choosing an $m \in \mathbb{N}$ that satisfies

$$\mathrm{e}^{R(n_0+l_0)s_0} < m < \mathrm{e}^{Rn_0s_0+n_0\epsilon^*}, \tag{13}$$

which is possible due to (10). From (8), (9), and the right-hand side of (13) we obtain

$$m\mathrm{e}^{-\gamma} = \mathrm{e}^{-2n_0\epsilon^*+(\log m - Rn_0s_0)} < \mathrm{e}^{-n_0\epsilon^*} < \frac{\epsilon}{4}. \tag{14}$$

(iv) *Feinstein's lemma.* Let us put $t_0 = n_0s_0$. Based on the probability measure μ_0 given in (2) we define the product measures

$$\mu_- = \bigotimes_{k=1}^{\infty} \langle\mu_0\rangle_{-ks_0}, \qquad \mu_0' = \bigotimes_{k=0}^{n_0-1} \langle\mu_0\rangle_{ks_0}, \qquad \mu_+ = \bigotimes_{k=n_0}^{\infty} \langle\mu_0\rangle_{ks_0}$$

on $\mathcal{X}_-^0$, $\mathcal{X}_0^{t_0}$, and $\mathcal{X}_{t_0}^+$, respectively. Further, we define for any $x_0 \in X_0^{t_0}$ and $B \in \mathcal{Y}_0^{t_0}$

$$\bar{\kappa}(x_0, B) = \int_{X_-^0 \times X_{t_0}^+} \kappa\big((x_-, x_0, x_+), [B]\big)\, \mathrm{d}\mu_- \otimes \mu_+(x_-, x_+)$$

using for $x \in X$ the representation as 3-tupel $x = (x_-, x_0, x_+) \in X_-^0 \times X_0^{t_0} \times X_{t_0}^+$. According to (A.3.iii) $\bar{\kappa}$ is a Markov kernel from $(X_0^{t_0}, \mathcal{X}_0^{t_0})$ to $(Y_0^{t_0}, \mathcal{Y}_0^{t_0})$ and it induces together with μ_0' a probability measure on $\mathcal{X}_0^{t_0} \otimes \mathcal{Y}_0^{t_0}$, say $\mu\kappa_0'$, given for any $F \in \mathcal{X}_0^{t_0} \otimes \mathcal{Y}_0^{t_0}$ by

$$\mu\kappa_0'(F) = \int_{X_0^{t_0}} \bar{\kappa}(x_0, F_{x_0}) \, \mathrm{d}\mu_0'(x_0).$$

From the definition of $\mu\kappa$, $\alpha_0^{n_0}$, $\beta_0^{n_0}$, and $\bar{\kappa}$ as well as from the product structure of μ and part (A.8.i) of Fubini's theorem we obtain that $\mu\kappa_0'$ is the distribution of $(\alpha_0^{n_0}, \beta_0^{n_0})$. Furthermore, μ_0' is the distribution of $\alpha_0^{n_0}$ and ν_0' is the distribution of $\beta_0^{n_0}$, where ν_0' denotes the marginal measure of $\mu\kappa_0'$ on $\mathcal{Y}_0^{t_0}$. It follows that the function $f^{(n_0)}$ used to define the set G_γ in (12) is the $\mu_0' \otimes \nu_0'$-density of $\mu\kappa_0'$ and

$$\mu\kappa_0'\big(G_\gamma^{\mathrm{c}}\big) = \mu\kappa\big((\alpha_0^{n_0}, \beta_0^{n_0}) \in G_\gamma^{\mathrm{c}}\big). \tag{15}$$

Consider the sets A_0 and A_{n_0} defined in (3). Since the outer μ_0-measure of A_0 is equal to 1 the outer μ_0'-measure of A_{n_0} is also equal to 1 (see (A.11.iii)). Consequently, according to Corollary 8.3 there exist

$$u_1', u_2', \ldots, u_m' \in A_{n_0} \qquad \text{and} \qquad \hat{B}_1, \hat{B}_2, \ldots, \hat{B}_m \in \mathcal{Y}_0^{t_0}$$

such that

$$\bar{\kappa}(u_i', \hat{B}_i^{\mathrm{c}}) < m\mathrm{e}^{-\gamma} + \mu\kappa_0'\big(G_\gamma^{\mathrm{c}}\big) < \frac{\epsilon}{2}, \tag{16}$$

hold for all $i \in \{1, 2, \ldots, m\}$, where γ and m are chosen as in (12) and (13). The upper bound of $\epsilon/2$ is obtained from (11), (14), and (15).

(v) *Code construction.* If the expectation of a random variable is less than some constant, then the probability that the random variable has values less than this constant is positive. Since $\bar{\kappa}(u_i', \hat{B}_i^{\mathrm{c}})$ is the expectation of $\kappa\big((\cdot, u_i', \cdot), [\hat{B}_i^{\mathrm{c}}]\big)$ we obtain from (16)

$$\mu_- \otimes \mu_+\Big(\kappa\big((\cdot, u_i', \cdot), [\hat{B}_i^{\mathrm{c}}]\big) < \epsilon/2\Big) > 0.$$

This implies together with the specific structure of μ_-, μ_+, and μ_0 that for all $i \in \{1, 2, \ldots, m\}$ there exists a $\hat{u}_i \in A$ coinciding on $(0, n_0 s_0]$ with u_i' such that

$$\kappa(\hat{u}_i, [\hat{B}_i^{\mathrm{c}}]) < \frac{\epsilon}{2},$$

where A is given in (3). The stationarity of the channel κ then implies

$$\kappa(\hat{u}_i, [\hat{B}_i^{\mathrm{c}}]) = \kappa\big(\theta_{l_0 s_0}(\hat{u}_i), \theta_{l_0 s_0}([\hat{B}_i])\big) < \frac{\epsilon}{2}, \tag{17}$$

where $\theta_{l_0 s_0}$ denotes the shift operator defined in Paragraph 1.2.

Let u_i denote the projection of $\theta_{l_0 s_0}(\hat{u}_i)$ to X_0^b and B_i the projection of $\theta_{l_0 s_0}([\hat{B}_i])$ to Y_0^b, where b is given by

$$b = (n_0 + l_0) s_0.$$

The code

$$\mathcal{C} = \big\{(u_i, B_i), i \in \{1, 2, \dots, m\}\big\}$$

has a block length b and a code size $m > \mathrm{e}^{Rb}$ due to (13). Each decoding set satisfies $B_i \in \mathcal{Y}_0^b$ and $[B_i] = \theta_{l_0 s_0}([\hat{B}_i]) \in [\mathcal{Y}_{l_0 s_0}^b]$. Because $u_i \in A_{n_0+l_0}$ and $A_0 \subset E_{s_0}$ we have for each codeword $u_i \in E_b$, since we assume that the family $\mathcal{E}$ of input constraints satisfies the regularity condition (3.1.4). Furthermore, we assume that the channel κ is causal and asymptotically input-memoryless for the set E'' so that (17) and the comment below (4) imply for all $i \in \{1, 2, \dots, m\}$

$$\kappa\big(x_i, [B_i^c]\big) < \epsilon,$$

for all $x_i \in E''$ coinciding on $(0, b]$ with u_i. The maximal decoding error $\varrho_{\max}(E'')$ is therefore bounded by ϵ. Thus, $\mathcal{C}$ is a $(b, E_b, E'', \mathrm{e}^{Rb}, \epsilon)$-code, which completes the proof. □

(9.3) Theorem (*Weak converse). Assume that the channel κ is stationary and the family $\mathcal{E}$ of input constraints satisfies the regularity condition (3.1.4).*

If the information rate capacity C of κ for the constraints $\mathcal{E}$ is finite, then for any $\rho > 0$ there exists a constant $\epsilon^* = \epsilon^*(\rho, C) \in (0, 1)$, such that for any $\epsilon < \epsilon^*$ and block length $b \in T_+$ there does not exist a $(b, E_b, U_b^*, \mathrm{e}^{(C+\rho)b}, \epsilon)$-code, where U_b^* is defined in (3.4.1).

(9.4) Remark. Since we have $U_b^* \subset E_b^* \subset E''$ for the sets specified in (3.1.1) and (3.1.3) the weak converse holds if U_b^* is replaced by E_b^* or E''.

Clearly, the weak converse does only make sense for finite information rate capacity. It holds in particular under the conditions of Theorem 9.1. The combination of the coding theorem and converse establishes the operational significance of the information rate capacity. Note that the coding theorem is a statement in the sense of an optimistic coding capacity (Ahlswede, 2014, p. 176), which means for code rates below C the existence of "good" codes is guaranteed for infinitely many block lengths, however, not necessarily for all block lengths above a certain minimum value. In contrast, the more commonly considered pessimistic capacity concept means that once we have found a good code, then it is guaranteed that for all larger block lengths good codes exist (Ahlswede, 2014, p. 176). For a corresponding version of Theorem 9.1 the second part of (9.1.i) would read as follows: *"... , then for any $\rho \in (0, C)$ and $\epsilon \in (0, 1)$ there exists a $b_0 \in T_+$ such that for all block lengths $b \geq b_0$ there exists a $(b, E_b, E'', \mathrm{e}^{(C-\rho)b}, \epsilon)$-code."* The reformulation of (9.1.ii) would be similar. Now, if the limit superior in the definition of C in (5.1.1) is actually a limit, then the previous statement is indeed valid and C is also equal to the pessimistic coding capacity. This follows from the fact that the quantity s_0 chosen in the beginning of the proof of Theorem 9.1 can be replaced by any $\tilde{s}_0 \geq s_0$.

Proof. (i) *Setup.* Let us fix some $\rho > 0$, a block length $b \in T_+$, and an integer $m \geq \mathrm{e}^{(C+\rho)b}$. Further, let

$$\mathcal{C}(b, E_b) = \big\{(u_i, B_i), i \in \{1, 2, \dots, m\}\big\}$$

be some block code satisfying the input constraint E_b. For the channel κ consider the notation of Definition 2.3 and assume that μ_0 denotes the probability measure on $\mathcal{X}_0^b$ defined by

$$\mu_0 = \sum_{i=1}^{m} p_i \delta_{u_i}, \tag{1}$$

where δ_{u_i} denotes the Dirac measure on $\mathcal{X}_0^b$ for the codeword u_i. Without loss of generality we choose $p_i = 1/m$. Based on μ_0 we construct the product measures

$$\mu_- = \bigotimes_{k\in\mathbb{N}} \langle\mu_0\rangle_{-kb}, \qquad \mu_+ = \bigotimes_{k\in\mathbb{N}} \langle\mu_0\rangle_{kb}, \qquad \mu = \mu_- \otimes \mu_0 \otimes \mu_+ \tag{2}$$

on $\mathcal{X}_-^0$, $\mathcal{X}_b^+$, and $\mathcal{X}$, respectively. Subsequently, we consider the input-output probability space $(X \times Y, \mathcal{X} \otimes \mathcal{Y}, \mu\kappa)$ with the measure $\mu\kappa$ induced by the b-i.i.d. input probability measure μ and the channel κ.

Let $\{A_1, A_2, \dots, A_m\}$ be a partition of X_0^b with $A_i \in \mathcal{X}_0^b$ such that $u_i \in A_i$ holds.[19] Then, $\{[A_1], [A_2], \dots, [A_m]\}$ is a partition of $X \times Y$ with $[A_i] \in [\mathcal{X}_0^b]$, where the projections for the inverse images are defined here on $X \times Y$. Correspondingly, we obtain from the decoding sets the partition $\{[B_1], [B_2], \dots, [B_m]\}$ of $X \times Y$ with $[B_i] \in [\mathcal{Y}_0^b]$. These partitions on $X \times Y$ generate finite and therefore completely atomic σ-algebras (see Paragraph A.5), denoted subsequently by $\mathcal{A}$ and $\mathcal{B}$.

(ii) *Applying Fano's inequality.* At first, we have

$$\begin{aligned} I(\mathcal{A};\mathcal{B}) &\le I\big([\mathcal{X}_0^b];[\mathcal{Y}_0^b]\big) \\ &\le C_b \\ &\le bC, \end{aligned} \tag{3}$$

where the first inequality follows from $\mathcal{A} \subset [\mathcal{X}_0^b]$ and $\mathcal{B} \subset [\mathcal{Y}_0^b]$ together with (4.7.ii). Since μ belongs to the class of probability measures w. r. t. which C_b is defined in Definition 5.1, we obtain the second inequality. We assume a stationary channel κ and a family of input constraints $\mathcal{E}$ satisfying the regularity condition (3.1.4). Therefore, we can apply Lemma 5.3 and obtain the last inequality from (5.3.1).

If p_e is given by

$$p_e = \sum_{i=1}^{m} \mu\kappa\big([A_i] \cap [B_i^c]\big), \tag{4}$$

then we obtain

$$\begin{aligned} h_m(p_e) &\ge H(\mathcal{A}|\mathcal{B}) \\ &= \log m - I(\mathcal{A};\mathcal{B}) \\ &\ge \log m - bC \\ &= \left(1 - \frac{C}{\log(m)/b}\right) \log m \\ &\ge \left(1 - \frac{C}{C+\rho}\right) \log m. \end{aligned} \tag{5}$$

Fano's inequality given in Lemma 8.6 yields the first inequality, where h_m is defined in (8.6.1). The subsequent equality holds due to the first equality in (4.7.v) and the last relation in (4.7.vii),

[19] We assume that the σ-algebra $\mathcal{X}_0^b$ is large enough so that it is possible to find such a partition. Using a partition instead of directly considering the codewords u_i is required because in general we do *not* have $\{u_i\} \in \mathcal{X}_0^b$.

because $\mu\kappa([A_i]) = \mu_0(A_i) = 1/m$. The next inequality follows from (3) and the last from the assumption $m \geq \mathrm{e}^{(C+\rho)b}$.

Assume that we restrict the function h_m to the interval $[0, (m-1)/m]$ and denote the corresponding inverse by h_m^{-1}. Then h_m^{-1} is a well-defined, monotonically increasing function on $[0, \log m]$ with values in $[0, (m-1)/m]$. The following chain of inequalities holds

$$0 < h_{m-1}^{-1}\big(a\log(m-2)\big) < h_m^{-1}\big(a\log(m-1)\big) < h_m^{-1}\big(a\log m\big) \tag{6}$$

for all $0 < a \leq 1$ and $m \geq 4$. If $m = 2$, then the two terms in the middle have to be omitted and if $m = 3$, then only the left of these terms has to be omitted. See, e. g., (Mittelbach, 2012, p. 67) for details. Applying h_m^{-1} to (5) and using (6) yields

$$p_e \geq h_{m_0}^{-1}\left(\left[1 - \frac{C}{C+\rho}\right]\log 2\right) > 0, \tag{7}$$

where $m_0 = 2$ if $m = 2$ and $m_0 = 3$ if $m > 2$. We have also used $(1 - C/(C+\rho)) \in (0,1)$.

(iii) *Evaluation of error probabilities.* We define for any $x_0 \in X_0^b$ and $B \in \mathcal{Y}$

$$\bar{\kappa}(x_0, B) = \int_{X_-^0 \times X_b^+} \kappa\big((x_-, x_0, x_+), B\big)\, \mathrm{d}\mu_- \otimes \mu_+(x_-, x_+) \tag{8}$$

using the measures defined in (2) and the representation of an element $x \in X$ as 3-tupel $x = (x_-, x_0, x_+) \in X_-^0 \times X_0^b \times X_b^+$. To have a distinction to previously considered inverse images w. r. t. projections we denote by $[A_i]'$ and $[B_i^{\mathrm{c}}]'$ the inverse images of A_i and B_i^{c} w. r. t. the projections defined on X and Y, respectively. Then we have

$$\begin{aligned}
\mu\kappa\big([A_i] \cap [B_i^{\mathrm{c}}]\big) &= \mu\kappa\big([A_i]' \times [B_i^{\mathrm{c}}]'\big) \\
&= \int_{[A_i]'} \kappa(x, [B_i^{\mathrm{c}}]')\, \mathrm{d}\mu(x) \\
&= \int_{A_i} \bar{\kappa}(x_0, [B_i^{\mathrm{c}}]')\, \mathrm{d}\mu_0(x_0) \\
&= \int_{A_i} \frac{1}{m}\bar{\kappa}(x_0, [B_i^{\mathrm{c}}]')\, \mathrm{d}\delta_{u_i}(x_0) \\
&= \frac{1}{m}\bar{\kappa}(u_i, [B_i^{\mathrm{c}}]'),
\end{aligned} \tag{9}$$

where the second equality is simply the definition of the channel input-output probability measure as given in Definition 2.1. From the specific product structure of μ defined in (2) we obtain the third equality using part (A.8.i) of Fubini's theorem. The definition of the set A_i and of the measure μ_0 in (1) yields the fourth equality. The last equality is the result of integrating w. r. t. a Dirac measure.

Inserting (9) into (4) and using (7) yields, that at least for one $i_0 \in \{1, 2, \ldots, m\}$ we have

$$\bar{\kappa}(u_{i_0}, [B_{i_0}^{\mathrm{c}}]') \geq \epsilon^* > 0, \tag{10}$$

where ϵ^* is defined as the middle expression of (7). Since $\kappa(\cdot, [B_{i_0}^{\mathrm{c}}]')$ is an $\mathcal{X}$-measurable function on X and because $X = X_-^0 \times X_0^b \times X_b^+$ and $\mathcal{X} = \mathcal{X}_-^0 \otimes \mathcal{X}_0^b \otimes \mathcal{X}_b^+$ we can consider

$\kappa\big((\cdot, u_{i_0}\cdot), [B^{\mathrm{c}}_{i_0}]'\big)$ as random variable on the probability space $(X^0_- \times X^+_b, \mathcal{X}^0_- \otimes \mathcal{X}^+_b, \mu_- \otimes \mu_+)$. If the expectation of a nonnegative random variable is greater or equal to some constant, then the probability that the random variable has values greater or equal to this constant is positive. Since $\bar\kappa(u_{i_0}, [B^{\mathrm{c}}_{i_0}]')$ is the expectation of $\kappa\big((\cdot, u_{i_0}, \cdot), [B^{\mathrm{c}}_{i_0}]'\big)$ we obtain from (10)

$$\mu_- \otimes \mu_+\Big(\kappa\big((\cdot, u_{i_0}, \cdot), [B^{\mathrm{c}}_{i_0}]'\big) \geq \epsilon^*\Big) > 0.$$

This implies together with the specific structure of μ_- and μ_+, that at least for one $x \in U^*_b \cap [u_{i_0}]'$, with U^*_b as defined in (3.4.1), we have

$$\kappa(x, [B^{\mathrm{c}}_{i_0}]') \geq \epsilon^* > 0.$$

Thus we have $\varrho_{\max}(U^*_b) \geq \epsilon^*$ for the maximal decoding error, which completes the proof. □

(9.5) Remark. The coding theorem and converse are stated for the maximal decoding error probability $\varrho_{\max}(\cdot)$ as defined in (3.4.2). Alternatively, we can consider the average decoding error probability. Based on the notation of Definition 3.4 it is given by

$$\bar\varrho(\cdot) = \frac{1}{|\mathcal{C}|}\sum_{i=1}^{|\mathcal{C}|} \varrho(u_i, \cdot).$$

It follows immediately that Theorem 9.1 also holds for the average decoding error probability. To verify Theorem 9.3 for $\bar\varrho(U^*_b)$ we have to have a closer look. Let us follow the proof of the weak converse up to inequality (10). Then we observe that even

$$\frac{1}{m}\sum_{i=1}^{m} \bar\kappa(u_i, [B^{\mathrm{c}}_i]') \geq \epsilon^* > 0$$

holds. Similar to the last part of the preceding proof we can therefore conclude

$$\begin{aligned}\bar\varrho(U^*_b) &= \frac{1}{m}\sum_{i=1}^{m} \sup_{(x_-, x_+)\in\tilde U} \kappa\big((x_-, u_i, x_+), [B^{\mathrm{c}}_i]'\big)\\ &\geq \sup_{(x_-, x_+)\in\tilde U} \frac{1}{m}\sum_{i=1}^{m} \kappa\big((x_-, u_i, x_+), [B^{\mathrm{c}}_i]'\big)\\ &\geq \epsilon^* > 0,\end{aligned}$$

where $\tilde U = \xi^0_-(U^*_b) \times \xi^+_b(U^*_b)$ with ξ_t denoting the projection introduced in Remark 5.2. This shows that Theorem 9.3 also holds for the average decoding error.

Making use of the derivations in this section we obtain the following lemma, which is useful, e. g., to prove the equality of the information rate capacity in Definitions 5.1 and 10.1.

(9.6) Lemma (*Lower bound for information rate capacity). Suppose the channel κ is stationary and the input constraints $\mathcal{E}$ satisfy the regularity condition (3.1.4). Assume that $\epsilon \in (0, 1/3)$ and $R > 0$ are constants and there exists a $(\hat b, E_{\hat b}, U^*_{\hat b}, \mathrm{e}^{R\hat b}, \epsilon)$-code*

$$\hat{\mathcal{C}} = \big\{(\hat u_i, \hat B_i), i \in \{1, 2, \ldots, \hat m\}\big\}$$

for some $\hat{b} \in T_+$, with $U^*_{\hat{b}}$ as defined in (3.4.1). Then the information rate capacity C of κ for the constraints $\mathcal{E}$ specified in Definition 5.1 satisfies the inequality

$$\left(1 - \frac{h_3(2\epsilon)}{\log 2}\right) R \,\leq\, C,$$

where the function h_3 is defined in (8.6.1).

Proof. First, we use the code $\hat{\mathcal{C}}$ to construct a $(b, E_b, U^*_b, \mathrm{e}^{Rb}, 2\epsilon)$-code

$$\mathcal{C} = \big\{(u_i, B_i), i \in \{1, 2, \ldots, m\}\big\}$$

with $b = 2\hat{b}$ and U^*_b as defined in (3.4.1), which satisfies in addition to the implicit inequality $m \geq \mathrm{e}^{Rb}$ the inequalities

$$4 \leq m \leq \mathrm{e}^{Rb} + 1. \tag{1}$$

This is achieved by considering the Cartesian product $\hat{\mathcal{C}} \times \langle\hat{\mathcal{C}}\rangle_{\hat{b}}$ as a new code, which guarantees the left inequality. The maximal decoding error is bounded by 2ϵ and the code properties concerning the input constraints follow from the regularity condition (3.1.4). If necessary we throw away codewords to satisfy the right inequality. The decoding sets of the canceled codewords can be united, e. g., with the decoding set of the first codeword, which does not increase the decoding error.

Consider the proof of Theorem 9.3 and let the probability measures μ_0, μ_-, μ_+, and μ as well as the the σ-algebras $\mathcal{A}$ and $\mathcal{B}$ be defined as in part (i). From the arguments below (3) and (5) in part (ii) and the definition of p_e in (4) we obtain

$$\begin{aligned}\frac{1}{b} I\big([\mathcal{X}_0^b]; [\mathcal{Y}_0^b]\big) \;\geq\; \frac{1}{b} I(\mathcal{A}; \mathcal{B}) &= \frac{1}{b}\big(\log m - H(\mathcal{A}\,|\,\mathcal{B})\big) \\ &\geq \frac{1}{b}\big(\log m - h_m(p_e)\big) \\ &\geq \left(1 - \frac{h_3(p_e)}{\log 2}\right) R \end{aligned} \tag{2}$$

with h_m as defined in (8.6.1). For the last inequality we have used $m \geq \mathrm{e}^{Rb}$ and

$$\frac{h_m(p_e)}{b} \;\leq\; \frac{h_m(p_e)}{\log(m-1)} R \;\leq\; \frac{h_3(p_e)}{\log 2} R,$$

which holds due to (1) and the monotonicity of the middle term w. r. t. m.

The properties of the code $\mathcal{C}$ imply $\kappa(x, [B_i^{\mathrm{c}}]') \leq 2\epsilon$ for all $x \in U^*_b$ coinciding on $(0, b]$ with u_i, where $[B_i^{\mathrm{c}}]'$ is the inverse image of B_i^{c} w. r. t. the projection defined on Y. It follows together with the structure of μ_- and μ_+ that $\kappa\big((\cdot, u_i, \cdot), [B_i^{\mathrm{c}}]'\big) \leq 2\epsilon$ holds $\mu_- \otimes \mu_+$-almost surely. Then, from the definition of p_e and from (8) and (9) in part (iii) of the proof of Theorem 9.3 we obtain

$$p_e = \frac{1}{m} \sum_{i=1}^{m} \bar{\kappa}(u_i, [B_i^{\mathrm{c}}]') \leq \frac{1}{m} \sum_{i=1}^{m} 2\epsilon = 2\epsilon.$$

Since $\epsilon \in (0, 1/3)$ and h_3 is monotonically increasing on the interval $[0, 2/3]$ we obtain from (2)

$$\left(1 - \frac{h_3(2\epsilon)}{\log 2}\right) R \le \frac{1}{b} I\big([\mathcal{X}_0^b]; [\mathcal{Y}_0^b]\big). \tag{3}$$

Now consider the information rate

$$\bar{I}(\mu) = \lim_{s\to\infty} \frac{1}{s} I\big([\mathcal{X}_0^s]; [\mathcal{Y}_0^s]\big)$$

for which we have

$$\frac{1}{b} I\big([\mathcal{X}_0^b]; [\mathcal{Y}_0^b]\big) \le \bar{I}(\mu) \le C. \tag{4}$$

The first inequality follows with (E.2.5) and (5) in part (ii) of the proof of Theorem 9.1. The assumptions on κ and $\mathcal{E}$ allow to apply Lemma 5.3. Since we have $\mu \in \mathcal{P}_b$ with $\mathcal{P}_b$ defined in Definition 5.1 the identity (5.3.2) yields the second inequality. Combining (3) and (4) completes the proof. □

§ 10 Alternative Definition of Information Rate Capacity

The information rate capacity of Definition 5.1 is a characteristic of a channel with time structure, calculated for a family of input constraints. Due to the coding theorem and the weak converse given in Theorems 9.1 and 9.3 this parameter has an operational meaning under certain conditions on the channel and the input constraints. Kadota and Wyner (1972) introduced a different version of information rate capacity. We show that it must be equal to the information rate capacity specified in Definition 5.1 if it has an operational meaning in the sense of Theorem 9.1. Further, we derive Theorems 9.1 and 9.3 for Kadota and Wyner's information rate capacity if it is finite and discuss the limitations if it is infinite, thus demonstrating the advantage of Definition 5.1.

Subsequently, κ is a channel with time structure as introduced in Definition 2.3 and C is the information rate capacity of κ for the family $\mathcal{E} = \{E_s \subset X_0^s, s \in T_+\}$ of input constraints as defined in Definition 5.1.

(10.1) Definition (Information rate capacity, Kadota and Wyner (1972)). Consider the following modification of Definition 5.1. We substitute the set $\mathcal{P}_s$ of input probability measures by the set $\mathcal{P}'_s$, which is defined as $\mathcal{P}_s$ but the probability measure specified in (5.1.2) is replaced by some probability measure μ_0 for which the outer μ_0-measure of the constraint set E_s is equal to 1. We call

$$C' = \limsup_{s\to\infty} \frac{1}{s} C'_s \qquad \text{with} \qquad C'_s = \sup_{\mu\in\mathcal{P}'_s} I\big([\mathcal{X}_0^s]; [\mathcal{Y}_0^s]\big)$$

the information rate capacity of the channel κ for the constraints $\mathcal{E}$.

(10.2) Remark. In (Kadota and Wyner, 1972) the special case of a continuous-time channel with real-valued input and output signals is considered together with input constraints specified by functionals as in Example 3.2. Furthermore, the information rate capacity is defined based

on standard extensions (see Paragraph A.12) rather than directly on outer measures. However, this is equivalent for the following reasons. Consider some $s \in T_+$ and $\mu \in \mathcal{P}'_s$ and let μ_0 be the probability measure on $\mathcal{X}_0^s$ from which the product measure μ on $\mathcal{X}$ is constructed. To calculate the mutual information $I\big([\mathcal{X}_0^s];[\mathcal{Y}_0^s]\big)$ in Definition 10.1 the underlying probability space is the joint channel input-output space $(X \times Y, \mathcal{X} \otimes \mathcal{Y}, \mu\kappa)$. Based on the (completed)[20] standard extension $(X_0^s, \tilde{\mathcal{X}}_0^s, \tilde{\mu}_0^s)$ of the probability space $(X_0^s, \mathcal{X}_0^s, \mu_0^s)$ let us define

$$\tilde{\mathcal{X}} = \bigotimes_{k\in\mathbb{Z}} \langle \tilde{\mathcal{X}}_0^s \rangle_{ks} \qquad \text{and} \qquad \tilde{\mu} = \bigotimes_{k\in\mathbb{Z}} \langle \tilde{\mu}_0^s \rangle_{ks}.$$

In Kadota and Wyner's capacity definition, instead of $I\big([\mathcal{X}_0^s];[\mathcal{Y}_0^s]\big)$ the mutual information $I\big([\tilde{\mathcal{X}}_0^s];[\mathcal{Y}_0^s]\big)$ is used, where the underlying probability space is the extended input-output space $(X\times Y, \tilde{\mathcal{X}}\otimes\mathcal{Y}, \tilde{\mu}\kappa)$. Here, $\tilde{\mu}\kappa$ denotes the probability measure on $\tilde{\mathcal{X}}\otimes\mathcal{Y}$ induced by the channel κ and the probability measure $\tilde{\mu}$ on $\tilde{\mathcal{X}}$. It can be shown, that these mutual informations are equal, roughly speaking, because the σ-algebras $\tilde{\mathcal{X}}_0^s$ and $\mathcal{X}_0^s$ differ only by $\tilde{\mu}_0^s$-nullsets. Standard extensions are considered for measurability reasons. However, in the situations relevant for the coding theorem it is advantageous to use the more direct definition with outer measures, because the derivations are more transparent.

(10.3) Theorem (*Results applicable to C and C'). The information rate capacities C and C' of κ for the constraints $\mathcal{E}$, as defined in Definitions 5.1 and 10.1, satisfy the inequality*

$$C \le C'. \tag{1}$$

Lemma 5.3 holds literally if C_s, C, and $\mathcal{P}$ are replaced by C'_s, C', and $\mathcal{P}' = \bigcup_{s\in T_+} \mathcal{P}'_s$ with $\mathcal{P}'_s$ as defined in Definition 10.1. If $C'<\infty$, then Theorems 9.1 and 9.3 hold literally if C is replaced by C'.

Proof. Let $s \in T_+$ and assume that μ_0 is a probability measure on $\mathcal{X}_0^s$ as defined in (5.1.2). Every set $A \in \mathcal{X}_0^s$ containing the constraint set E_s has μ_0-measure 1 so that the outer μ_0-measure of E_s is equal to 1. Thus the sets $\mathcal{P}_s$ and $\mathcal{P}'_s$ of input probability measures in Definition 5.1 and 10.1 satisfy $\mathcal{P}_s \subset \mathcal{P}'_s$. Therefore, we have $C_s \le C'_s$ and $C \le C'$.

To show that Lemma 5.3 holds with C_s, C, $\mathcal{P}_s$, and $\mathcal{P}$ replaced by C'_s, C', $\mathcal{P}'_s$, and $\mathcal{P}'$ we follow with these substitutions exactly the original proof given in Paragraph E.2. We only change the part concerning the probability measure μ_0 between (E.2.2) and (E.2.4). Now μ_0 is such that the outer μ_0-measure of the constraint set E_{s_0} is equal to 1. Therefore the outer μ'_0-measure of the set $\times_{k=0}^{n-1}\langle E_{s_0}\rangle_{ks_0}$ is equal to 1 for $\mu'_0 = \bigotimes_{k=0}^{n-1}\langle\mu_0\rangle_{ks_0}$ (see (A.11.iii)). The monotonicity of outer measures (see (A.11.i)) and the regularity condition (3.1.4) imply that the outer μ'_0-measure of the constraint set E_{ns_0} is also equal to 1. For the product measure μ in the original proof we then have $\mu \in \mathcal{P}_{ns_0}$ and we can continue with the modified form of (E.2.4).

Due to the inequality $C \le C'$ Theorem 9.3 obviously holds for C'. To derive Theorem 9.1 with C replaced by C' given C' is finite we can adopt the original proof with the following modifications. We assume that $C' < \infty$ and substitute C by C'. The probability measure μ_0 on $\mathcal{X}_0^{s_0}$ considered in the original proof is now a suitable probability measure for which the outer μ_0-measure of the constraint set E_{s_0} is equal to 1. Since this μ_0 belongs to a larger class of probability measures the inequality (6) and the last inequality in (5) on page 44 are not longer

[20] To complete a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ the σ-algebra $\mathcal{F}$ is extended by all subsets of P-nullsets.

valid. However, (E.2.7) holds with C_{ns_0} and C replaced by C'_{ns_0} and C' due to the comments in the previous paragraph so that we have

$$\frac{1}{ns_0} I(\alpha_0^n; \beta_0^n) \le \frac{1}{ns_0} C'_{ns_0} \le C'.$$

This implies

$$\frac{1}{s_0} \bar{I}(\alpha; \beta) \le C' \tag{1}$$

and due to the assumption $C' < \infty$ the information rate $\bar{I}(\alpha; \beta)$ is finite and we can apply the ergodic theorem of information theory as in the original proof of Theorem 9.1. The rest of the proof is identical. We just have to replace in part (iv) and (v) the sets A_0, A_n, and A defined in (3) on page 44 by the sets E_{s_0},

$$\bigtimes_{k=0}^{n-1} \langle E_{s_0} \rangle_{ks_0}, \qquad \text{and} \qquad E^*_{s_0} = \bigtimes_{k\in\mathbb{Z}} \langle E_{s_0} \rangle_{ks_0}. \qquad \square$$

(10.4) Remark. If the information rate capacity C' is infinite, then we cannot use it in the last part of the previous proof to conclude that the information rate $\bar{I}(\alpha; \beta)$ is finite. Let us give some comments on the proof of Kadota and Wyner (1972) in case of infinite C'. Assume that $\xi = \{\xi_t, t \in T\}$ and $\eta = \{\eta_t, t \in T\}$ are defined as in the proof of Theorem 9.1 and assume that the time axis is partitioned into segments of size s_0. We define the random sequences $\tilde{\alpha} = \{\tilde{\alpha}_k, k \in \mathbb{Z}\}$ and $\tilde{\beta} = \{\tilde{\beta}_k, k \in \mathbb{Z}\}$, where $\tilde{\alpha}_k$ and $\tilde{\beta}_k$ are given by

$$\tilde{\alpha}_{k+1} = \xi_{2ks_0}^{(2k+1)s_0} \qquad \text{and} \qquad \tilde{\beta}_{k+1} = \eta_{2ks_0}^{(2k+1)s_0}.$$

Instead of using the sequences $\alpha = \{\alpha_k, k \in \mathbb{Z}\}$ and $\beta = \{\beta_k, k \in \mathbb{Z}\}$ as defined in the proof of Theorem 9.1 the proof of Kadota and Wyner for infinite C' is based on the sequences $\tilde{\alpha}$ and $\tilde{\beta}$, i. e., only every second time slot is considered for the code construction.

Suppose ξ is an s_0-i.i.d. sequence and the pair sequence (ξ, η) is s_0-stationary. Then $\tilde{\alpha}$ is an i.i.d.-sequence and the pair sequence $(\tilde{\alpha}, \tilde{\beta})$ is stationary. Due to Corollary 4.14 the information rate $\bar{I}(\tilde{\alpha}; \tilde{\beta})$ exists and if

$$I(\xi_0^{s_0}; \eta_0^{s_0}) < \infty \qquad \text{and} \qquad I(\xi_-^0 \eta_-^0; \xi_{s_0}^+ \eta_{s_0}^+) < \infty, \tag{1}$$

then it is finite. Indeed, for all $n \in \mathbb{N}$ we have

$$\begin{aligned}
I(\tilde{\alpha}_0^n; \tilde{\beta}_0^n) &= I(\tilde{\alpha}_1; \tilde{\beta}_1) + I(\tilde{\alpha}_1^n; \tilde{\beta}_1^n) + I(\tilde{\alpha}_1^n; \tilde{\alpha}_1 \tilde{\beta}_1 \,|\, \tilde{\beta}_1^n) - I(\tilde{\alpha}_1; \tilde{\alpha}_1^n) + I(\tilde{\alpha}_1; \tilde{\beta}_1^n \,|\, \beta_1) \\
&= I(\tilde{\alpha}_1; \tilde{\beta}_1) + I(\tilde{\alpha}_0^{n-1}; \tilde{\beta}_0^{n-1}) + I(\tilde{\alpha}_1 \tilde{\beta}_1; \tilde{\alpha}_1^n \tilde{\beta}_1^n) - I(\tilde{\beta}_1; \tilde{\beta}_1^n) \\
&= nI(\tilde{\alpha}_1; \tilde{\beta}_1) + \sum_{k=1}^{n-1} I(\tilde{\alpha}_1 \tilde{\beta}_1; \tilde{\alpha}_1^{k+1} \tilde{\beta}_1^{k+1}) - \sum_{k=1}^{n-1} I(\tilde{\beta}_1; \tilde{\beta}_1^{k+1}) \\
&\le nI(\tilde{\alpha}_1; \tilde{\beta}_1) + (n-1) I(\tilde{\alpha}_1 \tilde{\beta}_1; \tilde{\alpha}_1^n \tilde{\beta}_1^n).
\end{aligned}$$

The first equality follows from applying the chain rule of mutual information given in (4.7.iv) several times. For the second equality we use the stationarity of $(\tilde{\alpha}, \tilde{\beta})$, the independence of $\tilde{\alpha}_1$

and $\tilde{\alpha}_1^n$ together with (4.7.i), and again the chain rule of mutual information. Repeating these steps for $I(\tilde{\alpha}_0^{n-1};\tilde{\beta}_0^{n-1})$, $I(\tilde{\alpha}_0^{n-2};\tilde{\beta}_0^{n-2})$ and so forth yields the third equality. The inequality then follows from the nonnegativity and monotonicity of the mutual information. Using the derived inequality yields for all $n \in \mathbb{N}$

$$\begin{aligned}\frac{1}{n} I(\tilde{\alpha}_0^n;\tilde{\beta}_0^n) &\le I(\tilde{\alpha}_1;\tilde{\beta}_1) + I(\tilde{\alpha}_1\tilde{\beta}_1;\tilde{\alpha}_1^n\tilde{\beta}_1^n)\\ &\le I(\xi_0^{s_0};\eta_0^{s_0}) + I(\xi_-^0\eta_-^0;\xi_{s_0}^+\eta_{s_0}^+),\end{aligned}$$

where the second inequality follows from the definition of the random variables $\tilde{\alpha}_k$ and $\tilde{\beta}_k$, the monotonicity of the mutual information, and the s_0-stationarity of the process (ξ,η). The assumption in (1) then implies that the information rate $\bar{I}(\tilde{\alpha};\tilde{\beta})$ is finite. That means, if condition (1) is satisfied, then the proof for infinite C' resembles that for finite C', when the sequences α and β are replaced by $\tilde{\alpha}$ and $\tilde{\beta}$.

To ensure (1) Kadota and Wyner assumed the channel to be ψ-mixing as defined in Definition 13.6 (in addition to the assumption of stationarity, causality, and asymptotic input-memorylessness). The ψ-mixing condition, however, is much more restrictive than the ergodicity required in the proof of Theorem 9.1 (see Theorems 13.9 and 13.10). Note, that the author could verify the corresponding derivations in (Kadota and Wyner, 1972) only for a more restrictive version of asymptotic input-memorylessness defined in (Mittelbach and Jorswieck, 2013, Def. IV.1). Under these conditions it is demonstrated in (Mittelbach and Jorswieck, 2013) that the ψ-mixing condition can be weakened to information regularity (see Definition 13.6), i. e., even for the approach of Kadota and Wyner the ψ-mixing condition is not required.

For discrete-time finite-alphabet channels Kadota and Wyner's information rate capacity C' of Definition 10.1 and the information rate capacity C of Definition 5.1 are obviously equal. The next theorem shows that the equality must hold in general if C' has an operational meaning in the sense of Theorem 9.1. Kemperman (1969, Sec. 4.3) gave comments on the memoryless channel (see Example 13.8) in a similar regard.

(10.5) Theorem (*Equality of C and C'). Suppose the channel κ is stationary and the input constraints $\mathcal{E}$ satisfy the regularity condition (3.1.4).*

(i) If C' is finite and for any $\rho \in (0, C')$ and $\epsilon \in (0,1)$ there exists a $(b, E_b, U_b^*, \mathrm{e}^{(C'-\rho)b}, \epsilon)$-code for some block length $b \in T_+$, with U_b^* as defined in (3.4.1), then we have

$$C = C'.$$

(ii) If for any $R > 0$ and $\epsilon \in (0,1)$ there exists a $(b, E_b, U_b^*, \mathrm{e}^{Rb}, \epsilon)$-code for some $b \in T_+$, then

$$C = C' = \infty.$$

(10.6) Remark. The set U_b^* in the formulation of Theorem 10.5 is the minimal requirement. Of course, it can be replaced by the set E'' of input signals defined in (3.1.3).

Proof. Under the assumptions in part (i) we obtain for any $\rho \in (0, C')$ and $\epsilon \in (0, 1/3)$ from (10.3.1) and Lemma 9.6

$$\left(1 - \frac{h_3(2\epsilon)}{\log 2}\right)(C' - \rho) \le C \le C'. \tag{1}$$

Since $h_3(2\epsilon)$ is arbitrarily small for vanishing ϵ the assertion of part (i) is shown.

Likewise, under the assumptions of part (ii) the inequalities in (1) hold with $(C' - \rho)$ replaced by R for arbitrary $R > 0$ and $\epsilon \in (0, 1/3)$, which proves the assertion of part (ii). Note that this can also be obtained with Theorem 9.3. □

Combining (10.5.i), the last assertion of Theorem 10.3, and the conditions of Theorem 9.1 we immediately obtain the following corollary.

(10.7) Corollary. *Suppose the channel κ is stationary, causal, asymptotically input-memoryless for the set E'', and totally ergodic for block-i.i.d. inputs, where E'' is defined in (3.1.3) based on the family $\mathcal{E}$ of input constraints. Further assume that $\mathcal{E}$ satisfies the regularity condition (3.1.4). If $C' < \infty$, then we have*

$$C = C'.$$

§11 Discussion of Results and Historical Notes

The main motivation for the first part of the thesis was to establish an abstract framework that allows us to formulate a general coding theorem for a point-to-point communication link. On the one hand, a central objective was to include continuous-time continuous-valued transmission models because in the literature it is payed much less attention to this case compared to discrete models. On the other hand, we aimed at a reduction to the essential channel properties required to prove the coding statements. The presentation of the results shows that once a suitable formulation is chosen the discrete- and the continuous-time case can be treated in exactly the same way. Essentially, the main contribution of this part of the thesis is a generalization of the work of Kadota and Wyner (1972), formulated in precise mathematical terms. The exposition includes the relevant material from probability and information theory in tailored form to allow a unified and transparent presentation. With regard to the generality of the information-theoretic models and tools the formulation is mainly influenced by the work of the Russian school of information theory, in particular by Kolmogorov (1956a), Dobrushin (1963), and Pinsker (1964). The statement of the channel coding theorem, however, follows the style of Ahlswede (2006) and Wolfowitz (1978) because it emphasizes the operational meaning of the results more clearly.

Relation to (Kadota and Wyner, 1972). The theorems in Section §9 generalize the work of Kadota and Wyner (1972) in a number of respects, namely in terms of channel model, input constraints, required channel properties, and definition of information rate capacity. Let us discuss the relevant extensions and modifications in relation to this work. Kadota and Wyner considered a continuous-time channel with real-valued input and output signals and formulated a coding theorem and converse for those channels. In contrast, the theorems in Section §9 apply to channels with time structure in general, i. e., to discrete- as well as continuous-time channels with completely arbitrary alphabets. A consequent measure-theoretic description, in particular the use of general product spaces allows this generalized and unified formulation. Please note that this could be achieved without using involved measure-theoretic concepts.

For Theorem 9.1 as well as the coding theorem of Kadota and Wyner the considered channel is required to be stationary, causal, and asymptotically input-memoryless. In addition, both theorems include a condition concerning the channel output memory, however, of fairly different quality. Kadota and Wyner used a property they called asymptotic output-memorylessness,

which is called ψ-mixing condition in Definition 13.6. This output memory condition is quite restrictive, in particular for the important class of Gaussian channels. In Theorem 13.10 we show that for Gaussian channels it is indeed equivalent to finite output memory. As a result, Kadota and Wyner's version of the theorem is for example not applicable to the simple stationary additive Gaussian noise channel with rational noise spectral density (see the discussion at the end of Paragraph 16.3).

The condition of total ergodicity for block-i.i.d. inputs used in Theorem 9.1 is much weaker than the ψ-mixing condition as demonstrated by Theorem 13.9. It is actually the weakest possible property that allows the application of Pinsker's ergodic theorem in the proof of the coding theorem, where at the relevant point a partition of the time axis into segments of size s_0 is considered together with an s_0-i.i.d. input probability measure. In Theorem 10.3 we show that the ergodicity condition is sufficient even when we employ Kadota and Wyner's information rate capacity of Definition 10.1, however, only if we additionally assume its finiteness. A derivation based on Definition 10.1 for infinite information rate capacity requires a stronger output memory condition. Kadota and Wyner used therefore the ψ-mixing condition but not only in this but unnecessarily in the finite capacity case. Actually, they do not differentiate between finite and infinite capacity in the formulation of the theorem. According to Remark 10.4 even with their approach ψ-mixing can be weakened to information regularity. However, using the characterization of information rate capacity given in Definition 5.1 allows us to prove Theorem 9.1 under the condition of total ergodicity for block-i.i.d. inputs with an identical approach for finite and infinite information rate capacity. This demonstrates the significance and advantage of introducing a modified version of information rate capacity. Section § 13 considers various sufficient conditions implying total ergodicity for block-i.i.d. inputs.

Gray and Ornstein (1979, p. 296) gave the comment that Kadota and Wyner's coding theorem should apply to ergodic channels. First we note that simple ergodicity is not restrictive enough. As shown we need the introduced form of total ergodicity because we need to partition the time axis into blocks. Gray and Ornstein considered a discrete-time channel with finite alphabets. In this special case the information rate capacity of Kadota and Wyner is automatically finite, which allows indeed a conclusion about the sufficiency of total ergodicity based on the approach of Kadota and Wyner, as rigorously shown in Theorem 10.3. However, for the setting of Kadota and Wyner and especially for abstract channels with time structure this conclusion is not straightforward and requires detailed analysis. In particular, we have to use the modified version of information rate capacity as introduced in Definition 5.1. Although in a somewhat different setting Pinsker (2007, p. 383)[21] also remarked that ergodicity of the channel in a certain sense allows to prove a coding theorem and he discussed an example with totally ergodic additive noise. However, additionally he has to assume a finite information rate capacity, which is not required for the approach we have taken. Note, the information rate capacity of discrete-time finite-alphabet channels is always finite. But if we consider continuous-time infinite-alphabet channels, the capacity can be infinite, even in reasonable situations, depending very much on the input constraint. See for example (Baker, 1978, p. 87) or (Baker and Ihara, 1991, p. 1314), where Gaussian channels are considered.

The weak converse in Theorem 9.3 applies to all stationary channels with time structure. No further channel properties are required. Kadota and Wyner (1972) do not give a differentiated formulation of the coding theorem and converse regarding channel properties. Again Gray and

[21] English translation, see (Pinsker, 1966) for Russian original

Ornstein (1979, p. 303) note that Kadota and Wyner's converse should hold for all stationary channels, which indeed can be shown using only the tools employed in the original publication.

Due to Definition 3.4 we have a convenient flexibility in taking the robustness of a block code w. r. t. past input signals into account (see Paragraph 3.5, Remarks 9.2 and 9.4). Input constraints are incorporated in Theorem 9.1 and Theorem 9.3 in an abstract form. In contrast to (Kadota and Wyner, 1972) we are able to accomplish this without using a standard extension (see Paragraph A.12) of the channel input σ-algebra. This is possible because we make use of outer measures directly and derive an adequate version of Feinstein's lemma in Corollary 8.3, which we believe is more transparent. In order to prove the theorems the constraints have to satisfy the regularity condition specified in (3.1.4). Kadota and Wyner characterized input constraints based on functionals as in Example 3.2. They do not mention any regularity condition although the example constraints they give satisfy the required condition. Further note that they also employ the monotonicity result of Corollary 4.14 but the proof in (Kadota and Wyner, 1972, Appendix II) is incorrect as demonstrated with the example in Paragraph 16.6. See Paragraph E.1 for a rigorous proof of a generalized statement. The mentioned monotonicity is also used in the derivations in (Kadota, 1973, 1978).

Finally, we remark that Kadota and Wyner (1972) also considered so-called incremental versions of the channel properties by using a σ-algebra of increments at the channel output. The motivation behind this extension is a mathematically rigorous treatment of continuous-time additive white noise. For the special case of real-valued output signals we can immediately apply this modification to the model used in this thesis, without any change in the proof of the coding theorem and converse.

Further historical notes. So far we have discussed the relation to the work of Kadota and Wyner (1972). We continue with more related work. First, we roughly trace the development of coding theorems and converses for channels with memory. For details on memory conditions see Section § 13. Khinchin (1957, Part II) was the first who rigorously established a coding theorem for channels with memory. He considered a discrete-time finite-alphabet stationary causal channel with finite input memory. Takano (1957) closed a gap in the proof of Khinchin by adding the condition of finite output memory. For the same model with identical assumptions Feinstein (1959) derived a coding theorem together with a weak converse and Wolfowitz (1960) added a strong converse. Feinstein extended his results also to channels with arbitrary output alphabets, so-called semi-continuous channels. See also the excellent books of Feinstein (1958) and Wolfowitz (1978) for corresponding results. Adler (1961) defined an ergodic-theoretic mixing condition for discrete-time channels and proved that it is sufficient for the ergodicity of the channel. Thereby, he showed that the coding theorems of Khinchin (1957, Part II) and Feinstein (1959) also hold, when the finite channel output memory condition is relaxed to the introduced mixing condition, which applies also to channels with infinite output memory. A further generalization of the coding theorem w. r. t. input memory and causality was achieved by Pfaffelhuber (1971), who considered channels with asymptotically decreasing input memory and anticipation. Gray and Ornstein (1979) proved a coding theorem under an even more relaxed condition on input memory and anticipation. This condition is called $\bar{d}$-continuity and is less restrictive than the asymptotic input-memorylessness required in the coding theorem derived in Section § 9. However, it is only applicable to discrete-time finite-alphabet channels and the coding theorem of Gray and Ornstein is proved for this special case. Another version of the theorem is derived in (Gray, 2011, Th. 14.1), where stationarity is weakened to asymptotic mean stationarity. But note, the result is a "one-shot" theorem, which assumes that the channel

is used only once, whereas we consider the repeated transmission of codewords. Kieffer (1981) introduced the class of stationary weakly continuous channels. He showed that it includes all stationary $\bar{d}$-continuous channels and still allows to prove a coding theorem. However, only in the sense of a joint source/channel coding theorem, whereas we focus on a pure channel coding theorem in this thesis.

An extension of Khinchin's results to discrete-time channels with continuous alphabets and finite capacity was obtained by Rosenblatt-Roth (1964, 1967). Jacobs (1962b) further developed the results of Rosenblatt-Roth (1964) and added a weak converse. A coding theorem and a weak converse for discrete-time channels with abstract alphabets were also derived by Wagner (1968), who considered the stationary memoryless case (see Example 13.8). In contrast to all previously mentioned references Wagner took input constraints into account and used Gallager's error exponent in his proof of the coding theorem, whereas the other authors used the maximal coding method, i. e., Feinstein's lemma. Very general results for discrete-time memoryless channels with abstract alphabets, including a strong converse, were derived by Augustin (1966) and Kemperman (1969). Dobrushin (1961, 1963) developed a generalized coding theorem for discrete-time channels from the point of view proposed by Kolmogorov (1956a). He showed that information stability of the channel is sufficient for the validity of the coding theorem. Ding (1962, 1964) proved that this theoretical condition is also necessary when the capacity expression of Dobrushin is employed. Since information stability is often difficult to check for concrete channels, the work on channels satisfying various memory properties has grown considerable. In (Gray and Ornstein, 1979) further historical remarks on channels with memory are given. See also (Kotz, 1966) as a recommendable survey paper for related references with a mathematical orientation. A general formula for the coding capacity of a discrete-time channel, including nonstationary and information unstable channels with abstract alphabets, is given by Verdú and Han (1994). The result is based on the so-called information spectrum method. The book of Han (2003) is a well-known reference following this approach.

Let us continue with historical remarks on coding theorems and converses for continuous-time channels. The classical information-theoretic work on continuous-time transmission models is devoted to the additive Gaussian noise channel. The main objective was to give rigorous derivations of and operational significance to Shannon's celebrated capacity formula "$W\log(1+P/NW)$", heuristically derived in (Shannon, 1948, Part IV). Ash (1963, 1964) proved a coding theorem and a weak converse for the stationary continuous-time Gaussian channel under an average energy constraint and Yoshihara (1964) obtained a strong converse. Wyner (1966) introduced different mathematical models for the case of strictly band-limited signals in white Gaussian noise and derived a coding theorem and a weak converse for each model. Gallager (1968, Ch. 8) introduced a widely accepted model of a continuous-time filtered channel with additive Gaussian noise incorporating certain input constraints in power and frequency. Following the approach of Holsinger (1964) he derived a capacity formula and a "one-shot" coding theorem, omitting the effect of interference between successive codewords. Cordaro and Wagner (1970) showed that the theorem still holds when the interference from previous codewords is taken into account. Wyner (1971) found weaker conditions and a more elementary proof for the result of Cordaro and Wagner. Baker (1978) and McKeague (1981) determined the information capacity of the stationary Gaussian channel subject to a generalized energy constrained. Baker (1987) and Baker and Ihara (1991) further developed these results such that they apply to practically relevant classes of input signals not fitting the model of Gallager (1968, Ch. 8). Baker (1991a,b) demonstrated the operational significance of the derived quantities by

showing that the information capacity is equal to the coding capacity under peak power constraints on the codewords. Furthermore, Baker (1991a,b) derived upper bounds on the coding capacity of continuous-time channels with additive non-Gaussian noise. A well-known book on information theory for continuous-time transmission models is (Ihara, 1993). The main aim of the book is to present all the material required to derive a coding theorem of the author for continuous-time Gaussian channels with feedback. The coding theorem is further developed in (Ihara, 1994, 1999).

The work of Kadota and Wyner (1972) is a generalization of (Pfaffelhuber, 1971) to continuous-time channels with real-valued alphabets. As described above in detail the coding theorem and converse derived in this thesis further generalize (Kadota and Wyner, 1972). In contrast to the previously listed references the theorems in Section § 9 apply to general continuous-time channel models, not only to those with additive noise or additive Gaussian noise. An essential step in the publications following the path of Gallager (1968, Ch. 8) is the representation of the continuous-time channel by an infinite series of parallel discrete-time channels. Following the approach of Kadota and Wyner (1972) we are able to avoid this transformation completely. Therefore, we can treat continuous-time channels with the same methods which are applicable to discrete-time channels. It is an inherent part of our model to take interference between successive codewords into account, which is an additional advantage compared to Gallager's approach. Furthermore, we incorporate input constraints in a general and flexible form. These are important arguments in favor of the path taken in this thesis. The channel properties we require to prove the coding theorem characterize a large class of practically relevant communication models. Indeed, causality is always physically justified in the context of transmission over time. If a channel code is to be used repeatedly, stationarity is a natural assumption and total ergodicity for block-i.i.d. inputs is a weak condition on the channel output memory. However, it can be difficult to verify the asymptotic input-memorylessness for specific channels or the condition is too strong in some applications because it is based on the total variation distance. Therefore, it is worth to investigate more relaxed conditions for the channel input memory in the context of abstract channels with time structure.

Relation to (Mittelbach, 2012). The starting point of this thesis was the diploma thesis (Mittelbach, 2012) of the author in mathematical stochastics. It contains intermediate results and was prepared to be suitable as basis for further extensions and generalizations. In the diploma thesis the author analyzed the results of Kadota and Wyner (1972) and worked out relevant details from probability and information theory with an emphasis on the stochastic background. As main individual contribution the proof of the monotonicity result in (Kadota and Wyner, 1972, Appendix II) was corrected. With regard to coded information transmission only the case of finite information rate capacity was studied, for which it was shown that total ergodicity is sufficient as channel output memory condition to prove the coding theorem of Kadota and Wyner. The converse was shown to hold for stationary channels. These results were formulated and derived in the special setting of the original publication, i. e., for continuous-time channels with real-valued input and output signals and an amplitude or average energy constraint. In addition, the approach of Kadota and Wyner was followed in terms of defining information rate capacity as well as using standard extensions to incorporate input constraints. Actually, some effort was undertaken to handle standard extensions rigorously, which are obsolete now in connection with the approach of this thesis. To possibly prove the coding theorem under the same relaxed output memory condition without the assumption of a finite information rate capacity was formulated as an important open problem. As shown and described above the problem was solved

in the present thesis for the general channel with time structure with arbitrary alphabets and abstract input constraints under an even more relaxed output memory condition. The proof is based on introducing an alternative version of the information rate capacity that allows simpler derivations of more general results. Deriving a mathematically convenient representation of the information rate capacity as in (5.3.2) was also one of the formulated open problems. With the identity in (5.3.2) we are able to prove Lemma 9.6, which, in turn, is used to prove Theorem 10.5, where the different versions of information rate capacity are related. The required fundamentals from probability and information theory necessarily parallel those used in the diploma thesis. However, the material on information-theoretic measures, tools, and models was completely reworked. It is presented in concise form and was extended such that it allows a unified, transparent, and tailored derivation of the coding results in the generalized context. Background material from probability and ergodic theory is summarized in Appendices A and B with extensions related to the derivations in this thesis.

We continue with the detailed analysis of memory conditions of channels. Modified versions of these conditions have already been introduced in (Mittelbach, 2012) but only to discuss a few basic relations.

Chapter III

Memory and Mixing Conditions

In contrast to characterizing finite memory there is a great variety of alternatives to model infinite memory. In this chapter, we analyze a selection of infinite memory conditions. We first introduce well-known memory conditions for probability measures and random processes called (strong) mixing conditions. Then we extend these memory conditions to channels with time structure. We formulate the memory conditions for channels in the same way as for processes, which allows us to exploit the connections to the rich field of strong mixing conditions efficiently. We show that the various memory conditions are not equivalent in general and that they form a hierarchy in terms of a sequence of implications. For the special case of a Gaussian channel we derive an important additional implication in the opposite direction. Furthermore, we study channels that transform an input probability measure with a certain memory property into an input-output probability measure with the corresponding memory property, i. e., we study the interplay between memory conditions of measures and channels.

§ 12 Mixing Conditions for Random Processes

Based on the dependence measures introduced in Section § 7 we define memory conditions for random processes and probability measures with time structure called mixing conditions. These conditions characterize in a sense to be specified the asymptotic independence of sufficiently time-separated events (random variables). The concepts are related to the ergodic-theoretic mixing conditions (see Appendix B) but, as seen later, they are of different quality. The main source for this section is the book of Bradley (2007), which contains many results from the large field of mixing conditions. In the literature on probability theory, mixing conditions play an important role in generalizing the central limit theorem to dependent random variables. In Paragraph 17.7, we consider an example in this direction from statistical signal processing as an application in connection with channels.

After defining relevant mixing conditions, we rank the conditions and show with detailed examples that they are not equivalent. For the important special case of a second order stationary Gaussian process, we characterize mixing conditions in terms of properties of the covariance function and the spectral measure or density. Finally, we give some simple but useful results for pair processes including a condition guaranteeing the existence of the information rate, which is relevant in the context of the ergodic theorem of information theory.

This section is the basis to extend the memory concepts to channels with time structure in Section § 13. In addition, the collected material is directly applicable to integration channels studied in Section § 15, which are composed of a channel function and a noise measure.

In the rest of this section we use the notation introduced in Paragraph 1.2 to denote by $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ the product measurable spaces generated by the families $\{(X_t, \mathcal{X}_t), t \in T\}$ and $\{(Y_t, \mathcal{Y}_t), t \in T\}$ of measurable spaces for which $(X_t, \mathcal{X}_t) = (X_0, \mathcal{X}_0)$ and $(Y_t, \mathcal{Y}_t) = (Y_0, \mathcal{Y}_0)$ for all $t \in T$. Further, μ denotes a probability measure on $\mathcal{X}$ and $\xi = \{\xi_t, t \in T\}$ denotes a

random process on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$, where the random variable ξ_t has values in the measurable space $(X_t, \mathcal{X}_t)$.

(12.1) Definition (α-, β-, ψ-mixing, information regularity, finite memory). The probability measure μ is called α-mixing if for all $s \in T$

$$\lim_{t\to\infty} \alpha\big([\mathcal{X}_-^s]; [\mathcal{X}_{s+t}^+]\big) = 0.$$

If the β-dependence coefficient (ψ-dependence coefficient) is employ instead, then μ is called β-mixing (ψ-mixing). It is called information regular if for all $s \in T$

$$\lim_{t\to\infty} I\big([\mathcal{X}_-^s]; [\mathcal{X}_{s+t}^+]\big) = 0.$$

We say that μ has finite memory if for all $s \in T$ there exists a $\tau(s) \in T_0$ such that

$$\psi\big([\mathcal{X}_-^s]; [\mathcal{X}_{s+\tau(s)}^+]\big) = 0.$$

We say the process ξ is α-mixing (β-, ψ-mixing, information regular, has finite memory), if the distribution P_ξ is α-mixing (β-, ψ-mixing, information regular, has finite memory).

(12.2) Remark. In view of Remark 7.6 we have the following equivalent form of the α-mixing condition for the random process ξ: The random process ξ is called α-mixing if for all $s \in T$

$$\lim_{t\to\infty} \alpha\big(\xi_-^s; \xi_{s+t}^+\big) = 0.$$

There are corresponding versions for β-mixing, ψ-mixing, information regularity, and finite memory. Due to (7.7.i), we further have: The random process ξ has finite memory if and only if for all $s \in T$ there exists a $\tau(s) \in T_0$ such that the random variables ξ_-^s and $\xi_{s+\tau(s)}^+$ are independent. Clearly, according to (7.7.i) and (4.7.i) any of the introduced dependence coefficients can be employed to define finite memory.

If the probability measure μ or the random process ξ are stationary and any of the defining relations holds for $s = 0$, then it holds for all $s \in T$. In this most relevant case the given definitions are identical to those in (Bradley, 2007, Def. 3.1, Def. 5.1). For the non-stationary case the definitions here are somewhat less restrictive. They are chosen to be consistent with the corresponding conditions for channels introduced in Definitions 13.3 and 13.6.

If ξ is a discrete time random process that forms a Markov chain (see (A.2.iii)), then we can apply (7.7.iv) to obtain for all $k \in \mathbb{Z}$ and $n \in \mathbb{N}$

$$\alpha\big(\xi_-^k; \xi_{k+n}^+\big) = \alpha(\xi_k; \xi_{k+n+1}). \tag{1}$$

The same holds for the β- and ψ-dependence coefficient as well as for the mutual information. For the latter we use (4.7.iv) and (4.7.i) to obtain this result. The identity (1) simplifies the analysis of mixing properties of Markov chains considerably because calculations can be reduced from infinite-dimensional to two-dimensional distributions.

The α-mixing condition was introduced by Rosenblatt (1956) and is also called strong mixing condition. To avoid confusion, we do not use this name because all of the given mixing conditions are also summarized under this term. In addition, the name strongly mixing is used in

ergodic theory for a related but different concept (see Definition B.3). Actually, the difference to this concept is that the convergence for the α-mixing condition is uniform in a certain sense.

The β-mixing condition was introduced by Volkonskii and Rozanov (1959) who attributed it to Kolmogorov and coined the name absolute regularity. The ψ-mixing condition is due to Philipp (1969). The information regularity condition was first studied by Volkonskii and Rozanov (1959, 1961) and is attributed there to M. S. Pinsker. The finite memory condition is often considered with time-independent memory length. Then this property is called m-dependence. It was introduced by Hoeffding and Robbins (1948) for sequences of random variables. The term finite memory is chosen by the author in accordance to the corresponding property for channels.

We consider the ψ-mixing condition because the corresponding form for channels is implicitly used by Kadota and Wyner (1972) to prove a coding theorem. We analyze the restrictiveness of this condition in Section § 13, particularly in connection with the special case of Gaussian channels. The classical and weakest α-mixing condition has important applications in the field of statistical signal processing (see Paragraph 17.7). Furthermore, we use it as a bridge between mixing properties based on dependence measures and mixing properties in the ergodic-theoretic sense. We are interested in the β-mixing condition because it is closely related to information regularity (especially in the Gaussian case), which, in turn, is interesting due to Remark 10.4 and Corollary 12.9.

The next theorem gives a hierarchy of the various mixing conditions. With Theorem B.7 in Appendix B we can further extend this sequence of implications up to ergodicity.

(12.3) Theorem (*Relations between mixing conditions). The mixing conditions satisfy the following sequence of implications:*

$$\textit{finite memory} \Longrightarrow \psi\textit{-mixing} \Longrightarrow \textit{information regular} \Longrightarrow \beta\textit{-mixing} \Longrightarrow \alpha\textit{-mixing} \overset{(*)}{\Longrightarrow} \textit{mixing}$$

The implication marked by (∗) holds for stationary probability measures (random processes).

Proof. Except for the last, the implications follow from (7.7.v), (7.7.vi), and Definition 12.1. That in the stationary case α-mixing implies mixing in the ergodic theoretic sense (see Definition B.3) is shown in (Bradley, 2007, 2.17 (a)-(d)). See also (Bradley, 2007, 5.22) for an extended hierarchy.□

For certain classes of random processes additional implications in the opposite direction hold. A stationary Markov chain, for example, which is irreducible, aperiodic, and mixing (in the ergodic-theoretic sense) is ψ-mixing if it has a finite alphabet and β-mixing if it has a countable alphabet (Bradley, 2007, Th. 7.7, Th. 7.14). As stated in Theorem 12.5 below, for stationary Gaussian random processes some of the mixing conditions are also equivalent. However, the next examples show, that in general the reversed implications in Theorem 12.3 are not valid.

(12.4) Example (Mixing conditions are not equivalent). We give simple examples to show the non-equivalence of the mixing conditions in Theorem 12.3. Subsequently, $\xi = \{\xi_k, k \in \mathbb{Z}\}$ denotes an i.i.d.-sequence of random variables. We consider Markov chains (see Paragraph A.2) obtained from a transformation of the sequence ξ, which is made explicit by introducing a function g. The examples are presented in a unified way to emphasize that the same building blocks and construction principles result in different mixing properties.

(i) *Finite memory vs. ψ-mixing.* Suppose the random variables ξ_k have values in the set $\{0, 1\}$, i. e., they are binary (as usual we take the power set as associated σ-algebra) such that ξ is a

Bernoulli sequence. The distribution is specified by

$$\mathrm{P}(\xi_0 = 0) = \frac{1}{2}(1+\epsilon) \qquad \text{and} \qquad \mathrm{P}(\xi_0 = 1) = \frac{1}{2}(1-\epsilon) \tag{1}$$

for some fixed $\epsilon \in (0,1)$. Let g be a function on $\{0,1\} \times \{0,1\}$ with values in $\{0,1\}$ given by

$$g(y,x) = (1-x)y + x(1-y) = y \oplus x$$

for all $x, y \in \{0,1\}$, where $\oplus$ denotes addition modulo 2. Let $\eta = \{\eta_k, k \in \mathbb{Z}\}$ be a sequence of binary random variables defined by the recurrence relation

$$\eta_k = g(\eta_{k-1}, \xi_k) = \eta_{k-1} \oplus \xi_k \tag{2}$$

$$= (1-\xi_k)\eta_{k-1} + \xi_k(1-\eta_{k-1}) \tag{3}$$

for all $k \in \mathbb{Z}$. That means, we randomly generate (independently from past and future) a 0 or 1 according to the probabilities in (1). If a 0 is generated, then η_k takes the same value as η_{k-1}. Otherwise, we flip the value of η_{k-1} and assign it to η_k.

We have the Markov chain $(\eta_-^{k-2} - \eta_{k-1} - \eta_k)$ for any $k \in \mathbb{Z}$ due to the i.i.d.-property of the sequence ξ and the defining relation (2). Therefore, η is a Markov chain (see (A.2.iii)). The representation in (3) shows that the distribution of η_k is the convex combination of the distributions of η_{k-1} and $(1-\eta_{k-1})$, because the binary ξ_k serves as switch between these two random variables. If η_{k-1} is uniformly distributed on $\{0,1\}$, then $(1-\eta_{k-1})$ is uniformly distributed, which implies η_k is uniformly distributed. Therefore, the uniform distribution is the stationary marginal distribution of the Markov chain.

We denote by K the Markov kernel from $(Y_0, \mathcal{Y}_0)$ to $(Y_1, \mathcal{Y}_1)$, characterizing the invariant transition probabilities of the Markov chain η, where $Y_0 = Y_1 = \{0,1\}$ and $\mathcal{Y}_0$ and $\mathcal{Y}_1$ are the corresponding power sets. From the distribution of ξ_k and from (3) we obtain

$$K(y_0, \cdot) = \frac{1}{2}(1+\epsilon)\delta_{y_0}(\cdot) + \frac{1}{2}(1-\epsilon)\delta_{(1-y_0)}(\cdot)$$

for all $y_0 \in Y_0$. Given $n \in \mathbb{N}$ the Markov kernel K_n from $(Y_0, \mathcal{Y}_0)$ to $(Y_n, \mathcal{Y}_n) = (Y_1, \mathcal{Y}_1)$ with

$$K_n(y_0, \cdot) = \frac{1}{2}(1+\epsilon^n)\delta_{y_0}(\cdot) + \frac{1}{2}(1-\epsilon^n)\delta_{(1-y_0)}(\cdot) \tag{4}$$

for all $y_0 \in Y_0$ characterizes the n-step transition probabilities. This is obtained from

$$\eta_n = \eta_0 \oplus \big(\xi_1 \oplus \xi_2 \oplus \ldots \oplus \xi_n\big) \tag{5}$$

and the i.i.d.-property of ξ, where (5) is the result of the repeated use of (2).

The Markov kernel in (4) and the uniform distribution of η_0 yield the joint distribution of (η_0, η_n). Using (7.6.6) we can directly calculate from this joint distribution the ψ-dependence coefficient

$$\psi(\eta_0; \eta_n) = \epsilon^n,$$

which converges to 0 as $n \to \infty$ because $\epsilon \in (0,1)$. In view of the comments in Remark 12.2 on simplifications of mixing conditions for Markov chains and stationarity processes it follows

that the stationary Markov chain η is ψ-mixing. However, $\psi(\eta_0;\eta_n)$ is positive for all $n \in \mathbb{N}$ so that η does not satisfy the finite memory condition. The Markov chain η is the same as that considered in (Bradley, 2007, Exm. 7.9) to illustrate various mixing properties.

(ii) *ψ-mixing vs. information regularity.* Assume now that the random variables ξ_k of the i.i.d.-sequence $\xi = \{\xi_k, k \in \mathbb{Z}\}$ are real-valued (as usual the σ-algebra associated to $\mathbb{R}$ is the corresponding Borel-σ-algebra) with expectation and variance

$$\mathrm{E}(\xi_0) = 0 \qquad \text{and} \qquad \mathrm{var}(\xi_0) = \sigma^2, \tag{6}$$

respectively, for some $\sigma^2 > 0$. Let ρ be a real constant satisfying $|\rho| < 1$ and $\rho \neq 0$. Further, let g denote the real-valued function on $\mathbb{R} \times \mathbb{R}$ given by

$$g(y, x) = \rho y + x$$

for all $x, y \in \mathbb{R}$. We define the sequence $\eta = \{\eta_k, k \in \mathbb{Z}\}$ of real-valued random variables by the recurrence relation

$$\eta_k = g(\eta_{k-1}, \xi_k) = \rho\eta_{k-1} + \xi_k \tag{7}$$

for all $k \in \mathbb{Z}$. The sequence η is called an autoregressive (AR) process of order 1. It is actually the result of filtering the i.i.d.-sequence ξ ("white noise") by the most simple IIR-filter, i. e., a filter with one feedback link. See Paragraph C.3 in Appendix C for more details on AR processes (of order 1) and Example C.4 for a continuous-time version of this example.

The same arguments as in the previous example (i) show that η is a Markov chain. From (7) we obtain the explicit representation

$$\eta_k = \sum_{i=0}^{\infty} \rho^i \xi_{k-i}, \tag{8}$$

which is well defined because the series converges in mean square and almost surely (Brockwell and Davis, 2006, Prop. 3.1.1) due to the properties of the sequence ξ and because $|\rho| < 1$.

Using (8) we obtain for all $k, n \in \mathbb{Z}$ the expectation and covariance

$$\mathrm{E}(\eta_k) = 0 \qquad \text{and} \qquad \mathrm{cov}(\eta_k, \eta_{k+n}) = \frac{\sigma^2}{1-\rho^2}\rho^{|n|}. \tag{9}$$

This result is based on exchanging expectation and summation (dominated convergence), the i.i.d.-property of ξ together with (6), and the convergence of the involved geometric series for $|\rho| < 1$. As a consequence of (9) the sequence η is wide-sense stationary.

Let us additionally assume that ξ is a Gaussian random sequence. Then η is also a Gaussian random sequence because it is obtained by a linear transformation of ξ. Gaussian wide-sense stationary sequences are stationary so that η is a stationary Gaussian Markov chain. Please note, it even holds, that a stationary Gaussian random sequence forms a Markov chain if and only if it has a covariance function of the exponential form as given in (9) (see (Hida and Hitsuda, 2007, p. 30) and (Ihara, 1993, Th. 2.3.3)). According to (9) the correlation coefficient

$$\mathrm{cor}(\eta_0, \eta_n) = \rho^n$$

is nonzero for all $n \in \mathbb{N}$ because ρ is assumed to be nonzero. This implies, together with (7.7.vii),

$$\psi(\eta_0; \eta_n) \geq 2$$

for all $n \in \mathbb{N}$. Therefore, the sequence η is not ψ-mixing. However, the mutual information $I(\eta_0; \eta_n)$ is given by

$$I(\eta_0; \eta_n) = -\frac{1}{2} \log(1 - \rho^{2n}),$$

which is obtained, e. g., from (7.1.1) and Example 6.12. Just take (η_0, η_n) in Example 6.12 as the first random vector. The second random vector has the same marginals but independent components. Since $I(\eta_0; \eta_n)$ converges to 0 as $n \to \infty$ the stationary Markov chain η is information regular. The sequence η is an explicit example of a stationary Gaussian sequence with rational spectral density for which information regularity holds according to (12.5.i) and (12.5.iii).

(iii) *Information regularity vs. β-mixing.* Assume that $\xi = \{\xi_k, k \in \mathbb{Z}\}$ is the i.i.d.-sequence of binary random variables specified in example (i) with the parameter ϵ now taken from $[0, 1)$. Let $\zeta = \{\zeta_k, k \in \mathbb{Z}\}$ be another i.i.d.-sequence on $(\Omega, \mathcal{F}, \mathrm{P})$, which is independent of ξ. Suppose the random variables ζ_k have values in the interval $[0, 1]$ (equipped with the usual Borel-σ-algebra) and their distribution is given by the uniform distribution on $[0, 1]$ denoted by the measure λ. Let us define the function g on $[0, 1] \times \{0, 1\} \times [0, 1]$ with values in the interval $[0, 1]$ by

$$g(y, x, z) = (1 - x)y + xz$$

for all $x \in \{0, 1\}$ and $y, z \in [0, 1]$. With the recurrence relation

$$\eta_k = g(\eta_{k-1}, \xi_k, \zeta_k) = (1 - \xi_k)\eta_{k-1} + \xi_k \zeta_k \tag{10}$$

we define the sequence $\eta = \{\eta_k, k \in \mathbb{Z}\}$ of random variables. That means, we randomly generate (independent from past and future) a 0 or 1 according to the probabilities in (1). If a 0 is generated, then η_k takes the same value as η_{k-1}. Otherwise, we assign to η_k a value from the interval $[0, 1]$, randomly generated (independent from past and future and from the first random experiment) according to a uniform distribution.

The recurrence relation in (10) is similar to that in (3), only $(1-\eta_{k-1})$ is replaced by ζ_k. We can therefore apply the same arguments as used in example (i) to show that η is a Markov chain with stationary marginal distribution given by the uniform distribution on the interval $[0, 1]$. Here, we take the uniform distribution of the random variables ζ_k and the independence of ξ and ζ into account. If $Y_0 = Y_1 = [0, 1]$ and $\mathcal{Y}_0$ and $\mathcal{Y}_1$ denote the corresponding Borel-σ-algebras, then the Markov kernel K from $(Y_0, \mathcal{Y}_0)$ to $(Y_1, \mathcal{Y}_1)$ with

$$K(y_0, \cdot) = \frac{1}{2}(1 + \epsilon)\delta_{y_0}(\cdot) + \frac{1}{2}(1 - \epsilon)\lambda(\cdot)$$

for all $y_0 \in [0, 1]$ characterizes the invariant transition probabilities of the Markov chain η. This follows from (10) and the distributions of ξ_k and ζ_k. The n-step transition probabilities are described for all $n \in \mathbb{N}$ by the Markov kernel K_n from $(Y_0, \mathcal{Y}_0)$ to $(Y_n, \mathcal{Y}_n) = (Y_1, \mathcal{Y}_1)$, where

$$K_n(y_0, \cdot) = \left(\tfrac{1}{2}(1 + \epsilon)\right)^n \delta_{y_0}(\cdot) + \left(1 - \left(\tfrac{1}{2}(1 + \epsilon)\right)^n\right)\lambda(\cdot) \tag{11}$$

for all $y_0 \in Y_0$. Indeed, repeated use of (10) yields

$$\eta_n = \Big(\prod_{j=0}^{n-1} (1 - \xi_{n-j}) \Big) \eta_0 + \sum_{i=1}^{n} \Big(\prod_{j=0}^{n-1-i} (1 - \xi_{n-j}) \Big) \xi_i \zeta_i.$$

Therefore, we have $\eta_n = \eta_0$ if all $\xi_1, \xi_2, \ldots, \xi_n$ are 0 and otherwise we have $\eta_n = \zeta_i$ for some $i \in \{1, 2, \ldots, n\}$. Then the i.i.d.-property and the distribution of the sequence ξ and the distribution of ζ_i yield (11).

The uniform distribution of η_0 on $[0, 1]$ denoted by λ and the Markov kernel in (11) yield the joint distribution of (η_0, η_n) given by (apply (2.1.1))

$$\mathrm{P}_{\eta_0,\eta_n} = \Big(\tfrac{1}{2}(1+\epsilon) \Big)^n \lambda_{\hat{g}} + \Big(1 - \Big(\tfrac{1}{2}(1+\epsilon) \Big)^n \Big) \lambda \otimes \lambda, \tag{12}$$

where $\lambda_{\hat{g}}$ is the image measure of λ w. r. t. the function $\hat{g}$ given by $\hat{g}(y) = (y, y)$ for all $y \in [0, 1]$. Due to (7.6.2), the β-dependence coefficient is equal to half the total variation distance between the joint distribution and the product of the marginal distributions, which yields

$$\begin{aligned} \beta(\eta_0; \eta_n) &= \tfrac{1}{2} \big\| \mathrm{P}_{\eta_0,\eta_n} - \mathrm{P}_{\eta_0} \otimes \mathrm{P}_{\eta_n} \big\|_{\mathrm{tv}} \\ &= \big(\tfrac{1}{2}(1+\epsilon) \big)^n \tfrac{1}{2} \big\| \lambda_{\hat{g}} - \lambda \otimes \lambda \big\|_{\mathrm{tv}} \\ &= \big(\tfrac{1}{2}(1+\epsilon) \big)^n. \end{aligned}$$

For the last equality we used (7.6.3) and the fact that the upper bound in (6.7.2) is attained for the set $V = \{(y, y) : y \in [0, 1]\}$. Since $\beta(\eta_0; \eta_n)$ converges to zero for any $\epsilon \in [0, 1)$ as $n \to \infty$, the stationary Markov chain η is β-mixing. However, we have $\mathrm{P}_{\eta_0,\eta_n}(V) = \big(\tfrac{1}{2}(1+\epsilon) \big)^n > 0$ but $\mathrm{P}_{\eta_0} \otimes \mathrm{P}_{\eta_n}(V) = 0$ for all $n \in \mathbb{N}$. Therefore, $\mathrm{P}_{\eta_0,\eta_n}$ is not absolutely continuous w. r. t. $\mathrm{P}_{\eta_0} \otimes \mathrm{P}_{\eta_n}$ and according to Theorem 4.3 we have

$$I(\eta_0; \eta_n) = \infty$$

for all $n \in \mathbb{N}$. It follows that the Markov chain η is not information regular. This example is a slight generalization of (Bradley, 2007, Exm. 7.12), where the special case $\epsilon = 0$ was considered.

(iv) *β-mixing vs. α-mixing.* Let $\{\eta^{(i)}, i \in \mathbb{N}\}$ be an independent family of random sequences. Each sequence $\eta^{(i)} = \{\eta_k^{(i)}, k \in \mathbb{Z}\}$ consists of binary random variables on $(\Omega, \mathcal{F}, \mathrm{P})$, where the distribution of $\eta^{(i)}$ is equal to the distribution of the Markov chain constructed in example (i). For any $k \in \mathbb{Z}$, denote by $\eta_k^{(\cdot)}$ the i.i.d.-sequence $\eta_k^{(\cdot)} = \{\eta_k^{(i)}, i \in \mathbb{N}\}$ of uniformly distributed binary random variables.

We define the sequence $\eta = \{\eta_k, k \in \mathbb{Z}\}$ of random variables by

$$\eta_k = \sum_{i=1}^{\infty} \frac{1}{2^i} \eta_k^{(i)}, \tag{13}$$

which is well defined because the series converges in mean square and almost surely (Brockwell and Davis, 2006, Prop. 3.1.1). The sequence η is stationary because the sequences $\eta^{(i)}$ are stationary and the transformation of $\eta_k^{(\cdot)}$ into η_k is identical for all $k \in \mathbb{Z}$. According to (13) the random variables $\eta_k^{(1)}, \eta_k^{(2)}, \eta_k^{(3)}, \ldots$ are the digits of the binary expansion of the random

variable η_k, which has values in the interval $[0,1]$ (equipped with the usual Borel-σ-algebra). We observe that the binary expansion is unique except for countable many numbers. From the distribution of $\eta^{(i)}$ we obtain that η_k is uniformly distributed on the interval $[0,1]$. Except for P-nullsets we therefore have

$$\sigma(\eta_k) = \sigma\big(\eta_k^{(\cdot)}\big). \tag{14}$$

Together with the fact that the sequences $\eta^{(i)}$ are Markov chains we obtain from (A.2.i) and (A.2.ii) that η is a Markov chain.

From (14) and the first assertion in (7.7.ix) we obtain for all $n \in \mathbb{N}$

$$\beta(\eta_0;\eta_n) = \beta\big(\eta_0^{(\cdot)};\eta_n^{(\cdot)}\big) = 1,$$

since $\operatorname{cor}\big(\eta_0^{(1)},\eta_n^{(1)}\big) = \epsilon^n > 0$ for any $\epsilon \in (0,1)$ so that $\eta_0^{(1)}$ and $\eta_n^{(1)}$ are not independent. Therefore, the sequence η is not β-mixing. From (14) and the second assertion in (7.7.ix) we obtain for all $n \in \mathbb{N}$

$$\begin{aligned}\alpha(\eta_0;\eta_n) = \alpha\big(\eta_0^{(\cdot)};\eta_n^{(\cdot)}\big) &\le \frac{1}{4}\sup_{i\in\mathbb{N}}|\operatorname{cor}(\eta_0^{(i)},\eta_n^{(i)})| \\ &= \frac{1}{4}\operatorname{cor}\big(\eta_0^{(1)},\eta_n^{(1)}\big) = \frac{1}{4}\epsilon^n.\end{aligned}$$

Consequently, $\alpha(\eta_0;\eta_n)$ converges to 0 as $n \to \infty$, i. e., the stationary Markov chain η is α-mixing. This example is taken from (Bradley, 2007, Exm. 7.16).

(v) *α-mixing vs. mixing (in the ergodic-theoretic sense).* Assume that $\xi = \{\xi_k, k \in \mathbb{Z}\}$ is the i.i.d.-sequence of binary random variables specified in example (i) with parameter $\epsilon = 0$. Let g denote the function on $[0,1] \times \{0,1\}$ with values in the interval $[0,1]$ given by

$$g(y,x) = \frac{1}{2}(y+x)$$

for all $x \in \{0,1\}$ and $y \in [0,1]$. We define the sequence $\eta = \{\eta_k, k \in \mathbb{Z}\}$ of random variables by the recurrence relation

$$\eta_k = g(\eta_{k-1},\xi_k) = \frac{1}{2}\big(\eta_{k-1} + \xi_k\big) \tag{15}$$

for all $k \in \mathbb{Z}$, which results in the explicit representation

$$\eta_k = \sum_{i=0}^{\infty} \frac{1}{2^{i+1}}\xi_{k-i}. \tag{16}$$

From the recurrence relation (15) we obtain in the same way as in example (i) that η is a Markov chain. The repeated use of (15) yields (16). The series is well defined for the same reasons as the one in (13). Likewise, as in example (iv) we have that η_k has values in the interval $[0,1]$ (equipped with the usual Borel-σ-algebra) and it is uniformly distributed on that interval. Further, we have

$$\sigma(\eta_k) = \sigma\big(\xi_-^k\big) \tag{17}$$

with the exception of P-nullsets, due to the same arguments used to derive (14).

Let X denote the (product) space of values of ξ, i. e., the set of two-sided binary sequences, and let Y denote the (product) space of values of η, i. e., the set of two-sided real-valued sequences with components in the interval $[0,1]$. We define the function $\hat{f}$ on X with values in Y by[22]

$$\hat{f}(x) = y = \{y_k, k \in \mathbb{Z}\}, \qquad y_k = \sum_{i=0}^{\infty} \frac{1}{2^{i+1}} x_{k-i} \tag{18}$$

for all $x = \{x_k, k \in \mathbb{Z}\} \in X$. In view of (16) we obviously have $\eta = \hat{f}(\xi)$ and therefore

$$\mathrm{P}_\eta = (\mathrm{P}_\xi)_{\hat{f}},$$

i. e., the distribution P_η of η is the image measure of the distribution P_ξ of ξ w. r. t. the function $\hat{f}$. Due to (B.13.i), the invariance of $\hat{f}$ and P_ξ implies the stationarity of P_η and therefore of η. According to Lemma B.11, P_ξ is mixing (in the ergodic-theoretic sense). Using assertion c) in (B.13.iii) yields that P_η and therefore η are mixing (in the ergodic-theoretic sense).

With the exception of P-nullsets we have $\sigma(\eta_0) \subset \sigma(\eta_n)$ for all $n \in \mathbb{N}$ using (17). The σ-algebra-based version of (7.7.ii) then implies

$$\alpha(\eta_0; \eta_0) \leq \alpha(\eta_0; \eta_n).$$

Since η_0 is uniformly distributed on the interval $[0,1]$, we have

$$\Big|\mathrm{P}\Big(\big\{\eta_0 \leq \tfrac{1}{2}\big\} \cap \big\{\eta_0 \leq \tfrac{1}{2}\big\}\Big) - \mathrm{P}\big(\eta_0 \leq \tfrac{1}{2}\big)\mathrm{P}\big(\eta_0 \leq \tfrac{1}{2}\big)\Big| = \frac{1}{4}.$$

Consequently, the definition of the α-dependence coefficent yields

$$\frac{1}{4} \leq \alpha(\eta_0; \eta_n).$$

Therefore, the stationary Markov chain η is not α-mixing. This example is taken from (Bradley, 2007, Exm. 2.15).

A summary of the preceding examples is given with the subsequent table. Recall, the independent i.i.d.-sequences $\xi = \{\xi_k, k \in \mathbb{Z}\}$ and $\zeta = \{\zeta_k, k \in \mathbb{Z}\}$ are transformed into the Markov chain $\eta = \{\eta_k, k \in \mathbb{Z}\}$ using the recurrence relation g. The last column lists the most restrictive mixing condition satisfied by η.

[22] The function $\hat{f}$ can be represented by component functions $\hat{f}_k$, defined on X_-^k by the right-hand side of (18). The function f_k is $\mathcal{X}_-^k/\mathcal{Y}_k$-measurable due to the results in Paragraph A.13, because the series in the definition converges for all $x_-^k = \{x_{k-i}, i \in \mathbb{N}_0\} \in X_-^k$. Then it follows from the derivations in Paragraph E.6 that the function $\hat{f}$ is $\mathcal{X}/\mathcal{Y}$-measurable.

	ξ, ζ	$g(y,x,z)$	η_k	mixing property η
(i)	ξ Bernoulli, $\mathrm{P}(\xi_0 = 0) = (1+\epsilon)/2$, $\epsilon \in (0,1)$	$(1-x)y + x(1-y)$	$g(\eta_{k-1}, \xi_k)$	ψ-mixing
(ii)	ξ Gaussian, $\mathrm{E}(\xi_0) = 0$	$\alpha y + x$, $\lvert\alpha\rvert < 1$, $\alpha \neq 0$	$g(\eta_{k-1}, \xi_k)$	information regular
(iii)	ξ as in (i), ζ with ζ_0 uniform on $[0,1]$	$(1-x)y + xz$	$g(\eta_{k-1}, \xi_k, \zeta_k)$	β-mixing
(iv)	$\{\xi^{(j)}, j \in \mathbb{N}\}$ independent family, $\xi^{(j)}$ as in (i)	$(1-x)y + x(1-y)$	$\sum_{j=1}^{\infty} \eta_k^{(j)}/2^j$, $\eta_k^{(j)} = g(\eta_{k-1}^{(j)}, \xi_k)$	α-mixing
(v)	ξ as in (i) with $\epsilon = 0$	$(y+x)/2$	$g(\eta_{k-1}, \xi_k)$	mixing (ergodic-theoretic)

Please note, in the examples (i), (ii), and (v) the Markov chain η is based on very similar recurrence relations. However, the resulting mixing properties are quite different. Also note, even though the sequence η is a stationary Markov chain with marginal distribution equal to the uniform distribution on the interval $[0,1]$ throughout the examples (iii), (iv), and (v) the mixing properties are of different quality.

The next theorem states that some of the mixing conditions are equivalent for the important special case of second order stationary Gaussian processes. For such processes we formulate properties of the covariance function and the spectral measure or density that imply or characterize certain mixing properties of the process. For details on second order random processes, spectral representation, and rational spectral densities including the terminology see Appendix C. Relevant material on mixing conditions in the ergodic-theoretic sense is given in Appendix B.

(12.5) Theorem (*Mixing conditions for stationary Gaussian processes). Let $\xi = \{\xi_t, t \in T\}$ be a real-valued second order stationary Gaussian process. In the continuous-time case suppose ξ is mean-square continuous.*

(i) The following mixing conditions are equivalent:

$$\begin{aligned} \textit{finite memory} &\iff \psi\textit{-mixing} \\ \textit{information regular} &\iff \beta\textit{-mixing} \\ \textit{weakly mixing} &\iff \textit{totally ergodic} \iff \textit{ergodic} \end{aligned}$$

(ii) The process ξ has finite memory if and only if there exists a $t_0 \in T_+$ such that for all $t \geq t_0$ the covariance function γ satisfies $\gamma(t) = 0$.

(iii) The process ξ is β-mixing if it has a rational spectral density.

(iv) For the process ξ to be α-mixing the existence of a spectral density is necessary. The discrete-time process ξ is α-mixing if it has a spectral density, which is continuous and positive on the whole interval $(-\pi, \pi]$. The continuous-time process ξ is α-mixing if it has a spectral density φ, which is uniformly continuous, positive on the whole real line, and satisfies the inequality

$$c_1/u^m \leq \varphi(u) \leq c_2/u^{m-1}$$

for all sufficiently large $u \in \mathbb{R}$ and some positive constants $c_1, c_2 \in \mathbb{R}$ and $m \in \mathbb{N}$.

(v) The process ξ is mixing (in the ergodic-theoretic sense) if and only if the covariance function γ satisfies

$$\lim_{t\to\infty} \gamma(t) = 0.$$

In particular, ξ is mixing (in the ergodic-theoretic sense) if it has a spectral density.

(vi) The process ξ is weakly mixing if and only if the spectral measure σ is continuous, i. e., if $\sigma(\{u\}) = 0$ for all $u \in (-\pi, \pi]$ in the discrete- and for all $u \in \mathbb{R}$ in the continuous-time case.

Proof. Part (i). The first equivalence follows from (7.7.ii), (7.7.vii), and (A.6.i). See also (Bradley, 2007, Th. 9.7 (II)). The second equivalence was derived by Ibragimov and Rozanov (1970). See also (Ibragimov and Rozanov, 1978, p. 128) and (Bradley, 1983, Th. A).

In (Cornfeld et al., 1982, Sec. 8.2) it is shown that a continuous spectral measure is necessary for ergodicity. Together with the assertion in part (vi) we obtain, that ξ is ergodic if and only if it is weakly mixing. If ξ is weakly mixing, then it is totally ergodic due to Theorem B.7. For the continuous-time case this implication requires ξ to be continuous in the sense of Pinsker (see Definition B.5), which follows from the assumed mean-square continuity (see Remark B.6). According to Theorem B.7 total ergodicity implies ergodicity so that we can conclude total ergodicity and ergodicity are also equivalent.

Part (ii). This equivalence follows by the same arguments as the first equivalence in part (i).

Part (iii). In the discrete-time case the result follows from (Ibragimov and Rozanov, 1978, Ch. IV, Th. 8 and Lem. 6). In the continuous-time case the result is shown in (Pinsker, 1964, Th. 10.1.1). It also follows from (Ibragimov and Rozanov, 1978, Ch. IV, Th. 9).

Part (iv). The conditions for α-mixing are derived in (Kolmogorov and Rozanov, 1960).

Part (v). For the discrete-time case the characterization is shown in Cornfeld et al. (1982, Sec. 14.2, Th. 2) with a comment on pages 192 and 356 that the result carries over to the continuous-time case. An explicit proof of the characterization for the continuous-time case is given in Maruyama (1949, Th. 9 (ii)). That the existence of a spectral density is sufficient is shown in Cornfeld et al. (1982, p. 371) for the discrete-time case and in Itô (1944) and Maruyama (1949, Th. 8) for the continuous-time case.

Part (vi). The characterization is shown in Cornfeld et al. (1982, Sec. 14.2, Th. 1) for the discrete-time case and in Maruyama (1949, Th. 9 (i)) for the continuous-time case. □

(12.6) Remark. Please note that the required mean-square continuity in the continuous-time case is only a weak restriction, which already holds if the covariance function is continuous at $t = 0$ (see Paragraph C.1).

Not all the results collected in Theorem 12.5 are given in the most general form possible. One easily obtains, for example, that the first equivalence of (12.5.i) extends to vector-valued and complex-valued Gaussian processes. The results taken from (Maruyama, 1949), (Cornfeld et al., 1982), or (Ibragimov and Rozanov, 1978) directly apply to complex-valued processes. Ibragimov and Rozanov (1978, Ch. IV, Th. 8 and Th. 9) even give a characterization of information regularity in spectral terms and not just sufficient conditions. Also the sufficient conditions for α-mixing given by Kolmogorov and Rozanov (1960, Th. 4) are more general then those in Theorem 12.5. Furthermore, the result taken form (Pinsker, 1964, Th. 10.1.1) on rational spectral densities is formulated there for vector-valued processes. As shown in (Rosinski and Zak, 1997) the equivalence between weak mixing and ergodicity holds for all infinitely divisible processes.

If a probability measure on the product space $\mathcal{X} \otimes \mathcal{Y}$ possesses a specific mixing property, then the marginal measures on $\mathcal{X}$ and $\mathcal{Y}$ possess the same property. Based on this simple observation we obtain together with (4.7.vi) and the σ-algebra-based version of (7.7.iii) the following result.

(12.7) Lemma (*Mixing conditions for product measures). Let μ and ν be probability measures on the product-σ-algebras $\mathcal{X}$ and $\mathcal{Y}$, respectively. The product measure $\mu \otimes \nu$ is α-mixing (β-mixing, information regular, ψ-mixing, has finite memory) if and only if μ and ν are both α-mixing (β-mixing, information regular, ψ-mixing, have finite memory).*

(12.8) Remark. The corresponding formulation for random processes reads as follows. Let $\xi = \{\xi_t, t \in T\}$ and $\eta = \{\eta_t, t \in T\}$ be random processes on $(\Omega, \mathcal{F}, \mathrm{P})$. If ξ and η are independent, then the pair process (ξ, η) is α-mixing (β-mixing, information regular, ψ-mixing, has finite memory) if and only if ξ and η are both α-mixing (β-mixing, information regular, ψ-mixing, have finite memory).

In view of Lemma 4.12 and Theorem 8.4 we make the following observation for information regular processes.

(12.9) Corollary. *Let $\xi = \{\xi_t, t \in T\}$ and $\eta = \{\eta_t, t \in T\}$ be random processes on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ such that the pair process $(\xi, \eta) = \{(\xi_t, \eta_t), t \in T\}$ is stationary. In the continuous-time case assume that ξ and η are continuous in the sense of Pinsker (see Definition B.5). If ξ is information regular, then the information rate $\bar{I}(\xi; \eta)$ exists. If in addition the pair process (ξ, η) is ergodic and the information rate $\bar{I}(\xi; \eta)$ is finite, then (8.4.1) in the ergodic theorem of information theory holds.*

§ 13 Memory and Mixing Conditions for Channels

Output memory conditions imply the ergodicity of a channel with time structure as required in Theorem 9.1. Finite output memory is a classical condition of this type, first considered by Takano (1957), Feinstein (1958, Ch. 6), Wolfowitz (1960), and implicitly by Khinchin (1957, Part II, Ch. III). As a generalization Adler (1961) introduced infinite memory conditions based on mixing properties in the ergodic-theoretic sense. In this section we extend the dependence coefficient-based infinite memory conditions introduced for random processes in Section § 12 to channels with time structure. These mixing conditions lie between the finite memory condition and the mixing conditions of Adler (1961). On the one hand, it can be useful to have various sufficient conditions to verify the ergodicity of a channel. On the other hand, there are applications (see the comments preceding Paragraph 17.7 and the example therein) for which finite memory is not required but the memory condition of Adler (1961) is not restrictive enough.

Before we introduce the different channel output memory conditions, we define a finite input memory condition, which is partly required to ensure mixing properties of the induced channel input-output probability measure analyzed at the end of this section. Then we clarify how the various mixing conditions are related. We show for the important class of Gaussian channels that the ψ-mixing condition, which in general describes infinite output memory, is in fact equivalent to finite output memory. Further, we demonstrate how the most widely used model of a memoryless channel fits into the framework of this thesis. Throughout this section κ is a channel with time structure as introduced in Definition 2.3.

(13.1) Definition (Finite input memory). The channel κ has finite input memory if for all $s \in T$ there exists a $t_I(s) \in T_0$ such that for all $B \in \mathcal{Y}_s^+$ and $x, \tilde{x} \in X$ coinciding on $(s - t_I(s), \infty)$ we have

$$\kappa(x, [B]) = \kappa(\tilde{x}, [B]).$$

For given $s \in T$ we call the smallest possible $t_I(s)$ the input memory length at time s. If $t_I(s) = 0$ for all $s \in T$, then we say the channel is input-memoryless.

(13.2) Remark. If the channel κ is stationary and the defining relation of finite input memory holds for $s = 0$, then it holds for all $s \in T$ and the input memory length does not depend on s.

Obviously, a channel with finite input memory is asymptotically input-memoryless for the entire input signal set X. Similar to the definition of asymptotic input-memorylessness, we can define the finite input memory condition on a subspace $X' \subset X$ of input signals. However, we will not make use of this version.

There is a useful equivalent characterization of a channel with finite input memory given in (Kadota, 1972): The channel κ has finite input memory if for any $s \in T$ there exists a $t_I(s) \in T_0$ such that for all $B \in \mathcal{Y}_s^+$ the function $\kappa(\cdot, [B])$ is $[\mathcal{X}_{s-t_I(s)}^+]$-measurable. This equivalent definition is obtained in the same way as the alternative characterization of causality (see Remark 2.8).

(13.3) Definition (Finite output memory). The channel κ has finite output memory if for all $s \in T$ there exists a $t_o(s) \in T_0$ such that for any $B \in \mathcal{Y}_-^s$, $\hat{B} \in \mathcal{Y}_{s+t_o(s)}^+$, and $x \in X$ we have

$$\kappa(x, [B] \cap [\hat{B}]) = \kappa(x, [B])\kappa(x, [\hat{B}]).$$

For given $s \in T$ we call the smallest possible $t_o(s)$ the output memory length at time s. If $t_o(s) = 0$ for all $s \in T$, then we say the channel is output-memoryless.

(13.4) Definition (Mixing in the ergodic-theoretic sense). The channel κ is called mixing (in the ergodic-theoretic sense), if for all $x \in X$ and any two cylinder sets $B, \hat{B} \in \mathcal{Y}$

$$\lim_{t\to\infty} \big|\kappa\big(x, B \cap \theta_t(\hat{B})\big) - \kappa\big(x, B\big)\kappa\big(x, \theta_t(\hat{B})\big)\big| = 0. \tag{1}$$

Given $s \in T_+$ the channel κ is called s-weakly mixing (in the ergodic-theoretic sense) if for all $x \in X$ and any two cylinder sets $B, \hat{B} \in \mathcal{Y}$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \big|\kappa\big(x, B \cap \theta_{ks}(\hat{B})\big) - \kappa\big(x, B\big)\kappa\big(x, \theta_{ks}(\hat{B})\big)\big| = 0 \tag{2}$$

holds. If κ is s-weakly mixing for all $s \in T_+$, then it is called totally weakly mixing (in the ergodic-theoretic sense).

(13.5) Remark. In the classical work of Khinchin (1957, Part II, Ch. III), Takano (1957), Feinstein (1958, Ch. 6), Feinstein (1959), or Wolfowitz (1960) channels with finite input and output memory are considered for discrete-time finite alphabet models to generalize Shannon's original coding theorems. See also (Gray, 2011, Sec. 2.10) and (Kakihara, 1999, Sec. 3.1).

Adler (1961) introduced the mixing conditions in the ergodic-theoretic sense for discrete-time channels, which are similar to corresponding conditions for probability measures (see Definition B.3). See also (Gray, 2011, Sec. 2.11) or (Kakihara, 1999, Sec. 3.3). The introduced weak mixing conditions are modified versions such that they are suitable in the context of this thesis. We defined s-weakly mixing and totally weakly mixing channels, however, we have not defined a weakly mixing channel. In the discrete-time case it is natural to call a 1-weakly mixing channel weakly mixing. Because a convergent sequence and all of its subsequences have the same limit, a discrete-time channel is 1-weakly mixing if and only if it is totally weakly mixing so that a simpler definition is possible. In the continuous-time case it would be natural to define a weakly mixing channel based on an integral as in (B.3.3). However, the concept of a totally weakly mixing channel based on sums is more appropriate in connection with block coding theorems for continuous-time channels, where the time axis is partitioned into intervals of equal size (see the comments at the end of Remark B.4).

For a general channel with time structure the mixing conditions constitute the least restrictive explicit infinite output memory conditions implying the ergodicity of the channel (see Theorem 13.9) required in the coding theorem in Section §9. Recall that the ergodicity of channels is defined indirectly in (2.7.ii).

Usually, the phrase "for all $x \in X$" can be replaced by "for almost all $x \in X$" where "almost" refers to probability measures from a class of relevant channel input probability measures. The definitions in (Adler, 1961) and (Kakihara, 1999, Sec. 3.3) have this form.

(13.6) Definition (α-mixing, β-mixing, ψ-mixing, information regular channels). The channel κ is called α-mixing if for all $s \in T$

$$\lim_{t\to\infty} \sup_{x\in X} \alpha\big([\mathcal{Y}_-^s]; [\mathcal{Y}_{s+t}^+] \,\big|\, x\big) = 0,$$

where $\alpha\big([\mathcal{Y}_-^s]; [\mathcal{Y}_{s+t}^+] \,\big|\, x\big)$ denotes the α-dependence coefficient of $[\mathcal{Y}_-^s]$ and $[\mathcal{Y}_{s+t}^+]$ given the underlying probability space is $(Y, \mathcal{Y}, \kappa(x,\cdot))$. If the β-dependence coefficient (ψ-dependence coefficient) is employ instead, then the channel is called β-mixing (ψ-mixing).

The channel κ is called information regular if for all $s \in T$

$$\lim_{t\to\infty} \sup_{x\in X} I\big([\mathcal{Y}_-^s]; [\mathcal{Y}_{s+t}^+] \,\big|\, x\big) = 0,$$

where $I\big([\mathcal{Y}_-^s]; [\mathcal{Y}_{s+t}^+] \,\big|\, x\big)$ denotes the mutual information between $[\mathcal{Y}_-^s]$ and $[\mathcal{Y}_{s+t}^+]$ given the underlying probability space is $(Y, \mathcal{Y}, \kappa(x,\cdot))$.

(13.7) Remark. Finite channel output memory means, sufficiently time-separated output events are independent given a fixed input. The mixing conditions of Definition 13.6, in turn, mean that for given input, future and present outputs are asymptotically independent from outputs remote in the past. As the conditions formulated in Definition 13.4, they characterize channels with infinite output memory, but of a different type. The properties in Definition 13.4 are inspired from ergodic theory, whereas the conditions in Definition 13.6 are based on the dependence measures introduced in Section §7. They are similar to the mixing concepts for probability measures and random processes defined in Definition 12.1. Note that the mixing properties for channels considered in (Mittelbach, 2012, (4.27.iv)) are defined in a different way.

Takano (1974) defined the α-mixing condition for discrete-time channels and used the term strong mixing. Kadota and Wyner (1972) introduced ψ-mixing channels under the name asymptotically output-memoryless channels. We reformulated this property using the ψ-dependence coefficient. An essential advantage of using this representation is that one can exploit more easily the connections to the rich field of strong mixing conditions. By changing the measure of dependence we obtain further classes of channels with infinite output memory. From the long list of dependence coefficients considered in Bradley (2007), we selected those that are of interest in connection with the thesis (see comments at the end of Remark 12.2). The names of the defined channel output memory properties are chosen according to the corresponding mixing properties for random processes.

Note that we have the following equivalent definition of finite output memory: The channel κ has finite output memory if for all $s \in T$ there exists a $t_o(s) \in T_0$ such that for all $t \geq t_o(s)$ we have

$$\sup_{x \in X} \psi\big([\mathcal{Y}_-^s]; [\mathcal{Y}_{s+t}^+] \,\big|\, x\big) = 0.$$

The equivalence follows from the σ-algebra based versions of (7.7.ii) and (7.7.i). Recall that the ψ-dependence coefficient is 0 if and only if the considered σ-algebras (random variables) are independent. Because the same holds for the α- and β-dependence coefficient and for the mutual information, these dependence measures can be used as well to characterize finite output memory. The asymptotic versions in Definition 13.6 are natural generalizations. As will be seen, these infinite output memory conditions are no longer equivalent.

For stationary channels we have the following simplification: If the defining relation of finite output memory (α-mixing, β-mixing, ψ-mixing, information regularity) holds for $s = 0$, then it holds for all $s \in T$. Therefore, the memory length for a stationary finite output memory channel does not depend on s. As discussed in Remark 13.5 the modification of the definitions "for almost all $x \in X$" is also possible.

(13.8) Example (Memoryless channel). Assume that κ is a discrete-time channel, i. e., we have $T = \mathbb{Z}$. Let $k \in T$ be a time index and $a_- \in X_-^{k-1}$ and $a_+ \in X_k^+$ be arbitrary but fixed one-sided input sequences. We define for all $x_k \in X_k$ and $B_k \in \mathcal{Y}_k$

$$\kappa_k(x_k, B_k) = \kappa\big((a_-, x_k, a_+), [B_k]\big).$$

Then κ_k is a Markov kernel from $(X_k, \mathcal{X}_k)$ to $(Y_k, \mathcal{Y}_k)$ due to (A.3.iii). Assume that the channel κ is causal, then κ_k does not depend on the choice of a_+. Further assume that κ is input-memoryless, i. e., it has finite input memory with input memory length of 0 for all time indices. Then κ_k does not depend on the choice of a_-. In addition, we assume that κ is output-memoryless, i. e., it has finite output memory with output memory length of 0 for all time indices. Then for all $x \in X$, $l \in T_+$, and $B_k \in \mathcal{Y}_k$ for $k \in \{-l, -(l-1), \ldots, l\}$ we have

$$\begin{aligned} \kappa\big(x, [B_{-l} \times B_{-(l-1)} \times \ldots \times B_l]\big) &= \prod_{k=-l}^{l} \kappa(x, [B_k]) \\ &= \prod_{k=-l}^{l} \kappa_k(x_k, B_k), \end{aligned}$$

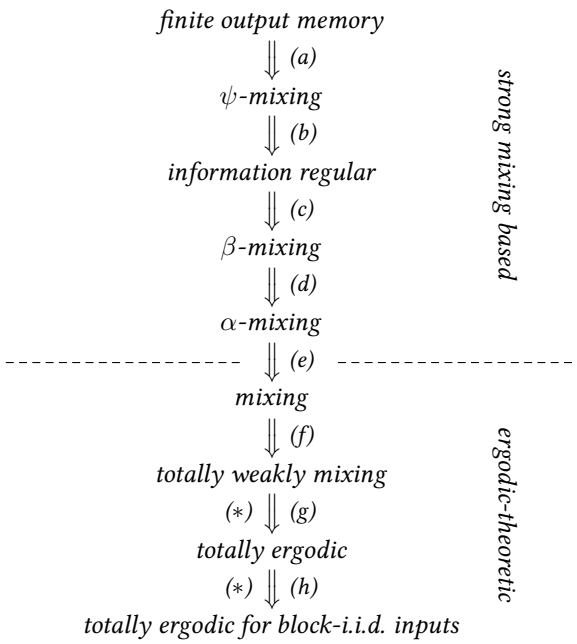
where the first equality is due to the output-memorylessness and the second equality due to the input-memorylessness and the causality. If κ is also stationary, then we have

$$\kappa\big(x, [B_{-l} \times B_{-(l-1)} \times \ldots \times B_l]\big) = \prod_{k=-l}^{l} \kappa_0(x_k, B_k). \tag{1}$$

A channel satisfying (1) for all x, l, and B_k is called a stationary memoryless channel. Usually, (1) is used as defining relation. We emphasize, that using just the single term "memoryless" means the combination of input-memoryless, output-memoryless, and causal. Due to its simplicity, this special case, often with finite alphabets, is the most widely used model in the information-theoretic literature.

The following hierarchy relates the various channel output memory conditions. See Theorem 12.3 for corresponding relations between mixing conditions for random processes. Except for (g), the examples in Paragraphs 16.2 to 16.4 show that the converse implications are not true in general. A channel that is totally ergodic but not totally weakly mixing remains to be found.

(13.9) Theorem (*Relations between mixing conditions). The following implications between mixing channels hold.*



The implications marked by (∗) hold for stationary channels.

Proof. The implications are shown one by one. (a) follows from the representation of the finite output memory condition given in Remark 13.7.

(b), (c), (d) follow from the inequalities in (7.7.v) and (7.7.vi).

To show (e), we proceed similarly to (Mittelbach, 2012, Lem. 4.18). Assume that the channel κ is α-mixing. Then for all $s \in T$ and $x \in X$ we have

$$\lim_{t\to\infty} \alpha\big([\mathcal{Y}_-^s]; [\mathcal{Y}_{s+t}^+] \,\big|\, x\big) = 0. \tag{1}$$

Let $\epsilon > 0$ and assume that $F, G \in \mathcal{Y}$ are cylinder sets. Then there exists an $s \in T$, such that $F \in [\mathcal{Y}_-^s]$. In addition, there exists a $t_0 \in T_+$ such that for all $t \in T_+$ we have $\theta_{t_0+t}(G) \in [\mathcal{Y}_{s+t}^+]$. According to (1) there exists a $t_1 \in T_+$ such that for all $t \geq t_1$

$$\big|\kappa\big(x, F \cap \theta_{t_0+t}(G)\big) - \kappa\big(x, F\big)\kappa\big(x, \theta_{t_0+t}(G)\big)\big| \leq \epsilon.$$

Since $\epsilon > 0$ was chosen arbitrary and $F, G \in \mathcal{Y}$ are arbitrary cylinder sets we obtain (e).

Now assume that κ is mixing in the ergodic-theoretic sense. Let $s \in T_+$ be arbitrary and assume that $B, \hat{B} \in \mathcal{Y}$ are arbitrary cylinder sets. Then the limit in (13.4.1) is 0, in particular if only shifts by multiples of s are considered. Since the limit of a convergent sequence is identical to the limit of its Cesàro means, (13.4.2) follows. This shows implication (f).

If κ is totally weakly mixing, then it is s-weakly mixing for all $s \in T_+$. That s-weakly mixing implies s-ergodicity for an s-stationary channel is shown in the same way as Adler (1961) has shown that weakly mixing implies ergodicity in the stationary discrete-time case. Therefore implication (g) holds.

Finally, implication (h) follows from the definitions in (2.7.ii) and the fact that an s-i.i.d. probability measure is s-stationary and s-ergodic. □

The least restrictive condition in Theorem 13.9 is what we actually need to proof the coding theorem in Section § 9. In contrast, Kadota and Wyner (1972) formulated and proved their coding theorem for ψ-mixing channels. The next theorem was given by the author in (Mittelbach and Jorswieck, 2013). It shows that for the important class of Gaussian channels (see Example 2.5) the ψ-mixing condition does actually not generalize the finite output memory condition. The first equivalence in (12.5.i) is the corresponding result for random processes. A practically relevant example of a Gaussian channel with infinite output memory is given in Paragraph 16.3. This demonstrates that the ψ-mixing condition is quite restrictive and that the coding theorem of Kadota and Wyner excludes important channel models.

(13.10) Theorem (*ψ-mixing Gaussian channels). If κ is a ψ-mixing Gaussian channel, then it has finite output memory.*

Proof. Let $\{\eta_t, t \in T\}$ denote the family of coordinate projections on the channel output space, where η_t is the projection from Y to Y_t. If κ is ψ-mixing, then according to Definition 13.6 there exists for any $s \in T$ a $t_o(s) \in T_0$, such that for all $t \geq t_o(s)$ we have

$$\sup_{x\in X} \psi\big(\eta_-^s; \eta_{s+t}^+ \,\big|\, x\big) < 2.$$

This implies together with (7.7.ii), (7.7.vii), and (A.6.i) that η_-^s and η_{s+t}^+ are independent as random variables on the probability space $(Y, \mathcal{Y}, \kappa(x, \cdot))$ for all $x \in X$. Thus, we obtain with (7.7.i)

$$\sup_{x\in X} \psi\big(\eta_-^s; \eta_{s+t}^+ \,\big|\, x\big) = 0.$$

The assertion then follows from the representation of the finite output memory condition given in Remark 13.7. □

For the mixing properties in the ergodic-theoretic sense it is known from Adler (1961), that a mixing channel transforms a mixing input probability measure into a mixing input-output probability measure. The next theorem formulates related results for the mixing properties given in Definition 12.1. Roughly speaking, the mixing property of the input probability measure is preserved if the channel possesses (at least) the corresponding mixing property, is causal, and satisfies an input memory condition.

Recall that a stationary probability measure satisfying one of the mixing conditions is ergodic due to Theorem 12.3 and Theorem B.7. In view of Corollary 12.9 we therefore obtain that Theorem 13.11 has applications in the context of Theorem 8.4, i. e., in applying the ergodic theorem of information theory. The results are also useful to analyze mixing properties of a cascade of channels considered in Section § 14.

In addition, for applications in the field of statistics or statistical signal processing it is important to know whether a probability measure (random process) satisfies a mixing condition that is more restrictive than mixing in the ergodic-theoretic sense. With Theorem 13.11 it is now possible to apply results from these fields to problems related to the processing of channel output signals. In Paragraph 17.7 we discuss an important example regarding the Fourier transform of the channel output signal. Further applications are commented before Paragraph 17.7.

(13.11) Theorem (*Mixing properties of channel input-output probability measure). Let μ be a probability measure on the input space of the channel κ and suppose $\mu\kappa$ denotes the corresponding channel input-output probability measure. Then we have the following implications: If μ has the property in the first column and κ the properties in the second column, then $\mu\kappa$ has the mixing property in the third column.*

	μ	κ	$\mu\kappa$
(i)	*finite memory*	*finite output memory, causal, finite input memory*†	*finite memory*
(ii)	ψ*-mixing*	ψ*-mixing, causal, finite input memory*†	ψ*-mixing*
(iii)	*information regular*	*information regular, causal, finite input memory*†	*information regular*
(iv)	β*-mixing*	β*-mixing, causal, asymptotically input-memoryless*†	β*-mixing*
(v)	α*-mixing*	α*-mixing, causal, asymptotically input-memoryless*†	α*-mixing*
(vi)	*stationary and mixing*	*stationary and mixing*	*stationary and mixing*
(vii)	*stationary and totally weakly mixing*	*stationary and totally weakly mixing*	*stationary and totally weakly mixing*

The superscript † *denotes we additionally assume that the possibly time-varying channel input memory lengths (for fixed tolerance ϵ) are bounded. If asymptotic input-memorylessness is assumed, then for the whole input signal set X.*

(13.12) Remark. From the hierarchies given in Theorem 12.3 and Theorem 13.9 it follows that if we replace either the mixing property of the input probability measure μ or of the channel κ by a less restrictive mixing property in one of the statements, then the input-output probability measure $\mu\kappa$ satisfies the less restrictive mixing condition. Further note, that if $\mu\kappa$ satisfies a mixing condition, then this mixing condition is in particular satisfied by the corresponding channel output probability measure.

In (13.11.ii) and (13.11.iii) the finite input memory condition can be replaced by the input memory condition given in (Mittelbach and Jorswieck, 2013, Def. IV.1), which includes the finite memory case but is more restrictive than asymptotic input-memorylessness as defined in (2.7.iv). However, to keep the proof length reasonable we restricted ourselves to the finite memory condition for these statements. For the same reason we assumed in (13.11.iv) and (13.11.v) that the channel κ is asymptotically input-memoryless for the whole input signal set X. It is sufficient to have this property for a set of input signals, which has (outer) μ-measure equal to 1.

The conditions in Theorem 13.11 on the input measure μ and the channel κ are sufficient but do not have to be necessary. An example illustrating this fact is discussed in Paragraph 17.1 below (17.1.4). The result in (13.11.v) is a generalization of (Takano, 1974, Th. 3), where discrete-time channels with finite input memory are considered. Similar to the derivations in the proof below we can generalize (Takano, 1974, Th. 2) to asymptotically input-memoryless channels.

Proof. Throughout the proof let $s \in T$ and $\epsilon > 0$ be arbitrary but fixed. The notation of Definitions 12.1 and 13.6 is used freely.

Part (i). From μ having finite memory it follows that there exists a $t_1 \in T_+$ such that for all $A_1 \in [\mathcal{X}_-^s]$ and $A_2 \in [\mathcal{X}_{s+t_1}^+]$

$$\mu(A_1 \cap A_2) = \mu(A_1)\mu(A_2). \tag{1}$$

Based on the assumption that κ has finite input memory with bounded memory lengths, we define $t_2 = \sup_{\tau\in T} t_I(\tau) < \infty$, where $t_I(\tau)$ denotes the input memory length at time τ. Due to the output-memorylessness of κ there exists a $t \geq t_1 + t_2$ such that for all $x \in X$, $B_1 \in [\mathcal{Y}_-^s]$, and $B_2 \in [\mathcal{Y}_{s+t}^+]$

$$\kappa(x, B_1 \cap B_2) = \kappa(x, B_1)\kappa(x, B_2). \tag{2}$$

For any $F \in [\mathcal{X}_-^s \otimes \mathcal{Y}_-^s]$ and $G \in [\mathcal{X}_{s+t}^+ \otimes \mathcal{Y}_{s+t}^+]$ we have

$$\begin{aligned}
\mu\kappa(F \cap G) &= \int_X \kappa(x, F_x \cap G_x)\,\mathrm{d}\mu(x) \\
&= \int_X \kappa(x, F_x)\kappa(x, G_x)\,\mathrm{d}\mu(x) \\
&= \int_X \kappa(x, F_x)\,\mathrm{d}\mu(x) \int_X \kappa(x, G_x)\,\mathrm{d}\mu(x) \\
&= \mu\kappa(F)\mu\kappa(G).
\end{aligned}$$

The definition of the channel input-output probability measure in Definition 2.1 gives the first and last equality. The second equality is due to (2). The first factor under the integral is $[\mathcal{X}_-^s]$-measurable since κ is causal (see Remark 2.8). The second factor is $[\mathcal{X}_{s+t_1}^+]$-measurable due to

the finite memory assumption (see Remark 13.2). Together with (1) and part (A.8.i) of Fubini's theorem we obtain the third equality. Therefore, we have

$$\alpha\big([\mathcal{X}_-^s \otimes \mathcal{Y}_-^s]; [\mathcal{X}_{s+t}^+ \otimes \mathcal{Y}_{s+t}^+]\big) = 0 = \psi\big([\mathcal{X}_-^s \otimes \mathcal{Y}_-^s]; [\mathcal{X}_{s+t}^+ \otimes \mathcal{Y}_{s+t}^+]\big),$$

where the second equality follows from the σ-algebra based version of (7.7.i). Since $s \in T$ was chosen arbitrarily the assertion is proved.

Part (ii). Since μ is ψ-mixing, there exists a $t_1 \in T_+$ such that

$$\psi\big([\mathcal{X}_-^s]; [\mathcal{X}_{s+t_1}^+]\big) \leq \epsilon. \tag{3}$$

Let t_2 be chosen as in the proof of part (i) above. Since κ is assumed to be ψ-mixing, there exists a $t \geq t_1 + t_2$ such that for all $x \in X$, $B_1 \in [\mathcal{Y}_-^s]$, and $B_2 \in [\mathcal{Y}_{s+t}^+]$

$$\big|\kappa(x, B_1 \cap B_2) - \kappa(x, B_1)\kappa(x, B_2)\big| \leq \epsilon\, \kappa(x, B_1)\kappa(x, B_2). \tag{4}$$

For any $F \in [\mathcal{X}_-^s \otimes \mathcal{Y}_-^s]$ and $G \in [\mathcal{X}_{s+t}^+ \otimes \mathcal{Y}_{s+t}^+]$ we have

$$\begin{aligned}
&\big|\mu\kappa(F \cap G) - \mu\kappa(F)\mu\kappa(G)\big| \\
&\quad = \left| \int_X \kappa(x, F_x \cap G_x)\, \mathrm{d}\mu(x) - \int_X \kappa(x, F_x)\, \mathrm{d}\mu(x) \int_X \kappa(x, G_x)\, \mathrm{d}\mu(x) \right| \\
&\quad \leq \left| \int_X \kappa(x, F_x \cap G_x)\, \mathrm{d}\mu(x) - \int_X \kappa(x, F_x)\kappa(x, G_x)\, \mathrm{d}\mu(x) \right| \qquad (5) \\
&\qquad + \left| \int_X \kappa(x, F_x)\kappa(x, G_x)\, \mathrm{d}\mu(x) - \int_X \kappa(x, F_x)\, \mathrm{d}\mu(x) \int_X \kappa(x, G_x)\, \mathrm{d}\mu(x) \right|, \qquad (6)
\end{aligned}$$

where we have used the definition of the channel input-output probability measure and the triangle inequality. The difference in (5) is bounded by

$$\begin{aligned}
&\int_X \big|\kappa(x, F_x \cap G_x) - \kappa(x, F_x)\kappa(x, G_x)\big|\, \mathrm{d}\mu(x) \\
&\quad \leq \epsilon \int_X \kappa(x, F_x)\kappa(x, G_x)\, \mathrm{d}\mu(x) \qquad (7) \\
&\quad \leq \epsilon \left| \int_X \kappa(x, F_x)\kappa(x, G_x)\, \mathrm{d}\mu(x) - \int_X \kappa(x, F_x)\, \mathrm{d}\mu(x) \int_X \kappa(x, G_x)\, \mathrm{d}\mu(x) \right| \qquad (8) \\
&\qquad + \epsilon \int_X \kappa(x, F_x)\, \mathrm{d}\mu(x) \int_X \kappa(x, G_x)\, \mathrm{d}\mu(x) \\
&\quad \leq \epsilon\big(\psi\big([\mathcal{X}_-^s]; [\mathcal{X}_{s+t_1}^+]\big) + 1\big)\mu\kappa(F)\mu\kappa(G) \qquad (9) \\
&\quad \leq \epsilon\big(\epsilon + 1\big)\mu\kappa(F)\mu\kappa(G),
\end{aligned}$$

where (7) follows from (4). Then we apply again the triangle inequality and the second inequality of (7.7.viii) to obtain (9). This is possible due to the causality of κ and the finite input memory assumption, which imply the $[\mathcal{X}_-^s]$-measurability of $\kappa(\cdot, F_\cdot)$ and the $[\mathcal{X}_{s+t_1}^+]$-measurability of $\kappa(\cdot, G_\cdot)$. The last inequality is a consequence of (3).

The difference in (6) is bounded just as the difference in (8) by applying the second inequality of (7.7.viii). Collecting terms yields

$$\big|\mu\kappa(F\cap G)-\mu\kappa(F)\mu\kappa(G)\big|\le\epsilon\big(\epsilon+2\big)\mu\kappa(F)\mu\kappa(G),$$

which implies

$$\psi\big([\mathcal{X}_-^s\otimes\mathcal{Y}_-^s];[\mathcal{X}_{s+t}^+\otimes\mathcal{Y}_{s+t}^+]\big)\le\epsilon\big(\epsilon+2\big).$$

Since $s\in T$ and $\epsilon>0$ were chosen arbitrarily the assertion is proved.

Part (iii). Since μ is information regular, there exists a $t_1\in T_+$ such that

$$I\big([\mathcal{X}_-^s];[\mathcal{X}_{s+t_1}^+]\big)\le\epsilon. \tag{10}$$

Let t_2 be chosen as in the proof of part (i) above. Since κ is assumed to be information regular, there exists a $t\ge t_1+t_2$ such that

$$\sup_{x\in X} I\big([\mathcal{Y}_-^s];[\mathcal{Y}_{s+t}^+]\,\big|\,x\big)\le\epsilon. \tag{11}$$

It is convenient to continue with the families $\{\xi_t,t\in T\}$ and $\{\eta_t,t\in T\}$ of coordinate projections on the channel input-output space, where ξ_t is the projection from $X\times Y$ to X_t and η_t is the projection from $X\times Y$ to Y_t. Applying the random variable-based version of (4.7.ii) and of the chain rule given in (4.7.iv) we obtain

$$\begin{aligned}
I(\xi_-^s\eta_-^s;\xi_{s+t}^+\eta_{s+t}^+)\le\quad & I(\xi_-^s\eta_-^s;\xi_{s+t_1}^+\eta_{s+t}^+)\\
=\quad & I(\xi_-^s;\xi_{s+t_1}^+) && (12)\\
&+I(\eta_-^s;\xi_{s+t_1}^+\,|\,\xi_-^s) && (13)\\
&+I(\xi_-^s;\eta_{s+t}^+\,|\,\xi_{s+t_1}^+) && (14)\\
&+I(\eta_-^s;\eta_{s+t}^+\,|\,\xi_-^s\xi_{s+t_1}^+). && (15)
\end{aligned}$$

The right-hand side of (12) is equal to the left-hand side of (10) and therefore bounded by ϵ. Due to the causality of κ we have the Markov chain $(\eta_-^s-\xi_-^s-\xi_s^+)$. Since κ has finite input memory we have the Markov chain $(\xi_-^{s+t_1}-\xi_{s+t_1}^+-\eta_{s+t_1+t_2}^+)$. Together with the random variable-based versions of (4.7.i) and (4.7.ii) we therefore obtain that (13) and (14) are equal to 0. For the conditional mutual information in (15) we have

$$\begin{aligned}
I(\eta_-^s;\eta_{s+t}^+\,|\,\xi_-^s\xi_{s+t_1}^+)&=I(\eta_-^s;\eta_{s+t}^+\,|\,\xi)\\
&=\int_X I\big([\mathcal{Y}_-^s];[\mathcal{Y}_{s+t}^+]\,\big|\,x\big)\,\mathrm{d}\mu(x)\le\epsilon,
\end{aligned}$$

where the second equality follows from the result in Example 4.9 and the inequality is due to (11). To obtain the first equality we apply the chain rule given in (4.7.iv), which yields

$$\begin{aligned}
I(\eta_-^s;\eta_{s+t}^+\,|\,\xi)=\quad & I(\eta_-^s;\eta_{s+t}^+\,|\,\xi_-^s\xi_{s+t_1}^+)\\
&+I(\xi_s^{s+t_1};\eta_-^s\eta_{s+t}^+\,|\,\xi_-^s\xi_{s+t_1}^+)-I(\xi_s^{s+t_1};\eta_-^s\,|\,\xi_-^s\xi_{s+t_1}^+)-I(\xi_s^{s+t_1};\eta_{s+t}^+\,|\,\xi_-^s\xi_{s+t_1}^+).
\end{aligned}$$

As a result of the combination of the causality and the finite input memory property of κ we have the Markov chain $\big(\xi_s^{s+t_1} - (\xi_-^s, \xi_{s+t_1}^+) - (\eta_-^s, \eta_{s+t_1+t_2}^+)\big)$. Applying (4.7.i) and (4.7.ii) then shows that all conditional mutual informations in the second row are 0. Collecting terms yields

$$I(\xi_-^s \eta_-^s; \xi_{s+t}^+ \eta_{s+t}^+) \le 2\epsilon.$$

Since $s \in T$ and $\epsilon > 0$ were chosen arbitrarily the assertion is proved.

Part (iv). Since μ is β-mixing, there exists a $t_1 \in T_+$ such that

$$\beta\big([\mathcal{X}_-^s]; [\mathcal{X}_{s+t_1}^+]\big) \le \epsilon. \tag{16}$$

Let $\dot\mu$ denote the (product) probability measure on $[\mathcal{X}_-^s \otimes \mathcal{X}_{s+t_1}^+]$ given for all $A_1 \in [\mathcal{X}_-^s]$ and $A_2 \in [\mathcal{X}_{s+t_1}^+]$ by

$$\dot\mu(A_1 \cap A_2) = \mu(A_1)\mu(A_2). \tag{17}$$

Using (7.6.2) and (7.6.4) we can rewrite (16) as

$$\|\mu - \dot\mu\|_{\mathrm{tv}} \le 2\epsilon, \tag{18}$$

where in (18) the restriction of μ on $[\mathcal{X}_-^s \otimes \mathcal{X}_{s+t_1}^+]$ is considered.

Based on the assumption that κ is asymptotically input-memoryless with bounded memory lengths for fixed ϵ, we define $t_2 = \sup_{\tau\in T} t_I(\tau,\epsilon) < \infty$, where $t_I(\tau,\epsilon)$ denotes the input memory length at time τ for the tolerance ϵ. Then for all $B \in [\mathcal{Y}_{s+t_1+t_2}^+]$ and $x, \tilde x \in X$ coinciding on $(s+t_1, \infty)$ we have

$$\big|\kappa(x,B) - \kappa(\tilde x, B)\big| \le \epsilon. \tag{19}$$

Since κ is assumed to be β-mixing, there exists a $t \ge t_1 + t_2$ such that

$$\sup_{x\in X} \beta\big([\mathcal{Y}_-^s]; [\mathcal{Y}_{s+t_1}^+]\,\big|\,x\big) \le \epsilon. \tag{20}$$

Let $\dot\kappa$ denote the (product) Markov kernel (see (A.3.iv)) from $(X, \mathcal{X})$ to $(Y, [\mathcal{Y}_-^s \otimes \mathcal{Y}_{s+t}^+])$ given for all $x \in X$, $B_1 \in [\mathcal{Y}_-^s]$, and $B_2 \in [\mathcal{Y}_{s+t}^+]$ by

$$\dot\kappa(x, B_1 \cap B_2) = \kappa(x, B_1)\kappa(x, B_2). \tag{21}$$

Based on (7.6.3) and (7.6.4) we can rewrite (20) as

$$\sup \big|\kappa(x,B) - \dot\kappa(x,B)\big| \le \epsilon, \tag{22}$$

where the supremum is taken w. r. t. all $x \in X$ and $B \in [\mathcal{Y}_-^s \otimes \mathcal{Y}_{s+t}^+]$.

Let $\dot\mu\dot\kappa$ denote the (product) probability measure on $[(\mathcal{X}_-^s \otimes \mathcal{Y}_-^s) \otimes (\mathcal{X}_{s+t}^+ \otimes \mathcal{Y}_{s+t}^+)]$ defined for all $G_1 \in [\mathcal{X}_-^s \otimes \mathcal{Y}_-^s]$ and $G_2 \in [\mathcal{X}_{s+t}^+ \otimes \mathcal{Y}_{s+t}^+]$ by

$$\dot\mu\dot\kappa(G_1 \cap G_2) = \mu\kappa(G_1)\mu\kappa(G_2). \tag{23}$$

To bring the asymptotic input-memorylessness into play let us fix an arbitrary $a \in X_s^{s+t_1}$. For all $x \in X$ we define $\tilde x(x) = (\pi_-^s(x), a, \pi_{s+t_1}^+(x))$, where π_-^s and $\pi_{s+t_1}^+$ denote the projections

from X to X_-^s and $X_{s+t_1}^+$, respectively. Then for any $F \in [(\mathcal{X}_-^s \otimes \mathcal{Y}_-^s) \otimes (\mathcal{X}_{s+t}^+ \otimes \mathcal{Y}_{s+t}^+)]$, we obtain by the triangle inequality

$$\begin{aligned}
|\mu\kappa(F) - \mu\dot{\kappa}(F)| &= \left| \int_X \kappa(x, F_x)\, \mathrm{d}\mu(x) - \mu\dot{\kappa}(F) \right| \\
&\leq \left| \int_X \kappa(x, F_x)\, \mathrm{d}\mu(x) - \int_X \dot{\kappa}(x, F_x)\, \mathrm{d}\mu(x) \right| && (24) \\
&\quad + \left| \int_X \dot{\kappa}(x, F_x)\, \mathrm{d}\mu(x) - \int_X \dot{\kappa}(\tilde{x}(x), F_x)\, \mathrm{d}\mu(x) \right| && (25) \\
&\quad + \left| \int_X \dot{\kappa}(\tilde{x}(x), F_x)\, \mathrm{d}\mu(x) - \int_X \dot{\kappa}(\tilde{x}(x), F_x)\, \mathrm{d}\dot{\mu}(x) \right| && (26) \\
&\quad + \left| \int_X \dot{\kappa}(\tilde{x}(x), F_x)\, \mathrm{d}\dot{\mu}(x) - \mu\dot{\kappa}(F) \right|. && (27)
\end{aligned}$$

The difference in (24) is bounded by

$$\int_X \left| \kappa(x, F_x)\, \mathrm{d}\mu(x) - \dot{\kappa}(x, F_x) \right| \mathrm{d}\mu(x) \leq \epsilon$$

due to (22) and the difference in (25) is bounded by

$$\int_X \left| \dot{\kappa}(x, F_x)\, \mathrm{d}\mu(x) - \dot{\kappa}(\tilde{x}(x), F_x) \right| \mathrm{d}\mu(x) \leq \epsilon$$

due to (19), the definition of $\dot{\kappa}$ in (21), and the causality of κ. Using (6.10.v) and (18) we can upper bound (26) by

$$\|\mu - \dot{\mu}\|_{\mathrm{tv}} \leq 2\epsilon.$$

We can apply (6.10.v) in this form because $\dot{\kappa}(\tilde{x}(\cdot), F_\cdot)$ is $[\mathcal{X}_-^s \otimes \mathcal{X}_{s+t_1}^+]$-measurable and bounded by 1. The measurability follows from the definition of $\tilde{x}(\cdot)$, the choice of the set F and from (A.3.iii). Finally, (27) is bounded by ϵ. Indeed, for any $G \in [\mathcal{X}_-^s \otimes \mathcal{Y}_-^s]$ and $G' \in [\mathcal{X}_{s+t}^+ \otimes \mathcal{Y}_{s+t}^+]$ we have

$$\begin{aligned}
&\left| \mu\dot{\kappa}(G \cap G') - \int_X \dot{\kappa}(\tilde{x}(x), (G \cap G')_x)\, \mathrm{d}\dot{\mu}(x) \right| \\
&= \left| \mu\kappa(G)\mu\kappa(G') - \int_X \kappa(\tilde{x}(x), G_x)\kappa(\tilde{x}(x), G'_x)\, \mathrm{d}\dot{\mu}(x) \right| && (28) \\
&= \left| \mu\kappa(G) \int_X \kappa(x, G'_x)\, \mathrm{d}\mu(x) - \int_X \kappa(\tilde{x}(x), G_x)\, \mathrm{d}\mu(x) \int_X \kappa(\tilde{x}(x), G'_x)\, \mathrm{d}\mu(x) \right| && (29) \\
&= \left| \int_X \kappa(x, G'_x)\, \mathrm{d}\mu(x) - \int_X \kappa(\tilde{x}(x), G'_x)\, \mathrm{d}\mu(x) \right| \mu\kappa(G) && (30) \\
&\leq \mu\kappa(G) \int_X \left| \kappa(x, G'_x)\, \mathrm{d}\mu(x) - \kappa(\tilde{x}(x), G'_x) \right| \mathrm{d}\mu(x) && (31) \\
&\leq \mu\kappa(G)\epsilon. && (32)
\end{aligned}$$

From the definitions of $\dot{\kappa}$ and $\mu\dot{\kappa}$ in (21) and (23) we obtain (28). The second summand in (29) follows from the definition of $\dot{\mu}$ in (17), the $[\mathcal{X}_-^s]$-measurability of $\kappa(\tilde{x}(\cdot), G_\cdot)$, the $[\mathcal{X}_{s+t_1}^+]$-measurability of $\kappa(\tilde{x}(\cdot), G'_\cdot)$, and part (A.8.i) of Fubini's theorem. The measurability, in turn, follows from the definition of $\tilde{x}(\cdot)$ and the choice of G and G'. The first factor of the second summand in (29) is equal to $\mu\kappa(G)$ due to the causality of κ, which yields (30). Then (32) is obtained from (31) using (19), i. e., the asymptotic input-memorylessness of κ. Based on these derivation we obtain for (27) the upper bound ϵ for all sets $F \in [(\mathcal{X}_-^s \otimes \mathcal{Y}_-^s) \otimes (\mathcal{X}_{s+t}^+ \otimes \mathcal{Y}_{s+t}^+)]$.

Collecting terms yields

$$\big|\mu\kappa(F) - \dot{\mu}\kappa(F)\big| \leq 5\epsilon,$$

which implies together with (7.6.3) and (7.6.4)

$$\beta\big([\mathcal{X}_-^s \otimes \mathcal{Y}_-^s]; [\mathcal{X}_{s+t}^+ \otimes \mathcal{Y}_{s+t}^+]\big) \leq 5\epsilon.$$

Since $s \in T$ and $\epsilon > 0$ were chosen arbitrarily the assertion is proved.

Part (v). Since μ is α-mixing, there exists a $t_1 \in T_+$ such that

$$\alpha\big([\mathcal{X}_-^s]; [\mathcal{X}_{s+t_1}^+]\big) \leq \epsilon. \tag{33}$$

Let t_2 be chosen as in the proof of part (iv). Then (19) holds due to the asymptotic input-memorylessness of κ. Since κ is assumed to be α-mixing, there exists a $t \geq t_1 + t_2$ such that for all $x \in X$, $B_1 \in [\mathcal{Y}_-^s]$, and $B_2 \in [\mathcal{Y}_{s+t}^+]$

$$\big|\kappa(x, B_1 \cap B_2) - \kappa(x, B_1)\kappa(x, B_2)\big| \leq \epsilon. \tag{34}$$

Let $a \in X_-^{s+t_1}$ be arbitrary but fixed. For all $x \in X$ we define $\tilde{x}(x) = (a, \pi_{s+t_1}^+(x))$, where $\pi_{s+t_1}^+$ denotes the projection from X to $X_{s+t_1}^+$. Then for any $F \in [\mathcal{X}_-^s \otimes \mathcal{Y}_-^s]$ and $G \in [\mathcal{X}_{s+t}^+ \otimes \mathcal{Y}_{s+t}^+]$ we obtain by the triangle inequality

$$\begin{aligned}
&\big|\mu\kappa(F \cap G) - \mu\kappa(F)\mu\kappa(G)\big| \\
= &\left|\int_X \kappa(x, F_x \cap G_x)\,\mathrm{d}\mu(x) - \int_X \kappa(x, F_x)\,\mathrm{d}\mu(x) \int_X \kappa(x, G_x)\,\mathrm{d}\mu(x)\right| \\
\leq &\left|\int_X \kappa(x, F_x \cap G_x)\,\mathrm{d}\mu(x) - \int_X \kappa(x, F_x)\kappa(x, G_x)\,\mathrm{d}\mu(x)\right| && (35)\\
&+\left|\int_X \kappa(x, F_x)\kappa(x, G_x)\,\mathrm{d}\mu(x) - \int_X \kappa(x, F_x)\kappa(\tilde{x}(x), G_x)\,\mathrm{d}\mu(x)\right| && (36)\\
&+\left|\int_X \kappa(x, F_x)\kappa(\tilde{x}(x), G_x)\,\mathrm{d}\mu(x) - \int_X \kappa(x, F_x)\,\mathrm{d}\mu(x) \int_X \kappa(\tilde{x}(x), G_x)\,\mathrm{d}\mu(x)\right| && (37)\\
&+\left|\int_X \kappa(\tilde{x}(x), G_x)\,\mathrm{d}\mu(x) - \int_X \kappa(x, G_x)\,\mathrm{d}\mu(x)\right| \int_X \kappa(x, F_x)\,\mathrm{d}\mu(x) && (38)
\end{aligned}$$

The difference in (35) is bounded by

$$\int_X \big|\kappa(x, F_x \cap G_x) - \kappa(x, F_x)\kappa(x, G_x)\big|\,\mathrm{d}\mu(x) \leq \epsilon$$

due to (34). The difference in (36) is bounded by

$$\int_X \kappa(x, F_x)\big|\kappa(x, G_x) - \kappa(\tilde{x}(x), G_x)\big| \,\mathrm{d}\mu(x) \leq \epsilon \int_X \kappa(x, F_x) \,\mathrm{d}\mu(x) \leq \epsilon$$

due to the asymptotic input-memorylessness of κ and (19). An upper bound for (37) is given by

$$4\alpha\big([\mathcal{X}_-^s]; [\mathcal{X}_{s+t_1}^+]\big) \leq 4\epsilon$$

using (7.7.viii) and (33). It is possible to apply (7.7.viii) in this form because $\kappa(\cdot, F_\cdot)$ is $[\mathcal{X}_-^s]$-measurable, $\kappa(\tilde{x}(\cdot), G_\cdot)$ is $[\mathcal{X}_{s+t_1}^+]$-measurable, and both random variables are bounded by 1. The measurability is a consequence of the causality of κ, the definition of $\tilde{x}(\cdot)$, and the choice of the sets F and G. Finally, we can bound (38) similar to (36) by ϵ. Collecting terms yields

$$\big|\mu\kappa(F \cap G) - \mu\kappa(F)\mu\kappa(G)\big| \leq 7\epsilon,$$

which implies

$$\alpha\big([\mathcal{X}_-^s \otimes \mathcal{Y}_-^s]; [\mathcal{X}_{s+t}^+ \otimes \mathcal{Y}_{s+t}^+]\big) \leq 7\epsilon.$$

Since $s \in T$ and $\epsilon > 0$ were chosen arbitrarily the assertion is proved.

Part (vi). This result is due to Adler (1961) for the discrete-time case. There are no significant changes in the proof for the continuous-time case.

Part (vii). For all $s \in T_+$ the proof that an s-weakly mixing input probability measure together with an s-weakly mixing channel induces an s-weakly mixing input-output probability measure is identical to that given in (Adler, 1961) or (Kakihara, 1999, p. 140) for the discrete-time case. □

Chapter IV

Channel Model Revisited

In this chapter, we consider aspects that are useful to analyze concrete channel models. We derive results for cascade channels that allow to conclude properties of a complex model from properties of basic building blocks. Furthermore, we study integration channels, for which the channel model can be decomposed into a deterministic and a random part, which is possible for many physically relevant channel models.

§ 14 Cascade Channels

In practical scenarios the physical channel often consists of multiple components connected in cascade. For example, the transmitted signal is altered by multipath fading (channel 1), then it is disturbed by additive noise at the receiver (channel 2), which quantizes the noisy signal values (channel 3). In this section, we derive properties of a cascade of channels based on the properties of component channels. These results allow to obtain properties of the overall channel by studying the properties of simpler individual blocks, which is quite useful in applications.

(14.1) Definition (Cascade channels). Let $(X, \mathcal{X})$, $(U, \mathcal{U})$, and $(Y, \mathcal{Y})$ be arbitrary measurable spaces. Assume that $\dot{\kappa}$ is a channel with input space $(X, \mathcal{X})$ and output space $(U, \mathcal{U})$ and $\ddot{\kappa}$ is a channel with input space $(U, \mathcal{U})$ and output space $(Y, \mathcal{Y})$. The cascade of $\dot{\kappa}$ and $\ddot{\kappa}$ is defined as the channel κ with input space $(X, \mathcal{X})$ and output space $(Y, \mathcal{Y})$ given by

$$\kappa(x, B) = \int_U \ddot{\kappa}(\cdot, B) \, \mathrm{d}\dot{\kappa}(x, \cdot) \tag{1}$$

for all $x \in X$ and $B \in \mathcal{Y}$.

The channel κ in Definition 14.1 is obtained by connecting the channels $\dot{\kappa}$ and $\ddot{\kappa}$ in cascade as illustrated in Figure 3. Due to (A.3.ii) κ is indeed a channel (Markov kernel). The definition is canonical. It is also given in Gray (2011).

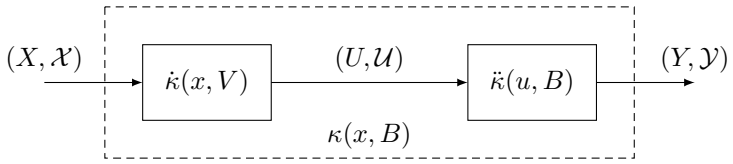


Figure 3: Cascade channel.

The next theorem considers properties relevant for Theorem 9.1.

(14.2) Theorem (*Properties of cascade channels with time structure). Consider the situation of Definition 14.1. Assume that $\dot\kappa$ and $\ddot\kappa$ (and therefore the cascade channel κ) are channels with time structure, where $(X,\mathcal{X})$ and $(Y,\mathcal{Y})$ are defined as in Definition 2.3. The product measurable space $(U,\mathcal{U})$ is defined correspondingly based on the family $\{(U_t,\mathcal{U}_t), t\in T\}$ of measurable spaces with $(U_t,\mathcal{U}_t) = (U_0,\mathcal{U}_0)$. Then we have the following implications: If $\dot\kappa$ has the property in the first column and $\ddot\kappa$ the property in the second column, then the cascade channel κ has the property in the third column.*

	$\dot\kappa$	$\ddot\kappa$	κ
(i)†	*(s-) stationary*	*(s-) stationary*	*(s-) stationary*
(ii) a)†	*s-stationary and s-ergodic (for s-i.i.d. inputs)*	*s-stationary and s-ergodic*	*s-stationary and s-ergodic (for s-i.i.d. inputs)*
b)	*stationary and totally ergodic (for block-i.i.d. inputs)*	*stationary and totally ergodic*	*stationary and totally ergodic (for block-i.i.d. inputs)*
(iii)	*causal*	*causal*	*causal*
(iv)	*asymptotically input-memoryless for $X'\subset X$*	*asymptotically input-memoryless‡ for $U'\subset U$*	*asymptotically input-memoryless for $X'\subset X$*

The superscript † denotes we assume that $s\in T$ in (i) *and $s\in T_+$ in* (ii). *The superscript ‡ denotes we additionally assume that U' is a product set, i. e., $U' = \times_{t\in T} U'_t$ with $U'_t\subset U_t$, and the outer $\dot\kappa(x,\cdot)$-measure of U' is equal to 1 for all $x\in X'$.*

(14.3) Remark. Note, the condition on $\ddot\kappa$ in (14.2.ii) is more restrictive than on $\dot\kappa$. The additional conditions on U' in (14.2.iv) ensure the signal set w. r. t. which the second channel is asymptotically input-memoryless is "large enough" w. r. t. the first channel $\dot\kappa$. The rectangular structure of U' is sufficient but not necessary. A simple example satisfying all conditions is $U' = U$.

Proof. Part (i). Assume that $s\in T$ and $\dot\kappa$ and $\ddot\kappa$ are s-stationary. Then for all $x\in X$ and $B\in\mathcal{Y}$ we have

$$\begin{aligned}
\kappa\big(\theta_s(x),\theta_s(B)\big) &= \int_U \ddot\kappa\big(\cdot,\theta_s(B)\big)\,\mathrm{d}\dot\kappa\big(\theta_s(x),\cdot\big)\\
&= \int_{\theta_{-s}(U)} \ddot\kappa\big(\theta_{-s}(\cdot),B\big)\,\mathrm{d}\dot\kappa\big(\theta_s(x),\cdot\big)\\
&= \int_U \ddot\kappa\big(\cdot,B\big)\,\mathrm{d}\Big(\dot\kappa\big(\theta_s(x),\cdot\big)\Big)_{\theta_{-s}}\\
&= \int_U \ddot\kappa\big(\cdot,B\big)\,\mathrm{d}\dot\kappa(x,\cdot)\\
&= \kappa(x,B),
\end{aligned}$$

i. e., the channel κ is s-stationary according to (2.7.i). The first and last equality are due to the defining relation for cascade channels in (14.1.1). The second equality follows from the s-stationarity of $\ddot{\kappa}$ and the shift invariance of U. Applying the substitution rule for integrals, we obtain the third equality, where $(\dot{\kappa}(\theta_s(x),\cdot))_{\theta_{-s}}$ denotes the distribution of θ_{-s}, considered as random variable on the probability space $(U,\mathcal{U},\dot{\kappa}(\theta_s(x),\cdot))$. For the fourth equality, we have used the s-stationarity of $\dot{\kappa}$. From what is shown the assertion, where s-stationarity is replaced by stationarity follows immediately. A similar proof is given in (Gray, 2011, Lem. 2.10).

Part (ii). Let $s \in T_+$ and assume that $\dot{\kappa}$ as well as $\ddot{\kappa}$ are s-stationary and s-ergodic. Further, let μ be an s-stationary s-ergodic probability measure on $\mathcal{X}$. We consider the product measurable space $(X \times U \times Y, \mathcal{X} \otimes \mathcal{U} \otimes \mathcal{Y})$ and the joint measure ν on $\mathcal{X} \otimes \mathcal{U} \otimes \mathcal{Y}$, which is induced by μ and the cascade of the channels $\dot{\kappa}$ and $\ddot{\kappa}$, i. e., ν is defined for any $C \in \mathcal{X} \otimes \mathcal{U} \otimes \mathcal{Y}$ by

$$\nu(C) = \int_X \left[\int_U \ddot{\kappa}\big(u, (C_x)_u\big) \, \mathrm{d}\dot{\kappa}(x,u) \right] \mathrm{d}\mu(x),$$

where $(C_x)_u$ is the u-section of C_x and C_x is the x-section of C. The marginal measure of ν on $\mathcal{X} \otimes \mathcal{U}$ is equal to the input-output probability measure $\mu\dot{\kappa}$ of the channel $\dot{\kappa}$ and by construction we have the Markov chain $([\mathcal{X}] - [\mathcal{U}] - [\mathcal{Y}])$. Thus, the measure ν is also given for any $C \in \mathcal{X} \otimes \mathcal{U} \otimes \mathcal{Y}$ by

$$\nu(C) = \int_{X \times U} \ddot{\kappa}\big(u, (C_x)_u\big) \, \mathrm{d}\mu\dot{\kappa}(x,u), \tag{1}$$

where $\ddot{\kappa}$ is considered here as Markov-kernel from $(X \times U, \mathcal{X} \otimes \mathcal{U})$ to $(Y, \mathcal{Y})$. The probability measure $\mu\dot{\kappa}$ is s-stationary and s-ergodic because the measure μ and the channel $\dot{\kappa}$ are s-stationary and s-ergodic. Together with the s-stationarity and s-ergodicity of the channel $\ddot{\kappa}$ and the representation in (1) we obtain that ν is s-stationary and s-ergodic. Since the input-output probability measure $\mu\kappa$ of the cascade channel κ is equal to the marginal measure of ν on $\mathcal{X} \otimes \mathcal{Y}$, it is also s-stationary and s-ergodic. Therefore, according to (2.7.ii) we have shown that the cascade channel κ is s-stationary and s-ergodic if the channels $\dot{\kappa}$ and $\ddot{\kappa}$ are both s-stationary and s-ergodic. The assertion, where the s-ergodicity of $\dot{\kappa}$ and κ are weakened to s-ergodicity for s-i.i.d. inputs is shown in the same way. Above, we just have to consider an s-i.i.d. input probability measure μ. From what is shown, the assertions regarding total ergodicity (total ergodicity for block-i.i.d. inputs) follow immediately.

Part (iii). Assume that $\dot{\kappa}$ and $\ddot{\kappa}$ are causal. Then for all $t \in T$, $B \in \mathcal{Y}_-^t$, and $x, \tilde{x} \in X$ coinciding on $(-\infty, t]$ we have

$$\begin{aligned} \kappa(x,[B]) &= \int_U \ddot{\kappa}\big(\cdot,[B]\big) \, \mathrm{d}\dot{\kappa}(x,\cdot) \\ &= \int_U \ddot{\kappa}\big(\cdot,[B]\big) \, \mathrm{d}\dot{\kappa}(\tilde{x},\cdot) \\ &= \kappa\big(\tilde{x},[B]\big), \end{aligned}$$

i. e., the channel κ is causal according to (2.7.iii). The first and last equality are due to the defining relation for cascade channels in (14.1.1). From the causality of $\ddot{\kappa}$ it follows that $\ddot{\kappa}(u,[B])$ does only depend on the coordinates of u in the time period $(-\infty, t]$. Therefore, only the marginal measure of $\dot{\kappa}(x,\cdot)$ on $\mathcal{U}_-^t$ is relevant for integrating the function $\ddot{\kappa}(\cdot,[B])$. Since the causality of $\dot{\kappa}$ implies the equality of the measures $\dot{\kappa}(x,\cdot)$ and $\dot{\kappa}(\tilde{x},\cdot)$ on $[\mathcal{U}_-^t]$, we obtain the second equality.

Part (iv). Let $\epsilon > 0$ and $s \in T$. We assume that $\ddot{\kappa}$ is asymptotically input-memoryless for the signal set U'. Then there exists a $t_2 = t_2(\epsilon, s) \in T_0$, such that for all $B \in \mathcal{Y}_s^+$ and $u, \tilde{u} \in U'$ coinciding on $(s - t_2, \infty)$ we have

$$\left|\ddot{\kappa}(u, [B]) - \ddot{\kappa}(\tilde{u}, [B])\right| < \frac{\epsilon}{3}. \tag{2}$$

Further, we assume that $\dot{\kappa}$ is asymptotically input-memoryless for the signal set X' so that there exists a $t_1 = t_1(\epsilon, s - t_2) \in T_0$, such that for all $x, \tilde{x} \in X'$ coinciding on $(s - t_2 - t_1, \infty)$

$$\|\dot{\nu}_x - \dot{\nu}_{\tilde{x}}\|_{\mathrm{tv}} < \frac{\epsilon}{3}, \tag{3}$$

holds, where $\dot{\nu}_x$ and $\dot{\nu}_{\tilde{x}}$ denote the marginal measures of $\dot{\kappa}(x, \cdot)$ and $\dot{\kappa}(\tilde{x}, \cdot)$ on $\mathcal{U}_{s-t_2}^+$. This formulation is obtained using the characterization of the total variation distance given in (6.7.1).

Let us fix some $B \in \mathcal{Y}_s^-$ and $x, \tilde{x} \in X'$ coinciding on $(s - t_2 - t_1, \infty)$. To bring the asymptotic input-memorylessness into play we fix an arbitrary element $a \in U_-^{s-t_2}$ from the projection of U' to $U_-^{s-t_2}$. Further, let $a \in U_-^{s-t_2}$ be some element from the projection of U' to $U_-^{s-t_2}$. For an element $u \in U$ we make use of the representation $u = (u_-, u_+) \in U_-^{s-t_2} \times U_{s-t_2}^+$ and obtain from the triangle inequality

$$\begin{aligned}
\left|\kappa\big(x, [B]\big) - \kappa\big(\tilde{x}, [B]\big)\right| = {} & \left| \int_U \ddot{\kappa}\big(u, [B]\big)\, \mathrm{d}\dot{\kappa}(x, u) - \int_U \ddot{\kappa}\big(u, [B]\big)\, \mathrm{d}\dot{\kappa}(\tilde{x}, u) \right| \\
\leq {} & \left| \int_U \ddot{\kappa}\big(u, [B]\big)\, \mathrm{d}\dot{\kappa}(x, u) - \int_U \ddot{\kappa}\big((a, u_+), [B]\big)\, \mathrm{d}\dot{\kappa}(x, u) \right| && (4) \\
& + \left| \int_U \ddot{\kappa}\big((a, u_+), [B]\big)\, \mathrm{d}\dot{\kappa}(x, u) - \int_U \ddot{\kappa}\big((a, u_+), [B]\big)\, \mathrm{d}\dot{\kappa}(\tilde{x}, u) \right| && (5) \\
& + \left| \int_U \ddot{\kappa}\big((a, u_+), [B]\big)\, \mathrm{d}\dot{\kappa}(\tilde{x}, u) - \int_U \ddot{\kappa}\big(u, [B]\big)\, \mathrm{d}\dot{\kappa}(\tilde{x}, u) \right|. && (6)
\end{aligned}$$

Assume that we have $U' = \bigtimes_{t \in T} U'_t$ with $U'_t \subset U_t$ and the outer $\dot{\kappa}(x, \cdot)$-measure of U' is equal to 1. Then (4) is bounded above by

$$\int_U \left| \ddot{\kappa}\big((u_-, u_+), [B]\big) - \ddot{\kappa}\big((a, u_+), [B]\big) \right| \mathrm{d}\dot{\kappa}\big(x, (u_-, u_+)\big) \leq \frac{\epsilon}{3}$$

due to (2) and (A.11.ii). If further the outer $\dot{\kappa}(\tilde{x}, \cdot)$-measure of U' is equal to 1, then by the same arguments we obtain that (6) is bounded above by $\frac{\epsilon}{3}$. In (5) we can replace $\dot{\kappa}(x, \cdot)$ and $\dot{\kappa}(\tilde{x}, \cdot)$ by the marginal measures $\dot{\nu}_x$ and $\dot{\nu}_{\tilde{x}}$ since the integrands only depend on u_+. Therefore, we can rewrite and upper-bound (5) by

$$\left| \int_{U_{s-t_2}^+} \ddot{\kappa}\big((a, u_+), [B]\big)\, \mathrm{d}\dot{\nu}_x(u_+) - \int_{U_{s-t_2}^+} \ddot{\kappa}\big((a, u_+), [B]\big)\, \mathrm{d}\dot{\nu}_{\tilde{x}}(u_+) \right| \leq \|\dot{\nu}_x - \dot{\nu}_{\tilde{x}}\|_{\mathrm{tv}} \leq \frac{\epsilon}{3}.$$

The first inequality follows either from (6.10.v) or from the data processing inequality (6.4.ii), which holds in particular for the total variation distance. We have implicitly used that $\ddot{\kappa}((a, \cdot), \cdot)$ is a Markov-kernel from $(U_{s-t_2}^+, \mathcal{U}_{s-t_2}^+)$ to $(Y, \mathcal{Y})$, which holds due to (A.3.iii). From (3) we

obtain the second inequality. Combining all inequalities we have for fixed $\epsilon > 0$, $s \in T$, arbitrary $B \in \mathcal{Y}_s^-$, and $x, \tilde{x} \in X'$ coinciding on $(s - t_2 - t_1, \infty)$

$$\big|\kappa\big(x, [B]\big) - \kappa\big(\tilde{x}, [B]\big)\big| \leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

Thus, according to (2.7.iv) the cascade channel κ is asymptotically input-memoryless for the signal set X' under the assumptions of (14.2.iv). □

The next theorem provides sufficient conditions that allow to verify memory properties of the overall channel by analyzing memory properties of individual components.

(14.4) Theorem (*Memory and mixing properties of cascade channels with time structure). Consider a cascade channel with time structure as in Theorem 14.2. Then we have the following implications: If $\dot{\kappa}$ has the property in the first column and $\ddot{\kappa}$ the properties in the second column, then the cascade channel κ has the mixing property in the third column.*

	$\dot{\kappa}$	$\ddot{\kappa}$	κ
(i)	*finite input memory*	*finite input memory*	*finite input memory*
(ii)	*finite output memory*	*finite output memory, causal, finite input memory*†	*finite output memory*
(iii)	*ψ-mixing*	*ψ-mixing, causal, finite input memory*†	*ψ-mixing*
(iv)	*information regular*	*information regular, causal, finite input memory*†	*information regular*
(v)	*β-mixing*	*β-mixing, causal, asymptotically input-memoryless*†	*β-mixing*
(vi)	*α-mixing*	*α-mixing, causal, asymptotically input-memoryless*†	*α-mixing*

The superscript † denotes we additionally assume that the possibly time-varying channel input memory lengths (for fixed tolerance ϵ) are bounded. If asymptotic input-memorylessness is assumed, then for the whole input signal set X.

(14.5) Remark. Theorem 14.4 is closely related to Theorem 13.11, actually the statements (14.4.ii)–(14.4.vi) correspond to the statements (13.11.i)–(13.11.v) with $\dot{\kappa}$ in the role of μ, $\ddot{\kappa}$ in the role of κ, and the cascade channel in the role of the marginal measure of $\mu\kappa$ on the channel output space. The comments given in Remark 13.12 hold in an analogous manner.

The proof of (14.4.i) works in the same way as the proof of (14.2.iii) and is therefore omitted. The proofs of (14.4.ii)–(14.4.vi) are also omitted because they can be easily obtained from the proofs of (13.11.i)–(13.11.v). This is possible because for any $x \in X$ the probability measure $\kappa(x, \cdot)$ related to the cascade channel κ is equal to the channel output measure of $\ddot{\kappa}$, when $\dot{\kappa}(x, \cdot)$ is the input measure.

Obtaining similar results for the ergodic-theoretic mixing properties on the basis of the proof of Theorem 13.11 is not possible because there the stationarity of the input probability measure

is assumed but the probability measure $\dot\kappa(x,\cdot)$ is usually not stationary. However, the following holds. If $\dot\kappa$ and $\ddot\kappa$ are both stationary and mixing (in the ergodic-theoretic sense) and μ is a stationary and mixing probability measure (in the ergodic-theoretic sense) on the input space of $\dot\kappa$, then the input-output probability measure $\mu\kappa$ of the cascade channel κ is stationary and mixing (in the ergodic-theoretic sense). Indeed, the input-output probability measure $\mu\dot\kappa$ of the channel $\dot\kappa$ is stationary and mixing due to (13.11.vi). Then using the Markov chain argument as in the proof of (14.2.ii) and again (13.11.vi) shows the assertion. A corresponding result holds for the totally weakly mixing property.

§15 Integration Channels

In this section, we consider so-called integration channels, for which the output is the result of a deterministic mapping applied to the channel input and a random source of noise. Noise means any kind of unwanted or unavoidable disturbance of the input, e. g. additive noise, fading, interference, etc. If a functional relation is known, characterizing the input-output behavior of a transmission system and the random impairment of the transmission, then the integration channel is the communication model of choice. This is the case in a large number of practically relevant situations as demonstrated in Sections §16 and §17 by various examples.

After introducing integration channels, we specify conditions for the channel function and the noise source that imply properties of the integration channel relevant in connection with coding theorems or signal processing applications. These results demonstrate the advantage of separating the channel model into a deterministic and a random part to analyze channel properties. We further consider a cascade of integration channels and a useful representation based on component functions, which is helpful in connection with verifying measurability properties of the channel.

(15.1) Definition (Integration channel). Let $(X,\mathcal{X})$, $(Y,\mathcal{Y})$, and $(Z,\mathcal{Z})$ be measurable spaces, suppose λ is a probability measure on $\mathcal{Z}$, and f is an $\mathcal{X}\otimes\mathcal{Z}/\mathcal{Y}$-measurable function on $X\times Z$ with values in Y. The channel κ with input space $(X,\mathcal{X})$ and output space $(Y,\mathcal{Y})$, defined by

$$\kappa(x,B)=\lambda_{f(x,\cdot)}(B) \tag{1}$$

for any $x\in X$ and $B\in\mathcal{Y}$, is called an integration channel with channel function f and noise measure λ. By $\lambda_{f(x,\cdot)}$ we denote the distribution of the random variable $f(x,\cdot)$.

The integration channel is illustrated in Figure 4.

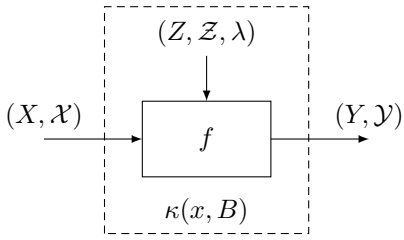


Figure 4: Integration channel.

(15.2) Remark. From the identity

$$\mathbb{1}_B\big(f(x,z)\big) = \mathbb{1}_{\{f(x,\cdot)\in B\}}(z)$$

for all $(x,z) \in X \times Z$ we easily obtain

$$\kappa(x,B) = \lambda_{f(x,\cdot)}(B) = \lambda\big(f(x,\cdot) \in B\big) \tag{1}$$

$$= \int_Z \mathbb{1}_B\big(f(x,z)\big)\,\mathrm{d}\lambda(z). \tag{2}$$

This integral form motivates the name integration channel, which is adopted from Nakamura (1975). Alternatively, the channel is called channel with a noise source.

That κ is indeed a Markov kernel is verified as follows. For any input $x \in X$ the function $f(x,\cdot)$ is $\mathcal{Z}/\mathcal{Y}$-measurable, i. e., it is a random variable on the probability space $(Z,\mathcal{Z},\lambda)$. Its distribution is $\kappa(x,\cdot)$ so that $\kappa(x,\cdot)$ is a probability measure for all $x \in X$. Further, for all sets $B \in \mathcal{Y}$ the composition $\mathbb{1}_B(f)$ is $\mathcal{X}\otimes\mathcal{Z}$-measurable since f is $\mathcal{X}\otimes\mathcal{Z}/\mathcal{Y}$-measurable and $\mathbb{1}_B$ as function on Y is $\mathcal{Y}$-measurable. Part (A.8.i) of Fubini's theorem then yields for all $B \in \mathcal{Y}$ the $\mathcal{X}$-measurability of

$$\kappa(\cdot,B) = \int_Z \mathbb{1}_B\big(f(\cdot,z)\big)\,\mathrm{d}\lambda(z).$$

Let us define the function g on $X \times Z$ with values in $X \times Y$ by

$$g(x,z) = (x, f(x,z))$$

for all $(x,z) \in X \times Z$. Then g is $\mathcal{X}\otimes\mathcal{Z}/\mathcal{X}\otimes\mathcal{Y}$-measurable because f is $\mathcal{X}\otimes\mathcal{Z}/\mathcal{Y}$-measurable, $\mathcal{X}\otimes\mathcal{Y} = \sigma(\mathcal{X}\times\mathcal{Y})$, and $g^{-1}(A\times B) = (A\times Z)\cap f^{-1}(B)$ for all $A \in \mathcal{X}$ and $B \in \mathcal{Y}$. Suppose μ is a probability measure on $\mathcal{X}$. Then g can be considered as random variable on the probability space $(X\times Z, \mathcal{X}\otimes\mathcal{Z}, \mu\otimes\lambda)$, where its distribution $(\mu\otimes\lambda)_g$ is a probability measure on $\mathcal{X}\otimes\mathcal{Y}$. If $\mu\kappa$ denotes the input-output probability measure on $\mathcal{X}\otimes\mathcal{Y}$ induced by μ and the channel κ, then we have the equality

$$\mu\kappa = (\mu\otimes\lambda)_g. \tag{3}$$

Indeed, for any $C \in \mathcal{X}\otimes\mathcal{Y}$ we have

$$\begin{aligned}
\mu\kappa(C) &= \int_X \kappa(x, C_x)\,\mathrm{d}\mu(x) \\
&= \int_X \left[\int_Z \mathbb{1}_{\{f(x,\cdot)\in C_x\}}(z)\,\mathrm{d}\lambda(z)\right]\mathrm{d}\mu(x) \\
&= \int_{X\times Z} \mathbb{1}_{\{f(x,\cdot)\in C_x\}}(z)\,\mathrm{d}\mu\otimes\lambda(x,z) \\
&= \mu\otimes\lambda\Big(\big\{(x,z)\in X\times Z : (x,f(x,z))\in C\big\}\Big) \\
&= \mu\otimes\lambda\big(g\in C\big),
\end{aligned}$$

where the third equality follows from part (A.8.i) of Fubini's theorem. This identity is used by Baker (1976, 1978, 1979, 1983) to define integration channels. Similar to the previous derivations, we obtain

$$\nu = (\mu \otimes \lambda)_f, \tag{4}$$

where ν denotes the marginal measure of $\mu\kappa$ on $\mathcal{Y}$.

Nakamura (1975) studied the ergodicity of discrete-time integration channels and their information capacity for finite alphabets. These results are reproduced in (Kakihara, 1999, Sec. 4.1). Baker (1976, 1978, 1979, 1983) analyzed the mutual information and the information capacity of integration channels for more general spaces with a focus on Hilbert spaces and Gaussian channels. The primitive channels of Neuhoff and Shields (1979, 1982a,b,c) are special stationary discrete-time discrete alphabet integration channels with finite input and output memory and an i.i.d. noise source. Primitive channels are used to approximate channels with a certain type of infinite input and output memory.

For later reference let us fix the notation for integration channels with time structure.

(15.3) Example (Integration channel with time structure). Let κ be an integration channel with channel function f and noise measure λ. Adopting the notation of Paragraph 1.2 we assume that $(X, \mathcal{X})$, $(Y, \mathcal{Y})$, and $(Z, \mathcal{Z})$ in Definition 15.1 are product measurable spaces generated by the families $\{(X_t, \mathcal{X}_t), t \in T\}$, $\{(Y_t, \mathcal{Y}_t), t \in T\}$, and $\{(Z_t, \mathcal{Z}_t), t \in T\}$ of measurable spaces for which we have $(X_t, \mathcal{X}_t) = (X_0, \mathcal{X}_0)$, $(Y_t, \mathcal{Y}_t) = (Y_0, \mathcal{Y}_0)$, $(Z_t, \mathcal{Z}_t) = (Z_0, \mathcal{Z}_0)$ for all $t \in T$. Then the integration channel κ is a channel with time structure as introduced in Definition 2.3. For this channel the noise measure λ is also called noise source.

The next theorem expresses properties of an integration channel with time structure in terms of properties of the channel function and the noise source. We consider properties relevant in connection with coding (Theorem 9.1 and 9.3) and signal processing applications (e. g. Paragraph 17.7). The theorem is useful, in particular in concrete practical examples, because it allows to verify channel properties by verifying properties of the channel function and the noise source separately. Regarding ergodicity as required in Theorem 9.1, the theorem demonstrates another advantage of the integration channel formulation: The indirectly formulated ergodicity property of the channel given in (2.7.ii) can be expressed by a direct ergodicity condition on the noise source.

(15.4) Theorem (*Properties of integration channels with time structure). Let κ be an integration channel with time structure as in Example 15.3 with channel function f and noise source λ. Then we have the following implications: If f has the property in the first column and λ the property in the second column, then the integration channel κ has the property in the third column.*

	f	λ	κ
(i)†	*(s-) invariant*	*(s-) stationary*	*(s-) stationary*
(ii)	$[\mathcal{X}_-^t] \otimes \mathcal{Z}/[\mathcal{Y}_-^t]$-*measurable all* $t \in T$		*causal*
(iii)‡	$[\mathcal{X}_{s-t(s)}^+] \otimes \mathcal{Z}/[\mathcal{Y}_s^+]$-*measurable all* $s \in T$		*finite input memory*

		f	λ	κ
(iv)	*a)*†	*s-invariant*	*s-stationary and s-ergodic*	*s-stationary and s-ergodic for s-i.i.d. inputs*
	b)	*invariant*	*stationary and totally ergodic*	*stationary and totally ergodic for block-i.i.d. inputs*
	c)†	*s-invariant*	*s-stationary and s-weakly mixing*	*s-stationary and s-ergodic*
	d)	*invariant*	*stationary and totally weakly mixing*	*stationary and totally ergodic*
(v)‡	*a)*	*$\mathcal{X} \otimes [\mathcal{Z}_-^{s+t_1(s)}]/[\mathcal{Y}_-^s]$-measurable and $\mathcal{X} \otimes [\mathcal{Z}_s^+]/[\mathcal{Y}_{s+t_2(s)}^+]$-measurable all $s \in T$*	*finite memory*	*finite output memory*
	b)	*as in a)*	*ψ-mixing*	*ψ-mixing*
	c)	*as in a)*	*information regular*	*information regular*
	d)	*as in a)*	*β-mixing*	*β-mixing*
	e)	*as in a)*	*α-mixing*	*α-mixing*

The superscript † denotes we assume that $s \in T$ in (i) *and $s \in T_+$ in* (iv)*. The superscript ‡ denotes we assume that for all $s \in T$ there exist $t(s) \in T_0$ or $t_1(s), t_2(s) \in T_0$, respectively, such that measurability of the requested form is given.*

Proof. Part (i). Let $s \in T$ and assume that the channel function f is s-invariant and the noise source λ is s-stationary. Then for all $x \in X$ and $B \in \mathcal{Y}$ we have

$$\begin{aligned}\kappa\big(\theta_s(x), \theta_s(B)\big) &= \lambda\big(f\big(\theta_s(x), \theta_s(\cdot)\big) \in \theta_s(B)\big)\\ &= \lambda\big(\theta_s\big(f(x,\cdot)\big) \in \theta_s(B)\big)\\ &= \lambda\big(f(x,\cdot) \in B\big)\\ &= \kappa(x, B).\end{aligned}$$

The first and last equality are due to the defining relation for integration channels in (15.2.1). The second equality follows from the s-invariance of the channel function f and the third from the s-stationarity of the noise source λ. Thus the channel κ is s-stationary according to (2.7.i). The assertion, where s-stationarity and s-invariance is replaced by stationarity and invariance is now evident.

Part (ii). Let $t \in T$, $B \in \mathcal{Y}_-^t$ and assume that f is $[\mathcal{X}_-^t] \otimes \mathcal{Z}/[\mathcal{Y}_-^t]$-measurable. Then $\mathbb{1}_{[B]}(f)$ is $[\mathcal{X}_-^t] \otimes \mathcal{Z}$-measurable and the integral representation (15.2.2) together with part (A.8.i) of Fubini's theorem imply the $[\mathcal{X}_-^t]$-measurability of $\kappa(\cdot, [B])$. Using the alternative characterization of causality given in Remark 2.8 yields the assertion.

Part (iii). Using the characterization of finite input memory given in Remark 13.2 the proof is similar to that of part (ii) and is therefore omitted.

Part (iv). Let $s \in T_+$ and assume that f is s-invariant. Consider the function g introduced in Remark 15.2. The s-invariance of f immediately implies the s-invariance of g. Suppose μ is a probability measure on the channel input space. If μ is an s-i.i.d. probability measure and λ is s-stationary and s-ergodic, then the product measure $\mu \otimes \lambda$ is s-stationary and s-ergodic due to (B.15.i) and Lemma B.11. Correspondingly, if μ is s-stationary and s-ergodic and λ is s-stationary and s-weakly mixing, then $\mu \otimes \lambda$ is s-stationary and s-ergodic, again according to (B.15.i). Applying (B.13.i) shows that the image measure $(\mu \otimes \lambda)_g$ is s-stationary and s-ergodic. From the identity (15.2.3) follows that the channel input-output probability measure $\mu\kappa$ is s-stationary and s-ergodic. According to the definition in (2.7.ii), this implies assertions a) and c). From what is shown, the remaining assertions b) and d) are evident.

Part (v). We prove assertion b) and a). The remaining assertions are shown in the same way. Only the dependence measures have to be replaced. These proofs are therefore omitted.

Assume that $\epsilon > 0$ and $s \in T$. By assumption there exists a $t_1 = t_1(s) \in T_0$ such that f is $\mathcal{X} \otimes [\mathcal{Z}_-^{s+t_1}]/[\mathcal{Y}_-^s]$-measurable. From λ being ψ-mixing we obtain that there exists a $\tau = \tau(s+t_1) \in T_0$ such that

$$\psi\big([\mathcal{Z}_-^{s+t_1}]; [\mathcal{Z}_{s+t_1+\tau}^+]\big) \leq \epsilon. \tag{1}$$

Again by assumption there exists a $t_2 = t_2(s+t_1+\tau) \in T_0$ such that f is $\mathcal{X} \otimes [\mathcal{Z}_{s+t_1+\tau}^+]/[\mathcal{Y}_{s+t_o}^+]$-measurable with $t_o = t_1 + \tau + t_2$.

Let $x \in X$ be arbitrary but fixed. Further, let $\{\eta_t, t \in T\}$ denote the family of coordinate projections on the channel output space, where η_t is the projection from Y to Y_t, considered as random variable on the probability space $(Y, \mathcal{Y}, \kappa(x,\cdot))$. Then we have for all $t \geq t_o$

$$\begin{aligned} \psi\big([\mathcal{Y}_-^s]; [\mathcal{Y}_{s+t}^+] \,\big|\, x\big) &= \psi\big(\eta_-^s; \eta_{s+t}^+\big) \\ &= \psi\Big(\eta_-^s\big(f(x,\cdot)\big); \eta_{s+t}^+\big(f(x,\cdot)\big)\Big) \\ &\leq \psi\big([\mathcal{Z}_-^{s+t_1}]; [\mathcal{Z}_{s+t_1+\tau}^+]\big) \\ &\leq \epsilon. \end{aligned} \tag{2}$$

For the second equality we have used the identity (7.6.5) and the defining relation of integration channels given in (15.1.1), i. e., $\kappa(x,\cdot) = \lambda_{f(x,\cdot)}$. The subsequent inequality follows from the previously stated measurability properties of f and the monotonicity of the ψ-dependence coefficient (σ-algebra based version of (7.7.ii)). The last inequality is due to (1). Because the derived inequality holds for all $x \in X$ and $\epsilon > 0$ was chosen arbitrary we have shown that the integration channel κ is ψ-mixing. Note that alternatively we can use the integral representation given in (15.2.2) together with (7.7.viii) to obtain this result.

In view of Definition 12.1 and the representation of the finite output memory condition given in Remark 13.7 we obtain assertion a) by repeating the above derivations with $\epsilon = 0$. □

(15.5) Remark. Following the proof it seems that without change of the arguments the condition for the noise source λ in implication c) of (15.4.iv) can be weakened in the following way: For any s-stationary and s-ergodic channel input probability measure μ the noise source λ is such that the product measure $\mu \otimes \lambda$ is s-stationary and s-ergodic. However, for this to hold the noise source λ has to be s-weakly mixing due to the second part of Remark B.16.

From the identity $\mu\kappa = (\mu \otimes \lambda)_g$ together with (B.13.iii) and (B.15.ii) we easily obtain the following: If the channel function is s-invariant and the noise source is s-stationary and s-weakly mixing, then the channel input-output probability measure is s-stationary and s-weakly mixing for all s-stationary and s-weakly mixing input probability measures. However, we cannot conclude that an s-invariant channel function f together with an s-stationary s-weakly mixing noise source λ yields an s-weakly mixing channel as introduced in Definition 13.4. Even though f is s-invariant, $f(x, \cdot)$ is usually not, so that for fixed input x the probability measure $\kappa(x, \cdot) = \lambda_{f(x,\cdot)}$ results from a noninvariant transformation of the noise source. Therefore, we cannot use that material from ergodic theory, which requires stationarity. The same applies to the mixing condition (in the ergodic-theoretic sense).

In particular, the conditions in (15.4.v) are not necessary in general. There are integration channels, where the interplay between the noise source and the channel function is such that the integration channel satisfies a mixing condition from (15.4.v), even though the channel function does not satisfy the measurability condition formulated there. Using the sequence considered in (12.4.ii) together with the channel in Paragraph 16.2 yields a simple example of this type.

The mixing properties of the integration channel depend on both, the channel function and the noise source and regarding the channel function only the z-coordinate is relevant. Aside from the invariance condition for the channel function, the ergodicity of the integration channel depends solely on the noise source. Furthermore, whether or not an integration channel is causal or has finite input memory solely depends on the channel function for which only the x-coordinate is relevant. For the less restrictive asymptotic input-memorylessness, for which Theorem 15.4 does not provide conditions, the situation is more complicated. Assume that $s \in T$, $B \in [\mathcal{Y}_s^+]$, and $x, \tilde{x} \in X$. Then we have

$$\begin{aligned}
&\big|\kappa(x, B) - \kappa(\tilde{x}, B)\big| \\
&= \big|\lambda\big(f(x, \cdot) \in B\big) - \lambda\big(f(\tilde{x}, \cdot) \in B\big)\big| \\
&= \big|\lambda\big(\{f(x, \cdot) \in B\} \cap \{f(\tilde{x}, \cdot) \in B\}^{\mathrm{c}}\big) - \lambda\big(\{f(\tilde{x}, \cdot) \in B\} \cap \{f(x, \cdot) \in B\}^{\mathrm{c}}\big)\big| \qquad (1) \\
&\leq \lambda\big(\{f(x, \cdot) \in B\} \triangle \{f(\tilde{x}, \cdot) \in B\}\big). \qquad (2)
\end{aligned}$$

If the channel has finite input memory, then the two intersections in (1) and the symmetric difference in (2) are empty if x and $\tilde{x}$ coincide on $(s - t, \infty)$ for sufficiently large t. Therefore, the properties of the noise source do not matter in this case. However, if the channel has infinite input memory, then these sets are not empty for $x \neq \tilde{x}$. The sets are determined by the channel function and their probability mass is determined by the noise source. Thus, for asymptotically input-memoryless integration channels the noise source and the channel function have to interact in the right way.

Theorem 13.11 formulates conditions on the channel and the input probability measure such that the input-output probability measure satisfies a certain mixing condition. Together with Theorem 15.4 we can translate the conditions on the channel into conditions on the channel function and the noise source. Alternatively, these conditions can be obtained as follows. If the channel input probability measure and the noise source are both α-mixing (β-mixing, information regular, ψ-mixing, have finite memory), then the product measure $\mu \otimes \lambda$ is α-mixing (β-mixing, information regular, ψ-mixing, has finite memory) due to Lemma 12.7. If in addition the channel function satisfies the measurability conditions specified in (15.4.ii), (15.4.iii), and (15.4.v), then the input-output probability measure $\mu\kappa = (\mu \otimes \lambda)_g$ is α-mixing (β-mixing, information regular, ψ-mixing, has finite memory).

(15.6) Components of a channel function with time structure. For practical applications we want to make use of the convenient representation of a channel function with time structure by component functions. Consider the integration channel with time structure from Example 15.3. Assume that $0 \le u_x, u_z, v_x, v_z \in \overline{T}$ are fixed. For any $w \in T$ let f_w be a function on $X_{w-u_x}^{w+v_x} \times Z_{w-u_z}^{w+v_z}$ with values in Y_w. We define for all $s \le t \in \overline{T}$ the function f_s^t on $X_{s-u_x}^{t+v_x} \times Z_{s-u_z}^{t+v_z}$ with values in Y_s^t by

$$f_s^t\big(x_{s-u_x}^{t+v_x}, z_{s-u_z}^{t+v_z}\big) = y_s^t = \big\{f_w\big(x_{w-u_x}^{w+v_x}, z_{w-u_z}^{w+v_z}\big), w \in J\big\}$$

for all $x_{s-u_x}^{t+v_x} = \{x_w\} \in X_{s-u_x}^{t+v_x}$ and $z_{s-u_z}^{t+v_z} = \{z_w\} \in Z_{s-u_z}^{t+v_z}$, where the index set J is defined as in Paragraph 1.2 with u, v replaced by s, t.[23] Thus f_s^t is specified by the component functions f_w. We assume that f_w is $\mathcal{X}_{w-u_x}^{w+v_x} \otimes \mathcal{Z}_{w-u_z}^{w+v_z}/\mathcal{Y}_w$-measurable. Then f_s^t is $\mathcal{X}_{s-u_x}^{t+v_x} \otimes \mathcal{Z}_{s-u_z}^{t+v_z}/\mathcal{Y}_s^t$-measurable as shown in Paragraph E.6 in Appendix E.

We define the channel function f of the integration channel with time structure by $f = f_{-\infty}^{\infty}$. The output of this channel at time $w \in T$ is the result of a deterministic mapping f_w applied to the channel input in the time period J_x and the random noise in the time period J_z. The index set J_x is defined as the set J in Paragraph 1.2 with u, v replaced by $w - u_x, w + v_x$ and correspondingly the set J_z with u, v replaced by $w - u_z, w + v_z$. Note that it would be possible that u_x, u_z, v_x, v_z vary with w, however, we will not require this generalization.

Most of the examples of practical relevance considered later are defined as specified above. A particularly useful fact of this representation is that the $\mathcal{X} \otimes \mathcal{Z}/\mathcal{Y}$-measurability of the channel function is implied by the measurability property of its component functions. Furthermore, there are simple conditions such that the properties of the channel function required in Theorem 15.4 are satisfied. If we have $f_{w+s}(\cdot) = f_w(\langle\cdot\rangle_{-s})$ for all $w \in T$, then the channel function is s-invariant and if we have $f_w(\cdot) = f_0(\langle\cdot\rangle_{-w})$ for all $w \in T$, then the channel function is invariant. If $v_x = 0$, then the condition in (15.4.ii) is satisfied, i. e., the channel is causal. If u_x is finite, then the condition in (15.4.iii) is satisfied, i. e., the channel has finite input memory with memory length u_x. The measurability condition in (15.4.v) is satisfied if u_z and v_z are finite. Often, in typical examples we even have $u_z = v_z = 0$.

(15.7) Cascade of integration channels. The cascade of two integration channels is an integration channel. Indeed, consider the situation of Definition 14.1, where the channel κ is obtained by connecting the channels $\dot{\kappa}$ and $\ddot{\kappa}$ in cascade as illustrated in Figure 3. Assume that $\dot{\kappa}$ is an integration channel with channel function $\dot{f}$ and noise measure space $(\dot{Z}, \dot{\mathcal{Z}}, \dot{\lambda})$. Further, assume that $\ddot{\kappa}$ is an integration channel with channel function $\ddot{f}$ and noise measure space $(\ddot{Z}, \ddot{\mathcal{Z}}, \ddot{\lambda})$. The cascade channel κ is an integration channel with noise measure space $(Z, \mathcal{Z}, \lambda)$ given by

$$(Z, \mathcal{Z}, \lambda) = (\dot{Z} \times \ddot{Z}, \dot{\mathcal{Z}} \otimes \ddot{\mathcal{Z}}, \dot{\lambda} \otimes \ddot{\lambda}) \tag{1}$$

and channel function f given by

$$f\big(x, \dot{z}, \ddot{z}\big) = \ddot{f}\big(\dot{f}\big(x, \dot{z}\big), \ddot{z}\big) \tag{2}$$

[23] We assume that u_x, u_z, v_x, v_z, s, t are taken from the extended time index set $\overline{T}$, i. e., they can be infinite. The notation using the set J allows a correct treatment of infinite interval boundaries. We use the usual conventions $\pm\infty + \mathrm{c} = \pm\infty$ for any real constant c and $\pm\infty \pm \infty = \pm\infty$.

for all $(x, \dot{z}, \ddot{z}) \in X \times \dot{Z} \times \ddot{Z}$. Indeed, we have

$$\begin{aligned}\kappa(x, B) &= \int_U \ddot{\kappa}\big(u, B\big)\, \mathrm{d}\dot{\kappa}(x, u) \\ &= \int_U \left[\int_{\ddot{Z}} \mathbb{1}_B\big(\ddot{f}(u, \ddot{z})\big)\, \mathrm{d}\ddot{\lambda}(\ddot{z}) \right] \mathrm{d}\dot{\lambda}_{\dot{f}(x,\cdot)}(u) \\ &= \int_{\dot{Z}} \left[\int_{\ddot{Z}} \mathbb{1}_B\big(\ddot{f}\big(\dot{f}(x, \dot{z}), \ddot{z}\big)\big)\, \mathrm{d}\ddot{\lambda}(\ddot{z}) \right] \mathrm{d}\dot{\lambda}(\dot{z}) \\ &= \int_{\dot{Z}\times\ddot{Z}} \mathbb{1}_B\big(f\big(x, \dot{z}, \ddot{z}\big)\big)\, \mathrm{d}\dot{\lambda} \otimes \ddot{\lambda}(\dot{z}, \ddot{z}).\end{aligned}$$

The first equality is actually the defining relation of a cascade channel given in (14.1.1). Using the defining relation in (15.1.1) for the integration channel $\dot{\kappa}$ and the integral representation in (15.2.2) for the integration channel $\ddot{\kappa}$ yields the second equality. The third equality is due to the substitution rule for integrals. Finally, the last equality is the result of applying part (A.8.i) of Fubini's theorem. Figure 5 shows the cascade of integration channels and the resulting joint integration channel.

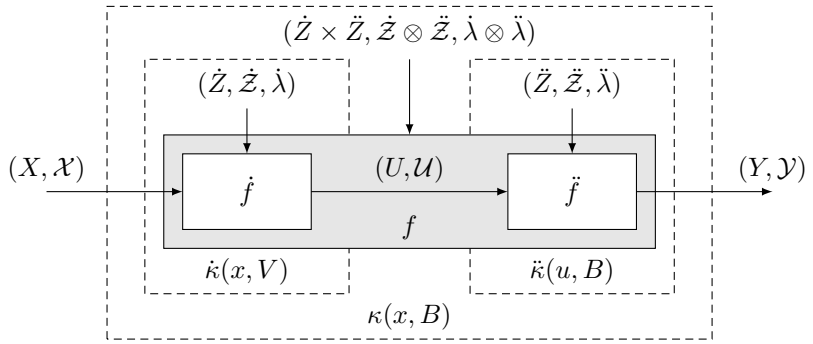


Figure 5: Cascade of integration channels.

Chapter V

Examples and Applications

§16 Basic Examples

In this section, we discuss some basic examples, which either serve as building blocks for more complex models or illustrate theoretical facts from the previous chapters. First, we consider some special integration channels, namely the deterministic channel, the purely random channel, the additive noise channel, and the multiplicative noise channel. In particular, the additive noise channel is used as nontrivial example to demonstrate that the ergodicity and mixing conditions for channels considered in Section §13 are not equivalent. We then continue with a channel that induces a concave sequence of mutual informations. Thereby, we give a counterexample to a convexity argument used in a proof of Kadota and Wyner (1972, Appendix II). Furthermore, we discuss the memory properties of a generalization of the Gilbert-Elliott channel, which is a simple model of a transmission over a time-varying state-dependent channel. Finally, with a simple method we construct nonergodic channels based on ergodic ones. To those channels the coding theorem considered in this thesis does not apply.

In the following examples κ is an integration channel as specified in Definition 15.1 with channel function f and noise measure λ. The next two integration channels are, in a sense, two extreme cases: the deterministic and the purely random channel.

(16.1) Deterministic channel. Suppose $\hat{f}$ is an $\mathcal{X}/\mathcal{Y}$-measurable function on X with values in Y and the channel function f is given for all $(x, z) \in X \times Z$ by

$$f(x, z) = \hat{f}(x),$$

i. e., it depends only on the input coordinate. Then the integration channel κ is called deterministic channel (see (Gray, 2011, Sec. 2.5)). Due to (15.2.2) we have for any $x \in X$ and $B \in \mathcal{Y}$

$$\begin{aligned}\kappa(x, B) &= \int_Z \mathbb{1}_B\big(\hat{f}(x)\big)\,\mathrm{d}\lambda(z)\\ &= \mathbb{1}_B\big(\hat{f}(x)\big)\int_Z \mathrm{d}\lambda(z)\\ &= \mathbb{1}_B\big(\hat{f}(x)\big) = \delta_{\hat{f}(x)}(B),\end{aligned}$$

where $\delta_y(\cdot)$ denotes the Dirac measure on $\mathcal{Y}$ for the element $y \in Y$. That means the probability is 1 if the channel output $\hat{f}(x)$ corresponding to the input x lies in the set B and 0 if it is not the case. The noise measure does not have any impact. (Actually, there is no noise disturbing the input.) If μ is a channel input probability measure, then using (15.2.3) we obtain for the input-output probability measure $\mu\kappa$

$$\mu\kappa = \mu_{\hat{g}},$$

where the function $\hat{g}$ is given for all $x \in X$ by

$$\hat{g}(x) = (x, \hat{f}(x)).$$

Properties. Let us reconsider the results of Theorem 15.4 for a deterministic channel κ with time structure. If $\hat{f}$ is (s-) invariant, then κ is (s-) stationary. If $\hat{f}$ is $[\mathcal{X}_-^t]/[\mathcal{Y}_-^t]$-measurable for all $t \in T$, then κ is causal and if for all $s \in T$ there exists a $t(s) \in T_0$ such that $\hat{f}$ is $[\mathcal{X}_{s-t(s)}^+]/[\mathcal{Y}_s^+]$-measurable, then κ has finite input memory. Further, for any $x \in X$ and $B_1, B_2 \in \mathcal{Y}$ we have

$$\begin{aligned}\kappa(x, B_1 \cap B_2) &= \mathbb{1}_{B_1 \cap B_2}\big(\hat{f}(x)\big) \\ &= \mathbb{1}_{B_1}\big(\hat{f}(x)\big)\mathbb{1}_{B_2}\big(\hat{f}(x)\big) \\ &= \kappa(x, B_1)\kappa(x, B_2),\end{aligned}$$

which implies that any deterministic channel with time structure is output-memoryless. In view of Theorem 13.9 the channel therefore satisfies all mixing conditions considered there and if $\hat{f}$ is (s-) invariant, then also all ergodicity properties considered in Theorem 15.4. Given $s \in T$ let $B \in [\mathcal{Y}_s^+]$ and assume that $x, \tilde{x} \in X$. Then we have

$$\begin{aligned}\big|\kappa(x, B) - \kappa(\tilde{x}, B)\big| &= \big|\delta_{\hat{f}(x)}(B) - \delta_{\hat{f}(\tilde{x})}(B)\big| \\ &= \begin{cases} 0 & \text{if} \quad \big(\hat{f}(x) \in B \text{ and } \hat{f}(\tilde{x}) \in B\big) \text{ or } \big(\hat{f}(x) \notin B \text{ and } \hat{f}(\tilde{x}) \notin B\big) \\ 1 & \text{else} \end{cases}\end{aligned}$$

which implies that a deterministic channel with time structure is asymptotically input-memoryless if and only if it has finite input memory.

When we consider practical examples of deterministic channels with time structure, we want to use a component representation of $\hat{f}$ similar to that of the channel function f in Paragraph 15.6. We proceed in an analog manner to define the function $\hat{f}$ by $\hat{f} = \hat{f}_{-\infty}^{\infty}$ based on $\hat{f}_s^t$ and the component functions $\hat{f}_w$ with parameters u_x and v_x. The definitions are as in Paragraph 15.6 with f replaced by $\hat{f}$ and all z-coordinates omitted. The comments given there regarding properties of the component functions implying certain channel properties apply similarly. The measurability result derived in Paragraph E.6 does apply to $\hat{f}$ as well.

(16.2) Purely random channel. Suppose $\check{f}$ is a $\mathcal{Z}/\mathcal{Y}$-measurable function on Z with values in Y and the channel function f is given for all $(x, z) \in X \times Z$ by

$$f(x, z) = \check{f}(z), \tag{1}$$

i. e., it depends only on the noise coordinate. Then the integration channel κ is called purely random channel (see (Gray, 2011, Sec. 2.4)). Due to (15.1.1) we have for any $x \in X$ and $B \in \mathcal{Y}$

$$\kappa(x, B) = \lambda_{\check{f}}(B) \tag{2}$$

i. e., the probability of an output event does not depend on the channel input. Therefore, it is also called constant channel (see (Kakihara, 1999, p. 123)). If μ is a channel input probability

measure, then the input-output probability measure $\mu\kappa$ is given for all $C \in \mathcal{X} \otimes \mathcal{Y}$ by

$$\begin{aligned}\mu\kappa(C) &= \int_X \kappa(x, C_x)\,\mathrm{d}\mu(x) \\ &= \int_X \lambda_{\check{f}}(C_x)\,\mathrm{d}\mu(x) \\ &= \mu \otimes \lambda_{\check{f}}(C),\end{aligned} \tag{3}$$

so the joint measure is indeed a product measure.

Note that there are integration channels whose channel function f cannot be represented as in (1), however, for which the probability measure $\lambda_{f(x,\cdot)}$ is identical for all $x \in X$. Those channels are in fact purely random because they are equivalent to a purely random channel of the form specified above. An example is given at the end of Paragraph 16.5.

Properties. A purely random channel with time structure is always causal and input-memoryless. The remaining properties of the channel considered in Theorem 15.4 hold if the probability distribution $\lambda_{\check{f}}$ has the properties required there for λ. If $(Z, \mathcal{Z}) = (Y, \mathcal{Y})$ and $\check{f}$ is the identity map we even have due to (2) that the purely random channel is s-weakly mixing (totally weakly mixing, mixing in the ergodic-theoretic sense, α-mixing, β-mixing, information regular, ψ-mixing, has finite output memory) if and only if the noise source λ is s-weakly mixing (totally weakly mixing, mixing in the ergodic-theoretic sense, α-mixing, β-mixing, information regular, ψ-mixing, has finite memory). Based on this equivalence we obtain with Example 12.4 that the reversed implications in (a), (b), (c), (d), and (e) of Theorem 13.9 are not true in general. Using (12.5.v) and (12.5.vi) we can construct a probability measure which is weakly mixing (due to (b) in Theorem B.7 this is equivalent to totally weakly mixing in the discrete-time case) but not mixing in the ergodic-theoretic sense. Maruyama (1949, Th. 11) gives a concrete example of this type. Thus the reversed implication in (f) of Theorem 13.9 does not hold in general. However, the considered purely random channel is stationary and totally weakly mixing if and only if it is stationary and totally ergodic. This follows from Remark B.16 and the relation in (3) by similar arguments as used in Paragraph 16.4. The purely random channel is therefore not appropriate to demonstrate that the reversed implication in (g) of Theorem 13.9 does not hold in general.

The given examples are trivial in the sense that a purely random channel is a trivial channel. Further nontrivial examples showing that the channel mixing conditions are not equivalent are discussed in Paragraph 16.3.

In the following examples the integration channel κ has time structure as in Example 15.3. We specify the channel function by component functions as in Paragraph 15.6 and the end of Paragraph 16.1. We consider only invariant channel functions, which are determined by the component function f_0 (or $\hat{f}_0$ for deterministic channels). Furthermore, we assume that the channels are causal and have finite input memory so that we have for the parameters u_x and v_x of the component function: $v_x = 0$ and u_x is finite. Whenever the set of real numbers or vectors is considered, then we associate with it the corresponding Borel-σ-algebra.

(16.3) Additive noise. We assume real-valued input and output signals as well as real-valued noise samples, i. e., we have $X_0 = Y_0 = Z_0 = \mathbb{R}$ as input, output, and noise alphabets. The component f_0 of the invariant channel function f is defined on $X_0 \times Z_0$, i. e., we have $u_x = v_x = u_z = v_z = 0$. Its values in Y_0 are given by

$$y_0 = f_0(x_0, z_0) = x_0 + z_0 \tag{1}$$

for all $(x_0, z_0) \in X_0 \times Z_0$. It is a basic measure-theoretic fact, that f_0 is $\mathcal{X}_0 \otimes \mathcal{Z}_0/\mathcal{Y}_0$-measurable.

Properties. Because the channel function is invariant, the channel itself is stationary if the noise source λ is stationary. The channel function also satisfies the other conditions considered in Theorem 15.4 so that the ergodicity and mixing properties of the additive noise channel depend completely on the noise source. It even holds, that the additive noise channel is α-mixing (β-mixing, information regular, ψ-mixing, has finite output memory) if and only if the noise source is α-mixing (β-mixing, information regular, ψ-mixing, has finite memory). To obtain this result we observe that $f_0^{-1}(x_0, \mathcal{Y}_0) = \mathcal{Z}_0$ holds for all $x_0 \in X_0$, which implies equality in (2) in the proof of Theorem 15.4 (the same holds for the other dependence measures). This equality, in turn, follows from the supplement at the end of Paragraph E.6 in Appendix E.

Based on this equivalence it is now easy to find nontrivial examples showing that in (a), (b), (c), and (d) of Theorem 13.9 the reversed implication is not true in general. We simply take the probability measures (distributions of the random processes) considered in Example 12.4 as the noise source of the additive noise channel.[24]

Some further comments on an example discussed in (Mittelbach and Jorswieck, 2013) are in order. If the additive noise is second order stationary Gaussian with a rational spectral density (see Paragraph C.3), then we have a Gaussian channel in the sense of Example 2.5, which is information regular due to (12.5.iii) and the previously derived equivalence. If the spectral density is truly rational, i. e., has an autoregressive part, then the covariance function is not concentrated in a finite interval so that the corresponding Gaussian channel is not ψ-mixing due to (12.5.ii) and the first equivalence in (12.5.i). Recall that we have already shown in Theorem 13.10 that ψ-mixing Gaussian channels have finite output memory. Taking the additive noise as in (12.4.ii) for the discrete-time case or as in Example C.4 for the continuous-time case gives specific Gaussian channels, which are information regular but not ψ-mixing. Stationary Gaussian noise with a rational spectral density results from passing stationary white Gaussian noise through a (well-behaved) linear filter. It is therefore relevant to model various practical situations. This example demonstrates that the original formulation of the coding theorem of Kadota and Wyner (1972) for ψ-mixing channels does not apply to important channel models, which are covered by Theorem 9.1.

To generalize the additive noise channel to m-dimensional real vector-valued signals, we simply have to use the alphabets $X_0 = Y_0 = Z_0 = \mathbb{R}^m$ and vector addition in (1). Of course, any set for which an addition is declared, can serve as alphabet.

(16.4) Totally ergodic vs. totally ergodic for block i.i.d. inputs. Suppose we consider the additive noise channel introduced in Paragraph 16.3 for discrete time and with the following modifications. We assume binary alphabets, i. e., $X_0 = Y_0 = Z_0 = \{0, 1\}$, equipped with the corresponding power set as σ-algebra and instead of ordinary addition the operation is addition modulo 2, i. e., for the component function f_0 we have

$$y_0 = f_0(x_0, z_0) = x_0 + z_0 \mod 2$$

for all $(x_0, z_0) \in X_0 \times Z_0$.

[24] The constructed random sequences in Example 12.4 have values in smaller spaces with smaller σ-algebras. However, a "translation" into real-valued sequences equipped with the usual Borel-σ-algebra is straightforward for the following reason. If $(\Omega_1, \mathcal{F}_1, \mathrm{P}_1)$ is a probability space and $(\Omega_2, \mathcal{F}_2)$ is a measurable space such that $\Omega_1 \in \mathcal{F}_2$ and $\mathcal{F}_1 = \mathcal{F}_2 \cap \Omega_1$, then P_2 with $\mathrm{P}_2(F_2) = \mathrm{P}_1(F_2 \cap \Omega_1)$ for all $F_2 \in \mathcal{F}_2$ is a probability measure on $\mathcal{F}_2$, which is equal to P_1 when restricted to $\mathcal{F}_1$.

Since the channel function f is invariant it follows from d) in (15.4.iv) and (b) in Theorem B.7 that the integration channel κ is stationary and totally ergodic if the noise source λ is stationary and weakly mixing. In the present case even the converse is true: The channel κ is stationary and totally ergodic if and only if the noise source λ is stationary and weakly mixing.

Indeed, the function $f_0(x_0, \cdot)$ is bijective for all $x_0 \in X_0$, which implies the function $f(x, \cdot)$ is bijective for all $x \in X$ due to the specific structure of f. Therefore, the function g defined in Remark 15.2 on the basis of the channel function f is also bijective. Because we have finite alphabets equipped with the power sets as σ-algebras the product input, output, and noise spaces $(X, \mathcal{X})$, $(Y, \mathcal{Y})$, and $(Z, \mathcal{Z})$ are Polish spaces, i. e., separable completely metrizable topological spaces, were the corresponding Borel-σ-algebras are equal to the product-σ-algebras $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ (see (Cohn, 1980, Prop. 8.1.3 and Prop. 8.1.5)). Together with the bijectivity and $\mathcal{X} \otimes \mathcal{Z}/\mathcal{X} \otimes \mathcal{Y}$-measurability of g we obtain from the measurability theorem of Kuratowski (see (Cohn, 1980, Th. 8.3.7)) that the inverse function g^{-1} is $\mathcal{X} \otimes \mathcal{Y}/\mathcal{X} \otimes \mathcal{Z}$-measurable. Part a) of (B.13.ii) together with Remark B.14 therefore imply that the channel input-output probability measure $\mu\kappa = (\mu \otimes \lambda)_g$ is s-ergodic if and only if the product $\mu \otimes \lambda$ of the channel input measure μ and the noise measure λ is s-ergodic. From a) in (B.15.i), (b) in Theorem B.7, Remark B.16, and the definition in (2.7.ii) we therefore obtain that the channel κ is stationary and totally ergodic if and only if the noise measure λ is stationary and weakly mixing. This characterization is a version of (Nakamura, 1975, p. 217, Cor. 2).

Assume that we take as noise measure of this channel the distribution of the random sequence constructed in Example B.8, which is stationary and totally ergodic but not weakly mixing. Then the channel κ is stationary and totally ergodic for block i.i.d. inputs due to b) in (15.4.iv). However, from the previously derived equivalence it follows that κ is not totally ergodic. Therewith we have a simple binary but nontrivial example showing that total ergodicity for block i.i.d. inputs is indeed less restrictive than total ergodicity, i. e., the reversed implication of (h) in Theorem 13.9 is not true in general.

(16.5) Multiplicative noise. Consider the setting as in Paragraph 16.3 with the difference that the component function f_0 is defined by

$$y_0 = f_0(x_0, z_0) = z_0 \cdot x_0 \tag{1}$$

for all $(x_0, z_0) \in X_0 \times Z_0$. Then we have a channel with real scalar input and output signals and multiplicative noise. It is a simple measure-theoretic result, that f_0 is $\mathcal{X}_0 \otimes \mathcal{Z}_0/\mathcal{Y}_0$-measurable. Alternatively, we can consider the alphabets $X_0 = \mathbb{R}^m$, $Y_0 = \mathbb{R}^n$, and $Z_0 = \mathbb{R}^{n \times m}$. In this case the product symbol in (1) represents matrix multiplication of the input vector with the noise matrix. The channel is known as multiple-input multiple-output channel.

Properties. As for the additive noise channel stationarity, ergodicity and mixing properties of the channel are implied by the corresponding property of the (multiplicative) noise source because the channel is causal, input-memoryless, and the channel function is invariant.

A multiplicative noise channel can be a purely random channel of the type mentioned below (16.2.3). Assume that we have scalar inputs, outputs and noise samples and the input alphabet is reduced to the set $X_0 = \{-1, 1\}$ equipped with the corresponding power set as σ-algebra. Further, assume that we have discrete time and the noise measure λ is such that the projections $\{\zeta_k, k \in \mathbb{Z}\}$ from Z to Z_k are independent Gaussian random variables with $\mathrm{E}(\zeta_k) = 0$. Then the resulting channel is a purely random channel.

The additive and multiplicative noise channel can have either discrete- or continuous-time. The components of the channel function operate only on the current input and noise sample. In the next example, the component function depends also on previous inputs.

(16.6) Modulo 2 addition of consecutive binary inputs. Let κ be a stationary deterministic discrete-time channel with binary input and output alphabets, i. e., $X_0 = Y_0 = \{0,1\}$ with $\mathcal{X}_0$ and $\mathcal{Y}_0$ being the corresponding power sets. The component $\hat{f}_0$ of the input-output mapping $\hat{f}$ is defined on $X_{-1} \times X_0$ and its values in Y_0 are given by

$$y_0 = \hat{f}_0(x_{-1}, x_0) = x_{-1} + x_0 \mod 2$$

for all $(x_{-1}, x_0) \in X_{-1} \times X_0$. In other words the current channel output results from the modulo 2 addition of the current and the previous input. That $\hat{f}_0$ is $\mathcal{X}_{-1} \otimes \mathcal{X}_0/\mathcal{Y}_0$-measurable is trivial because we consider the power set as σ-algebra on the domain of $\hat{f}_0$.

Properties. This stationary, causal channel has finite input memory and, as any deterministic channel, is output-memoryless. For a continuous-time version of this channel see (Mittelbach, 2012, Exm. 4.37). We are interested in this channel for the following reason. Let μ_0 be a probability measure on $\mathcal{X}_0$ with $\mu_0(\{1\}) = q$, i. e., a Bernoulli distribution with parameter $q \in [0,1]$. We assume that the channel input probability measure μ on $\mathcal{X}$ is given by

$$\mu = \bigotimes_{k\in\mathbb{Z}} \langle \mu_0 \rangle_k .$$

By $\{\xi_k, k \in \mathbb{Z}\}$ and $\{\eta_k, k \in \mathbb{Z}\}$ we denote the sequences of coordinate projections on the channel input-output space, where ξ_k is the projection from $X \times Y$ to X_k and η_k is the projection from $X \times Y$ to Y_k. On the one hand the channel satisfies the conditions of the coding theorem of Kadota and Wyner (1972) and the pair sequence (ξ, η) satisfies the conditions of Corollary 4.14. On the other hand the sequence $\{I(\xi_0^n; \eta_0^n), n \in \mathbb{N}\}$ of mutual informations is strictly concave in n for all $q \in (0, 1/2) \cup (1/2, 1)$, as shown by the author in (Mittelbach, 2012, Rmk. 2.27, Exm. 2.28). This demonstrates the insufficiency of the convexity argument used in (Kadota and Wyner, 1972, Appendix II) to prove the monotonicity of the sequence $\{n^{-1} I(\xi_0^n; \eta_0^n), n \in \mathbb{N}\}$.

In the remainder of this section κ is a channel with time structure as introduced in Definition 2.3, not necessarily an integration channel.

(16.7) Channels with state. We consider a discrete-time channel and in addition to the input product space $(X, \mathcal{X})$ and the output product space $(Y, \mathcal{Y})$ let $(U, \mathcal{U})$ be a product space generated by the family $\{(U_k, \mathcal{U}_k), k \in \mathbb{Z}\}$ of measurable spaces with $(U_k, \mathcal{U}_k) = (U_0, \mathcal{U}_0)$. In this model U_0 represents the set of possible channel states.

Assume that $\hat{\kappa}_0$ is a Markov kernel from $(U_0 \times X_0, \mathcal{U}_0 \otimes \mathcal{X}_0)$ to $(Y_0, \mathcal{Y}_0)$. Let us define for all $u = \{u_k, k \in \mathbb{Z}\} \in U$, $x = \{x_k, k \in \mathbb{Z}\} \in X$, $l \in \mathbb{N}$, and $B_k \in \mathcal{Y}_k$ with $k \in \{-l, -(l-1), \ldots, l\}$

$$\hat{\kappa}\big(u, x, [B_{-l} \times B_{-(l-1)} \times \ldots \times B_l]\big) = \prod_{k=-l}^{l} \hat{\kappa}_0(u_k, x_k, B_k).$$

Then, similar to (A.3.iv), $\hat{\kappa}$ can be uniquely extended to a Markov kernel from $(U \times X, \mathcal{U} \otimes \mathcal{X})$ to $(Y, \mathcal{Y})$ denoted also by $\hat{\kappa}$. According to Example 13.8 this Markov kernel represents a stationary

memoryless channel with input space $(U \times X, \mathcal{U} \otimes \mathcal{X})$ and output space $(Y, \mathcal{Y})$. Given $\hat{\lambda}$ is a probability measure on $\mathcal{U}$ we define the discrete-time channel κ with input space $(X, \mathcal{X})$ and output space $(Y, \mathcal{Y})$ by

$$\kappa(x, B) = \int_U \hat{\kappa}(u, x, B) \, \mathrm{d}\hat{\lambda}(u) \tag{1}$$

for all $x \in X$ and $B \in \mathcal{Y}$.

In state $u_0 \in U_0$ the transmission of a symbol is characterized by the channel $\hat{\kappa}_0(u_0, \cdot, \cdot)$. If all channel states are known for the transmission of a sequence of symbols, then the transmission is memoryless as represented by $\hat{\kappa}$. The state sequence is drawn randomly according to the probability measure $\hat{\lambda}$. The resulting channel κ between input and output is thus given by (1).

Properties. Since $\hat{\kappa}$ is memoryless we obtain with (1) that κ is causal and input-memoryless. Since $\hat{\kappa}$ is stationary we obtain also with (1) that κ is stationary if the distribution $\hat{\lambda}$ of the state process is stationary. Whether or not the channel satisfies one of the output memory conditions depends solely on $\hat{\lambda}$. The channel κ is α-mixing (β-mixing, information regular, ψ-mixing, has finite output memory), if the state distribution $\hat{\lambda}$ is α-mixing (β-mixing, information regular, ψ-mixing, has finite memory). Because $\hat{\kappa}$ is memoryless the derivations of these results are identical to the proofs of (13.11.i) to (13.11.v) with $\hat{\kappa}$ in the role of the channel, $\hat{\lambda}$ in the role of the input probability measure, and κ in the role of the input-output probability measure. Regarding the mixing properties in the ergodic-theoretic sense the comment in the second paragraph of Remark 15.5 applies.

The input-output probability measure $\mu\kappa$ induced by the channel κ and a probability measure μ on $\mathcal{X}$ is given for any $C \in \mathcal{X} \otimes \mathcal{Y}$ by

$$\begin{aligned} \mu\kappa(C) &= \int_X \kappa(x, C_x) \, \mathrm{d}\mu(x) \\ &= \int_{U \times X} \hat{\kappa}(u, x, C_x) \, \mathrm{d}\hat{\lambda} \otimes \mu(u, x) \end{aligned}$$

due to (1) and part (A.8.i) of Fubini's theorem. Since $\hat{\kappa}$ is stationary and memoryless it is totally ergodic due to Theorem 13.9 so that any s-stationary and s-ergodic probability measure $\hat{\mu}$ on $\mathcal{U} \otimes \mathcal{X}$ induces together with $\hat{\kappa}$ an s-stationary and s-ergodic probability measure $\hat{\mu}\hat{\kappa}$ on $\mathcal{U} \otimes \mathcal{X} \otimes \mathcal{Y}$. Therefore we obtain with a) in (B.15.i), Lemma B.11, and (a) and (b) in Theorem B.7: If the state distribution $\hat{\lambda}$ is stationary and totally ergodic, then the channel κ is stationary and totally ergodic for block i.i.d. inputs. If $\hat{\lambda}$ is stationary and weakly mixing, then κ is stationary and totally ergodic.

A classical special case of the described model is the Gilbert-Elliott channel (Gallager, 1968, pp. 98, 99). It has binary input and output alphabets, i. e., $X_0 = Y_0 = \{0, 1\}$, and two possible states, i. e., $U_0 = \{0, 1\}$. The channel $\hat{\kappa}_0(0, \cdot, \cdot)$ in state 0 is a binary symmetric channel with a low error probability and the channel $\hat{\kappa}_0(1, \cdot, \cdot)$ in state 1 is a binary symmetric channel with high error probability. During transmission the binary state process switches randomly between these two channels according to the distribution $\hat{\lambda}$, that is equal to the distribution of the Markov chain constructed in (12.4.i). The channel models the transmission of bits over a "good" or a "bad" channel, where the random channel selection is governed by a Markov chain. If the parameter ϵ, characterizing the transition probabilities of the Markov chain in (12.4.i), satisfies $\epsilon \in (0, 1)$, then the Gilbert-Elliott channel is ψ-mixing.

We conclude the section with an example of theoretical interest.

(16.8) Averaged channels. With this example we demonstrate that it is easy to construct channels for which central results of this thesis, e. g., Theorem 9.1 do not apply. Assume that $\dot{\kappa}$ and $\ddot{\kappa}$ are channels with time structure as in Definition 2.3, both with input signal space $(X, \mathcal{X})$ and output signal space $(Y, \mathcal{Y})$. For some constant $\alpha \in (0, 1)$ we define for all $x \in X$ and $B \in \mathcal{Y}$

$$\bar{\kappa}(x, B) = \alpha\dot{\kappa}(x, B) + (1 - \alpha)\ddot{\kappa}(x, B).$$

Then $\bar{\kappa}$ is a channel (see (A.3.v)) with input space $(X, \mathcal{X})$ and output space $(Y, \mathcal{Y})$, called averaged channel. The name is adopted from Ahlswede (1968), who attributed the introduction of averaged channels to Jacobs (1962a). In (Han, 2003, Sec. 3.3) those channels are called mixed channels.

We easily verify: If $\dot{\kappa}$ and $\ddot{\kappa}$ are both stationary (causal, asymptotically input-memoryless, have finite input memory), then $\bar{\kappa}$ is stationary (causal, asymptotically input-memoryless, has finite input memory). However, ergodicity is critical. If μ is a probability measure on the channel input space, then we have

$$\mu\bar{\kappa} = \alpha\mu\dot{\kappa} + (1 - \alpha)\mu\ddot{\kappa}, \tag{1}$$

where $\mu\bar{\kappa}$, $\mu\dot{\kappa}$, and $\mu\ddot{\kappa}$ denote the channel input-output probability measures induced by μ together with $\bar{\kappa}$, $\dot{\kappa}$, and $\ddot{\kappa}$, respectively. Suppose for some $s \in T_+$ the channels $\dot{\kappa}$ and $\ddot{\kappa}$ are s-stationary and s-ergodic. Further suppose $\dot{\kappa}$ and $\ddot{\kappa}$ are distinct in the sense that there exists at least one s-stationary and s-ergodic probability measure μ on $\mathcal{X}$ such that $\mu\dot{\kappa} \neq \mu\ddot{\kappa}$. If μ is such a probability measure, then $\mu\dot{\kappa}$ and $\mu\ddot{\kappa}$ are s-stationary and s-ergodic according to the definition in (2.7.ii). However, due to Lemma B.9 and (1) the probability measure $\mu\bar{\kappa}$ is not s-ergodic, which implies that the averaged channel $\bar{\kappa}$ is not s-ergodic. This observation directly extends to averages of channels that are totally ergodic (for block-i.i.d. inputs). Even the averages of memoryless channels (see Example 13.8) are not ergodic. Averaged channels are well known for their non-ergodicity.

§ 17 Signal Processing, Composed Models

Does a certain signal processing operation, applied to the input or output of a channel, have an impact on relevant properties of the channel, e. g., does it modify the input or output memory? A useful approach to answer this question is to represent the signal processing operation as a deterministic channel. Then we can consider the original channel together with the signal processing as cascade channel, to which we can apply the results from Section § 14. In this section we formulate typical examples such as linear filtering, quantization, and thresholding in terms of the channel framework used in this thesis. We identify properties of these deterministic channels and then consider a random version of a linear filter, which can serve as model of a wireless communication link. Then we study some aspects of connecting channels in cascade with a particular focus on channels composed of a linear filter and an additive noise channel. For these cascade channels we are able to verify the asymptotic input-memorylessness in some special cases, even though the filter as a single component does not satisfy this property. At the end of this section we list applications from various fields. In particular, we discuss memory properties that allow to apply the central limit theorem to the calculation of the Fourier transform of stationary sequences.

(17.1) Time-invariant filter. Suppose κ is a discrete-time deterministic channel. We characterize the input-output relation of the deterministic channel with the function $\hat{f}$ as introduced in Paragraph 16.1. The time index set T is equal to $\mathbb{Z}$ and the input and output spaces $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ are product spaces as introduced in Definition 2.3 for channels with time-structure. We assume real-valued inputs and outputs, i. e., $X_0 = Y_0 = \mathbb{R}$, where both sets are equipped with the usual Borel-σ-algebra. Suppose $\hat{f}$ is defined by component functions $\hat{f}_k$ with parameters u_x and v_x as specified at the end of Paragraph 16.1 and in Paragraph 15.6. We assume that $\hat{f}_k(\cdot) = \hat{f}_0(\langle\cdot\rangle_{-k})$ for all $k \in \mathbb{Z}$ such that $\hat{f}$ is invariant. The specifications so far imply stationarity and output-memorylessness of κ. We further assume that $v_x = 0$ so that κ is causal.

The channel represents a causal time-invariant discrete-time filter, where the current output value is the result of filtering the current and past input values by (shifted versions of) the function $\hat{f}_0$. The filter can be nonlinear in general, however, let us restrict ourselves to the linear case. First assume that $u_x = n$ is a finite positive integer. Then $\hat{f}_0$ is defined on X_{-n}^0 with values in Y_0 by

$$\hat{f}_0(x_{-n}^0) = \sum_{i=0}^{n-1} a_i x_{-i}$$

for all $x_{-n}^0 = (x_{-n+1}, \ldots, x_{-1}, x_0) \in X_{-n}^0$, where $a_0, a_1, \ldots, a_{n-1} \in \mathbb{R}$ denote the filter coefficients. Because $\hat{f}_0$ is continuous, it is a basic measure-theoretic fact that it is $\mathcal{X}_{-n}^0/\mathcal{B}(\mathbb{R})$-measurable. The filter function depends on $n-1$ past inputs so that the channel has finite input memory with a memory length of $n-1$.

Given u_x is infinite, then $\hat{f}_0$ is defined on X_-^0. Assume that $\{a_i, i \in \mathbb{N}_0\}$ is a fixed sequence of filter coefficients $a_i \in \mathbb{R}$. We define

$$\hat{f}_0(x_-^0) = \lim_{n\to\infty} \sum_{i=0}^{n-1} a_i x_{-i} =: \sum_{i=0}^{\infty} a_i x_{-i} \tag{1}$$

for all $x_-^0 \in X'$, where

$$X' := \Big\{ x_-^0 = \{x_{-i}, i \in \mathbb{N}_0\} \in X_-^0 : \sum_{i=0}^{\infty} a_i x_{-i} \text{ exists and is finite} \Big\}.$$

We extend this function on X' to a $\mathcal{X}_-^0/\mathcal{B}(\mathbb{R})$-measurable function on all of X_-^0. This is possible due to the derivations in Paragraph A.13, which also yield $X' \in \mathcal{X}_-^0$. If infinitely many filter coefficients are nonzero, then the filter function depends on infinitely many past inputs so that the channel has infinite input memory. From the discussion in Paragraph 16.1 on the input memory of deterministic channels we obtain that κ is *not* asymptotically input-memoryless.

We continue with the (more interesting) case of a linear time-invariant filter channel with infinite input memory. Assume that μ is a probability measure on the channel input space $(X, \mathcal{X})$ and μ_-^k denotes the marginal measure of μ on $\mathcal{X}_-^k$. Let $\xi = \{\xi_k, k \in \mathbb{Z}\}$ denote the sequence of input coordinate projections, where ξ_k is the projection from X to X_k. Suppose

$$\sup_{k\in\mathbb{Z}} \mathrm{E}|\xi_k| = \sup_{k\in\mathbb{Z}} \int_X |\xi_k(x)| \,\mathrm{d}\mu(x) < \infty \tag{2}$$

holds, which is for example the case when μ is stationary and $\mathrm{E}|\xi_0| < \infty$. Further, suppose the filter coefficients are absolutely summable, i. e.,

$$\sum_{i=0}^{\infty} |a_i| < \infty. \tag{3}$$

Then we have

$$\sum_{i=0}^{\infty} \mathrm{E}|a_i \xi_{k-i}| \leq \left(\sup_{i \in \mathbb{N}_0} \mathrm{E}|\xi_{k-i}| \right) \cdot \sum_{i=0}^{\infty} |a_i| < \infty,$$

for all $k \in \mathbb{Z}$. Together with the results in Paragraph A.13 this implies $\mu_-^k(\langle X' \rangle_k) = 1$. Thus, under the conditions (2) and (3) each component function $\hat{f}_k(\cdot) = \hat{f}_0(\langle \cdot \rangle_{-k})$ is μ_-^k-almost surely defined by a shifted version of (1) so that it does not matter how $\hat{f}_k$ is extended outside $\langle X' \rangle_k$. It follows that the channel function $\hat{f}$ is μ-almost surely defined. Given $\eta = \{\eta_k, k \in \mathbb{Z}\}$ is the sequence of output coordinate projections, where η_k denotes the projection from Y to Y_k, then we have

$$\eta = \hat{f}(\xi) \qquad \text{and} \qquad \eta_k = \hat{f}_k(\xi_-^k) = \sum_{i=0}^{\infty} a_i \xi_{k-i}, \tag{4}$$

where the last equality holds μ-almost surely.

Let us consider an example. Suppose a channel input probability measure μ is given such that $\xi = \{\xi_k, k \in \mathbb{Z}\}$ is an i.i.d.-sequence of Gaussian random variables with expectation $\mathrm{E}(\xi_0) = 0$ and finite variance $\mathrm{var}(\xi_0) = \sigma^2$. Further assume that the filter coefficients $\{a_i, i \in \mathbb{N}_0\}$ are given by $a_i = \rho^i$, where ρ is a fixed constant satisfying $|\rho| < 1$. Then we obtain

$$\sup_{k \in \mathbb{Z}} \mathrm{E}|\xi_k| = \mathrm{E}|\xi_0| = \sqrt{2\sigma^2/\pi} < \infty \qquad \text{and} \qquad \sum_{i=0}^{\infty} |a_i| = \frac{1}{1 - |\rho|} < \infty$$

using (Simon, 2006, eq. (2.4)) and the properties of the geometric series. Consequently, the right-hand side of (4) converges μ-almost surely to a finite value for all $k \in \mathbb{Z}$. The sequence $\eta = \{\eta_k, k \in \mathbb{Z}\}$ at the channel output results from linearly filtering the i.i.d. Gaussian input sequence ξ and we observe this example is identical to that given in (12.4.ii), formulated here in terms of a deterministic channel. Since ξ is an i.i.d.-sequence it is information regular and due to the derivations in (12.4.ii) the sequence η is information regular as well. We emphasize that η is information regular even though the channel κ is not asymptotically input-memoryless for the input signal set X'. Thus, the example illustrates that the input memory condition formulated in (13.11.iii) for κ to obtain an information regular output is not a necessary condition. This example can be extended to a more general situation: A second order stationary Gaussian process with rational spectral density is information regular according to (12.5.iii) and the second equivalence in (12.5.i). Such processes are ARMA processes and result from passing stationary white Gaussian noise through a (well-behaved) time-invariant linear filter (see Paragraph C.3). Thus, passing a stationary Gaussian process with rational spectral density through a linear filter is the same as passing stationary white Gaussian noise through a cascade of two linear filters, which is itself a linear filter. The output process is therefore stationary Gaussian with rational

spectral density. It follows that linearly filtering a stationary Gaussian process with rational spectral density results in an information regular Gaussian process, whether or not the filter has finite input memory.

As a second example assume that the channel input probability measure μ is such that ξ is an i.i.d.-sequence with $\mu(\xi_0 = 0) = \mu(\xi_0 = 1) = 1/2$ and the coefficients of the linear filter are given by $a_i = \frac{1}{2}(1/2)^i$ for $i \in \mathbb{N}_0$. The conditions (2) and (3) are satisfied so that we have μ-almost sure convergence on the right-hand side of (4) and we observe the situation is identical to that in example (12.4.v). Consequently, the output sequence η is not α-mixing and therefore, all the output memory conditions that are based on dependence measures are not satisfied. Comparing this infinite input memory channel with the one from the previous example shows: Even though both channels are identical (linear filter with coefficients from a geometric series) and both inputs have the same memory properties (i.i.d. inputs, i. e., there is no memory at all) the channel outputs have memory properties of rather different quality. This illustrates effects and difficulties for channels having infinite input memory. In Paragraph 17.5 we consider a linear time-invariant filter that is not asymptotically input-memoryless but in connection with additive noise the composed channel satisfies this condition.

(17.2) Quantization. We consider now a stationary deterministic channel with time structure. The input space $(X, \mathcal{X})$ is the product of an arbitrary input alphabet $(X_0, \mathcal{X}_0)$ and the output space $(Y, \mathcal{Y})$ is the product of the real alphabet $(Y_0, \mathcal{Y}_0) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The invariant channel function $\hat{f}$ between input and output is defined by component functions $\hat{f}_t(\cdot) = \hat{f}_0(\langle\cdot\rangle_{-t})$ for all $t \in T$. Let $\{A_1, A_2, \dots, A_n\}$ be a partition of X_0 with $A_i \in \mathcal{X}_0$ and let $b_1, b_2, \dots, b_n \in Y_0$ be distinct numbers. The component function $\hat{f}_0$ is defined on X_0 with values in Y_0 by

$$\hat{f}_0 = \sum_{i=1}^{n} b_i \mathbb{1}_{A_i}.$$

It represents the quantization of the values from the alphabet X_0. The $\mathcal{X}_0/\mathcal{Y}_0$-measurability of $\hat{f}_0$ (and therefore of $\hat{f}$) is assured as long as the sets A_i are taken from the σ-algebra $\mathcal{X}_0$. If the inputs are real scalars, then the A_i's are usually intervals and the b_i's are the center points of the intervals. If the inputs are real vectors, then the A_i's are for example Voronoi cells. In this case real vectors are considered as output alphabet and the b_i's are chosen as centroid of the corresponding cell. This stationary channel is causal, output-memoryless (as any deterministic channel), and input-memoryless such that it is memoryless (see Example 13.8).

(17.3) Thresholding. Assume that we have a stationary deterministic channel with time-structure as in Paragraph 17.2 with real inputs and outputs, i. e., $(X_0, \mathcal{X}_0) = (Y_0, \mathcal{Y}_0) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Let us fix some $\epsilon > 0$. The component function $\hat{f}_0$ of the invariant mapping $\hat{f}$ between input and output is defined on X_0 with values in Y_0 by

$$\hat{f}_0(x_0) = x_0 \mathbb{1}_{A_\epsilon}(x_0)$$

for all $x_0 \in X_0$, where the set A_ϵ is given by

$$A_\epsilon = \{x_0 \in X_0 : |x_0| \geq \epsilon\}.$$

The operation means all signal values below a certain constant (threshold) are set to 0. The threshold operation is applied for example in wireless sensor networks (Boche and Mönich,

2009). In such networks the sensors transmit only if the signal values are above a certain threshold, which saves energy. The function $\hat{f}_0$ is $\mathcal{X}_0/\mathcal{Y}_0$-measurable because $A_\epsilon \in \mathcal{X}_0$, the identity is measurable, and the pointwise product of measurable functions is measurable. The properties of this channel are identical to those of the quantization channel in Paragraph 17.2, i. e., it is stationary and memoryless.

(17.4) Time-variant and random filter. Consider the discrete-time deterministic channel κ specified at the beginning of Paragraph 17.1 with the difference, that the component functions $\hat{f}_k$ vary with the time index k, i. e., they are not necessarily shifted versions of $\hat{f}_0$. Then this is a model for a time-variant filter. For example, assume that $n \in \mathbb{N}_0$ is fixed and for all $k \in \mathbb{Z}$ the function $\hat{f}_k$ is defined on X_{k-n}^k with values in Y_k by

$$\hat{f}_k(x_{k-n}^k) = \sum_{i=0}^{n-1} a_{k,i} x_{k-i}$$

for all $x_{k-n}^k = (x_{k-n+1}, \dots, x_{k-1}, x_k) \in X_{k-n}^k$, where $a_{k,0}, a_{k,1}, \dots, a_{k,n-1} \in \mathbb{R}$ denote the filter coefficients, that vary (deterministically) with time index k. This linear time-variant filter channel is causal, output-memoryless, and has finite input memory with memory length $n-1$. However, the channel is not stationary and will therefore not be further considered.

We continue with filters that change randomly over time and restrict ourselves to the linear case. A wireless communication link, for example, with a randomly changing channel impulse response can be modeled by such a filter. Suppose κ is an integration channel with time structure as introduced in Example 15.3 and the channel function f is specified by component functions as in Paragraph 15.6. We adopt the real input and output alphabet and the discrete time axis from above and assume that the component functions satisfy $f_k(\cdot) = f_0(\langle\cdot\rangle_{-k})$ for all $k \in \mathbb{Z}$, where the shift $\langle\cdot\rangle_{-k}$ operates on the product of the input and noise (sub-) space. The corresponding channel function f is invariant by construction.

We assume that the components $(Z_k, \mathcal{Z}_k)$ of the noise sequence space $(Z, \mathcal{Z})$ are themselves product spaces given by

$$Z_k = \bigtimes_{i \in \mathbb{N}_0} Z_{k,i}, \qquad \mathcal{Z}_k = \bigotimes_{i \in \mathbb{N}_0} \mathcal{Z}_{k,i}, \tag{1}$$

where $Z_{k,i} = \mathbb{R}$ and $\mathcal{Z}_{k,i} = \mathcal{B}(\mathbb{R})$. We further assume that the component function f_0 is defined on $X_-^0 \times Z_0$ with values in Y_0. We define

$$f_0(x_-^0, z_0) = \lim_{n\to\infty} \sum_{i=0}^{n-1} z_{0,i} x_{-i} =: \sum_{i=0}^{\infty} z_{0,i} x_{-i} \tag{2}$$

for all $(x_-^0, z_0) \in U'$, where

$$U' := \Big\{ (x_-^0, z_0) = \{(x_{-i}, z_{0,i}), i \in \mathbb{N}_0\} \in X_-^0 \times Z_0 : \sum_{i=0}^{\infty} z_{0,i} x_{-i} \text{ exists and is finite} \Big\}.$$

We extend this function on U' to an $\mathcal{X}_-^0 \otimes \mathcal{Z}_0/\mathcal{B}(\mathbb{R})$-measurable function on all of $X_-^0 \times Z_0$. It is shown in Paragraph A.13 that this is always possible and also that $U' \in \mathcal{X}_-^0 \otimes \mathcal{Z}_0$ holds. For all $(x_-^0, z_0) \in U'$ the value $f(x_-^0, z_0)$ is the result of linearly filtering the input values $x_-^0 =$

$\{x_i, i \in \mathbb{N}_0\}$ with the coefficients $z_0 = \{z_{0,i}, i \in \mathbb{N}_0\}$. Subsequently, we specify conditions on the noise source and the channel input measure such that U' is a set of measure 1. The situation simplifies, when only finitely many, say n, filter coefficients are considered. Then the product space and σ-algebra in (1) are composed of n components, the set U' is not required, and f_0 is defined for all $(x_{-n}^0, z_0) \in X_{-n}^0 \times Z_0$ by the middle sum in (2) without the limit. For $n = 1$ we have the multiplicative noise channel introduced in Paragraph 16.5.

Let $\xi = \{\xi_k, \in \mathbb{Z}\}$, $\eta = \{\eta_k, \in \mathbb{Z}\}$, and $\zeta = \{\zeta_k, k \in \mathbb{Z}\}$ denote the sequences of coordinate projections on the input, output, and noise space, where ξ_k is the projection from X to X_k, η_k is the projection from Y to Y_k, and ζ_k is the projection from Z to Z_k. Further, let $\zeta_{k,i}$ denote the projection from Z to $Z_{k,i}$. In expressions containing several projections on different spaces it is understood that they are defined on the corresponding product space. Assume that the noise source λ describing the random variation of the filter coefficients is stationary and satisfies

$$\sum_{i=0}^{\infty} \mathrm{E}|\zeta_{0,i}| = \sum_{i=0}^{\infty} \int_Z |\zeta_{0,i}(z)| \, \mathrm{d}\lambda(z) < \infty. \tag{3}$$

This condition is the counterpart to (17.1.3) in the deterministic time-invariant case. If Z_k' denotes the set of absolutely summable filter coefficients at time index $k \in \mathbb{Z}$, i. e.,

$$Z_k' = \Big\{ z_k = \{z_{k,i}, i \in \mathbb{N}_0\} \in Z_k : \sum_{i=0}^{\infty} |z_{k,i}| < \infty \Big\},$$

then we have $\lambda_k(Z_k') = 1$, where λ_k denotes the marginal measure of λ on $\mathcal{Z}_k$. This follows from the conditions on the noise source λ and the results in Paragraph A.13.

Let μ be a probability measure on the channel input space $(X, \mathcal{X})$ and as in Paragraph 17.1 suppose it satisfies

$$\sup_{k \in \mathbb{Z}} \mathrm{E}|\xi_k| = \sup_{k \in \mathbb{Z}} \int_X |\xi_k(x)| \, \mathrm{d}\mu(x) < \infty, \tag{4}$$

which holds, for example, when μ is stationary and $\mathrm{E}|\xi_0| < \infty$. From conditions (3) and (4) together with

$$\begin{aligned} \mathrm{E}|\zeta_{k,i}\xi_{k-i}| &= \int_{X \times Z} |\zeta_{k,i}(z)\xi_{k-i}(x)| \, \mathrm{d}\mu \otimes \lambda(x, z) \\ &= \int_Z |\zeta_{k,i}(z)| \, \mathrm{d}\lambda(z) \int_X |\xi_{k-i}(x)| \, \mathrm{d}\mu(x) \\ &= \mathrm{E}|\zeta_{k,i}| \, \mathrm{E}|\xi_{k-i}| \end{aligned}$$

and the stationarity of λ we obtain

$$\begin{aligned} \sum_{i=0}^{\infty} \mathrm{E}|\zeta_{k,i}\xi_{k-i}| &= \sum_{i=0}^{\infty} \mathrm{E}|\zeta_{k,i}| \, \mathrm{E}|\xi_{k-i}| \\ &\le \Big(\sup_{i \in \mathbb{N}_0} \mathrm{E}|\xi_{k-i}| \Big) \cdot \sum_{i=0}^{\infty} \mathrm{E}|\zeta_{k,i}| < \infty. \end{aligned}$$

Then the results in Paragraph A.13 imply $\mu_-^k \otimes \lambda_k(\langle U' \rangle_k) = 1$, where μ_-^k is the marginal measure of μ on $\mathcal{X}_-^k$. Thus, under the conditions (3) and (4) each component function $f_k(\cdot) = f_0(\langle \cdot \rangle_{-k})$ is $\mu_-^k \otimes \lambda_k$-almost surely defined by a shifted version of (2) so that it does not matter how f_k is extended outside $\langle U' \rangle_k$. It follows that the channel function f is $\mu \otimes \lambda$-almost surely defined. At the channel output we have

$$\eta = f(\xi, \zeta) \qquad \text{and} \qquad \eta_k = f_k(\xi_-^k, \zeta_k) = \sum_{i=0}^{\infty} \zeta_{k,i} \xi_{k-i},$$

where the last equality holds $\mu \otimes \lambda$-almost surely.

The channel function f is invariant and the noise source λ is stationary. Therefore, the resulting channel κ is stationary. It is also causal and because the channel function f satisfies the measurability condition in (15.4.v) the mixing property of the noise source λ implies the corresponding mixing property of the channel κ. With a finite number of filter coefficients the channel has finite input memory. Verifying the asymptotic input-memorylessness of a random linear time-variant filter with an infinite number of filter coefficients can be difficult. In Paragraph 17.6 we derive the asymptotic input-memorylessness of a filter of this type when it is combined with an additive noise channel.

With the channels defined in this section and Section § 16 we can build composed models. For example, connecting the multiplicative and additive noise channel from Paragraphs 16.3 and 16.5 results in a cascade channel, usually referred to as flat fading channel. From the properties identified in Paragraphs 16.3 and 16.5 for the single components we easily obtain properties of the composed channel by applying Theorems 14.2 and 14.4.

We observe that a deterministic memoryless channel has no impact on the memory properties of a channel, when connected to the input or output. Thus, thresholding at the channel input or quantization at the channel output preserves the memory properties of the original channel. The situation is different when the memoryless channel is not deterministic. Subsequently, we demonstrate the impact of an additive noise channel on the input memory, when connected with a linear filter. The example also illustrates the limitations of concluding properties of a composed channel from the properties of its single components. In particular the first part of the next example is formulated such that it can be easily extended to other composed channels.

(17.5) Linear time-invariant filter and additive noise. Let κ be a discrete-time channel with the input and output product spaces $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$, respectively. Assume that the channel consists of two components connected in cascade as illustrated in Figure 6. Suppose the first component is a linear time-invariant filter $\hat{f}$ as specified in Paragraph 17.1 with absolutely summable filter coefficients $\{a_i, i \in \mathbb{N}_0\}$. Let the second component $\dot{\kappa}$ be an additive noise channel as introduced in Paragraph 16.3 with noise measure space $(Z, \mathcal{Z}, \lambda)$ and channel function $\dot{f}$. As in Paragraphs 16.3 and 17.1 we assume that all alphabets are equal to the real line so that input, output, and noise signals are real-valued sequences.

According to Paragraph 15.7 the composed channel κ is an integration channel with channel function f given by

$$f(x, z) = \dot{f}\big(\hat{f}(x), z\big)$$

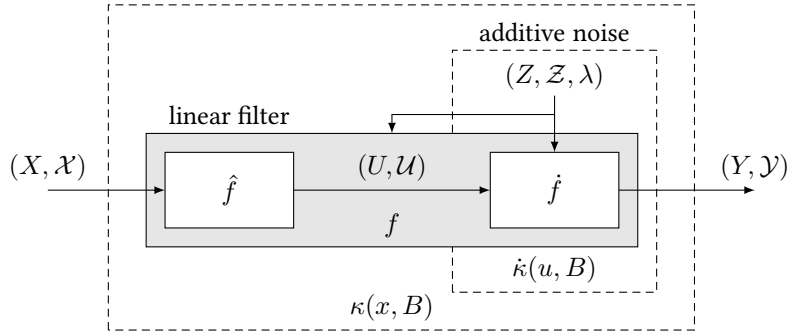


Figure 6: Cascade of linear time-invariant filter and additive noise channel.

for all $x \in X$ and $z \in Z$. The channel function f is characterized by component functions $f_k(\cdot) = f_0(\langle\cdot\rangle_{-k})$, $k \in \mathbb{Z}$, where f_0 is defined on $X_-^0 \times Z_0$ with values in Y_0 as

$$f_0(x_-^0, z_0) = \dot{f}_0\big(\hat{f}_0(x_-^0), z_0\big)$$

for all $x_-^0 \in X_-^0$ and $z_0 \in Z_0$. Here, $\hat{f}_0$ and $\dot{f}_0$ are the defining component functions of $\hat{f}$ and $\dot{f}$, respectively. Since the linear time-invariant filter is a deterministic channel, the noise source of the composed integration channel κ is equal to that of the additive noise channel $\dot{\kappa}$.

We can apply Theorem 15.4 to obtain properties of the composed integration channel κ. Alternatively, from the properties identified in Paragraphs 16.3 and 17.1 for the linear filter and the additive noise channel we can conclude properties of the channel κ using Theorems 14.2 and 14.4. For example, κ is causal because both components are causal. Since the linear time-invariant filter is output-memoryless and the additive noise channel is causal and input-memoryless the output memory of κ is determined by the output memory of the additive noise channel. Therefore, it completely depends on the noise source. The stationarity of the noise source implies the stationarity of κ because $\hat{f}$ and $\dot{f}$ are invariant. However, drawing conclusions about the input memory of κ based on (14.2.iv) is not possible in general because the linear filter is not asymptotically input-memoryless if it has an infinite number of nonzero filter coefficients.

Subsequently, we consider an example for which we can verify the asymptotic input-memorylessness of the composed channel κ, even though the linear filter does not satisfy this property. Actually, the basis for this effect is the data processing inequality given in (6.4.ii) due to which the total variation distance between two probability measures is at most decreased when the measures are passed through a channel. Let $\zeta = \{\zeta_k, k \in \mathbb{Z}\}$ denote the sequence of coordinate projections on the noise sample space, where ζ_k denotes the projection from Z to Z_k. Assume that the noise source λ is such that ζ is a second order stationary Gaussian sequence with

$$\mathrm{E}(\zeta_k) = 0, \qquad \mathrm{cov}(\zeta_j, \zeta_k) = \sigma^2 \rho^{|j-k|}, \tag{1}$$

for all $j, k \in \mathbb{Z}$, where ρ and σ^2 are real constants satisfying $|\rho| < 1$ and $\sigma^2 > 0$. According to Example C.4 or (12.4.ii) the sequence ζ is an AR(1) process. If $\rho = 0$, then[25] ζ is an i.i.d.-sequence and κ is output-memoryless. If $\rho \neq 0$, then κ has infinite output memory. It is actually information regular (but not ψ-mixing) due to the results in Paragraph 16.3.

[25] In case $\rho = 0$, we put $0^0 := 1$ in (1).

Suppose X'' denotes the set of all input sequences bounded by the finite constant $\mathrm{c} > 0$, i. e.,

$$X'' = \big\{x = \{x_k, k \in \mathbb{Z}\} \in X : \sup_{k\in\mathbb{Z}} |x_k| \le \mathrm{c}\big\}. \tag{2}$$

Since we assume that (17.1.3) holds (absolutely summable filter coefficients) the filter function $\hat{f}$ is well-defined on X'' because the series in (17.1.1) converges for all $x \in X''$. The channel κ is asymptotically input-memoryless for the signal set X'' if for all $\epsilon > 0$ there exists an $m = m(\epsilon) \in \mathbb{N}_0$ such that for all $x, \tilde{x} \in X''$ coinciding on $(-m, \infty)$ we have

$$\|\kappa_0(x,\cdot) - \kappa_0(\tilde{x},\cdot)\|_{\mathrm{tv}} \le \epsilon, \tag{3}$$

where $\kappa_0(x,\cdot)$ denotes the marginal measure of $\kappa(x,\cdot)$ on $\mathcal{Y}_0^+$. It is sufficient to verify this condition because κ is stationary and X'' is shift-invariant.

Let $\eta = \{\eta_k, k \in \mathbb{Z}\}$ denote the sequence of coordinate projections on the channel output space, where η_k denotes the projection from Y to Y_k. If $x \in X''$ is fixed, then η is a sequence of random variables on the probability space $(Y, \mathcal{Y}, \kappa(x,\cdot))$. It is a Gaussian sequence with the same covariance function as the noise sequence ζ. The expectations are given by

$$\mathrm{E}(\eta_k) = \mathrm{E}(\zeta_k) + \hat{f}_k(x_-^k) = \sum_{i=0}^{\infty} a_i x_{k-i}$$

for all $k \in \mathbb{Z}$. The random sequence η_0^+ is equal to that considered in (6.12.iii) with the covariance function specified in (6.12.10), which allows us to use the results derived there. The distribution of η_0^+ is equal to $\kappa_0(x,\cdot)$.

Given $x, \tilde{x} \in X''$ we define

$$d_k = \hat{f}_k(x_-^k) - \hat{f}_k(\tilde{x}_-^k).$$

Let us fix some $m \in \mathbb{N}_0$. Then for all $k \in \mathbb{N}$ and $x, \tilde{x} \in X''$ coinciding on $(-m, \infty)$ we have

$$|d_k| = \left| \sum_{i=k+m}^{\infty} a_i (x_{k-i} - \tilde{x}_{k-i}) \right| \le 2\mathrm{c} \sum_{i=k+m}^{\infty} |a_i|, \tag{4}$$

where the inequality holds because the sequences of X'' are bounded by the constant c. Applying Pinsker's inequality (6.10.i) we obtain

$$\|\kappa_0(x,\cdot) - \kappa_0(\tilde{x},\cdot)\|_{\mathrm{tv}} \le \sqrt{2D\big(\kappa_0(x,\cdot)\|\kappa_0(\tilde{x},\cdot)\big)} \tag{5}$$

and using (6.12.5) together with the results from (6.12.iii) yields

$$D\big(\kappa_0(x,\cdot)\|\kappa_0(\tilde{x},\cdot)\big) \le \frac{2}{\sigma^2(1-\rho^2)} \sum_{k=1}^{\infty} d_k^2. \tag{6}$$

Combining (4), (5), and (6) yields

$$\|\kappa_0(x,\cdot) - \kappa_0(\tilde{x},\cdot)\|_{\mathrm{tv}} \le \frac{4\mathrm{c}}{\sqrt{\sigma^2(1-\rho^2)}} \sqrt{\sum_{k=1}^{\infty} \left(\sum_{i=k+m}^{\infty} |a_i| \right)^2}. \tag{7}$$

The right-hand side of (7) is a function of m and when it is possible to make it arbitrarily small by increasing m, then the channel κ is asymptotically input-memoryless for the signal set X''.

We provide three examples of filter coefficients. First assume that we have

$$a_i = q^i, \tag{8}$$

for all $i \in \mathbb{N}_0$, where q is a real constant satisfying $|q| < 1$. Evaluating the right-hand side of (7) based on the properties of the geometric series yields for the coefficients in (8)

$$\|\kappa_0(x,\cdot) - \kappa_0(\tilde{x},\cdot)\|_{\mathrm{tv}} \leq \frac{4\mathrm{c}|q|}{(1-|q|)\sqrt{\sigma^2(1-\rho^2)(1-q^2)}}\,|q|^m. \tag{9}$$

As a second example assume that the filter coefficients are given by

$$a_i = iq^i, \tag{10}$$

for all $i \in \mathbb{N}_0$, where $|q| < 1$. Then we obtain by similar calculations

$$\|\kappa_0(x,\cdot) - \kappa_0(\tilde{x},\cdot)\|_{\mathrm{tv}} \leq \frac{4\mathrm{c}|q|}{(1-|q|)\sqrt{\sigma^2(1-\rho^2)(1-q^2)}}\sqrt{A+Bm+m^2}\,|q|^m, \tag{11}$$

where

$$A = \frac{1+q^2(1-|q|)^2-|q|^3(2-|q|^3)}{(1-q^2)^2(1-|q|)^2} \qquad \text{and} \qquad B = \frac{2(1-|q|^3)}{(1-q^2)(1-|q|)}.$$

Third, when the filter coefficients are given by

$$a_i = \frac{1}{(i+1)(i+2)}, \tag{12}$$

for all $i \in \mathbb{N}_0$ we obtain by evaluating the right-hand side of (7)

$$\|\kappa_0(x,\cdot) - \kappa_0(\tilde{x},\cdot)\|_{\mathrm{tv}} \leq \frac{4\mathrm{c}}{\sqrt{\sigma^2(1-\rho^2)}}\sqrt{\Psi'(m+2)}, \tag{13}$$

where $\Psi'(\cdot)$ denotes the trigamma function as defined in (Olver et al., 2010, Sec. 5.15). To derive (13) we have used (Prudnikov et al., 1986, 4.1.4.5) and (Olver et al., 2010, 5.15.1).

We can make the right-hand side of (9) or (11) or (13) arbitrarily small by increasing m. Therefore, the channel κ with additive Gaussian noise as specified in (1) and filter coefficients given by (8) or (10) or (12) is asymptotically input-memoryless for the input signal set X''. The choice in (8) and (10) is representative for all linear filters with coefficients whose magnitude decays exponentially for sufficiently large indices. The choice in (12) is representative for linear filters with coefficients whose magnitude decays quadraticly for sufficiently large indices.

Due to the data processing inequality given in (6.4.ii) the relative entropy is at most decreased by a (deterministic or random) transformation. Therefore, the above results can be extended to situations with non-Gaussian noise given the noise can be represented as a transformation of the considered Gaussian noise. When the Gaussian noise has a covariance function different from (1), then the upper bound in (6) has to be replaced. In the situation of the present example

the relative entropy is calculated with the expression in (6.12.2). The fact that (6.12.2) is the so-called Mahalanobis distance can be used to derive an adequate upper bound.

In this paragraph, we analyzed an example, where the output of a linear time-invariant filter is connected to the input of an additive noise channel. Assume that we connect these components in reversed order. Then we have a different channel. However, due to the linearity of the filter this channel is equivalent to the cascade of the components in the original order, but with an additive noise process, which is the result of passing the original noise through the linear filter. Thus, we can analyze the channel as above but have to replace the additive noise by its filtered version. As an example assume that the additive noise is i.i.d. Gaussian with zero mean and variance $\dot{\sigma}^2$. Further suppose the linear filter, connected to the output of the additive noise channel, has coefficients $a_i = q^i, i \in \mathbb{N}_0$, for some constant q with $|q| < 1$. According to (12.4.ii) the filtered noise is a stationary Gaussian AR(1) process with zero mean. Therefore, the composed channel is equivalent to the example analyzed above with stationary Gaussian noise specified in (1) and filter coefficients specified in (8). In (1) we only have to replace σ^2 by $\dot{\sigma}^2/(1-q^2)$ and ρ by q. It follows that this cascade channel is information regular and asymptotically input-memoryless for the input signal set X''. An interesting aspect of this example is that the single components are output-memoryless but the composed channel has infinite output memory. However, having the filter as first and the additive noise as second component results in an output-memoryless channel. This illustrates the impact of the order of the components on the memory properties of the composed channel.

(17.6) Random linear filter and additive noise. Next we modify the example from Paragraph 17.5 by substituting the linear time-invariant deterministic filter for a linear time-variant random filter. Consider the cascade of integration channels illustrated in Figure 5 on page 101 and let us adopt all the notation from Paragraph 15.7. Suppose the first component $\dot{\kappa}$ with channel function $\dot{f}$ and noise measure space $(\dot{Z}, \dot{\mathcal{Z}}, \dot{\lambda})$ is a linear time-variant random filter as defined in Paragraph 17.4. Further suppose the second component $\ddot{\kappa}$ with channel function $\ddot{f}$ and noise measure space $(\ddot{Z}, \ddot{\mathcal{Z}}, \ddot{\lambda})$ is an additive noise channel as introduced in Paragraph 16.3. As in Paragraphs 16.3 and 17.4 we assume that all spaces are products generated from the set of real numbers and the time axis is discrete. The composed channel κ is an integration channel with channel function f given by (15.7.1) and noise measure space $(Z, \mathcal{Z}, \lambda)$ given by (15.7.2). Based on the component functions $\dot{f}_0$ and $\ddot{f}_0$ of $\dot{f}$ and $\ddot{f}$, respectively, we define the function f_0 on $X_-^0 \times \dot{Z}_0 \times \ddot{Z}_0$ with values in Y_0 by

$$f_0(x_-^0, \dot{z}_0, \ddot{z}_0) = \ddot{f}_0\big(\dot{f}_0(x_-^0, \dot{z}_0), \ddot{z}_0\big)$$

for all $x_-^0 \in X_-^0$, $\dot{z}_0 \in \dot{Z}_0$, and $\ddot{z}_0 \in \ddot{Z}_0$. The channel function f is then characterized by the component functions $f_k(\cdot) = f_0(\langle\cdot\rangle_{-k})$, $k \in \mathbb{Z}$.

We can use Theorems 14.2 and 14.4 to obtain properties of the composed channel κ based on the properties identified in Paragraphs 16.3 and 17.4 for the filter and the additive noise component. Alternatively, we can apply Theorem 15.4 to derive properties of the integration channel κ. Clearly, we have causality of κ. Because the channel function f is invariant the stationarity of κ follows from the stationarity of the composed noise source $\lambda = \dot{\lambda} \otimes \ddot{\lambda}$, which, in turn, follows from the stationarity of the noise sources $\dot{\lambda}$ and $\ddot{\lambda}$. Since the channel function f satisfies the measurability condition in (15.4.v) the mixing property of the composed noise source $\lambda = \dot{\lambda} \otimes \ddot{\lambda}$ implies the corresponding mixing property of the channel κ. According to Lemma 12.7 mixing properties of the product source $\dot{\lambda} \otimes \ddot{\lambda}$ follow from corresponding mixing properties of the

individual sources $\dot{\lambda}$ and $\ddot{\lambda}$. If the number of filter coefficients is finite, then the channel has finite input memory because the additive noise channel is memoryless. An infinite number of filter coefficients results in infinite input memory. Whether or not the channel is asymptotically input-memoryless depends on the specific situation. A positive example is considered next.

Suppose X'' denotes the set of all input sequences bounded by the finite constant $\mathrm{c} > 0$ as defined in (17.5.2). Let $\dot{\zeta} = \{\dot{\zeta}_k, k \in \mathbb{Z}\}$ and $\ddot{\zeta} = \{\ddot{\zeta}_k, k \in \mathbb{Z}\}$ denote the sequences of coordinate projections on the noise spaces, where $\dot{\zeta}_k$ is the projection from $\dot{Z}$ to $\dot{Z}_k$ and $\ddot{\zeta}_k$ is the projection from $\ddot{Z}$ to $\ddot{Z}_k$. Further, let $\dot{\zeta}_{k,i}$ denote the projection from $\dot{Z}$ to $\dot{Z}_{k,i}$. Assume that the noise source $\dot{\lambda}$ describing the random variation of the filter coefficients is such that $\dot{\zeta}$ is an i.i.d.-sequence and $\{\dot{\zeta}_{0,i}, i \in \mathbb{N}_0\}$ is a sequence of independent Gaussian random variables with

$$\sum_{i=0}^{\infty} \mathrm{E}\big(\dot{\zeta}_{0,i}^2\big) < \infty, \tag{1}$$

$$\mathrm{E}(\dot{\zeta}_{0,i}) = 0, \qquad \mathrm{var}(\dot{\zeta}_{0,i}) = \dot{\sigma}_i^2, \tag{2}$$

where $\dot{\sigma}_i^2 > 0$ for all $i \in \mathbb{N}_0$. Further assume that the noise source $\ddot{\lambda}$ describing the additive noise is such that $\ddot{\zeta}$ is an i.i.d.-sequence of Gaussian random variables with

$$\mathrm{E}(\ddot{\zeta}_0) = 0, \qquad \mathrm{var}(\ddot{\zeta}_0) = \ddot{\sigma}^2, \tag{3}$$

where $\ddot{\sigma}^2 > 0$ is a finite constant. Since the noise sources $\dot{\lambda}$ and $\ddot{\lambda}$ are memoryless the composed noise source λ is memoryless. Therefore, the channel κ is output memoryless.

Let $\eta = \{\eta_k, k \in \mathbb{Z}\}$ denote the sequence of coordinate projections on the channel output space, where η_k denotes the projection from Y to Y_k. If $x \in X''$ is fixed, then η is a sequence of independent Gaussian random variables on the probability space $(Y, \mathcal{Y}, \kappa(x, \cdot))$. The random variable η_k is then given by

$$\eta_k = \dot{f}_k\big(x_-^k, \dot{\zeta}_k\big) \quad + \ddot{\zeta}_k \tag{4}$$

$$= \sum_{i=0}^{\infty} x_{k-i}\dot{\zeta}_{k,i} + \ddot{\zeta}_k. \tag{5}$$

Due to the assumptions on $\dot{\lambda}$ and the boundedness of the inputs from the set X'' we can use the results from the last part of Paragraph A.13, which imply that the first summand in (4) is equal to the first summand in (5) $\dot{\lambda}$-almost surely and in mean square. Further we can apply (A.13.10) to calculate the expectation

$$\mathrm{E}(\eta_k) = \sum_{i=0}^{\infty} x_{k-i}\mathrm{E}(\dot{\zeta}_{k,i}) + \mathrm{E}(\ddot{\zeta}_k) = 0 \tag{6}$$

of η_k and (A.13.11) to calculate the variance

$$\mathrm{var}(\eta_k) = \mathrm{E}(\eta_k^2) = \mathrm{E}\big(\dot{f}_k(x_-^k, \dot{\zeta}_k)^2\big) + \mathrm{E}(\ddot{\zeta}_k^2) \tag{7}$$

$$= \sum_{i=0}^{\infty} x_{k-i}^2\mathrm{E}(\dot{\zeta}_{k,i}^2) + \mathrm{E}(\ddot{\zeta}_k^2)$$

$$= \sum_{i=0}^{\infty} x_{k-i}^2\dot{\sigma}_i^2 + \ddot{\sigma}^2 < \infty. \tag{8}$$

The second equality in (7) holds because $\dot f_k(x_-^k, \dot\zeta_k)$ and $\ddot\zeta_k$ are independent as random variables on the composed noise space $(Z, \mathcal{Z}, \lambda)$. In (6) and (8) we have used (2) and (3) and the stationarity of the noise sources $\dot\lambda$ and $\ddot\lambda$. The distribution of the sequence η_0^+ is equal to $\kappa_0(x,\cdot)$, the marginal measure of $\kappa(x,\cdot)$ on $\mathcal{Y}_0^+$.

As in Paragraph 17.5 we verify the asymptotic input-memorylessness of the composed channel κ using Pinsker's inequality. Given $x, \tilde x \in X''$ we define

$$a_k = \sum_{i=0}^{\infty} x_{k-i}^2 \dot\sigma_i^2 + \ddot\sigma^2, \qquad b_k = \sum_{i=0}^{\infty} \tilde x_{k-i}^2 \dot\sigma_i^2 + \ddot\sigma^2. \tag{9}$$

Based on (6.12.7) and the results in (6.12.iii) we obtain the relative entropy

$$D\big(\kappa_0(x,\cdot)\|\kappa_0(\tilde x,\cdot)\big) = \frac{1}{2}\sum_{k=1}^{\infty}\left(\log\left(\frac{b_k}{a_k}\right) + \frac{a_k}{b_k} - 1\right). \tag{10}$$

Let us fix some $m \in \mathbb{N}_0$. For all $k \in \mathbb{N}$ and $x, \tilde x \in X''$ coinciding on $(-m, \infty)$ we obtain from (9) by basic calculations

$$\left(1 + \frac{1}{\ddot\sigma^2}\sum_{i=k+m}^{\infty} \tilde x_{k-i}^2 \dot\sigma_i^2\right)^{-1} \le \frac{a_k}{b_k} \le \left(1 + \frac{1}{\ddot\sigma^2}\sum_{i=k+m}^{\infty} x_{k-i}^2 \dot\sigma_i^2\right). \tag{11}$$

Combining (10) and (11) and applying Pinsker's inequality as in (17.5.5) yields

$$\begin{aligned}\|\kappa_0(x,\cdot) - \kappa_0(\tilde x,\cdot)\|_{\mathrm{tv}} &\le \sqrt{\sum_{k=1}^{\infty}\left(\log\left(1 + \frac{\mathrm{c}^2}{\ddot\sigma^2}\sum_{i=k+m}^{\infty}\dot\sigma_i^2\right) + \frac{\mathrm{c}^2}{\ddot\sigma^2}\sum_{i=k+m}^{\infty}\dot\sigma_i^2\right)} \\ &\le \sqrt{\frac{2\mathrm{c}^2}{\ddot\sigma^2}\sum_{k=1}^{\infty}\left(\sum_{i=k+m}^{\infty}\dot\sigma_i^2\right)},\end{aligned} \tag{12}$$

where we have also used $\log(x+1) \le x$ and that the sequences from X'' are bounded by the constant c. According to the version of the definition given next to (17.5.3) the channel κ is asymptotically input-memoryless for the signal set X'' if it is possible to make (12) arbitrarily small by increasing m.

We consider three examples of filter coefficients. Suppose we have

$$\dot\sigma_i^2 = q^i \tag{13}$$

for all $i \in \mathbb{N}_0$, where $0 < q < 1$ is some constant. Then we obtain with (12)

$$\|\kappa_0(x,\cdot) - \kappa_0(\tilde x,\cdot)\|_{\mathrm{tv}} \le \sqrt{\frac{2\mathrm{c}^2 q}{\ddot\sigma^2(1-q)^2}}\; q^{\frac{m}{2}}, \tag{14}$$

based on the limit of the geometric series. Next, if we have

$$\dot\sigma_i^2 = iq^i \tag{15}$$

for all $i \in \mathbb{N}_0$, where $0 < q < 1$, then

$$\|\kappa_0(x,\cdot) - \kappa_0(\tilde{x},\cdot)\|_{\mathrm{tv}} \le \sqrt{\frac{2\mathrm{c}^2 q}{\ddot{\sigma}^2(1-q)^2}\left(\frac{1+q}{1-q}+m\right)}\, q^{\frac{m}{2}} \tag{16}$$

follows similarly. Finally, evaluating (12) given

$$\dot{\sigma}_i^2 = \frac{1}{(i+1)(i+2)(i+3)} \tag{17}$$

for all $i \in \mathbb{N}_0$ yields

$$\|\kappa_0(x,\cdot) - \kappa_0(\tilde{x},\cdot)\|_{\mathrm{tv}} \le \frac{\mathrm{c}}{\sqrt{\ddot{\sigma}^2}}\,\frac{1}{\sqrt{m+2}}, \tag{18}$$

where we have used (Prudnikov et al., 1986, 4.1.5.3 and 5.1.6.5).

The specified channel κ represents a random filter combined with additive memoryless zero mean Gaussian noise. The filter has an infinite number of independent zero mean Gaussian coefficients, which change independently over time. When the variance of the filter coefficients is given by (13) or (15) or (17), then we can make the right-hand side of (14) or (16) or (18) arbitrarily small by increasing m. Thus, with these parameters the channel κ is asymptotically input-memoryless for the input signal set X''. The channel can model a wireless communication link with an infinite channel impulse response. The choice in (13) and (15) is representative for an exponentially decreasing power delay profile. The choice in (17) is representative for a power delay profile decreasing with the power of 3.

When we introduce dependence only between the random filter coefficients $\{\dot{\zeta}_{k,i}, i \in \mathbb{N}_0\}$, then the relative entropy is still given by (10). However, the variances in (9) are different because the calculation is based on (A.13.9) rather than (A.13.11). As a consequence the inequalities in (11) have to be adapted accordingly. When the Gaussian filter coefficients change dependently over time or the additive Gaussian noise has memory, then (10) is not valid anymore and the calculation of the relative entropy is based on (6.12.6), which can be difficult but worth to analyze.

Let us continue with a list of some more applications. For discrete-time finite-alphabet channels Takano (1974, Th. 5) derived a central limit theorem for information densities under a mixing condition that lies between ψ-mixing and β-mixing. Using the same conditions Takano (1977) proved a law of iterated logarithm for information densities and demonstrated how this result can be applied to analyze convergence rates of decoding error probabilities. Zhang and Weissman (2005) derived an asymptotically universal denoiser for a discrete-time causal input-memoryless channel, which has finite alphabets and satisfies the α-mixing condition. We conclude this section with a more detailed example from statistical signal processing. Assume that we want to apply a theorem to the output of a channel with time structure that relies on some mixing condition but a priori only mixing properties of the channel and the channel input are known. In this situation, we can apply Theorem 13.11, which formulates conditions for the channel and the channel input that guarantee a certain mixing property of the channel output. Next, we give an example in this regard, where mixing in the ergodic-theoretic sense is not sufficient. There are many more applications, where mixing conditions play an important role. To name just one more, Samson (2000) proved concentration inequalities for random sequences satisfying a mixing condition that lies between ψ-mixing and β-mixing. Marton (2003) obtained further results in this direction.

(17.7) Fourier transform of stationary sequences. Suppose $\eta = \{\eta_k, k \in \mathbb{Z}\}$ is a real-valued stationary second-order random sequence on $(\Omega, \mathcal{F}, \mathrm{P})$ with $\mathrm{E}(\eta_0) = 0$. We define the Fourier transform for n components of the sequence η by

$$\vartheta_n(u) = \sum_{k=1}^{n} \eta_k \mathrm{e}^{\jmath k u}, \quad u \in (-\pi, \pi].$$

The Fourier transform and the related quantity $|\vartheta_n(u)|^2/(2\pi n)$, $u \in \mathbb{R}$, called periodogram, are fundamental tools to solve problems of statistical inference for stationary time series (Priestley, 1981a, Ch. 6), (Brockwell and Davis, 2006, Ch. 10). The estimation of the spectral density, for example, relies on the distribution of the periodogram, which usually does not have a closed form so that the analysis is based on the asymptotic distribution as $n \to \infty$. Next, we give a result in this direction derived recently by Peligrad and Wu (2010). Note that the original formulation is somewhat more general than the version given below. We assume that the interval $(-\pi, \pi]$ is equipped with the corresponding Borel-σ-algebra and the Lebesgue measure λ.

(i) *Central limit theorem for Fourier transforms of stationary processes.* Suppose the so-called past tail σ-algebra $\mathcal{F}_{\mathrm{past}} := \bigcap_{n=1}^{\infty} \sigma(\eta_-^{-n})$ of η is trivial, i. e., contains only sets of P-measure 0 or 1. Then the random sequence η has a spectral density, say φ, and for almost all $u \in (-\pi, \pi]$ we have

$$\lim_{n\to\infty} \frac{\mathrm{E}|\vartheta_n(u)|^2}{2\pi n} = \varphi(u),$$

and

$$\frac{1}{\sqrt{n}}\big(\mathrm{Re}(\vartheta_n(u)), \mathrm{Im}(\vartheta_n(u))\big) \xrightarrow[(n\to\infty)]{\text{in distribution}} \sqrt{\pi\varphi(u)}\big(\phi_1, \phi_2\big) \qquad (\text{under P}),$$

$$\frac{1}{2\pi n}|\vartheta_n(u)|^2 \xrightarrow[(n\to\infty)]{\text{in distribution}} \varphi(u)\phi_3 \qquad (\text{under P}),$$

where ϕ_1 and ϕ_2 are independent and identically distributed Gaussian random variables with mean 0 and variance 1 and ϕ_3 is an exponentially distributed random variable with mean 1. Furthermore, for almost all $u_1, u_2 \in (-\pi, \pi]$ with $u_1 \neq u_2$ the random variables

$$\frac{1}{\sqrt{n}}\vartheta_n(u_1) \qquad \text{and} \qquad \frac{1}{\sqrt{n}}\vartheta_n(u_2)$$

are asymptotically independent, i. e., asymptotically the process $\big\{\vartheta_n(u)/\sqrt{n},\ u \in (-\pi, \pi]\big\}$ is white Gaussian noise as $n \to \infty$.

This central limit theorem justifies for a class of stationary sequences what is commonly presumed in frequency-domain analysis of stationary time series: the Fourier transform of stationary processes are asymptotically independent Gaussian. Results in this direction under different dependence conditions have a long history. The theorem of Peligrad and Wu (2010) improves the result in Wu (2005). Rosenblatt (1985, Ch. 5, Th. 3) gives a version based on the α-mixing condition (see (Peligrad and Wu, 2010) for more references). In fact, a trivial past tail σ-algebra implies mixing in the ergodic-theoretic sense and is implied by α-mixing (Bradley, 2007, 2.17). Thus, the given central limit theorem applies in particular to real-valued stationary second-order random sequences satisfying the α-mixing condition.

(ii) *Fourier transform of channel output.* Let us now consider a discrete-time channel κ as introduced in Definition 2.3 with real-valued input and output signals. Assume that the channel

input probability measure is μ such that the input-output probability space is $(X \times Y, \mathcal{X} \otimes \mathcal{Y}, \mu\kappa)$. By $\xi = \{\xi_k, k \in \mathbb{Z}\}$ and $\eta = \{\eta_k, k \in \mathbb{Z}\}$ we denote the sequences of coordinate projections on the input-output space, where ξ_k is the projection from $X \times Y$ to X_k and η_k is the projection from $X \times Y$ to Y_k. We are interested now in conditions on the input sequence ξ and the channel κ, which allow to apply the central limit theorem of part (i) to the channel output sequence η. Due to Lemma 2.9 and Theorem 13.11 we have: If ξ is stationary and α-mixing and κ is stationary, α-mixing, causal, and asymptotically input-memoryless, then η is stationary and α-mixing. Thus, under these conditions we can apply part (i) if η has zero mean and finite second moments. The additive and multiplicative noise channel defined in Paragraphs 16.3 and 16.5 allow to apply this results, if the noise sources are α-mixing. The same holds for the channel with state from Paragraph 16.7, if the state distribution is α-mixing. If bounded input sequences are considered, then we can apply the result also to the composed channels studied in Paragraphs 17.5 and 17.6.

Summary and Open Problems

Summary. In the first part, of this thesis we generalized a coding theorem and a converse of Kadota and Wyner (1972) to abstract channels with time structure. As a main contribution we proved the coding theorem for a significantly weaker condition on the channel output memory, called total ergodicity for block-i.i.d. inputs. We achieved this result mainly by introducing an alternative characterization of information rate capacity. We showed that the ψ-mixing condition (asymptotic output-memorylessness), used by Kadota and Wyner, is quite restrictive, in particular for the important class of Gaussian channels. In fact, we proved that for Gaussian channels the ψ-mixing condition is equivalent to finite output memory. Moreover, we derived a weak converse for all stationary channels with time structure. Intersymbol interference as well as input constraints were taken into account in a flexible way. Due to the direct use of outer measures and a derivation of an adequate version of Feinstein's lemma we were able to avoid the standard extension of the channel input σ-algebra and obtained a more transparent derivation. We aimed at a presentation from an operational perspective and considered an abstract framework, which enabled us to treat discrete- and continuous-time channels in a unified way.

In the second part, we systematically analyzed infinite output memory conditions for abstract channels with time structure. We exploited the connections to the rich field of strongly mixing random processes to derive a hierarchy for the nonequivalent infinite channel output memory conditions in terms of a sequence of implications. The ergodic-theoretic memory condition used in the proof of the coding theorem and the ψ-mixing condition employed by Kadota and Wyner (1972) were shown to be part of this taxonomy. In addition, we specified conditions for the channel, under which memory properties of a random process are invariant when the process is passed through the channel.

In the last part, we analyzed cascade and integration channels with regard to mixing conditions as well as properties required in the context of the coding theorem. The results are useful to study many physically relevant channel models and allow a component-based analysis of the overall channel. We considered a number of examples including composed models and deterministic as well as random filter channels. Finally, an application of strong mixing conditions from statistical signal processing involving the Fourier transform of stationary random sequences was discussed and a list of further applications was given.

Open problems. Naturally, a variety of open problems remains. We list some of them, which would be interesting for future work. The material in this thesis is prepared in the hope to serve as a suitable starting point for further generalizations and extensions of the given results.
Chapter I.

- To prove the monotonicity of the sequence $\{n^{-1}I(\xi_0^n;\eta_0^n), n \in \mathbb{N}\}$ in Corollary 4.14 Kadota and Wyner (1972) argued that it is convex. However, in Paragraph 16.6 we constructed a strictly concave example. Under what general conditions the sequence is indeed convex?
- The information rate capacity is introduced in Definition 5.1 as limit superior. In Lemma 5.3 we derived two representations as supremum, given the channel is stationary and the input constraints satisfy the regularity condition in (3.1.4). What channel properties imply that the limit superior is in fact a limit?

- The information rate capacity is defined with respect to an almost sure input constraint. Kadota (1973) showed for his channel model and a certain input constraint that the considered version of information rate capacity does not change its value, when the almost sure constraint is replaced by a corresponding average constraint. We conjecture that a similar result holds for the information rate capacity of Definition 5.1, when the conditions of Theorem 9.1 are satisfied.
- We introduced in Definition 7.2 the ψ-variation and proved in Paragraph E.3 that it satisfies the data processing inequality (7.3.ii) if integration channels are considered. We conjecture that the ψ-variation satisfies this inequality for an arbitrary channel.

Chapter II

- The coding theorem as formulated in Theorem 9.1 was derived for asymptotically input-memoryless channels, a condition based on the total variation distance. This condition is sometimes too strong or it is difficult to verify it for specific channels. Therefore, it is worth investigating more relaxed conditions on the channel input memory. Gray and Ornstein (1979) introduced a weaker input memory condition, called $\bar{d}$-continuity, for the special case of a discrete-time finite-alphabet channel and proved a coding theorem. A generalization within the abstract framework of this thesis would be interesting for future work.
- We proved with Theorem 9.3 a weak converse for stationary channels with time structure having arbitrary alphabets. Deriving a corresponding strong converse is a possible further research direction. Augustin (1966) considered abstract memoryless channels in this regard. To investigate a generalization within the framwork of this thesis to channels with memory would be interesting.
- We have shown in Theorem 10.5, that Kadota and Wyner's information rate capacity given in Definition 10.1 is equal to the version introduced in Definition 5.1 if it has an operational meaning in the sense of Theorem 9.1. To find a purely analytic proof without reference to the coding theorem would be a desirable result.

Chapter III and Chapter IV

- That the reversed implications in (e) and (f) of Theorem 13.9 do not hold in general has been shown only in Paragraph 16.2 for a purely random channel. It remains to find "true" channels illustrating this aspect. An example showing that the reversed implication in (g) of Theorem 13.9 does not hold in general is still missing.
- It might be possible to derive some of the implications in Theorems 13.11, 14.4 and 15.4 under weaker conditions.

Chapter V

- It is straightforward to rigorously extend the discrete-time examples in Paragraphs 17.1 and 17.4 to continuous-time multipath channels with a countable number of taps. A rigorous extension within the framework of integration channels to general continuous-time linear filters would be desirable. An extension of the examples in Paragraphs 17.5 and 17.6 in this regard would include the analysis of the relative entropy between continuous-time random processes.

Appendix

A Basics of Probability and Measure Theory

In this section we collect frequently used definitions and results from probability and measure theory in a form suitable for the purposes of the thesis. Recommendable books on the fundamentals of probability and measure theory are (Shiryaev, 1995; Billingsley, 1995; Bauer, 1995, 2001; Ash, 2000; Elstrodt, 2005). A compact reference on the foundations of modern probability is (Kallenberg, 2002).

(A.1) Independence. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and S be any index set. The family $\{A_s, s \in S\}$ of sets $A_s \in \mathcal{F}$ is called independent, if for any finite set $J \subset S$ we have

$$\mathrm{P}\Big(\bigcap_{s\in J} A_s \Big) = \prod_{s\in J} \mathrm{P}(A_s).$$

The family $\{\mathcal{A}_s, s \in S\}$ of set systems $\mathcal{A}_s \subset \mathcal{F}$ is called independent, if any family $\{A_s, s \in S\}$ of sets $A_s \in \mathcal{A}_s$ is independent. The family $\{\xi_s, s \in S\}$ of random variables on $(\Omega, \mathcal{F}, \mathrm{P})$ is called independent, if $\{\sigma(\xi_s), s \in S\}$ is an independent family of σ-algebras. As a reference see (Bauer, 1995, §6, §7).

(A.2) Markov chain and Markov process. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be sub-σ-algebras of $\mathcal{F}$. We say $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ form a Markov chain or Markov triplet in this order (see (Bradley, 2007, Def. 7.1)) if for all $C \in \mathcal{C}$

$$\mathrm{P}(C \,|\, \mathcal{B} \vee \mathcal{A}) = \mathrm{P}(C \,|\, \mathcal{B}) \qquad \text{P-a.s.}$$

In this case we also write $(\mathcal{A} - \mathcal{B} - \mathcal{C})$. We have $(\mathcal{A} - \mathcal{B} - \mathcal{C})$ if and only if we have $(\mathcal{C} - \mathcal{B} - \mathcal{A})$. Furthermore, $(\mathcal{A} - \mathcal{B} - \mathcal{C})$ holds if and only if for all $A \in \mathcal{A}$ and $C \in \mathcal{C}$

$$\mathrm{P}(A \cap C \,|\, \mathcal{B}) = \mathrm{P}(A \,|\, \mathcal{B}) \mathrm{P}(C \,|\, \mathcal{B}) \qquad \text{P-a.s.}$$

(see (Loève, 1978, Sec. 28.3)). Then we also say $\mathcal{A}$ and $\mathcal{C}$ are conditionally independent given $\mathcal{B}$.

(i) Suppose S is some index set and $\{\mathcal{A}_s, s \in S\}$, $\{\mathcal{B}_s, s \in S\}$, and $\{\mathcal{C}_s, s \in S\}$ are families of sub-σ-algebras of $\mathcal{F}$ such that $\{\mathcal{A}_s \vee \mathcal{B}_s \vee \mathcal{C}_s, s \in S\}$ is an independent family of σ-algebras and $(\mathcal{A}_s - \mathcal{B}_s - \mathcal{C}_s)$ holds for all $s \in S$. Then we have the Markov chain

$$\Big(\bigvee_{s\in S} \mathcal{A}_s - \bigvee_{s\in S} \mathcal{B}_s - \bigvee_{s\in S} \mathcal{C}_s \Big).$$

This result is taken from (Bradley, 2007, A701 (IV)).

(ii) If $(\mathcal{A} - \mathcal{B} - \mathcal{C})$ is a Markov chain of sub-σ-algebras of $\mathcal{F}$ and $\mathcal{A} \doteq \mathcal{A}'$, $\mathcal{B} \doteq \mathcal{B}'$, and $\mathcal{C} \doteq \mathcal{C}'$ holds for sub-σ-algebras $\mathcal{A}'$, $\mathcal{B}'$, and $\mathcal{C}'$ of $\mathcal{F}$, then we have $(\mathcal{A}' - \mathcal{B}' - \mathcal{C}')$. Here $\mathcal{A} \doteq \mathcal{A}'$ means for all $A \in \mathcal{A}$ there exists a $A' \in \mathcal{A}'$ with $\mathrm{P}(A \triangle A') = 0$ and vice versa. This result is taken from (Bradley, 2007, A701 (V)).

(iii) We say the random variables ξ, η, and ζ on $(\Omega, \mathcal{F}, \mathrm{P})$ form a Markov chain in this order if we have $\big(\sigma(\xi) - \sigma(\eta) - \sigma(\zeta)\big)$. In this case we also write $\big(\xi - \eta - \zeta\big)$ and say ξ and ζ are conditionally independent given η. If $\xi = \{\xi_k, k \in \mathbb{Z}\}$ is a sequence of random variables on $(\Omega, \mathcal{F}, \mathrm{P})$, then ξ is called a Markov chain if $\big(\xi_-^{k-2} - \xi_{k-1} - \xi_k\big)$ holds for all $k \in \mathbb{Z}$. More generally, if $\xi = \{\xi_t, t \in T\}$ is a stochastic process on $(\Omega, \mathcal{F}, \mathrm{P})$, where ξ_t has values in $(X_t, \mathcal{X}_t)$, then ξ is called Markov process, if for all $s, t \in T$ with $s < t$ and $A \in \mathcal{X}_t$

$$\mathrm{P}(\xi_t \in A | \sigma(\xi_-^s)) = \mathrm{P}(\xi_t \in A | \sigma(\xi_s)) \qquad \mathrm{P}\text{-a.s.}$$

See for example (Ihara, 1993, Def. 2.1.2) or (Revuz and Yor, 1999, Ch. III) and (Gikhman and Skorokhod, 1974, pp. 159–163), but note that in the latter two references more restrictive definitions based on Markov kernels are considered.

(A.3) Markov kernel. Assume that $(\Omega_1, \mathcal{F}_1)$, $(\Omega_2, \mathcal{F}_2)$, and $(\Omega_3, \mathcal{F}_3)$ are measurable spaces. The function $K : \Omega_1 \times \mathcal{F}_2 \to [0, 1]$ is called a Markov-kernel from $(\Omega_1, \mathcal{F}_1)$ to $(\Omega_2, \mathcal{F}_2)$

• if $K(\cdot, F_2)$ is $\mathcal{F}_1$-measurable for all $F_2 \in \mathcal{F}_2$ and

• if $K(\omega_1, \cdot)$ is a probability measure on $\mathcal{F}_2$ for all $\omega_1 \in \Omega_1$. Markov kernels are considered for example in (Bauer, 1995, §36) or (Gikhman and Skorokhod, 1974, Ch. II, §4), where the term stochastic kernel is used.

(i) If P_1 is a probability measure on $\mathcal{F}_1$, then the set function P on $\mathcal{F}_1 \otimes \mathcal{F}_2$ defined by

$$\mathrm{P}(F) = \int_{\Omega_1} K(\omega_1, F_{\omega_1}) \, \mathrm{dP}_1(\omega_1), \qquad F \in \mathcal{F}_1 \otimes \mathcal{F}_2,$$

is a probability measure, where F_{ω_1} denotes the ω_1-section of the set F. This result is derived in (Ash, 1972, Sec. 2.6.2).

(ii) If L is another Markov-kernel, now from $(\Omega_2, \mathcal{F}_2)$ to $(\Omega_3, \mathcal{F}_3)$, then M with

$$M(\omega_1, F_3) = \int_{\Omega_2} L(\cdot, F_3) \, \mathrm{d}K(\omega_1, \cdot), \quad \omega_1 \in \Omega_1, \ F_3 \in \mathcal{F}_3$$

is a Markov-kernel from $(\Omega_1, \mathcal{F}_1)$ to $(\Omega_3, \mathcal{F}_3)$, called cascade Markov-kernel. This result follows from (Gikhman and Skorokhod, 1974, Th. 1, p. 76).

(iii) If $(\Omega_1, \mathcal{F}_1)$ is a product space, i. e., $(\Omega_1, \mathcal{F}_1) = (\Omega_1' \times \Omega_1'', \mathcal{F}_1' \otimes \mathcal{F}_1'')$, then for any $\omega_1' \in \Omega_1'$ $K((\omega_1', \cdot), \cdot)$ is a Markov-kernel from $(\Omega_1'', \mathcal{F}_1'')$ to $(\Omega_2, \mathcal{F}_2)$. If P_1' is a probability measure on $\mathcal{F}_1'$, then $\tilde{K}$ with

$$\tilde{K}(\omega_1'', F_2) = \int_{\Omega_1'} K\big((\omega_1', \omega_1''), F_2\big) \, \mathrm{dP}_1'(\omega_1'), \quad \omega_1'' \in \Omega_1'', \ F_2 \in \mathcal{F}_2$$

is a Markov-kernel from $(\Omega_1'', \mathcal{F}_1'')$ to $(\Omega_2, \mathcal{F}_2)$ and is called induced Markov-kernel. These results are obtained with (Bauer, 2001, Lem. 23.5) and part (A.8.i) of Fubini's theorem.

(iv) Assume that $(\Omega_2, \mathcal{F}_2)$ is a product space of the form $(\Omega_2, \mathcal{F}_2) = (\Omega_2' \times \Omega_2'', \mathcal{F}_2' \otimes \mathcal{F}_2'')$ and for all $\omega_1 \in \Omega_1$, $F_2' \in \mathcal{F}_2'$, and $F_2'' \in \mathcal{F}_2''$ we define

$$\dot{K}(\omega_1, F_2' \times F_2'') = K(\omega_1, F_2' \times \Omega_2'') K(\omega_1, \Omega_2' \times F_2'').$$

When we extend $\dot{K}(\omega_1, \cdot)$ for any $\omega_1 \in \Omega_1$ in the usual way to a probability measure on $\mathcal{F}_2$, then $\dot{K}$ is a Markov kernel from $(\Omega_1, \mathcal{F}_1)$ to $(\Omega_2, \mathcal{F}_2)$, and is called product Markov kernel. One can show that the set system

$$\mathcal{D} = \left\{ F_2 \in \mathcal{F}_2 : \dot{K}(\cdot, F_2) \text{ is } \mathcal{F}_1\text{-measurable} \right\}$$

is the smallest Dynkin system containing $\mathcal{F}_2' \times \mathcal{F}_2''$. Since $\mathcal{F}_2' \times \mathcal{F}_2''$ is closed w. r. t. finite intersections it follows from the monotone class theorem that $\mathcal{D}$ is equal to $\mathcal{F}_2$. Thus $\dot{K}(\cdot, F_2)$ is $\mathcal{F}_1$-measurable for all $F_2 \in \mathcal{F}_2$ so that $\dot{K}$ is indeed a Markov kernel.

(v) Let $\alpha \in (0, 1)$ be a constant and assume that K_1 and K_2 are both Markov kernels from $(\Omega_1, \mathcal{F}_1)$ to $(\Omega_2, \mathcal{F}_2)$. We define for all $\omega_1 \in \Omega_1$ and $F_2 \in \mathcal{F}_2$

$$\bar{K}(\omega_1, F_2) = \alpha K_1(\omega_1, F_2) + (1 - \alpha) K_2(\omega_1, F_2).$$

Then $\bar{K}$ is a Markov kernel from $(\Omega_1, \mathcal{F}_1)$ to $(\Omega_2, \mathcal{F}_2)$ called averaged Markov kernel. $\bar{K}$ is indeed a Markov kernel because the convex combination of probability measures is a probability measure and the convex combination of measurable functions is measurable.

(A.4) Special measures. Assume that $(\Omega, \mathcal{F})$ is an arbitrary measurable space. If $\omega \in \Omega$ is some fixed element, then δ_ω denotes the Dirac measure on $\mathcal{F}$ for the element ω. It is defined by

$$\delta_\omega(F) = \begin{cases} 1 & \text{if } \omega \in F \\ 0 & \text{if } \omega \notin F \end{cases}$$

for any $F \in \mathcal{F}$. The measure λ is called counting measure on $\mathcal{F}$, if it is defined by

$$\lambda(F) = |F|,$$

for any $F \in \mathcal{F}$. If $(\Omega, \mathcal{F}) = (\mathbb{Z}, 2^{\mathbb{Z}})$, then λ can be written as

$$\lambda = \sum_{k \in \mathbb{Z}} \delta_k.$$

This particular counting measure can be used to write sums as integrals, e. g., if $x = \{x_t, t \in \mathbb{Z}\}$ denotes a two-sided real sequence and $A = \{1, 2, \ldots, n\}$, then we have

$$\int_{\mathbb{Z}} x_t \, \mathrm{d}\lambda(t) = \sum_{t \in \mathbb{Z}} x_t \qquad \text{and} \qquad \int_A x_t \, \mathrm{d}\lambda(t) = \sum_{t=1}^{n} x_t.$$

Consider now the real measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The measure λ on $\mathcal{B}(\mathbb{R})$ is called (one-dimensional) Lebesgue measure, if it assigns to any interval $(a, b]$ the value

$$\lambda((a, b]) = b - a,$$

where $a \leq b \in \mathbb{R}$. Please refer to Bauer (2001, p. 12, §6) for more details.

(A.5) Finite and completely atomic σ-algebras. Assume that $(\Omega, \mathcal{F}, \mathrm{P})$ is a probability space and $\mathcal{A}$ is a sub-σ-algebra of $\mathcal{F}$. A set $A \in \mathcal{A}$ is called an atom of $\mathcal{A}$ if $\mathrm{P}(A) > 0$ and if for all $B \in \mathcal{A}$ with $B \subset A$ we either have $\mathrm{P}(B) = 0$ or $\mathrm{P}(A \triangle B) = 0$. The σ-algebra $\mathcal{A}$ is called completely atomic if there are finite or countable infinite atoms $A_1, A_2, \ldots$ of $\mathcal{A}$ such that $\mathrm{P}(A_i \cap A_j) = 0$ for $i \neq j$ and $\mathrm{P}\big(\bigcup_{i=1}^{\infty} A_i\big) = 1$. For such a completely atomic σ-algebra there exists a partition $\{B_1, B_2, \ldots\}$ of Ω, where B_i is an atom of $\mathcal{A}$ with $\mathrm{P}(A_i \triangle B_i) = 0$.

As an example consider the probability measure P is given by

$$\mathrm{P} = \sum_{i=1}^{m} p_i \delta_{a_i},$$

where m is some positive integer, δ_{a_i} is the Dirac measure on $\mathcal{F}$ for the element $a_i \in \Omega$, and p_i is a positive constant such that $\sum_{i=1}^{m} p_i = 1$. Then the σ-algebra $\mathcal{F}$ (and $\mathcal{A}$) is completely atomic. For simplicity let us assume that $\mathcal{F}$ is large enough such that there exist sets $A_1, A_2, \ldots, A_m \in \mathcal{F}$ with $a_i \in A_i$ and $a_j \notin A_i$ for $i \neq j$. Then A_i is an atom of $\mathcal{F}$.

A set $A \in \mathcal{A}$ is called a minimal nonempty set of $\mathcal{A}$, if there does not exist a set $B \in \mathcal{A}$ such that $B \subset A$ other than $B = A$ or $B = \varnothing$. Assume that $\mathcal{A}$ is a finite σ-algebra, i. e., the number of sets in $\mathcal{A}$ is finite. Then there exists a partition $\{A_1, A_2, \ldots, A_m\}$ of Ω, consisting of minimal nonempty sets $A_i \in \mathcal{A}$. The set A_i is either a nullset or an atom of $\mathcal{A}$. Thus, every finite σ-algebra is completely atomic. Since every σ-algebra generated by a finite partition is finite, it is also completely atomic. If ξ is a discrete random variable on $(\Omega, \mathcal{F}, \mathrm{P})$, then the σ-algebra $\sigma(\xi)$ is completely atomic. For more details see (Bradley, 2007, Secs. 0.7, A040, A042, A051).

(A.6) Gaussian distribution. Let ξ be a real random variable defined on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. We call ξ a Gaussian random variable if its distribution is either the Dirac measure δ_a at a point $a \in \mathbb{R}$ or has density

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(x-a)^2}{2\sigma^2} \right), \qquad x \in \mathbb{R},$$

w. r. t. the Lebesgue measure for some $a \in \mathbb{R}$ and $\sigma > 0$. In the latter case the measure ξ is called nondegenerate. The real n-dimensional random vector $(\xi_1, \xi_2, \ldots, \xi_n)$ on $(\Omega, \mathcal{F}, \mathrm{P})$ is called a Gaussian random vector if

$$a_1\xi_1 + a_2\xi_2 + \ldots + a_n\xi_n$$

is a Gaussian random variable for all $a_1, a_2, \ldots, a_n \in \mathbb{R}$. Let $\xi = \{\xi_t, t \in T\}$ with $\xi_t = (\xi_{t,1}, \ldots, \xi_{t,m})$ be a real m-dimensional vector process on $(\Omega, \mathcal{F}, \mathrm{P})$. Then ξ is called a Gaussian process if $(\xi_{t_1,k_1}, \xi_{t_2,k_2}, \ldots, \xi_{t_n,k_n})$ is a Gaussian random vector for all $n \in \mathbb{N}$, $t_1, t_2, \ldots, t_n \in T$, and $k_1, k_2, \ldots, k_n \in \{1, 2, \ldots, m\}$. The definitions are based on (Bogachev, 1998, Def. 1.1.1) and (Sasvári, 2013, 2.1.1, Th. 1.10.4). Further in-depth references on Gaussian random variables and processes are (Ibragimov and Rozanov, 1978) and (Hida and Hitsuda, 2007).

(i) Let $\xi = \{\xi_t, t \in T\}$ with $\xi_t = (\xi_{t,1}, \ldots, \xi_{t,n})$ be an n-dimensional Gaussian vector process and assume that $T_1, T_2 \subset T$ are disjoint sets. If $\mathrm{cor}(\xi_{t_1,k_1}, \xi_{t_2,k_2}) = 0$ for all $t_1 \in T_1$, $t_2 \in T_2$, and $k_1, k_2 \in \{1, 2, \ldots, n\}$, then the random variables $\xi_{T_1} = \{\xi_t, t \in T_1\}$ and $\xi_{T_2} = \{\xi_t, t \in T_2\}$ are independent. See (Bradley, 2007, A901) for a generalized version of this result, which is based on the finite-dimensional result in (Dudley, 2002, Th. 9.5.14).

(A.7) Density, absolute continuity, singularity, Lebesgue decomposition. Details to the results in this paragraph are given in (Bauer, 2001, § 17). Let $(\Omega, \mathcal{F})$ be a measurable space and assume that μ and ν are measures on $\mathcal{F}$.

(i) *Measures with density.* If there is a nonnegative $\mathcal{F}$-measurable function f on Ω such that

$$\nu(F) = \int_F f \,\mathrm{d}\mu$$

holds for all $F \in \mathcal{F}$, then f is called μ-density of ν. If g is a real $\mathcal{F}$-measurable function on Ω and ν has μ-density f, then

$$\int_F g \,\mathrm{d}\nu = \int_F gf \,\mathrm{d}\mu$$

holds for all $F \in \mathcal{F}$ for which one of the integrals exists.

(ii) *Absolute continuity and singularity.* The measure ν is called absolutely continuous w. r. t. the measure μ, written as $\nu \ll \mu$, if every μ-nullset is a ν-nullset. If ν has a μ-density, then it is absolutely continuous w. r. t. μ. On the other hand, according to the Radon-Nikodym theorem, if ν is absolutely continuous w. r. t. μ and μ is σ-finite[26], then ν has a μ-density, which is finite μ-almost everywhere if and only if ν is σ-finite. Since the density is μ-almost everywhere uniquely determined it can be chosen finite everywhere exactly when ν is σ-finite. The density is also called Radon-Nikodym derivative or Radon-Nikodym density of ν w. r. t. μ and is often denoted by $\mathrm{d}\nu/\,\mathrm{d}\mu$.

The measure ν is called singular w. r. t. the measure μ if a μ-nullset N exists, such that N^{c} is a ν-nullset. The relation is symmetric and the singularity of ν w. r. t. μ means, there exists a μ-nullset N, such that $\nu(F) = \nu(F \cap N)$ holds for all $F \in \mathcal{F}$.

(iii) *Lebesgue's decomposition theorem.* If μ and ν are σ-finite, then, due to Lebesgue's decomposition theorem, there exists a μ-nullset N, such that $\nu(\cdot \cap N^{\mathrm{c}}) \ll \mu$ holds, where the measure $\nu(\cdot \cap N^{\mathrm{c}})$ and therefore the decomposition

$$\nu = \nu(\cdot \cap N^{\mathrm{c}}) + \nu(\cdot \cap N) \tag{1}$$

are unique. It follows with (ii) that $\nu(\cdot \cap N^{\mathrm{c}})$ has a μ-density, which is finite everywhere and that $\nu(\cdot \cap N)$ is singular w. r. t. μ, since $\nu(N^{\mathrm{c}} \cap N) = \nu(\varnothing) = 0$. If $\nu \ll \nu$ holds, then the singular part in (1) is zero, since we can choose $N = \varnothing$ and because the decomposition is unique. If ν is singular w. r. t. μ, then the absolutely continuous part in (1) is zero, which follows from the definition of singularity and the uniqueness of the decomposition.

(A.8) Theorem (*Fubini's theorem). Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be measurable spaces. Suppose μ is a σ-finite measure[26] on $\mathcal{F}_1$ and ν is a σ-finite measure on $\mathcal{F}_2$. Further suppose f is a real-valued or numerical $\mathcal{F}_1 \otimes \mathcal{F}_2$-measurable function on the product space $\Omega_1 \times \Omega_2$.*

(i) If f is nonnegative, then

$$\omega_1 \mapsto \int_{\Omega_2} f(\omega_1, \omega_2) \,\mathrm{d}\nu(\omega_2), \quad \omega_1 \in \Omega_1, \tag{1}$$

[26] See footnote 15 on page 28.

is a nonnegative $\mathcal{F}_1$-measurable function on Ω_1. Correspondingly,

$$\omega_2 \mapsto \int_{\Omega_1} f(\omega_1, \omega_2)\, \mathrm{d}\mu(\omega_1), \quad \omega_2 \in \Omega_2, \tag{2}$$

is a nonnegative $\mathcal{F}_2$-measurable function on Ω_2. We further have

$$\begin{aligned}\int_{\Omega_1 \times \Omega_2} f \,\mathrm{d}\mu \otimes \nu &= \int_{\Omega_1} \left[\int_{\Omega_2} f(\omega_1, \omega_2)\, \mathrm{d}\nu(\omega_2) \right] \mathrm{d}\mu(\omega_1) \\ &= \int_{\Omega_2} \left[\int_{\Omega_1} f(\omega_1, \omega_2)\, \mathrm{d}\mu(\omega_1) \right] \mathrm{d}\nu(\omega_2).\end{aligned}$$

(ii) If f is $\mu \otimes \nu$-integrable, then $f(\omega_1, \cdot)$ is ν-integrable for μ-almost all $\omega_1 \in \Omega_1$ and

$$A := \big\{\omega_1 \in \Omega_1 : f(\omega_1, \cdot) \text{ is not } \nu\text{-integrable}\big\} \in \mathcal{F}_1.$$

Correspondingly, $f(\cdot, \omega_2)$ is μ-integrable for ν-almost all $\omega_2 \in \Omega_2$ and

$$B := \big\{\omega_2 \in \Omega_2 : f(\cdot, \omega_2) \text{ is not } \mu\text{-integrable}\big\} \in \mathcal{F}_2.$$

Then the function defined by (1) *is μ-integrable over A^{c} and the function defined by* (2) *is ν-integrable over B^{c}. We further have*[27]

$$\begin{aligned}\int_{\Omega_1 \times \Omega_2} f \,\mathrm{d}\mu \otimes \nu &= \int_{A^{\mathrm{c}}} \left[\int_{\Omega_2} f(\omega_1, \omega_2)\, \mathrm{d}\nu(\omega_2) \right] \mathrm{d}\mu(\omega_1) \\ &= \int_{B^{\mathrm{c}}} \left[\int_{\Omega_1} f(\omega_1, \omega_2)\, \mathrm{d}\mu(\omega_1) \right] \mathrm{d}\nu(\omega_2).\end{aligned}$$

(iii) If for f one of the following integrals is finite

$$\int_{\Omega_1 \times \Omega_2} |f| \,\mathrm{d}\mu \otimes \nu,$$

$$\int_{\Omega_1} \left[\int_{\Omega_2} |f(\omega_1, \omega_2)|\, \mathrm{d}\nu(\omega_2) \right] \mathrm{d}\mu(\omega_1), \quad \int_{\Omega_2} \left[\int_{\Omega_1} |f(\omega_1, \omega_2)|\, \mathrm{d}\mu(\omega_1) \right] \mathrm{d}\nu(\omega_2),$$

then all integrals are finite and equal, f is $\mu \otimes \nu$-integrable, and the statements of (ii) *are valid.*

This form of Fubini's theorem is taken form (Elstrodt, 2005, Ch. V, Satz 2.1). See (Bauer, 2001, Th. 23.6, Cor. 23.7) for a similar formulation. Note that the theorem also holds for quasi-integrable functions. A version for complete product spaces can be found in (Elstrodt, 2005, Ch. V, Satz 2.4) or (Rudin, 1987, Th. 8.8).

(A.9) Theorem (*Approximation theorem). Suppose $(\Omega, \mathcal{F})$ is a measurable space with $\mathcal{F} = \sigma(\mathcal{A})$, where $\mathcal{A}$ is an algebra. Let P and Q be probability measures on $\mathcal{F}$. Then there exists for any $\epsilon > 0$ and $F \in \mathcal{F}$ a set $A \in \mathcal{A}$ such that*

$$\mathrm{P}(A \mathbin{\triangle} F) \le \epsilon \qquad \textit{and} \qquad \mathrm{Q}(A \mathbin{\triangle} F) \le \epsilon$$

hold simultaneously.

[27] Note that this result is usually stated with A^{c} replaced by Ω_1 and B^{c} replaced by Ω_2. A justification is given in (Elstrodt, 2005, p. 176).

This generalized form of the approximation theorem is taken from (Dobrushin, 1963, p. 356). A version with a finite number of probability measures is given in (Bradley, 2007, A056).

(A.10) Lemma (*Factorization lemma). Let Ω be a set and $(\Omega', \mathcal{F}')$ be a measurable space. Assume that g is a function on Ω with values in Ω' and f is a real-valued (numerical) function on Ω. The function f is $\sigma(g)$-measurable if and only if there exists a real-valued (numerical) $\mathcal{F}'$-measurable function h on Ω' such that*

$$f = h(g).$$

If f is nonnegative and $\sigma(g)$-measurable, then there exists such a function h, which is nonnegative.

This lemma is taken from (Bauer, 2001, Lem. 11.7).

(A.11) Outer measure. Suppose $(\Omega, \mathcal{F})$ is a measurable space and μ is a finite measure on $\mathcal{F}$, i. e., a measure satisfying $\mu(\Omega) < \infty$, for example a probability measure. Then we call the set function μ^*, defined for all $G \subset \Omega$ by

$$\mu^*(G) = \inf_{G \subset F \in \mathcal{F}} \mu(F),$$

the outer measure generated by μ or outer μ-measure. This form is suitable in the context of the thesis and is adopted from (Doob, 1947, p. 20).

(i) We obviously have for all $A \subset B \subset \Omega$

$$\mu^*(A) \leq \mu^*(B)$$

i. e., the outer μ-measure is monotone.

(ii) If f is a nonnegative $\mathcal{F}$-measurable function on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ satisfying $f(\omega) \leq \mathrm{c}$ for all $\omega \in A$, where c is a nonnegative constant and $A \subset \Omega$ is a set with outer P-measure equal to 1. Then we have

$$\mathrm{E}(f) = \int_\Omega f \,\mathrm{dP} \leq \mathrm{c}.$$

Proof: Assume that

$$\int_\Omega f \,\mathrm{dP} > \mathrm{c}$$

holds. Then we have $\mathrm{P}(f > \mathrm{c}) > 0$. Since f is $\mathcal{F}$-measurable, $\{f > \mathrm{c}\} \in \mathcal{F}$ holds and because $f(\omega) \leq \mathrm{c}$ for all $\omega \in A$ we have $\{f > \mathrm{c}\} \subset A^{\mathrm{c}}$. Furthermore, for all $B \in \mathcal{F}$ with $B \subset A^{\mathrm{c}}$ we have $\mathrm{P}(B) = 0$ because the outer P-measure of A is equal to 1. This implies $\mathrm{P}(f > \mathrm{c}) = 0$, which is a contradiction to the initial assumption. Therefore, the assertion must be true.

(iii) Assume that $(\Omega_1, \mathcal{F}_1, \mathrm{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathrm{P}_2)$ are two probability spaces. If $A \subset \Omega_1$ has outer P_1-measure equal to 1 and $B \subset \Omega_2$ has outer P_2-measure equal to 1, then $A \times B \subset \Omega_1 \times \Omega_2$ has outer $(\mathrm{P}_1 \otimes \mathrm{P}_2)$-measure equal to 1.

Proof. Suppose the outer $\mathrm{P}_1 \otimes \mathrm{P}_2$-measure of $A \times B$ is less than 1. Due to the definition of the outer measure there exists a set $G \in \mathcal{F}_1 \otimes \mathcal{F}_2$ with $A \times B \subset G$ and $\mathrm{P}_1 \otimes \mathrm{P}_2(G) < 1$. Let p denote the $\mathcal{F}_1$-measurable function on Ω_1 defined by

$$p(\omega_1) = \mathrm{P}_2(G_{\omega_1}), \qquad \omega_1 \in \Omega_1,$$

where G_{ω_1} denotes the ω_1-section of G. Then we have

$$\int_{\Omega_1} p(\omega_1)\,\mathrm{d}\mathrm{P}_1(\omega_1) = \mathrm{P}_1 \otimes \mathrm{P}_2(G) < 1,$$

which implies

$$\mathrm{P}_1\big(p < 1\big) > 0. \tag{1}$$

For all $\omega_1 \in A$ we have $B \subset G_{\omega_1}$. Therefore and because the outer P_2-measure of B is equal to 1 we have $p(\omega_1) = 1$ for all $\omega_1 \in A$. This implies $\{p < 1\} \subset A^{\mathrm{c}}$ and we also have $\{p < 1\} \in \mathcal{F}_1$. Since the outer P_1-measure of A is equal to 1, any $E \in \mathcal{F}_1$ with $E \subset A^{\mathrm{c}}$ satisfies $\mathrm{P}_1(E) = 0$. Thus, we must have $\mathrm{P}_1\big(p < 1\big) = 0$, which is a contradiction to (1). Therefore, the initial assumption is false and the proof complete.

(A.12) Standard extension. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and assume that $A \subset \Omega$ is an arbitrary set with outer P-measure equal to 1. If $\tilde{\mathcal{F}} = \sigma(\mathcal{F} \cup \{A\})$ denotes the smallest σ-algebra containing $\mathcal{F}$ and A, then any set $\tilde{F} \in \tilde{\mathcal{F}}$ can be written in the form

$$\tilde{F} = F \cap A \,\cup\, G \cap A^{\mathrm{c}} \tag{1}$$

for some $F, G \in \mathcal{F}$. Based on this representation we define the set function $\tilde{\mathrm{P}}$ by

$$\tilde{\mathrm{P}}(\tilde{F}) = \mathrm{P}(F), \tag{2}$$

which is a probability measure on $\tilde{\mathcal{F}}$. The probability space $(\Omega, \tilde{\mathcal{F}}, \tilde{\mathrm{P}})$ is called the standard extension of $(\Omega, \mathcal{F}, \mathrm{P})$ w. r. t. the set A. This extension is nontrivial for $A \notin \mathcal{F}$ and it allows to consider probabilities related to the set A. The condition that the outer P-measure of A is equal to 1 is necessary and sufficient to ensure that $\tilde{\mathrm{P}}$ is well defined, although the decomposition in (1) may not be unique.

The measures $\tilde{\mathrm{P}}$ and P coincide on $\mathcal{F}$ and we have $\tilde{\mathrm{P}}(A) = 1$. From (1) and (2) we obtain $\tilde{\mathrm{P}}(\tilde{F}) = \tilde{\mathrm{P}}(F)$, i. e., to any set $\tilde{F} \in \tilde{\mathcal{F}}$ corresponds a set $F \in \mathcal{F}$ differing from $\tilde{F}$ by at most a set of $\tilde{\mathrm{P}}$-measure 0. Thus, the σ-algebra $\tilde{\mathcal{F}}$ is only a slight enlargement of $\mathcal{F}$ by $\tilde{\mathrm{P}}$-nullsets. It follows that if $\tilde{f}$ is a real $\tilde{\mathcal{F}}$-measurable function on Ω, then there exists a real $\mathcal{F}$-measurable function f on Ω with

$$\tilde{f} = f \qquad \tilde{\mathrm{P}}\text{-a.s.}$$

If Q is a second probability measure on $\mathcal{F}$ for which the set A has outer Q-measure 1, then we additionally have for the functions $\tilde{f}$ and f

$$\tilde{f} = f \qquad \tilde{\mathrm{Q}}\text{-a.s.},$$

where $\tilde{\mathrm{Q}}$ is the probability measure of the standard extension of $(\Omega, \mathcal{F}, \mathrm{Q})$ w. r. t. A.

See (Doob, 1953, Ch. II.2) or (Doob, 1947, 1940, 1937) for more details but note that standard extensions are considered there in connection with stochastic processes and complete probability spaces with a particular focus on so-called separable and measurable standard extensions. The basic properties of a standard extension, however, hold for an abstract probability space as considered here. See also (Mittelbach, 2012, Par. 1.23–1.27) for a summary of relevant facts of the given references.

(A.13) Measurability and convergence of random series. Assume that $(X, \mathcal{X})$ is an arbitrary measurable space and $\{\xi_k, k \in \mathbb{N}\}$ is a sequence of real-valued $\mathcal{X}/\mathcal{B}(\mathbb{R})$-measurable functions on X. Then for any $n \in \mathbb{N}$ the real-valued function η_n with

$$\eta_n(x) = \sum_{k=1}^{n} \xi_k(x) \tag{1}$$

for all $x \in X$ is $\mathcal{X}/\mathcal{B}(\mathbb{R})$-measurable. The numerical functions $\hat{\eta}$ and $\check{\eta}$ with

$$\hat{\eta}(x) = \liminf_{n\to\infty} \eta_n(x) \qquad \text{and} \qquad \check{\eta}(x) = \limsup_{n\to\infty} \eta_n(x)$$

for all $x \in X$ are $\mathcal{X}/\mathcal{B}(\bar{\mathbb{R}})$-measurable and the set

$$X' = \big\{x \in X : \hat{\eta}(x) = \check{\eta}(x)\big\} \cap \big\{x \in X : \hat{\eta}(x) \text{ and } \check{\eta}(x) \text{ are finite}\big\}$$

satisfies $X' \in \mathcal{X}$. Thus, the set of all $x \in X$ for which $\lim_{n\to\infty} \eta_n(x) =: \sum_{k=1}^{\infty} \xi_k(x)$ exists and is finite is contained in $\mathcal{X}$. Due to the $\mathcal{X}/\mathcal{B}(\mathbb{R})$-measurability of ξ_k the nonnegative function $|\xi_k|$ is $\mathcal{X}/\mathcal{B}(\mathbb{R})$-measurable. Therefore, replacing ξ_k by $|\xi_k|$ in (1) yields as before that the set

$$X'' = \big\{x \in X : \sum_{k=1}^{\infty} |\xi_k(x)| \text{ is finite}\big\}$$

is contained in $\mathcal{X}$. We further have $X'' \subset X'$ since absolutely convergent sequences are in particular convergent. See, e. g., (Cohn, 1980, Sec. 2.1) or (Halmos, 1974, § 19, § 20) for details to these standard measure-theoretic facts.

Let us define the real-valued function η_∞ on X' and the nonnegative function $\tilde{\eta}_\infty$ on X'' by

$$\eta_\infty(x) = \sum_{k=1}^{\infty} \xi_k(x), \quad x \in X' \qquad \text{and} \qquad \tilde{\eta}_\infty(x) = \sum_{k=1}^{\infty} |\xi_k(x)|, \quad x \in X''. \tag{2}$$

Because the restriction of η_n to X' is $\mathcal{X}'/\mathcal{B}(\mathbb{R})$-measurable, where $\mathcal{X}' = \{A \cap X' : A \in \mathcal{X}\}$ denotes the trace-σ-algebra of $\mathcal{X}$ w. r. t. X', the function η_∞ is $\mathcal{X}'/\mathcal{B}(\mathbb{R})$-measurable. Due to (Dudley, 2002, Th. 4.2.5) there exists an extension of η_∞ to a real-valued function on all of X, which is $\mathcal{X}/\mathcal{B}(\mathbb{R})$-measurable. Correspondingly, $\tilde{\eta}_\infty$ can be extended to a nonnegative numerical function on all of X, which is $\mathcal{X}/\mathcal{B}(\bar{\mathbb{R}})$-measurable. A natural possibility is to put $\tilde{\eta}_\infty(x) = \infty$ for all $x \notin X''$.

Assume that μ is a probability measure on $\mathcal{X}$. Let us define the function Ψ on $\mathbb{N} \times X$ by

$$\Psi(k, x) = \xi_k(x)$$

for all $k \in \mathbb{N}$ and $x \in X$. Then Ψ and $|\Psi|$ are $2^{\mathbb{N}} \otimes \mathcal{X}/\mathcal{B}(\mathbb{R})$-measurable because

$$\{\Psi \in B\} = \bigcup_{k\in\mathbb{N}} \{k\} \times \{\xi_k \in B\}$$

holds for all $B \in \mathcal{B}(\mathbb{R})$. With Ψ and the counting measure λ on $\mathbb{N}$ introduced in Paragraph A.4 we obtain from part (A.8.i) of Fubini's theorem

$$\begin{aligned}\sum_{k=1}^{\infty} \mathrm{E}|\xi_k| &= \int_{\mathbb{N}} \left[\int_X |\Psi(k,x)| \, \mathrm{d}\mu(x) \right] \mathrm{d}\lambda(k) \\ &= \int_X \left[\int_{\mathbb{N}} |\Psi(k,x)| \, \mathrm{d}\lambda(k) \right] \mathrm{d}\mu(x) \\ &= \int_X \left[\sum_{k=1}^{\infty} |\xi_k(x)| \right] \mathrm{d}\mu(x) \\ &= \mathrm{E}\left(\sum_{k=1}^{\infty} |\xi_k| \right).\end{aligned} \tag{3}$$

Under the condition

$$\sum_{k=1}^{\infty} \mathrm{E}|\xi_k| = \sum_{k=1}^{\infty} \int_X |\xi_k(x)| \, \mathrm{d}\mu(x) < \infty \tag{4}$$

we obtain from part (A.8.iii) of Fubini's theorem that $|\Psi|$ is $\lambda \otimes \mu$-integrable. Then part (A.8.ii) of Fubini's theorem implies $|\Psi(\cdot, x)|$ is λ-integrable for μ-almost all $x \in X$, and therefore we have $\mu(X'') = 1$. Due to $X'' \subset X'$ we also have $\mu(X') = 1$. Thus, if (4) holds, then η_∞ and $\tilde{\eta}_\infty$ are random variables on $(X, \mathcal{X}, \mu)$, defined μ-almost surely by (2) so that it does not matter how they are defined outside X' or X'', respectively. Due to part (A.8.iii) of Fubini's theorem also Ψ is $\lambda \otimes \mu$-integrable under condition (4). Thus part (A.8.ii) of Fubini's theorem applies and as in (3) we obtain that summation and expectation can be exchanged, i. e.,

$$\sum_{k=1}^{\infty} \mathrm{E}(\xi_k) = \mathrm{E}\left(\sum_{k=1}^{\infty} \xi_k \right) < \infty, \tag{5}$$

where we use the series on the right-hand side as synonym for η_∞, which is justified by the almost sure convergence. The finiteness of the expectation is due to the integrability of Ψ.

In addition to the almost sure convergence we sometimes require the convergence in mean square. Assume that

$$\sum_{k=1}^{\infty} \sqrt{\mathrm{E}(\xi_k^2)} < \infty \tag{6}$$

holds. Due to Jensen's inequality (6) implies (4) so that $\eta_n = \sum_{k=1}^{n} \xi_k$ converges μ-almost surely to $\sum_{k=1}^{\infty} \xi_k$ as $n \to \infty$ and (5) holds. Moreover, (6) implies mean square convergence of η_n to $\sum_{k=1}^{\infty} \xi_k$, i. e.,

$$\lim_{n \to \infty} \mathrm{E}\left(\eta_n - \sum_{k=1}^{\infty} \xi_k \right)^2 = 0. \tag{7}$$

Indeed, for all $m < n$ we have

$$\begin{aligned} \mathrm{E}\big(\eta_n - \eta_m\big)^2 = \mathrm{E}\bigg(\sum_{k=m+1}^{n} \xi_k\bigg)^2 &= \sum_{i,j=m+1}^{n} \mathrm{E}(\xi_i \xi_j) \\ &\leq \sum_{i,j=m+1}^{n} \sqrt{\mathrm{E}(\xi_i^2)\mathrm{E}(\xi_j^2)} \\ &= \bigg(\sum_{i=m+1}^{n} \sqrt{\mathrm{E}(\xi_i^2)}\bigg)^2, \end{aligned} \tag{8}$$

where we have applied the Cauchy-Schwarz inequality. Condition (6) implies (8) converges to 0 as $m, n \to \infty$. Therefore, η_n converges in mean square by the Cauchy criterion. Given a sequence of random variables converges almost surely and in mean square, then both limits are equal almost surely by Fatou's lemma so that we finally have (7).

Due to the mean square convergence, we have

$$\mathrm{E}\bigg(\sum_{k=1}^{\infty} \xi_k\bigg)^2 = \lim_{n\to\infty} \mathrm{E}(\eta_n^2) = \lim_{n\to\infty} \sum_{i,j=1}^{n} \mathrm{E}(\xi_i \xi_j), \tag{9}$$

where the limit on the right-hand side is finite because we can bound the sum as in (8) and because condition (6) holds. The equality in (9) is actually the reason why we are interested in mean square convergence. When the ξ_k's are uncorrelated, then

$$\sum_{i,j=m+1}^{n} \mathrm{E}(\xi_i \xi_j) = \sum_{i=m+1}^{n} \mathrm{E}(\xi_i^2)$$

so that the mean square convergence of the η_n's already follows from $\sum_{k=1}^{\infty} \mathrm{E}(\xi_k^2) < \infty$, a condition implied by and therefore weaker than (6). In that case we have

$$\mathrm{E}\bigg(\sum_{k=1}^{\infty} \xi_k\bigg) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{E}(\xi_i), \tag{10}$$

$$\mathrm{E}\bigg(\sum_{k=1}^{\infty} \xi_k\bigg)^2 = \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{E}(\xi_i^2) < \infty, \tag{11}$$

where the series on the left-hand side denote the mean square limit. Note that in the present situation (4) is not required for (10) to hold. If the ξ_k's are even independent with $\mathrm{E}(\xi_k) = 0$, then we also have μ-almost sure convergence. This result is given, e. g., in (Shiryaev, 1995, Ch. IV.2, Th. 1). For more details on convergence in mean square see (Jazwinski, 1970, Ch. 3.3).

Usually, we use the previous results in the situation, where $(X, \mathcal{X})$ is the product measurable space generated by the sequence $\{(X_k, \mathcal{X}_k), k \in \mathbb{N}\}$ of measurable spaces with $(X_k, \mathcal{X}_k) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then the function ξ_k is typically given by

$$\xi_k(x) = a_k x_k$$

for all $x = \{x_k, k \in \mathbb{N}\} \in X$, where a_k is some real constant.

B Ergodicity and Mixing in the Ergodic-Theoretic Sense

This section provides a tailored selection of basic definitions and facts from ergodic theory relevant for the thesis. Some of the results are proved, either because of their special form or to demonstrate a typical way of reasoning. There is a large literature on ergodic theory. Among the standard books are (Halmos, 1956; Cornfeld et al., 1982; Walters, 1982; Petersen, 1983). References with an information-theoretic orientation are (Billingsley, 1965), (Kakihara, 1999, Ch. II), and (Gray, 2009, Secs. 7, 8, 10).

Subsequently, we use the notation introduced in Paragraph 1.2 to denote by $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ the product measurable spaces generated by the families $\{(X_t, \mathcal{X}_t), t \in T\}$ and $\{(Y_t, \mathcal{Y}_t), t \in T\}$ of measurable spaces for which $(X_t, \mathcal{X}_t) = (X_0, \mathcal{X}_0)$ and $(Y_t, \mathcal{Y}_t) = (Y_0, \mathcal{Y}_0)$ for all $t \in T$. Further, $\xi = \{\xi_t, t \in T\}$ is a random process on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$, where the random variable ξ_t has values in the measurable space $(X_t, \mathcal{X}_t)$. We consider the discrete- and the continuous-time case in parallel so that the definitions and results apply to both, $T = \mathbb{Z}$ or $T = \mathbb{R}$, unless stated otherwise.

(B.1) Definition (Stationarity, ergodicity, i.i.d.). Let μ be a probability measure on $\mathcal{X}$ and assume that $s \in T$. Then μ is called s-stationary if

$$\mu(A) = \mu(\theta_s(A))$$

holds for all $A \in \mathcal{X}$ and s-ergodic if

$$\mu(A) = 0 \qquad \text{or} \qquad \mu(A) = 1 \tag{1}$$

holds for all s-invariant sets $A \in \mathcal{X}$. The probability measure μ is called (strictly) stationary if it is s-stationary for all $s \in T$. It is called totally ergodic if it is s-ergodic for all $s \in T_+$. Further, μ is called ergodic if (1) holds for all invariant sets $A \in \mathcal{X}$.

Assume that $s \in T_+$ and let μ_k be a probability measure on $\mathcal{X}_{ks}^{(k+1)s}$ for all $k \in \mathbb{Z}$. Then the product measure

$$\mu = \bigotimes_{k \in \mathbb{Z}} \mu_k$$

on $\mathcal{X}$ is also called an s-independent probability measure. If $\mu_k = \mu_0$ for all $k \in \mathbb{Z}$, then the s-stationary, s-independent probability measure μ is also called s-i.i.d. (i. e., s-independent and s-identically distributed) probability measure.

The random process ξ is called s-stationary (stationary, s-ergodic, (totally) ergodic, s-independent, s-i.i.d.), if the distribution P_ξ of ξ is s-stationary (stationary, s-ergodic, (totally) ergodic, s-independent, s-i.i.d.). If $T = \mathbb{Z}$ and P_ξ is a 1-i.i.d. probability measure, then ξ is called i.i.d.-process or sequence, i. e., sequence of independent and identically distributed random variables.

(B.2) Remark. Stationarity and ergodicity are standard concepts. The definition of s-stationarity and s-ergodicity is adopted from Berger (1968, p. 256/257). Then s-i.i.d. probability measures are natural special cases. We use the term total ergodicity as in (Gray, 2009, Sec. 7.8). Synonyms are complete ergodicity (Pinsker, 2007, p. 384) or block ergodicity (Berger, 1968, p. 257). It is reasonable to alternatively call s-independent (and in particular s-i.i.d.) probability measures also s-memoryless.

Let $s \in T_+$ and assume that the probability measure μ on $\mathcal{X}$ is s-stationary. Then μ is s-ergodic if and only if

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu\big(\theta_{ks}(A) \cap B\big) = \mu(A)\mu(B)$$

holds for all $A, B \in \mathcal{X}$.

If we have discrete time, then stationarity and 1-stationarity are equivalent as well as ergodicity and 1-ergodicity. In the continuous-time case a stationary probability measure μ is ergodic if and only if

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \mu\big(\theta_t(A) \cap B\big)\, \mathrm{d}t = \mu(A)\mu(B)$$

holds for all $A, B \in \mathcal{X}$. If $\mathcal{X} = \sigma(\mathcal{G})$ for a family $\mathcal{G}$ of subsets of X, which is closed w. r. t. finite intersections, then it is sufficient to consider only sets $A, B \in \mathcal{G}$. The same holds for the given characterization of s-ergodicity. The characterizations of (s-) ergodicity for (s-) stationary probability measures is standard. See for example (Pinsker, 1964, p. 70).

(B.3) Definition (Mixing in the ergodic-theoretic sense). Let μ be a probability measure on $\mathcal{X}$. If μ is stationary and

$$\lim_{t\to\infty} \big|\mu\big(\theta_t(A) \cap B\big) - \mu(A)\mu(B)\big| = 0 \tag{1}$$

holds for all $A, B \in \mathcal{X}$, then it is called (strongly) mixing (in the ergodic-theoretic sense).

Let $s \in T_+$ and assume that the probability measure μ is s-stationary. Then μ is called s-weakly mixing (in the ergodic-theoretic sense) if

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \big|\mu\big(\theta_{ks}(A) \cap B\big) - \mu(A)\mu(B)\big| = 0 \tag{2}$$

holds for all $A, B \in \mathcal{X}$. If μ is stationary and s-weakly mixing for all $s \in T_+$, then it is called totally weakly mixing.

If $T = \mathbb{Z}$, i. e., in the discrete-time case, a stationary probability measure is called weakly mixing, if it is 1-weakly mixing. If $T = \mathbb{R}$, i. e., in the continuous-time case, a stationary probability measure μ is called weakly mixing if

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \big|\mu\big(\theta_s(A) \cap B\big) - \mu(A)\mu(B)\big|\, \mathrm{d}s = 0 \tag{3}$$

holds for all $A, B \in \mathcal{X}$.

The random process ξ is called mixing in the ergodic-theoretic sense (s-weakly mixing, (totally) weakly mixing), if its distribution P_ξ has this property.

(B.4) Remark. These definitions are standard. If $\mathcal{X} = \sigma(\mathcal{G})$ for a family $\mathcal{G}$ of subsets of X, which is closed w. r. t. finite intersections, then it is sufficient to verify (B.3.1) or (B.3.2) or (B.3.3) only for sets $A, B \in \mathcal{G}$ to prove the (weak) mixing property. See for example (Pinsker, 1964, p. 71) or (Petersen, 1983, p. 58) in this regard.

Please note, that the s-weakly mixing condition is based on sums but the weakly mixing condition in the continuous-time case is based on integrals. Also note that the mixing properties of probability measures and random processes are considered here only in connection with stationarity. Some non-stationary versions are given, e. g., in (Kakihara, 1999, p. 99).

Total ergodicity is a concept relevant in connection with block coding theorems because in this context we partition the time axis into segments of equal size. Theorem B.7 below shows that totally weakly mixing probability measures are always totally ergodic, without additional assumptions. In contrast, in the continuous-time case weakly mixing probability measures have to satisfy certain continuity conditions to be totally ergodic. This is the reason why we introduced the easier to handle totally weakly mixing concept.

(B.5) Definition (Continuity). Let μ be a stationary probability measure on $\mathcal{X}$ and assume that $T = \mathbb{R}$. Then μ is called continuous in the sense of Pinsker if for all $A \in \mathcal{X}$ we have

$$\lim_{t \to 0} \mu(\theta_t(A) \triangle A) = 0.$$

The stationary continuous-time random process $\xi = \{\xi_t, t \in \mathbb{R}\}$ is called continuous in the sense of Pinsker if its distribution P_ξ is continuous in the sense of Pinsker.

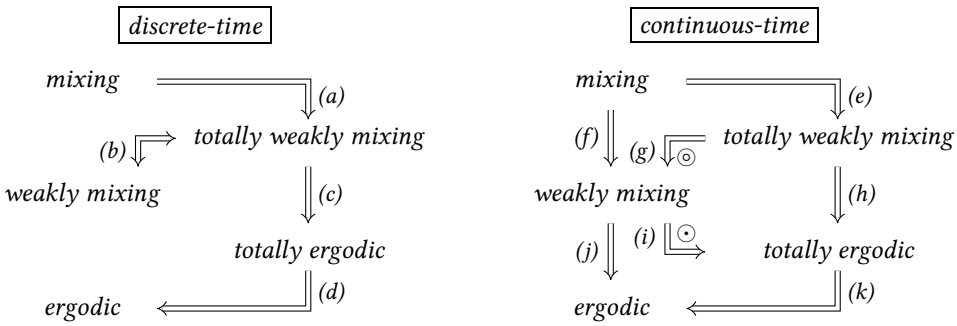
(B.6) Remark. This special form of continuity is introduced in (Pinsker, 1964, p. 70), which motivates the name that is adopted from (Pursley, 1977).

Suppose the space $(X_0, \mathcal{X}_0)$ of values of the random variables ξ_t is a separable metric space equipped with the corresponding Borel-σ-algebra. Then Pursley (1977) has shown that the stationary continuous-time random process $\xi = \{\xi_t, t \in \mathbb{R}\}$ is continuous in the sense of Pinsker if and only if it is continuous in probability (see (Gikhman and Skorokhod, 1974, p. 168)). In particular, a real-valued stationary continuous-time random process ξ is continuous in the sense of Pinsker if and only if for all $\epsilon > 0$

$$\lim_{t \to \infty} \mathrm{P}(|\xi_0 - \xi_t| > \epsilon) = 0. \tag{1}$$

For example, continuity in mean square (see Paragraph C.1) is sufficient for (1) to hold.

(B.7) Theorem (*Relations between ergodicity and mixing). Ergodicity and mixing (in the ergodic-theoretic sense) are related in the following way.*



All implications involving a mixing condition require stationarity. The implication marked by $\odot$ requires continuity in the sense of Pinsker. The implication marked by $\circledcirc$ requires a continuity condition specified in the proof below.

Proof. We verify the implications one by one. (a) and (e) hold because the Cesàro mean of a convergent sequence has the same limit as the sequence itself.

The downward implication of (b) holds by definition. The upward implication follows from the fact that a 1-weakly mixing probability measure is k-weakly mixing for all $k \in \mathbb{N}$ (see (Mittelbach, 2012, Lem. 4.16)).

The implications (c), (h), and (j) hold because based on the characterizations given in Remark B.2 we obtain that an (s-) stationary (s-) weakly mixing probability measure is (s-) ergodic.

Since an invariant set is s-invariant for all $s \in T$ implications (d) and (k) hold per definition.

Implication (f) is based on the following fact: For a nonnegative and bounded function h on $[0,\infty)$ with $\lim_{s\to\infty} h(s) = 0$ we have $\lim_{t\to\infty} \frac{1}{t}\int_0^t h(s)\,\mathrm{d}s = 0$.

To show (g) let μ be a stationary totally weakly mixing probability measure on $\mathcal{X}$, where $T = \mathbb{R}$. For all $A, B \in \mathcal{X}$ suppose $\mu(\theta_t(A) \cap B)$ is uniformly continuous as function of $t \in [0,\infty)$. Then for any $\epsilon > 0$ there exists a $\delta > 0$ such that for all $s \in [0,\infty)$ and $|t| < \delta$

$$\left|\mu(\theta_s(A) \cap B) - \mu(\theta_{s+t}(A) \cap B)\right| \leq \epsilon. \tag{1}$$

Let us fix $A, B \in \mathcal{X}$ and $\epsilon > 0$ and assume that $\delta > 0$ is chosen such that (1) holds. We can represent $t > 0$ as $t = n\delta + t_0$ for some integer n and $t_0 \in [0,\delta)$ and obtain

$$\begin{aligned}
&\frac{1}{t}\int_0^t \left|\mu(\theta_s(A) \cap B) - \mu(A)\mu(B)\right| \mathrm{d}s \\
&\qquad \leq \frac{1}{n\delta}\int_0^{(n+1)\delta} \left|\mu(\theta_s(A) \cap B) - \mu(A)\mu(B)\right| \mathrm{d}s \\
&\qquad = \frac{1}{n\delta}\sum_{k=0}^{n}\int_{k\delta}^{(k+1)\delta} \left|\mu(\theta_s(A) \cap B) - \mu(A)\mu(B)\right| \mathrm{d}s \\
&\qquad \leq \frac{1}{n\delta}\sum_{k=0}^{n}\int_{k\delta}^{(k+1)\delta} \Big(\left|\mu(\theta_{k\delta}(A) \cap B) - \mu(A)\mu(B)\right| + \epsilon\Big) \mathrm{d}s \\
&\qquad = \frac{1}{n}\sum_{k=0}^{n} \left|\mu(\theta_{k\delta}(A) \cap B) - \mu(A)\mu(B)\right| + \epsilon,
\end{aligned}$$

where the second inequality follows from (1). Therefore, we have

$$\begin{aligned}
0 \leq \lim_{t\to\infty} \frac{1}{t}\int_0^t &\left|\mu(\theta_s(A) \cap B) - \mu(A)\mu(B)\right| \mathrm{d}s \\
&\leq \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)\frac{1}{n+1}\sum_{k=0}^{n} \left|\mu(\theta_{k\delta}(A) \cap B) - \mu(A)\mu(B)\right| + \epsilon \\
&= \epsilon,
\end{aligned}$$

where the equality holds because μ is totally weakly mixing. Since ϵ was chosen arbitrary μ is weakly mixing under the introduced continuity assumption.

Finally, implication (i) is shown in (Berger, 1968, p. 271) for stationary probability measures that are continuous in the sense of Pinsker. □

(B.8) Example (Total ergodicity vs. weak mixing). We give an example, which shows that total ergodicity does not imply weak mixing. Assume that the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is chosen such that $\Omega = (0,1]$ is the unit interval equipped with the Borel-σ-algebra $\mathcal{F} = \mathcal{B}((0,1])$ and the Lebesgue measure $\mathrm{P} = \lambda$ on $(0,1]$. Let ϕ be a transformation of Ω into Ω given by

$$\phi(\omega) = \omega + a \mod 1$$

for all $\omega \in \Omega$, where $a \in (0,1]$ is a fixed irrational number. The transformation ϕ is $\mathcal{F}/\mathcal{F}$-measurable and invertible. The probability measure P is preserved under ϕ and the inverse ϕ^{-1} is $\mathcal{F}/\mathcal{F}$-measurable.

Suppose the set $X_0 = \{0,1\}$ is equipped with the power set as σ-algebra, denoted by $\mathcal{X}_0$. We define the random variable ξ_0 on $(\Omega, \mathcal{F}, \mathrm{P})$ with values in $(X_0, \mathcal{X}_0)$ by

$$\xi_0(\omega) = \begin{cases} 0 & \text{if } \omega \in (0, 1/2] \\ 1 & \text{otherwise} \end{cases}$$

for all $\omega \in \Omega$. Further, we consider the random sequence $\xi = \{\xi_k, k \in \mathbb{Z}\}$, with ξ_k given by

$$\xi_k(\omega) = \xi_0\big(\phi^k(\omega)\big)$$

for all $\omega \in \Omega$. Here ϕ^k denotes the k-fold application of ϕ if k is nonnegative and of ϕ^{-1} if k is negative. The sequence ξ of binary random variables is stationary and totally ergodic but not weakly mixing. This standard example has different equivalent representations and is known as irrational rotation of the unit circle. See (Walters, 1982, pp. 20, 29, 50), (Billingsley, 1965, Sec. 5) and (Berger, 1968, Appendix) for details.

A direct consequence of the extremal property of stationary ergodic probability measures is the following lemma.

(B.9) Lemma (*Average of stationary ergodic measures). Let $s \in T_+$ and assume that μ_1 and μ_2 are distinct (s-) stationary and (s-) ergodic probability measures on $\mathcal{X}$. Then for any $\alpha \in (0,1)$ the probability measure*

$$\bar{\mu} = \alpha\mu_1 + (1-\alpha)\mu_2$$

is (s-) stationary but not (s-) ergodic.

(B.10) Definition (Invariant function). Suppose f is an $\mathcal{X}/\mathcal{Y}$-measurable function on X with values in Y and assume that $s \in T$. Then f is called s-invariant if for all $x \in X$

$$f\big(\theta_s(x)\big) = \theta_s\big(f(x)\big),$$

where θ_s denotes the shift operator defined in Paragraph 1.2. The function f is called invariant if it is s-invariant for all $s \in T$.

(B.11) Lemma (*s-i.i.d. probability measures are mixing). Assume that $s \in T_+$. If μ is an s-i.i.d. probability measure on $\mathcal{X}$, then for all $A, B \in \mathcal{X}$ we have*

$$\lim_{k\to\infty} \mu\big(\theta_{ks}(A) \cap B\big) = \mu(A)\mu(B). \tag{1}$$

In particular, μ is s-stationary and s-weakly mixing.

(B.12) Remark. The proof of Lemma B.11 given below is similar to that of (Bradley, 2007, Prop. 2.8) and is based on standard techniques used in ergodic theory. Condition (B.11.1) is a version of (B.3.1), where only shifts that are multiples of s are considered. This is the adequate form for s-stationary probability measures. It is immediately clear, that an s-stationary probability measure satisfying (B.11.1) is s-weakly mixing.

Proof. Let $\epsilon > 0$ and $A, B \in \mathcal{X}$ be arbitrary. Due to the approximation theorem for probability measures (see Theorem A.9) there exist sets

$$F = \bigcup_{i=1}^{m} F_i, \qquad G = \bigcup_{j=1}^{n} G_j, \qquad F_i, G_j \in \left[\bigtimes_{k=-l}^{l} \mathcal{X}_{ks}^{(k+1)s} \right] \tag{1}$$

such that

$$\mu\big(A \triangle F\big) \le \frac{\epsilon}{4}, \qquad \mu\big(B \triangle G\big) \le \frac{\epsilon}{4}, \tag{2}$$

where the cylinder sets F_i are disjoint, the cylinder sets G_j are disjoint, and l, m, n are positive integers. Let k be a positive integer. Applying the triangle inequality we obtain

$$\begin{aligned} \big|\mu\big(\theta_{ks}(A) \cap B\big) - \mu(A)\mu(B)\big| \;\le\; & \big|\mu\big(\theta_{ks}(A) \cap B\big) - \mu\big(\theta_{ks}(F) \cap G\big)\big| && (3)\\ & + \big|\mu\big(\theta_{ks}(F) \cap G\big) - \mu(F)\mu(G)\big| && (4)\\ & + \big|\mu(F)\mu(G) - \mu(A)\mu(B)\big|. && (5) \end{aligned}$$

For the term on the right-hand side of (3) we have

$$\begin{aligned} \big|\mu\big(\theta_{ks}(A) \cap B\big) - \mu\big(\theta_{ks}(F) \cap G\big)\big| &\le \mu\big(\theta_{ks}(A) \triangle \theta_{ks}(F)\big) + \mu\big(B \triangle G\big) \\ &\le \mu\big(A \triangle F\big) + \mu\big(B \triangle G\big) \\ &\le \frac{\epsilon}{2}. \end{aligned}$$

The first inequality results from properties of the symmetric difference (see e. g. (Bradley, 2007, A053)). Since we assume that μ is an s-i.i.d. probability measure, it is s-stationary, which implies the second inequality and the last is due to (2). Similarly, we obtain for (5)

$$\begin{aligned} \big|\mu(F)\mu(G) - \mu(A)\mu(B)\big| &\le \mu\big(A \triangle F\big) + \mu\big(B \triangle G\big) \\ &\le \frac{\epsilon}{2}. \end{aligned}$$

Since F and G are unions of disjoint sets (4) is equal to

$$\left| \sum_{i=1}^{m} \sum_{j=1}^{n} \Big[\mu\big(\theta_{ks}(F_i) \cap G_j\big) - \mu(F_i)\mu(G_j) \Big] \right|. \tag{6}$$

The sets F_i and G_j are cylinder sets as given in (1) and μ is an s-i.i.d. probability measure. Therefore

$$\mu\big(\theta_{ks}(F_i) \cap G_j\big) = \mu\big(\theta_{ks}(F_i)\big)\mu(G_j) = \mu(F_i)\mu(G_j)$$

if $k > 2l$, i. e., (6) is equal to 0 if k is sufficiently large. Combining the derived inequalities we obtain for the left-hand side of (3)

$$\big|\mu\big(\theta_{ks}(A) \cap B\big) - \mu(A)\mu(B)\big| \leq \epsilon$$

for k sufficiently large and because $\epsilon > 0$ was chosen arbitrary we finally have

$$\lim_{k\to\infty} \mu\big(\theta_{ks}(A) \cap B\big) = \mu(A)\mu(B). \qquad \square$$

(B.13) Lemma (*Properties of image measures). Let μ be a probability measure on $\mathcal{X}$. Further suppose f is an $\mathcal{X}/\mathcal{Y}$-measurable function on X with values in Y and μ_f denotes the distribution of f. Then we have the following implications: If f has the property in the first column and μ the property in the second column, then the image measure μ_f has the property in the third column.*

		f	μ	μ_f
(i)†		*(s-) invariant*	*(s-) stationary*	*(s-) stationary*
(ii)	*a)*†	*(s-) invariant*	*(s-) ergodic*	*(s-) ergodic*
	b)	*invariant*	*totally ergodic*	*totally ergodic*
(iii)	*a)*†	*(s-) invariant*	*(s-) stationary and (s-) weakly mixing*	*(s-) stationary and (s-) weakly mixing*
	b)	*invariant*	*stationary and totally weakly mixing*	*stationary and totally weakly mixing*
	c)	*invariant*	*stationary and mixing*	*stationary and mixing*

The superscript † *denotes we assume that $s \in T$ in* (i) *and* (ii) *and $s \in T_+$ in* (iii).

Proof. Part (i). From the s-invariance of f we obtain $f^{-1}(\theta_s(B)) = \theta_s(f^{-1}(B))$ for all $B \in \mathcal{Y}$. If μ is s-stationary, then

$$\begin{aligned}\mu_f(B) &= \mu\big(f^{-1}(B)\big)\\ &= \mu\big(\theta_s(f^{-1}(B))\big)\\ &= \mu\big(f^{-1}(\theta_s(B))\big)\\ &= \mu_f\big(\theta_s(B)\big)\end{aligned}$$

for all $B \in \mathcal{Y}$, i. e., μ_f is s-stationary. From what is shown the assertion regarding stationarity is evident.

Part (ii). Assume that $B \in \mathcal{Y}$ is an (s-) invariant set. Then $f^{-1}(B)$ is (s-) invariant. If μ is (s-) ergodic, then

$$\mu_f(B) = \mu(f^{-1}(B)) = \ 0 \ \text{ or } \ 1,$$

i. e., μ_f is (s-) ergodic. The remaining assertion is now evident.

Part (iii). If f is invariant, then for all $A, B \in \mathcal{Y}$ and $t \in T$ we have

$$\begin{aligned}\mu_f(\theta_t(A) \cap B) &= \mu\big(f^{-1}(\theta_t(A)) \cap f^{-1}(B)\big) \\ &= \mu\big(\theta_t(f^{-1}(A)) \cap f^{-1}(B)\big)\end{aligned}$$

Given f is s-invariant the same holds with $t = ks$ for $k \in \mathbb{Z}, s \in T$. Together with the corresponding definitions this implies the assertions. □

(B.14) Remark. Under further conditions on the function f in Lemma B.13 we also obtain implications in the opposite direction. To be precise, assume that f is $\mathcal{X}/\mathcal{Y}$-measurable and invariant and in addition assume that f is bijective and the inverse function f^{-1} is $\mathcal{Y}/\mathcal{X}$-measurable. Then the probability measure μ is stationary and mixing if and only if the image measure μ_f is stationary and mixing. This equivalence directly follows from the previous proof and the additional assumptions on f. Similar equivalences hold for the remaining ergodicity and mixing properties considered in Lemma B.13.

(B.15) Lemma (*Properties of product measures). Let μ be a probability measures on $\mathcal{X}$ and ν be a probability measure on $\mathcal{Y}$. Then we have the following implications: If μ has the property in the first column and ν the property in the second column, then the product measure $\mu \otimes \nu$ has the property in the third column.*

		μ	ν	$\mu \otimes \nu$
(i)	*a)*†	*(s-) stationary and (s-) weakly mixing*	*(s-) stationary and (s-) ergodic*	*(s-) stationary and (s-) ergodic*
	b)	*stationary and totally weakly mixing*	*stationary and totally ergodic*	*stationary and totally ergodic*
(ii)	*a)*†	*(s-) stationary and (s-) weakly mixing*	*(s-) stationary and (s-) weakly mixing*	*(s-) stationary and (s-) weakly mixing*
	b)	*stationary and totally weakly mixing*	*stationary and totally weakly mixing*	*stationary and totally weakly mixing*
	c)	*stationary and mixing*	*stationary and mixing*	*stationary and mixing*

The superscript † *denotes we assume that* $s \in T_+$ *in* (i) *and* (ii).

Proof. We prove the first assertion of part (i) with the (s-) option. The remaining assertions are shown similarly. The given proof follows (Pinsker, 1964, p. 73) and illustrates the typical argumentation used to obtain such results.

First observe that $\mu \otimes \nu$ is s-stationary if μ and ν are s-stationary. Using Remark B.2 the assertion follows if for all $A_1, A_2 \in \mathcal{X}$ and $B_1, B_2 \in \mathcal{Y}$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu \otimes \nu\big(\theta_{ks}(A_1 \times B_1) \cap (A_2 \times B_2)\big) = \mu \otimes \nu\big(A_1 \times B_1\big)\mu \otimes \nu\big(A_2 \times B_2\big) \tag{1}$$

holds. With the identity

$$\begin{aligned}\mu \otimes \nu\big(\theta_{ks}(A_1 \times B_1) \cap (A_2 \times B_2)\big) &= \mu \otimes \nu\big((\theta_{ks}(A_1) \cap A_2) \times (\theta_{ks}(B_1) \cap B_2)\big) \\ &= \mu\big(\theta_{ks}(A_1) \cap A_2\big)\nu\big(\theta_{ks}(B_1) \cap B_2\big)\end{aligned}$$

we can write

$$\begin{aligned}\frac{1}{n} \sum_{k=0}^{n-1} \mu \otimes \nu\big(\theta_{ks}(A_1 \times B_1) \cap (A_2 \times B_2)\big) = \\ \frac{1}{n} \sum_{k=0}^{n-1} \big[\mu\big(\theta_{ks}(A_1) \cap A_2\big) - \mu(A_1)\mu(A_2)\big]\nu\big(\theta_{ks}(B_1) \cap B_2\big) \quad (2) \\ + \frac{1}{n} \sum_{k=0}^{n-1} \mu(A_1)\mu(A_2)\nu\big(\theta_{ks}(B_1) \cap B_2\big). \quad (3)\end{aligned}$$

The absolute value of (2) is bounded by

$$\frac{1}{n} \sum_{k=0}^{n-1} \big|\mu\big(\theta_{ks}(A_1) \cap A_2\big) - \mu(A_1)\mu(A_2)\big|. \tag{4}$$

Since we assume that μ is an s-stationary s-weakly mixing probability measure the Cesàro mean in (4) converges to 0 for $n \to \infty$ according to Definition B.3. The s-stationarity and s-ergodicity of ν imply together with Remark B.2 that (3) converges to

$$\mu(A_1)\mu(A_2)\nu(B_1)\nu(B_2) = \mu \otimes \nu(A_1 \times B_1)\mu \otimes \nu(A_2 \times B_2)$$

as $n \to \infty$. This yields in combination with the previous derivations that (1) holds, which completes the proof. □

(B.16) Remark. If λ is a probability measure on the product space $\mathcal{X} \otimes \mathcal{Y}$ having a property specified in Definition B.1 or Definition B.3, then the marginal measures on $\mathcal{X}$ and $\mathcal{Y}$ have the same property. Therefore, the implications in (B.15.ii) also hold in the opposite direction.

For the weak mixing condition there exists another characterization based on product measures: The probability measure μ on $\mathcal{X}$ is s-stationary (stationary) and s-weakly mixing (totally weakly mixing) if and only if the product measure $\mu \otimes \mu$ is s-stationary (stationary) and s-ergodic (totally ergodic). See (Walters, 1982, Th. 1.24) for details. Therefore, if a probability measure is ergodic but not weakly mixing (see Example B.8), then the product of two copies of this measure is not ergodic.

C Second Order Random Processes

In this section, $\xi = \{\xi_t, t \in T\}$ is either a discrete- or continuous-time random process, i. e., $T = \mathbb{Z}$ or $T = \mathbb{R}$. The random variables ξ_t are defined on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and have values either in the measurable space $(X_t, \mathcal{X}_t) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ or in $(X_t, \mathcal{X}_t) = (\mathbb{C}, \mathcal{B}(\mathbb{C}))$, i. e., we have a real- or complex-valued process. The material is presented for the complex case. It applies to the real case without change, except for straightforward simplifications.

The subsequent material on the spectral representation of the covariance function of second order processes is taken from (Sasvári, 2013, Secs. 2.1–2.3, 2.8, 2.9).

(C.1) Second order random process. We call $\xi = \{\xi_t, t \in T\}$ a second order random process, if $\mathrm{E}(|\xi_t|^2) < \infty$ holds for all $t \in T$. A second order random process ξ is called wide-sense stationary if $\mathrm{E}(\xi_t)$ is constant for all $t \in T$ and $\mathrm{E}(\xi_t \overline{\xi_{t+s}})$ is only a function of s but not of t, where $s, t \in T$. Here $\overline{z}$ denotes the complex conjugate of z. If ξ is a second order wide-sense stationary process, then we call

$$\gamma(t) = \mathrm{cov}(\xi_0, \xi_t) = \mathrm{E}\big((\xi_0 - \mathrm{c})(\overline{\xi_t - \mathrm{c}})\big), \quad t \in T,$$

the (auto)covariance function of ξ, where $\mathrm{c} = \mathrm{E}(\xi_0)$. If ξ is real-valued, then the covariance function is real-valued as well. A second order stationary process is wide-sense stationary. If a second order Gaussian process is wide-sense stationary, then it is stationary.

A continuous-time second order wide-sense stationary process ξ is called mean-square continuous (or strongly continuous) if

$$\lim_{t \to \infty} \mathrm{E}(|\xi_t - \xi_0|^2) = 0.$$

Mean-square continuity holds if and only if the covariance function is continuous at $t = 0$. In view of Remark B.6 mean-square continuity implies continuity in the sense of Pinsker as introduced in Definition B.5.

(C.2) Spectral representation of the covariance function. A continuous complex-valued function γ on $\mathbb{R}$ is the covariance function of a continuous-time mean-square continuous second order wide-sense stationary process if and only if it can be represented in the form

$$\gamma(t) = \int_{\mathbb{R}} \mathrm{e}^{\jmath t u} \, \mathrm{d}\sigma(u), \quad t \in \mathbb{R},$$

with some finite measure σ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

A complex-valued function γ on $\mathbb{Z}$ is the covariance function of a discrete-time second order wide-sense stationary process if and only if it can be represented in the form

$$\gamma(k) = \int_{(-\pi, \pi]} \mathrm{e}^{\jmath k u} \, \mathrm{d}\sigma(u), \quad k \in \mathbb{Z},$$

with some finite measure σ on $((-\pi, \pi], \mathcal{B}((-\pi, \pi]))$.

The measure σ is called the spectral measure of the second order wide-sense stationary process. If the process is real-valued, then the spectral measure is symmetric in the sense that $\sigma(B) = \sigma(-B)$ for all $B \in \mathcal{B}(\mathbb{R})$ or $B \in \mathcal{B}((-\pi, \pi])$, respectively.

Given the spectral measure is absolutely continuous w. r. t. the Lebesgue measure λ, then the corresponding density, say φ, is called the spectral density and we have

$$\gamma(t) = \int_{\mathbb{R}} \varphi(u) \mathrm{e}^{\jmath t u} \, \mathrm{d}\lambda(u), \quad t \in \mathbb{R}, \tag{1}$$

in the continuous-time case and

$$\gamma(k) = \int_{(-\pi,\pi]} \varphi(u) \mathrm{e}^{\jmath k u} \, \mathrm{d}\lambda(u), \quad k \in \mathbb{Z}, \tag{2}$$

in the discrete time case, respectively.

From (Sasvári, 2013, Ths. 1.3.6, 1.8.7, and 1.9.6) we obtain the following sufficient conditions for the existence of a spectral density. If the covariance function γ of a continuous-time mean-square continuous second order wide-sense stationary process satisfies

$$\int_{\mathbb{R}} |\gamma(t)| \, \mathrm{d}\lambda(t) < \infty,$$

then the spectral measure is absolutely continuous and the spectral density φ is given by

$$\varphi(u) = \frac{1}{2\pi} \int_{\mathbb{R}} \gamma(t) \mathrm{e}^{-\jmath t u} \, \mathrm{d}\lambda(t), \quad u \in \mathbb{R}.$$

If the covariance γ of a discrete-time second order wide-sense stationary process satisfies

$$\sum_{k=-\infty}^{\infty} |\gamma(k)| < \infty,$$

then the spectral measure is absolutely continuous and the spectral density φ is given by

$$\varphi(u) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) \mathrm{e}^{-\jmath k u}, \quad u \in (-\pi, \pi].$$

In (Ibragimov and Linnik, 1971, Sec. 16.7) conditions are given that are necessary and sufficient for the existence of a spectral density.

The following material on rational spectral densities and ARMA processes is based on (Ihara, 1993, Secs. 2.3, 2.5). For more details see the standard references on time series and ARMA processes (Box and Jenkins, 1976, Chs. I–III), (Brockwell and Davis, 2006, Chs. 3, 4), and (Priestley, 1981a, Chs. 3, 4). In the first two references the material is restricted to the discrete-time case, whereas in the latter reference also the continuous-time case is considered. For a treatment of multivariate ARMA models see (Priestley, 1981b).

(C.3) Rational spectral densities and ARMA processes. Let us define the polynomials A and B on $\mathbb{C}$ by

$$A(z) = a_0(z - a_1)(z - a_2) \dots (z - a_n) \tag{1}$$
$$B(z) = b_0 \, (z - b_1) \, (z - b_2) \dots (z - b_m) \tag{2}$$

for all $z \in \mathbb{C}$, where $a_0 b_0 \neq 0$ and the roots $a_i, b_l \in \mathbb{C}$ are required to satisfy $a_i \neq b_l$, i. e., the polynomials have no common roots.

Suppose $\xi = \{\xi_t, t \in T\}$ is a continuous-time mean-square continuous second order wide-sense stationary process with spectral density φ. If φ is given by

$$\varphi(u) = \frac{|B(ju)|^2}{|A(ju)|^2} = \frac{|b_0|^2 \prod_{l=1}^{m} |ju - b_l|^2}{|a_0|^2 \prod_{i=1}^{n} |ju - a_i|^2}, \quad u \in \mathbb{R},$$

with the additional assumptions $\mathrm{Re}(a_i) > 0$, $\mathrm{Re}(b_l) \geq 0$, and $m < n$ on the real parts of the roots and the degrees of the polynomials, then we say the continuous-time process ξ has a rational spectral density. The process ξ is also called autoregressive moving average (ARMA) process of order (n, m). In particular, if $B \equiv 1$ so that

$$\varphi(u) = \frac{1}{|A(ju)|^2} = \frac{1}{|a_0|^2 \prod_{i=1}^{n} |ju - a_i|^2}, \quad u \in \mathbb{R},$$

then ξ is called autoregressive (AR) process of order n.

Now suppose ξ is a discrete-time second order wide-sense stationary process with spectral density φ. If φ is given by

$$\varphi(u) = \frac{|B(\mathrm{e}^{ju})|^2}{|A(\mathrm{e}^{ju})|^2} = \frac{|b_0|^2 \prod_{l=1}^{m} |\mathrm{e}^{ju} - b_l|^2}{|a_0|^2 \prod_{i=1}^{n} |\mathrm{e}^{ju} - a_i|^2}, \quad u \in (-\pi, \pi], \tag{3}$$

with the additional assumptions $|a_i| < 1$ and $|b_l| \leq 1$, then we say the discrete-time process ξ has a rational spectral density. The process ξ is also called ARMA process of order (n, m). In particular, if $B \equiv 1$ so that

$$\varphi(u) = \frac{1}{|A(\mathrm{e}^{ju})|^2} = \frac{1}{|a_0|^2 \prod_{i=1}^{n} |\mathrm{e}^{ju} - a_i|^2}, \quad u \in (-\pi, \pi], \tag{4}$$

then ξ is called AR process of order n and if $A \equiv 1$ so that

$$\varphi(u) = |B(\mathrm{e}^{ju})|^2 = |b_0|^2 \prod_{l=1}^{m} |\mathrm{e}^{ju} - b_l|^2, \quad u \in (-\pi, \pi], \tag{5}$$

then ξ is called moving average (MA) process of order m.

When the polynomials A and B have no common roots, then the introduced rational spectral densities cannot be further reduced. The additional assumptions on the roots of A and B ensure consistency with the stationarity and the second moment condition on the process ξ. Furthermore, they allow in the discrete-time case a representation given next. Please note, in the literature different ways are used to specify ARMA models based on complex polynomials. This may result in exactly opposite stability conditions on the roots of the polynomials, e. g., they have to lie outside instead of inside the unit circle etc.

A discrete-time process with a spectral density as in (3), (4), or (5) is obtained in the following way. Let us rewrite the polynomials A and B by

$$\begin{aligned} A(z) &= a_0(z - a_1)(z - a_2) \dots (z - a_n) \\ &= -a_0\big(-z^n + \alpha_1 z^{n-1} + \alpha_2 z^{n-2} + \dots + \alpha_{n-1} z + \alpha_n\big) \qquad (6) \\ B(z) &= b_0(z - b_1)(z - b_2) \dots (z - b_m) \\ &= b_0\big(z^m + \beta_1 z^{m-1} + \beta_2 z^{m-2} + \dots + \beta_{m-1} z + \beta_m\big) \qquad (7) \end{aligned}$$

with suitable coefficients α_i and β_l. Suppose $\zeta = \{\zeta_k, k \in \mathbb{Z}\}$ is a second order random sequence with $\mathrm{E}(\zeta_k) = 0$ and $\mathrm{cov}(\zeta_k, \zeta_l) = 0$ for all $k \neq l$. If $\mathrm{var}(\zeta_k) = 2\pi|b_0|^2/|a_0|^2$ and the second order wide-sense stationary sequence $\xi = \{\xi_k, k \in \mathbb{Z}\}$ is given by

$$\begin{aligned}\xi_k = \quad & \alpha_1\xi_{k-1} + \alpha_2\xi_{k-2} + \ldots + \alpha_n\xi_{k-n} \\ & + \zeta_k + \beta_1\zeta_{k-1} + \beta_2\zeta_{k-2} + \ldots + \beta_m\zeta_{k-m},\end{aligned}$$

with the coefficients α_i, β_l taken from (6) and (7), then the spectral density of ξ is given by (3). This relation motivates the name ARMA process and shows that such a process is obtained by passing white noise (sequence of uncorrelated (or independent) random variables) through a linear filter. In particular, if we have $\mathrm{var}(\zeta_k) = 2\pi/|a_0|^2$ and the random variables ξ_k are defined only by the terms in the first row and the first term in the second row, then the spectral density of ξ is given by (4), i. e., we have an AR process. Correspondingly, if we have $\mathrm{var}(\zeta_k) = 2\pi|b_0|^2$ and the random variables ξ_k are defined by the terms in the second row, then the spectral density of ξ is given by (5), i. e., we have an MA process.

(C.4) Example (AR process of order 1). As a basic example let us consider a real-valued AR process of order 1. The defining polynomial A has the form

$$A(z) = a_0(z + a_1)$$

where a_0 and a_1 are real-valued. In the continuous-time case we have the additional assumption $a_1 > 0$. The spectral density is then given by

$$\varphi(u) = \frac{1}{a_0^2|ju - a_1|^2} = \frac{1}{a_0^2(u^2 + a_1^2)}, \quad u \in \mathbb{R},$$

and using (C.2.1) we obtain the corresponding covariance function

$$\gamma(t) = \frac{\pi}{a_0^2 a_1}\mathrm{e}^{-a_1|t|}, \quad t \in \mathbb{R}.$$

In the discrete-time case we assume that $|a_1| < 1$. The spectral density is then given by

$$\varphi(u) = \frac{1}{a_0^2|\mathrm{e}^{ju} - a_1|^2} = \frac{1}{a_0^2(1 - 2a_1\cos u + a_1^2)}, \quad u \in (-\pi, \pi],$$

and using (C.2.2) we obtain the corresponding covariance function

$$\gamma(k) = \frac{2\pi}{a_0^2(1 - a_1^2)}a_1^{|k|}, \quad k \in \mathbb{Z}.$$

We see that an AR process of order 1 has an exponentially decaying covariance function. A Gaussian continuous-time AR process of order 1 is called Ornstein-Uhlenbeck Brownian motition. This process has the nice property that it is a Markov process. In the discrete time case the AR process $\xi = \{\xi_k, k \in \mathbb{Z}\}$ of order 1 is obtained by the recurrence relation

$$\xi_k = a_1\xi_{k-1} + \zeta_k,$$

where $\zeta = \{\zeta_k, k \in \mathbb{Z}\}$ is the sequence of uncorrelated random variables introduced in Paragraph C.3 with $\mathrm{var}(\zeta_k) = 2\pi/a_0^2$. If ζ is not only a sequence of uncorrelated but independent random variables, then ξ is a Markov chain, irrespective of the resulting distribution of ξ.

D Further Mathematical Background

(D.1) Sub- and superadditive sequences. If a sequence $\{a_k, k \in \mathbb{N}\}$ of real numbers satisfies the inequality

$$a_{m+n} \leq a_m + a_n$$

for all $m, n \in \mathbb{N}$, then it is called subadditive. The sequence is called superadditive, if

$$a_{m+n} \geq a_m + a_n$$

holds for all $m, n \in \mathbb{N}$. It is useful to extend the usual definition of sub- and superadditivity also to sequences, whose elements might also be infinite. The corresponding sequences are called sub- or superadditive, if the above relations hold whenever the right-hand side is not undefined.

For a subadditive sequence $\{a_k, k \in \mathbb{N}\}$ of real numbers the limit

$$\lim_{k\to\infty} \frac{a_k}{k} \tag{1}$$

always exists and is given by

$$\lim_{k\to\infty} \frac{a_k}{k} = \inf_{k\in\mathbb{N}} \frac{a_k}{k}.$$

Correspondingly, if the sequence is superadditive, then the limit in (1) exists and is given by

$$\lim_{k\to\infty} \frac{a_k}{k} = \sup_{k\in\mathbb{N}} \frac{a_k}{k}.$$

This is Fekete's lemma proved, e. g., in (Steele, 1997, p. 3).

(D.2) Kac-Murdock-Szegö matrix. The Kac-Murdock-Szegö matrix is defined by

$$K(\rho) = \left(\rho^{|i-j|}\right)_{i,j=1}^{n} = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-2} & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-3} & \rho^{n-2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \dots & 1 & \rho \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & \rho & 1 \end{pmatrix}$$

for all $n \in \mathbb{N}$, where ρ is a real constant satisfying $|\rho| < 1$. The matrix $K(\rho)$ is a symmetric Toeplitz matrix and according to (Horn and Johnson, 1985, Sec. 7.2, Problems 12, 13) has inverse

$$K^{-1}(\rho) = \frac{1}{(1-\rho^2)} \begin{pmatrix} 1 & -\rho & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -\rho & 1+\rho^2 & -\rho & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & -\rho & 1 \end{pmatrix} \tag{1}$$

and determinant

$$\det\left(K(\rho)\right) = (1-\rho^2)^{n-1}. \tag{2}$$

E Proofs

(E.1) Proof of Lemma 4.13. In the subsequent proof, we use $\mathcal{A}_{m:n}$ as short hand notation for $\bigvee_{l=m}^{n} \mathcal{A}_l$. If $m > n$ and $\mathcal{A}_{m:n}$ is the conditioning σ-algebra in a conditional mutual information, then this denotes the corresponding unconditional mutual information, for example $I(\mathcal{B}_1; \mathcal{B}_2 \,|\, \mathcal{A}_{1:0}) = I(\mathcal{B}_1; \mathcal{B}_2)$.

(i) *Proof of (4.13.i).* First assume that $I(\mathcal{A}_{1:n}; \mathcal{B}_{1:n}) < \infty$ for all $n \in \mathbb{N}$. To show that the sequence $\{n^{-1} I(\mathcal{A}_{1:n}; \mathcal{B}_{1:n}), n \in \mathbb{N}\}$ is monotonically increasing, we show that

$$\frac{1}{n+1} I(\mathcal{A}_{1:n+1}; \mathcal{B}_{1:n+1}) - \frac{1}{n} I(\mathcal{A}_{1:n}; \mathcal{B}_{1:n}) \;\geq\; 0 \tag{1}$$

holds for all $n \in \mathbb{N}$. Assume that we have the identity

$$\begin{aligned} n\, I(\mathcal{A}_{1:n+1}; \mathcal{B}_{1:n+1}) - (n+1)\, I(\mathcal{A}_{1:n}; \mathcal{B}_{1:n}) = \\ \sum_{k=1}^{n} (n-k+1)\, I(\mathcal{A}_k; \mathcal{B}_{n+1} \,|\, \mathcal{B}_{1:n} \vee \mathcal{A}_{1:k-1}) \\ + \sum_{k=2}^{n+1} (k-1)\, I(\mathcal{A}_k; \mathcal{A}_1 \vee \mathcal{B}_1 \,|\, \mathcal{B}_{2:n+1} \vee \mathcal{A}_{2:k-1}). \end{aligned} \tag{2}$$

Then, dividing (2) by $n(n+1)$ and using the nonnegativity of the conditional mutual information given in (4.7.i) yields (1). To obtain (2) we first apply the chain rules given in (4.7.iv) repeatedly to obtain

$$\begin{aligned} & n\, I(\mathcal{A}_{1:n+1}; \mathcal{B}_{1:n+1}) \qquad\qquad - (n+1)\, I(\mathcal{A}_{1:n}; \mathcal{B}_{1:n}) \\ = \; & n \sum_{k=1}^{n+1} I(\mathcal{A}_k; \mathcal{B}_{1:n+1} \,|\, \mathcal{A}_{1:k-1}) - (n+1) \sum_{k=1}^{n} I(\mathcal{A}_k; \mathcal{B}_{1:n} \,|\, \mathcal{A}_{1:k-1}). \end{aligned} \tag{3}$$

We can rewrite (3) as

$$\begin{aligned} & \sum_{k=1}^{n} (n-k+1) \Big[I(\mathcal{A}_k; \mathcal{B}_{1:n+1} \,|\, \mathcal{A}_{1:k-1}) - I(\mathcal{A}_k; \mathcal{B}_{1:n} \,|\, \mathcal{A}_{1:k-1}) \Big] \\ + \; & \sum_{k=2}^{n+1} (k-1) \Big[I(\mathcal{A}_k; \mathcal{B}_{1:n+1} \,|\, \mathcal{A}_{1:k-1}) - I(\mathcal{A}_{k-1}; \mathcal{B}_{1:n} \,|\, \mathcal{A}_{1:k-2}) \Big], \end{aligned} \tag{4}$$

which can be seen by expanding the sums in (3) in the following form:

k			
1	$+\; n\, I(\mathcal{A}_1;\mathcal{B}_{1:n+1})$	$-\; n\, I(\mathcal{A}_1;\mathcal{B}_{1:n})$	$(*)$
2	$+\; I(\mathcal{A}_2;\mathcal{B}_{1:n+1}\,\vert\,\mathcal{A}_1)$	$-\; I(\mathcal{A}_1;\mathcal{B}_{1:n})$	$(\circ)$
	$+\,(n-1)\, I(\mathcal{A}_2;\mathcal{B}_{1:n+1}\,\vert\,\mathcal{A}_1)$	$-\,(n-1)\, I(\mathcal{A}_2;\mathcal{B}_{1:n}\,\vert\,\mathcal{A}_1)$	$(*)$
3	$+\; 2\, I(\mathcal{A}_3;\mathcal{B}_{1:n+1}\,\vert\,\mathcal{A}_{1:2})$	$-\; 2\, I(\mathcal{A}_2;\mathcal{B}_{1:n}\,\vert\,\mathcal{A}_1)$	$(\circ)$
	$+\,(n-2)\, I(\mathcal{A}_3;\mathcal{B}_{1:n+1}\,\vert\,\mathcal{A}_{1:2})$	$-\,(n-2)\, I(\mathcal{A}_3;\mathcal{B}_{1:n}\,\vert\,\mathcal{A}_{1:2})$	$(*)$
$\dots$	$\dots$	$-\; 3\, I(\mathcal{A}_3;\mathcal{B}_{1:n}\,\vert\,\mathcal{A}_{1:2})$	$(\circ)$
	$\dots$	$\dots$	
$n-1$	$+\,(n-2)\, I(\mathcal{A}_{n-1};\mathcal{B}_{1:n+1}\,\vert\,\mathcal{A}_{1:n-2})$	$\dots$	$(\circ)$
	$+\; 2\, I(\mathcal{A}_{n-1};\mathcal{B}_{1:n+1}\,\vert\,\mathcal{A}_{1:n-2})$	$-\; 2\, I(\mathcal{A}_{n-1};\mathcal{B}_{1:n}\,\vert\,\mathcal{A}_{1:n-2})$	$(*)$
n	$+\,(n-1)\, I(\mathcal{A}_n\;;\mathcal{B}_{1:n+1}\,\vert\,\mathcal{A}_{1:n-1})$	$-\,(n-1)\, I(\mathcal{A}_{n-1};\mathcal{B}_{1:n}\,\vert\,\mathcal{A}_{1:n-2})$	$(\circ)$
	$+\; I(\mathcal{A}_n\;;\mathcal{B}_{1:n+1}\,\vert\,\mathcal{A}_{1:n-1})$	$-\; I(\mathcal{A}_n\;;\mathcal{B}_{1:n}\,\vert\,\mathcal{A}_{1:n-1})$	$(*)$
$n+1$	$+\; n\, I(\mathcal{A}_{n+1};\mathcal{B}_{1:n+1}\,\vert\,\mathcal{A}_{1:n})$	$-\; n\, I(\mathcal{A}_n\;;\mathcal{B}_{1:n}\,\vert\,\mathcal{A}_{1:n-1})$	$(\circ)$

Applying the chain rules to the rows marked by $(*)$ yields

$$\begin{aligned}&(n-k+1)\Big[I(\mathcal{A}_k;\mathcal{B}_{1:n+1}\,\vert\,\mathcal{A}_{1:k-1}) - I(\mathcal{A}_k;\mathcal{B}_{1:n}\,\vert\,\mathcal{A}_{1:k-1})\Big]\\ &= (n-k+1)\;\, I(\mathcal{A}_k;\mathcal{B}_{n+1}\,\vert\,\mathcal{B}_{1:n}\vee\mathcal{A}_{1:k-1})\end{aligned}\tag{5}$$

for $k = 1, 2, \dots, n$. Applying the chain rules to the rows marked by $(\circ)$ yields

$$\begin{aligned}&(k-1)\Big[I(\mathcal{A}_k;\mathcal{B}_{1:n+1}\,\vert\,\mathcal{A}_{1:k-1}) - I(\mathcal{A}_{k-1};\mathcal{B}_{1:n}\,\vert\,\mathcal{A}_{1:k-2})\Big]\\ &= (k-1)\Big[I(\mathcal{A}_k;\mathcal{A}_1\vee\mathcal{B}_{1:n+1}\,\vert\,\mathcal{A}_{2:k-1}) - I(\mathcal{A}_k;\mathcal{A}_1\,\vert\,\mathcal{A}_{2:k-1}) - I(\mathcal{A}_k;\mathcal{B}_{2:n+1}\,\vert\,\mathcal{A}_{2:k-1})\Big]\\ &= (k-1)\Big[I(\mathcal{A}_k;\mathcal{A}_1\vee\mathcal{B}_1\,\vert\,\mathcal{B}_{2:n+1}\vee\mathcal{A}_{2:k-1}) - I(\mathcal{A}_k;\mathcal{A}_1\,\vert\,\mathcal{A}_{2:k-1})\Big]\\ &= (k-1)\;\, I(\mathcal{A}_k;\mathcal{A}_1\vee\mathcal{B}_1\,\vert\,\mathcal{B}_{2:n+1}\vee\mathcal{A}_{2:k-1})\end{aligned}\tag{6}$$

for $k = 2, 3, \dots, n+1$. For the first equality we have additionally used the condition (4.13.1). For (6) we have used

$$I(\mathcal{A}_k;\mathcal{A}_1\,\vert\,\mathcal{A}_{2:k-1}) = I(\mathcal{A}_{2:k};\mathcal{A}_1) - I(\mathcal{A}_{2:k-1};\mathcal{A}_1) = 0,$$

where the first equality follows again from the chain rule and the second holds due to the first relation in (4.7.i) in combination with the basic condition in the lemma that $\mathfrak{A} = \{\mathcal{A}_k, k \in \mathbb{N}\}$ is an independent family of σ-algebras. Now combining (3)–(6) yields the identity in (2). The shown monotonicity clearly implies

$$\bar{I}(\mathfrak{A};\mathfrak{B}) = \lim_{n\to\infty}\frac{1}{n} I(\mathcal{A}_{1:n};\mathcal{B}_{1:n}) = \sup_{n\geq 1}\frac{1}{n} I(\mathcal{A}_{1:n};\mathcal{B}_{1:n}).\tag{7}$$

The above derivations do not contain any indeterminate expression since from the initial assumption of the proof we have $I(\mathcal{A}_{1:n+1};\mathcal{B}_{1:n+1}) < \infty$, which implies together with (4.7.ii) and the chain rules the finiteness of all considered information quantities. Now let us consider

the case when the assumption $I(\mathcal{A}_{1:n};\mathcal{B}_{1:n})<\infty$ is not true for all $n\in\mathbb{N}$. Then, due to (4.7.ii) there exists an $n_0\in\mathbb{N}$ such that $I(\mathcal{A}_{1:n};\mathcal{B}_{1:n})<\infty$ for all $n<n_0$ and $I(\mathcal{A}_{1:n};\mathcal{B}_{1:n})=\infty$ for all $n\geq n_0$. In this situation inequality (1) is valid for all $n<n_0$, because for all $n<n_0-1$ we have the same situation as before and for $n=n_0-1$ we have $\infty\geq 0$. If $n\geq n_0$, then we have $n^{-1}I(\mathcal{A}_{1:n};\mathcal{B}_{1:n})=\infty$. (If $n_0=1$ or $n_0=2$ the previous discussion simplifies accordingly.) Thus, we have the monotonicity of the sequence $\{n^{-1}I(\mathcal{A}_{1:n};\mathcal{B}_{1:n}),n\in\mathbb{N}\}$. Finally, (7) still holds since both sides are infinite. This completes the proof of the first part of Lemma 4.13.

(ii) *Proof of (4.13.ii).* For $m,n\in\mathbb{N}$ we have

$$\begin{aligned}I(\mathcal{A}_{1:m+n};\mathcal{B}_{1:m+n}) &= I(\mathcal{A}_{1:m}\vee\mathcal{A}_{m+1:m+n};\mathcal{B}_{1:m}\vee\mathcal{B}_{m+1:m+n})\\ &\geq I(\mathcal{A}_{1:m};\mathcal{B}_{1:m})\;+\;I(\mathcal{A}_{m+1:m+n};\mathcal{B}_{m+1:m+n})\\ &= I(\mathcal{A}_{1:m};\mathcal{B}_{1:m})\;+\;I(\mathcal{A}_{1:n};\mathcal{B}_{1:n}).\end{aligned}$$

The inequality follows from (4.7.vi) in combination with the basic condition in the lemma that $\mathfrak{A}=\{\mathcal{A}_k,k\in\mathbb{N}\}$ is an independent family of σ-algebras. The last equality follows from condition (4.13.2). Thus, the sequence $\{I(\mathcal{A}_{1:n};\mathcal{B}_{1:n}),n\in\mathbb{N}\}$ is superadditive. If $I(\mathcal{A}_{1:n};\mathcal{B}_{1:n})<\infty$ for all $n\in\mathbb{N}$, then

$$\bar{I}(\mathfrak{A};\mathfrak{B})=\lim_{n\to\infty}\frac{1}{n}I(\mathcal{A}_{1:n};\mathcal{B}_{1:n})=\sup_{n\geq 1}\frac{1}{n}I(\mathcal{A}_{1:n};\mathcal{B}_{1:n}) \tag{8}$$

follows from Fekete's lemma given in Paragraph D.1. Otherwise, both sides of (8) are infinite. (See the discussion at the end of (i).) This completes the proof of the second part of Lemma 4.13.

(E.2) Proof of Lemma 5.3.

(i) *Proof of (5.3.1).* Let us define $\gamma=\sup_{s\in T_+}C_s/s$ and let $\epsilon>0$. If $\gamma<\infty$ we put $\rho=\gamma-\epsilon$ and if $\gamma=\infty$ let $\rho>0$ be arbitrary. Then due to the definition of γ and C_s in Definition 5.1 there exists in both cases an $s_0\in T_+$ and a $\mu\in\mathcal{P}_{s_0}$ such that

$$\rho\leq\frac{1}{s_0}I\big(\xi_0^{s_0};\eta_0^{s_0}\big). \tag{1}$$

Here we use the notation with the projections ξ_t and η_t introduced in Remark 5.2, which are random variables on the channel input-output probability space $(X\times Y,\mathcal{X}\otimes\mathcal{Y},\mu\kappa)$. Let us define the random sequences $\alpha=\{\alpha_k,k\in\mathbb{Z}\}$ and $\beta=\{\beta_k,k\in\mathbb{Z}\}$ with

$$\alpha_{k+1}=\xi_{ks_0}^{(k+1)s_0}\qquad\text{and}\qquad\beta_{k+1}=\eta_{ks_0}^{(k+1)s_0}.$$

By μ_0 we denote the measure on $\mathcal{X}_0^{s_0}$ from which μ is constructed as specified in Definition 5.1.

Since μ is an s_0-i.i.d. probability measure, $\{\alpha_k,k\in\mathbb{Z}\}$ is an i.i.d.-sequence of random variables. According to (2.7.i) the channel κ is s_0-stationary because it is assumed to be stationary. Together with the s_0-stationarity of μ we obtain from Lemma 2.9 the s_0-stationarity of the channel input-output probability measure $\mu\kappa$. Therefore, the pair sequence $\{(\alpha_k,\beta_k),k\in\mathbb{Z}\}$ of projections is stationary. Applying Corollary 4.14 yields

$$\frac{1}{s_0}I(\alpha_1;\beta_1)\leq\frac{1}{ns_0}I(\alpha_0^n;\beta_0^n) \tag{2}$$

for all $n\in\mathbb{N}$.

According to (5.1.2) the probability measure μ_0 has the form

$$\mu_0 = \sum_{i=1}^{m} p_i \delta_{a_i}$$

with $a_i \in E_{s_0}$. Therefore, the product measure $\mu_0' = \bigotimes_{k=0}^{n-1} \langle \mu_0 \rangle_{ks_0}$ has the form

$$\mu_0' = \sum_{j=1}^{m^n} p_j' \delta_{a_j'}, \tag{3}$$

where $p_j' = p_{i_0} \cdot p_{i_1} \cdot \ldots \cdot p_{i_{n-1}}$ and $a_j' = (a_{i_0}, a_{i_1}, \ldots, a_{i_{n-1}})$. Assume that we partition the time index set T into segments of size $s_0' = ns_0$ and use the measure μ_0' to construct the product measure μ' on $\mathcal{X}$ as in Definition 5.1. Then we clearly have $\mu = \mu'$. From $a_j' \in \times_{k=0}^{n-1} \langle E_{s_0} \rangle_{ks_0}$ and the assumption that the family $\mathcal{E}$ of input constraints satisfies the regularity condition (3.1.4) we obtain $\mu \in \mathcal{P}_{ns_0}$. This implies

$$\frac{1}{ns_0} I(\alpha_0^n; \beta_0^n) \leq \frac{1}{ns_0} C_{ns_0} \tag{4}$$

and in combination with (1) and (2) we obtain

$$\rho \leq \limsup_{s \to \infty} \frac{1}{s} C_s \leq \gamma,$$

where the second inequality is obvious from the definition of the limit superior. This proves (5.3.1) since ρ was chosen arbitrarily if $\gamma = \infty$ and since we defined $\rho = \gamma - \epsilon$ with $\epsilon > 0$ being arbitrary if $\gamma < \infty$.

(ii) *Existence of the information rate* $\bar{I}(\mu)$. Let us fix some $\mu \in \mathcal{P}$. Then $\mu \in \mathcal{P}_{s_0}$ for some $s_0 \in T_+$. Using the notation of part (i) we have

$$\lim_{n \to \infty} \frac{1}{ns_0} I(\xi_0^{ns_0}; \eta_0^{ns_0}) = \lim_{n \to \infty} \frac{1}{ns_0} I(\alpha_0^n; \beta_0^n) = \frac{1}{s_0} \bar{I}(\alpha; \beta),$$

where the existence of the information rate $\bar{I}(\alpha; \beta)$ follows from Corollary 4.14, which can be applied due to the properties of the random sequences α and β derived in part (i). Assume that $s \in (ns_0, (n+1)s_0]$, then we obtain with the first relation in (4.7.ii)

$$I(\alpha_0^n; \beta_0^n) \leq I(\xi_0^s; \eta_0^s) \leq I(\alpha_0^{n+1}; \beta_0^{n+1})$$

and with a simple calculation

$$\frac{n}{n+1} \cdot \frac{1}{ns_0} I(\alpha_0^n; \beta_0^n) \leq \frac{1}{s} I(\xi_0^s; \eta_0^s) \leq \frac{n+1}{n} \cdot \frac{1}{(n+1)s_0} I(\alpha_0^{n+1}; \beta_0^{n+1}).$$

Since the left- and the right-hand side converge for $n \to \infty$ to $\bar{I}(\alpha; \beta)/s_0$ we obtain in view of Remark 5.4

$$\begin{aligned} \bar{I}(\mu) = \bar{I}(\xi; \eta) &= \lim_{s \to \infty} \frac{1}{s} I(\xi_0^s; \eta_0^s) \\ &= \lim_{n \to \infty} \frac{1}{ns_0} I(\alpha_0^n; \beta_0^n) = \frac{1}{s_0} \bar{I}(\alpha; \beta). \end{aligned} \tag{5}$$

Note that this identity is true regardless of $\bar{I}(\alpha;\beta)$ being finite or infinite.

(iii) *Proof of (5.3.2).* Let us define $\gamma^* = \sup_{\mu\in\mathcal{P}} \bar{I}(\mu)$ and let $\epsilon > 0$. If $\gamma^* < \infty$ we put $\rho^* = \gamma^* - \epsilon$ and if $\gamma^* = \infty$ let $\rho^* > 0$ be arbitrary. Due to the definition of γ^* there exists in both cases an $s_0 \in T_+$ and a $\mu \in \mathcal{P}_{s_0} \subset \mathcal{P}$ such that

$$\rho^* \le \bar{I}(\mu) \le C. \tag{6}$$

Using (4) and the representation $C = \sup_{s\in T_+} C_s/s$ yields

$$\frac{1}{ns_0} I(\alpha_0^n;\beta_0^n) \le \frac{1}{ns_0} C_{ns_0} \le C \tag{7}$$

and together with (5) the second inequality in (6). In combination with the definition of ρ^* this implies the inequality

$$\gamma^* \le C. \tag{8}$$

Now consider the setup at the beginning of part (i). Then we have due to (1), (2), and (5)

$$\rho \le \bar{I}(\mu) \le \gamma^*$$

and from the definition of ρ follows

$$C \le \gamma^*. \tag{9}$$

Finally, combining (8) and (9) yields (5.3.2).

(E.3) Proof of (7.3.ii). Assume that the Markov kernel K has the form of an integration channel as specified in Definition 15.1 with channel function f and noise measure space $(Z, \mathcal{Z}, \lambda)$. Then we have

$$\begin{aligned}
\psi\big(\bar{\mathrm{P}}\big\|\bar{\mathrm{Q}}\big) &= \psi\big((\mathrm{P}\otimes\lambda)_f\big\|(\mathrm{Q}\otimes\lambda)_f\big)\\
&= \sup\left|\frac{\mathrm{P}\otimes\lambda(f\in F)}{\mathrm{Q}\otimes\lambda(f\in F)} - 1\right|\\
&= \sup\left|\frac{\mathrm{P}\otimes\lambda(G)}{\mathrm{Q}\otimes\lambda(G)} - 1\right|\\
&\le \psi(\mathrm{P}\otimes\lambda\|\mathrm{Q}\otimes\lambda)\\
&= \psi(\mathrm{P}\|\mathrm{Q}).
\end{aligned}$$

The first equality follows from the derivations in Remark 15.2, in particular from (15.2.4), and the second is due to the definition of the ψ-variation in Definition 7.2 and the definition of an image measure. The supremum is taken w. r. t. all $F \in \mathcal{Y}$ with $\mathrm{Q}\otimes\lambda(f\in F) > 0$. We rewrite this form in the next row as supremum w. r. t. all $G \in \sigma(f)$ with $\mathrm{Q}\otimes\lambda(G) > 0$. The inequality then follows from the definition of the ψ-variation and the $\mathcal{X}\otimes\mathcal{Z}/\mathcal{Y}$-measurability of the channel function f, which implies $\sigma(f) \subset \mathcal{X}\otimes\mathcal{Z}$. The last equality is due to (7.3.iii) and because

$$\frac{\mathrm{P}\otimes\lambda(A\times B)}{\mathrm{Q}\otimes\lambda(A\times B)} = \frac{\mathrm{P}(A)}{\mathrm{Q}(A)}$$

for all $A \in \mathcal{X}$, $B \in \mathcal{Z}$ with $\mathrm{Q}\otimes\lambda(A\times B) > 0$.

The second part of (7.3.ii) is a special case of the first because the function g can be considered as a deterministic channel (see Paragraph 16.1), which is a special integration channel.

(E.4) Proof of (7.3.iii). Let us define

$$\gamma := \sup \left| \frac{\mathrm{P}(A_1 \times A_2)}{\mathrm{Q}(A_1 \times A_2)} - 1 \right|,$$

where the supremum is taken w. r. t. all rectangles $A_1 \times A_2 \in \mathcal{X}_1 \times \mathcal{X}_2$ with $\mathrm{Q}(A_1 \times A_2) > 0$. Further, let us define

$$\tilde{\gamma} := \sup \left| \frac{\mathrm{P}(A)}{\mathrm{Q}(A)} - 1 \right|,$$

where the supremum is taken w. r. t. all $A \in \mathcal{A}$ with $\mathrm{Q}(A) > 0$. By $\mathcal{A}$ we denote the algebra generated by all rectangles $A_1 \times A_2 \in \mathcal{X}_1 \times \mathcal{X}_2$.

Obviously, we have

$$\gamma \leq \tilde{\gamma}. \tag{1}$$

Thus $\gamma = \infty$ implies $\tilde{\gamma} = \infty$. Now assume that $\gamma < \infty$. Then for all rectangles $F \in \mathcal{X}_1 \times \mathcal{X}_2$ we have

$$|\mathrm{P}(F) - \mathrm{Q}(F)| \leq \gamma \, \mathrm{Q}(F). \tag{2}$$

Assume that $A \in \mathcal{A}$ with $\mathrm{Q}(A) > 0$. Since $\mathcal{A}$ is an algebra generated from rectangles, there exist disjoint rectangles $F_1, F_2, \ldots, F_n \in \mathcal{X}_1 \times \mathcal{X}_2$ such that

$$A = \bigcup_{i=1}^{n} F_i.$$

Then we have

$$\begin{aligned} |\mathrm{P}(A) - \mathrm{Q}(A)| &= \left| \sum_{i=1}^{n} \mathrm{P}(F_i) - \sum_{i=1}^{n} \mathrm{Q}(F_i) \right| \\ &\leq \sum_{i=1}^{n} |\mathrm{P}(F_i) - \mathrm{Q}(F_i)| \\ &\leq \sum_{i=1}^{n} \gamma \mathrm{Q}(F_i) \\ &= \gamma \, \mathrm{Q}(A), \end{aligned}$$

where the first inequality follows from the triangle inequality and the second from (2). Thus

$$\tilde{\gamma} \leq \gamma$$

and together with (1) we have $\tilde{\gamma} = \gamma$.

We continue with the observation that

$$\tilde{\gamma} \leq \psi(\mathrm{P} \| \mathrm{Q}) \tag{3}$$

since $\mathcal{A} \subset \mathcal{X}_1 \otimes \mathcal{X}_2$. Thus, $\tilde{\gamma} = \infty$ implies $\psi(\mathrm{P}\|\mathrm{Q}) = \infty$. Now assume that $\tilde{\gamma} < \infty$. Then for all $A \in \mathcal{A}$ we have

$$|\mathrm{P}(A) - \mathrm{Q}(A)| \leq \tilde{\gamma}\, \mathrm{Q}(A). \tag{4}$$

Let $\delta > 0$ and $G \in \mathcal{X}_1 \otimes \mathcal{X}_2$ with $\mathrm{Q}(G) > 0$ be arbitrary. Due to the approximation theorem for probability measures (see Theorem A.9) there exists a set $A \in \mathcal{A}$ such that

$$|\mathrm{P}(G) - \mathrm{P}(A)| \leq \mathrm{P}(G \triangle A) \leq \delta \quad \text{and} \quad |\mathrm{Q}(G) - \mathrm{Q}(A)| \leq \mathrm{Q}(G \triangle A) \leq \delta \tag{5}$$

hold simultaneously. Then we have

$$\begin{aligned} |\mathrm{P}(G) - \mathrm{Q}(G)| &\leq |\mathrm{P}(G) - \mathrm{P}(A)| + |\mathrm{P}(A) - \mathrm{Q}(A)| + |\mathrm{Q}(A) - \mathrm{Q}(G)| \\ &\leq 2\delta + \tilde{\gamma}\, \mathrm{Q}(A) \\ &\leq 2\delta + \tilde{\gamma}\, \big(\mathrm{Q}(G) + \delta\big) \end{aligned}$$

due to the triangle inequality and due to (4) and (5). This implies

$$\begin{aligned} \left| \frac{\mathrm{P}(G)}{\mathrm{Q}(G)} - 1 \right| &\leq \frac{2\delta}{\mathrm{Q}(G)} + \tilde{\gamma} \frac{\mathrm{Q}(G) + \delta}{\mathrm{Q}(G)} \\ &= \frac{2 + \tilde{\gamma}}{\mathrm{Q}(G)}\, \delta + \tilde{\gamma} \end{aligned}$$

and because $\delta > 0$ was chosen arbitrary we can conclude

$$\psi(\mathrm{P}\|\mathrm{Q}) \leq \tilde{\gamma}.$$

Together with (3) and the first part of the proof we finally obtain

$$\gamma = \tilde{\gamma} = \psi(\mathrm{P}\|\mathrm{Q}).$$

(E.5) Proof of Corollary 8.3. Since the outer P_1-measure of the set A is equal to 1, the probability space $(\Omega_1, \mathcal{F}_1, \mathrm{P}_1)$ has a unique standard extension $(\Omega_1, \tilde{\mathcal{F}}_1, \tilde{\mathrm{P}}_1)$ w. r. t. A for which $A \in \tilde{\mathcal{F}}_1$ and $\tilde{\mathrm{P}}_1(A) = 1$ holds. Please refer to Paragraph A.12 for facts on standard extensions used in this proof. The subsequent notation is identical to Lemma 8.1 if $(\Omega_1, \mathcal{F}_1, \mathrm{P}_1)$ is used. If the space $(\Omega_1, \tilde{\mathcal{F}}_1, \tilde{\mathrm{P}}_1)$ is used instead, we additionally mark all quantities depending on this space with a tilde.

For the space $(\Omega_1, \tilde{\mathcal{F}}_1, \tilde{\mathrm{P}}_1)$ the assertion of Feinstein's lemma holds according to (8.1.3) for

$$\begin{aligned} \epsilon &= m\mathrm{e}^{-\gamma} + \tilde{\mathrm{P}}(\tilde{G}_\gamma^{\mathrm{c}}) + \tilde{\mathrm{P}}_1(A^{\mathrm{c}}) \\ &= m\mathrm{e}^{-\gamma} + \tilde{\mathrm{P}}\big(\tilde{f} \leq \mathrm{e}^{\gamma}\big), \end{aligned} \tag{1}$$

where $\tilde{f}$ is the $\tilde{\mathrm{P}}_1 \otimes \mathrm{P}_2$-density of $\tilde{\mathrm{P}}_a = \tilde{\mathrm{P}}(\cdot \cap \tilde{N}^{\mathrm{c}})$ and $\tilde{N}$ is the $\tilde{\mathrm{P}}_1 \otimes \mathrm{P}_2$-nullset from the Lebesgue decomposition of $\tilde{\mathrm{P}}$. The fact that the restriction of $\tilde{\mathrm{P}}_1$ to $\mathcal{F}_1$ is equal to P_1 implies $\tilde{\mathrm{P}}_2 = \mathrm{P}_2$, which we have already used in the preceding statement. It further implies that the restrictions of $\tilde{\mathrm{P}}$ and $\tilde{\mathrm{P}}_1 \otimes \mathrm{P}_2$ to $\mathcal{F}_1 \otimes \mathcal{F}_2$ are equal to P and $\mathrm{P}_1 \otimes \mathrm{P}_2$, respectively, which is used in the derivations below.

One can show that the product space $(\Omega_1 \times \Omega_2, \tilde{\mathcal{F}}_1 \otimes \mathcal{F}_2, \tilde{\mathrm{P}}_1 \otimes \mathrm{P}_2)$ is equal to the standard extension of the product space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathrm{P}_1 \otimes \mathrm{P}_2)$ w. r. t. the set $A \times \Omega_2$, which has outer $\mathrm{P}_1 \otimes \mathrm{P}_2$-measure 1. It follows on the one hand, that the set $\tilde{N}$ can be represented by

$$\tilde{N} = \hat{N} \cap (A \times \Omega_2) \cup \check{N} \cap (A^{\mathrm{c}} \times \Omega_2)$$

for suitable sets $\hat{N}, \check{N} \in \mathcal{F}_1 \otimes \mathcal{F}_2$, where $\hat{N}$ is a $\mathrm{P}_1 \otimes \mathrm{P}_2$-nullset. On the other hand, there exists a nonnegative $\mathcal{F}_1 \otimes \mathcal{F}_2$-measurable function $\hat{f}$ on $\Omega_1 \times \Omega_2$ such that we have

$$\tilde{f} = \hat{f} \qquad \tilde{\mathrm{P}}_1 \otimes \mathrm{P}_2\text{-a.s.} \tag{2}$$

Since one can further show that $(\Omega_1 \times \Omega_2, \tilde{\mathcal{F}}_1 \otimes \mathcal{F}_2, \tilde{\mathrm{P}})$ is equal to the standard extension of the space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathrm{P})$ w. r. t. the set $A \times \Omega_2$, which has outer P-measure 1, we also have

$$\tilde{f} = \hat{f} \qquad \tilde{\mathrm{P}}\text{-a.s.} \tag{3}$$

Once we have shown that the function $\hat{f}$ and the set $\hat{N}$ correspond to the function f and the set N in the Lebesgue decomposition of P in (8.1.1), we obtain

$$\begin{aligned} \mathrm{P}(G_\gamma^{\mathrm{c}}) &= \mathrm{P}(\hat{f} \le \mathrm{e}^\gamma) \\ &= \tilde{\mathrm{P}}(\hat{f} \le \mathrm{e}^\gamma) \\ &= \tilde{\mathrm{P}}(\tilde{f} \le \mathrm{e}^\gamma), \end{aligned}$$

where the second equality holds because the restriction of $\tilde{\mathrm{P}}$ to $\mathcal{F}_1 \otimes \mathcal{F}_2$ is equal to P and the last equality holds due to (3). Then together with (1) we obtain (8.3.1) and the corollary is proved.

Indeed, from (2) and the Lebesgue decomposition of $\tilde{\mathrm{P}}$ we obtain for any $F \in \mathcal{F}_1 \otimes \mathcal{F}_2$

$$\begin{aligned} \tilde{\mathrm{P}}(F \cap \tilde{N}^{\mathrm{c}}) &= \int_F \tilde{f} \, \mathrm{d}\tilde{\mathrm{P}}_1 \otimes \mathrm{P}_2 \\ &= \int_F \hat{f} \, \mathrm{d}\mathrm{P}_1 \otimes \mathrm{P}_2, \end{aligned} \tag{4}$$

where we have also used that $\hat{f}$ is $\mathcal{F}_1 \otimes \mathcal{F}_2$-measurable and that the restriction of $\tilde{\mathrm{P}}_1 \otimes \mathrm{P}_2$ to $\mathcal{F}_1 \otimes \mathcal{F}_2$ is equal to $\mathrm{P}_1 \otimes \mathrm{P}_2$. Furthermore, we have

$$\begin{aligned} \tilde{N}^{\mathrm{c}} &= \hat{N}^{\mathrm{c}} \cap (A \times \Omega_2) \cup \check{N}^{\mathrm{c}} \cap (A^{\mathrm{c}} \times \Omega_2) \\ F \cap \tilde{N}^{\mathrm{c}} &= F \cap \hat{N}^{\mathrm{c}} \cap (A \times \Omega_2) \cup F \cap \check{N}^{\mathrm{c}} \cap (A^{\mathrm{c}} \times \Omega_2) \end{aligned}$$

and therefore

$$\tilde{\mathrm{P}}(F \cap \tilde{N}^{\mathrm{c}}) = \mathrm{P}(F \cap \hat{N}^{\mathrm{c}}), \tag{5}$$

because $\tilde{\mathrm{P}}$ is also obtained as the standard extension of P w. r. t. $A \times \Omega_2$. Combining (4) and (5) yields for any $F \in \mathcal{F}_1 \otimes \mathcal{F}_2$

$$\mathrm{P}(F) = \int_F \hat{f} \, \mathrm{d}\mathrm{P}_1 \otimes \mathrm{P}_2 + \mathrm{P}(F \cap \hat{N}),$$

which is the unique Lebesgue decomposition of P. This completes the proof.

(E.6) Proof of the measurability of f_s^t defined in Paragraph 15.6. We have

$$
\begin{aligned}
(f_s^t)^{-1}(\mathcal{Y}_s^t) &= (f_s^t)^{-1}\Big(\bigotimes\nolimits_{w\in J}\mathcal{Y}_w\Big)\\
&= (f_s^t)^{-1}\Big(\sigma\Big(\bigcup\nolimits_{w\in J}\eta_w^{-1}(\mathcal{Y}_w)\Big)\Big) && (1)\\
&= \sigma\Big((f_s^t)^{-1}\Big(\bigcup\nolimits_{w\in J}\eta_w^{-1}(\mathcal{Y}_w)\Big)\Big) && (2)\\
&= \sigma\Big(\bigcup\nolimits_{w\in J}\xi_w^{-1}\Big(f_w^{-1}(\mathcal{Y}_w)\Big)\Big) && (3)\\
&\subseteq \sigma\Big(\bigcup\nolimits_{w\in J}\xi_w^{-1}\Big(\mathcal{X}_{w-u_x}^{w+v_x}\otimes\mathcal{Z}_{w-u_z}^{w+v_z})\Big)\Big) && (4)\\
&= \mathcal{X}_{s-u_x}^{t+v_x}\otimes\mathcal{Z}_{s-u_z}^{t+v_z}. && (5)
\end{aligned}
$$

By η_w we denote the projection from Y_s^t to Y_w and by ξ_w the projection from $X_{s-u_x}^{t+v_x}\otimes Z_{s-u_z}^{t+v_z}$ to $X_{w-u_x}^{w+v_x}\otimes Z_{w-u_z}^{w+v_z}$. The equality in (1) results from the definition of a product-σ-algebra. Since we can exchange $\sigma(\cdot)$ and $(\cdot)^{-1}$ we obtain (2). The identity $\eta_w(f_s^t)=f_w(\xi_w)$ yields (3) and the $\mathcal{X}_{w-u_x}^{w+v_x}\otimes\mathcal{Z}_{w-u_z}^{w+v_z}/\mathcal{Y}_w$-measurability of f_w for all $w\in J$ yields (4). For the equality in (5) we use again the definition of a product-σ-algebra.

Note that if we even have $f_w^{-1}(\mathcal{Y}_w)=\mathcal{X}_{w-u_x}^{w+v_x}\otimes\mathcal{Z}_{w-u_z}^{w+v_z}$, then equality holds in (4).

References

Adler, R. L. (1961). Ergodic and Mixing Properties of Infinite Memory Channels. *Proceedings of the American Mathematical Society 12*(6), 924–930.

Ahlswede, R. (1968). The Weak Capacity of Averaged Channels. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 11*, 61–73.

Ahlswede, R. (2006). On Concepts of Performance Parameters for Channels. In *General Theory of Information Transfer and Combinatorics*, 639–663. Springer.

Ahlswede, R. (2014). *Storing and Transmitting Data – Rudolf Ahlswede's Lectures on Information Theory 1.* Springer.

Ash, R. B. (1963). Capacity and Error Bounds for a Time-Continuous Gaussian Channel. *Information and Control 6*(1), 14–27.

Ash, R. B. (1964). Further Discussion of a Time-Continuous Gaussian Channel. *Information and Control 7*(1), 78–83.

Ash, R. B. (1965). *Information Theory.* Interscience.

Ash, R. B. (1972). *Real Analysis and Probability.* Academic Press.

Ash, R. B. (2000). *Probability and Measure Theory* (2nd ed.). Academic Press.

Augustin, U. (1966). Gedächtnisfreie Kanäle für diskrete Zeit. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 6*, 10–61.

Baker, C. R. (1976). Absolute Continuity and Applications to Information Theory. In *Probability in Banach Spaces*, Volume 526 of *Lecture Notes in Mathematics*, 1–11. Springer.

Baker, C. R. (1978). Capacity of the Gaussian Channel without Feedback. *Information and Control 37*(1), 70–89.

Baker, C. R. (1979). Calculation of the Shannon Information. *Journal of Mathematical Analysis and Applications 69*, 115–123.

Baker, C. R. (1983). Channel Models and Their Capacity. In P. K. Sen (Ed.), *Contributions to Statistics : Essays in Honour of Norman Lloyd Johnson*, 1–16. North-Holland.

Baker, C. R. (1987). Capacity of the Mismatched Gaussian Channel. *IEEE Transactions on Information Theory 33*(6), 802–812.

Baker, C. R. (1991a). Capacity of Dimension-Limited Channels. *Journal of Multivariate Analysis 37*(2), 239–258.

Baker, C. R. (1991b). Coding Capacity for a Class of Additive Channels. *IEEE Transactions on Information Theory 37*(2), 233–243.

Baker, C. R. and S. Ihara (1991). Information Capacity of the Stationary Gaussian Channel. *IEEE Transactions on Information Theory 37*(5), 1314–1326.

Barron, A. R. (1985). The Strong Ergodic Theorem For Densities: Generalized Shannon-McMillan-Breiman Theorem. *The Annals of Probability 13*(4), 1292–1303.

Bauer, H. (1995). *Probability Theory*. Walter de Gruyter.

Bauer, H. (2001). *Measure and Integration Theory*. Walter de Gruyter.

Berger, T. (1968). Rate Distortion Theory for Sources with Abstract Alphabets and Memory. *Information and Control 13*(3), 254–273.

Bharucha, B. H. (1969). A Posteriori Distributions and Detection Theory. *Information and Control 14*(1), 98–132.

Billingsley, P. (1965). *Ergodic Theory and Information*. Wiley.

Billingsley, P. (1995). *Probability and Measure* (3rd ed.). Wiley.

Boche, H. and U. J. Mönich (2009). Complete Characterization of Stable Bandlimited Systems Under Quantization and Thresholding. *IEEE Transactions on Signal Processing 57*(12), 4699–4710.

Bogachev, V. I. (1998). *Gaussian Measures*. AMS.

Box, G. E. P. and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Bradley, R. C. (1983). Information Regularity and the Central Limit Question. *Rocky Mountain Journal of Mathematics 13*(1), 77–97.

Bradley, R. C. (2007). *Introduction to Strong Mixing Conditions*, Volume 1. Kendrick Press.

Breiman, L. (1957). The Individual Ergodic Theorem of Information Theory. *The Annals of Mathematical Statistics 28*(3), 809–811.

Breiman, L. (1960). Correction to "The Individual Ergodic Theorem of Information Theory". *The Annals of Mathematical Statistics 31*(3), 809–810.

Brockwell, P. J. and R. A. Davis (2006). *Time Series: Theory and Methods* (2nd ed.). Springer.

Chintschin, A. J., D. K. Faddejew, A. N. Kolmogoroff, A. Rényi, and J. Balatoni (1967). *Arbeiten zur Informationstheorie I* (3rd ed.). Deutscher Verlag der Wissenschaften.

Cohn, D. L. (1980). *Measure Theory*. Birkhäuser.

Cordaro, J. T. and T. J. Wagner (1970). Intersymbol Interference on a Continuous-Time Gaussian Channel. *IEEE Transactions on Information Theory 16*(4), 422–429.

Cornfeld, I. P., S. V. Fomin, and Y. G. Sinai (1982). *Ergodic Theory*. Springer.

Csiszár, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences 8*, 85–108.

Csiszár, I. (1967). Information-Type Measures of Difference of Probability Distributions and Indirect Observations. *Studia Scientiarum Mathematicarum Hungarica 2*, 299–318.

Csiszár, I. (1978). Information Measures: A Critical Survey. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (held 1974)*, Volume B, 73–86.

Csiszár, I. and J. Körner (2011). *Information Theory: Coding Theorems for Discrete Memoryless Systems* (2nd ed.). Cambridge University Press.

Csiszár, I. and P. C. Shields (2004). *Information Theory and Statistics: A Tutorial*. Now.

Ding, H. G. (1962). On the Information Stability of a Sequence of Channels. *Theory of Probability and its Applications 7*(3), 258–269.

Ding, H. G. (1964). On Shannon Theorem and its Converse for Sequences of Communication Schemes in the Case of Abstract Random Variables. In *Transactions of the Third Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (held 1962)*, 285–332.

Dobruschin, R. L. (1963). *Arbeiten zur Informationstheorie IV: Allgemeine Formulierung des Shannonschen Hauptsatzes der Informationstheorie*. Deutscher Verlag der Wissenschaften.

Dobrushin, R. L. (1959). General Formulation of Shannon's Main Theorem in Information Theory. *Uspekhi Matematicheskih Nauk 14*(6(90)), 3–104. (in Russian).

Dobrushin, R. L. (1961). Mathematical Problems in the Shannon Theory of Optimal Coding of Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 211–252. University of California Press.

Dobrushin, R. L. (1963). General Formulation of Shannon's Main Theorem in Information Theory. In *AMS Translations, Series 2*, Volume 33, 323–438. AMS.

Doob, J. L. (1937). Stochastic Processes Depending on a Continuous Parameter. *Transactions of the American Mathematical Society 42*(1), 107–140.

Doob, J. L. (1940). Regularity Properties of Certain Families of Chance Variables. *Transactions of the American Mathematical Society 47*, 455–486.

Doob, J. L. (1947). Probability in Function Space. *Bulletin of the American Mathematical Society 53*, 15–30.

Doob, J. L. (1953). *Stochastic Processes*. Wiley.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press.

Elstrodt, J. (2005). *Maß- und Integrationstheorie* (4th ed.). Springer.

Fano, R. M. (1952). Class Notes for Transmission of Information, Course 6.574. Technical report, Massachusetts Institute of Technology.

Feinstein, A. (1954). A New Basic Theorem of Information Theory. *IRE Transactions on Information Theory PGIT-4*, 2–22.

Feinstein, A. (1958). *Foundations of Information Theory.* McGraw-Hill.

Feinstein, A. (1959). On the Coding Theorem and Its Converse for Finite-Memory Channels. *Information and Control 2*(1), 25–44.

Feldman, J. (1958). Equivalence and Perpendicularity of Gaussian Processes. *Pacific Journal of Mathematics 8*(4), 699–708.

Feldman, J. (1959). Correction to: Equivalence and Perpendicularity of Gaussian Processes. *Pacific Journal of Mathematics 9*(4), 1295–1296.

Gallager, R. G. (1968). *Information Theory and Reliable Communication.* Wiley.

Gelfand, I. M., A. M. Jaglom, A. N. Kolmogoroff, C. Tse-Pei, and I. P. Zaregradski (1958). *Arbeiten zur Informationstheorie II.* Deutscher Verlag der Wissenschaften.

Gelfand, I. M., A. N. Kolmogorov, and A. M. Yaglom (1956). On the General Definition of the Quantity of Information. *Doklady Akademii Nauk SSSR 111*(4), 745–748. (in Russian).

Gelfand, I. M. and A. M. Yaglom (1957). Calculation of the Amount of Information About a Random Function Contained in Another such Function. *Uspekhi Matematicheskih Nauk 12*(1(73)), 3–52. (in Russian).

Gelfand, I. M. and A. M. Yaglom (1959). Calculation of the Amount of Information About a Random Function Contained in Another such Function. In *AMS Translations, Series 2*, Volume 12, 199–246. AMS.

Gibbs, A. L. and F. E. Su (2002). On Choosing and Bounding Probability Metrics. *International Statistical Review 70*(3), 419–435.

Gikhman, I. I. and A. V. Skorokhod (1974). *The Theory of Stochastic Processes*, Volume I. Springer.

Gilardoni, G. L. (2010). On Pinsker's and Vajda's Type Inequalities for Csiszar's f-Divergences. *IEEE Transactions on Information Theory 56*(11), 5377–5386.

Girardin, V. (2005). On the Different Extensions of the Ergodic Theorem of Information Theory. In *Recent Advances in Applied Probability*, 163–179. Springer.

Gray, R. M. (2009). *Probability, Random Processes, and Ergodic Properties* (2 ed.). Springer.

Gray, R. M. (2011). *Entropy and Information Theory* (2nd ed.). Springer.

Gray, R. M. and J. C. Kieffer (1980). Mutual Information Rate, Distortion, and Quantization in Metric Spaces. *IEEE Transactions on Information Theory 26*(4), 412–422.

Gray, R. M. and D. S. Ornstein (1979). Block Coding for Discrete Stationary $\bar{d}$-Continuous Noisy Channels. *IEEE Transactions on Information Theory 25*(3), 292–306.

Halmos, P. R. (1956). *Lectures on Ergodic Theory*. Chelsea.

Halmos, P. R. (1974). *Measure Theory*. Springer.

Han, T. S. (2003). *Information-Spectrum Methods in Information Theory*. Springer.

Hida, T. and M. Hitsuda (2007). *Gaussian Processes*. AMS.

Hájek, J. (1958). On a Property of Normal Distributions of any Stochastic Process. *Czechoslovak Mathematical Journal 8*(83), 610–618. (in Russian).

Hájek, J. (1961). On a Property of Normal Distributions of any Stochastic Process. *Selected Translations in Mathematical Statistics and Probability 1*, 245–252.

Hoeffding, W. and H. Robbins (1948). The Central Limit Theorem for Dependent Random Variables. *Duke Mathematical Journal 15*(3), 773–780.

Holsinger, J. L. (1964). Digital Communication Over Fixed Time-Continuous Channels With Memory — With Special Application To Telephon Channels. Technical Report 430, Massachusetts Institute of Technology, Research Laboratory of Electronics.

Horn, R. A. and C. R. Johnson (1985). *Matrix Analysis*. Cambridge University Press.

Huang, J. and S. P. Meyn (2005). Characterization and Computation of Optimal Distributions for Channel Coding. *IEEE Transactions on Information Theory 51*(7), 2336–2351.

Ibragimov, I. A. and Y. V. Linnik (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff.

Ibragimov, I. A. and Y. A. Rozanov (1970). On the Connection Between Two Characteristics of Dependence of Gaussian Random Vectors. *Theory of Probability and its Applications 15*(2), 295–299.

Ibragimov, I. A. and Y. A. Rozanov (1978). *Gaussian Random Processes*. Springer.

Ihara, S. (1993). *Information Theory for Continuous Systems*. World Scientific.

Ihara, S. (1994). Coding Theorems for a Continuous-Time Gaussian Channel with Feedback. *IEEE Transactions on Information Theory 40*(6), 2041–2045.

Ihara, S. (1999). Information Transmission over Continuous-Time Gaussian Channels With Feedback. *Problems of Information Transmission 35*(1), 10–24.

Itô, K. (1944). On the Ergodicity of a Certain Stationary Process. *Proceedings of the Imperial Academy 20*(2), 54–55.

Jacobs, K. (1962a). Almost Periodic Channels. In *Colloquium on Combinatorial Methods in Probability Theory*, Aarhus, Denmark.

Jacobs, K. (1962b). Über Kanäle vom Dichtetypus. *Mathematische Zeitschrift 78*(1), 151–170.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*. Academic Press.

Jelinek, F. (1968). *Probabilistic Information Theory: Discrete and Memoryless Models.* McGraw-Hill.

Kadota, T. T. (1970). Generalization of Feinstein's Fundamental Lemma. *IEEE Transactions on Information Theory 16*(6), 791–792.

Kadota, T. T. (1972). Coding Theorem and its Converse for Continuous Incrementally Stationary Channels with Finite Incremental Memory. *IEEE Transactions on Information Theory 18*(4), 510–513.

Kadota, T. T. (1973). On the Capacity of Asymptotically Memoryless Continuous-Time Channels. *IEEE Transactions on Information Theory 19*(4), 556–557.

Kadota, T. T. (1974). On the Information Stability of Stationary Ergodic Processes. *SIAM Journal on Applied Mathematics 26*(1), 176–182.

Kadota, T. T. (1978). Sequential Operation of Communication Channels and the Capacity Using Variable-Length Codes. *Journal on Applied Mathematics 35*(1), 31–47.

Kadota, T. T. and A. D. Wyner (1972). Coding Theorem for Stationary, Asymptotically Memoryless, Continuous-Time Channels. *The Annals of Mathematical Statistics 43*(5), 1603–1611.

Kakihara, Y. (1999). *Abstract Methods in Information Theory.* World Scientific.

Kallenberg, O. (2002). *Foundations of Modern Probability* (2nd ed.). Springer.

Kemperman, J. H. B. (1969). On the Optimum Rate of Transmitting Information. *The Annals of Mathematical Statistics 40*(6), 2156–2177.

Kemperman, J. H. B. (1974). On the Shannon Capacity of an Arbitrary Channel. *Indagationes Mathematicae 77*(2), 101–115.

Khinchin, A. I. (1956). On the Fundamental Theorems of Information Theory. *Uspekhi Matematicheskih Nauk 11*(1), 17–75. (in Russian).

Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory.* Dover.

Kieffer, J. C. (1981). Block Coding for Weakly Continuous Channels. *IEEE Transactions on Information Theory 27*(6), 721–727.

Kolmogorov, A. N. (1956a). On the Shannon Theory of Information Transmission in the Case of Continuous Signals. *IRE Transactions on Information Theory 2*(4), 102–108.

Kolmogorov, A. N. (1956b). Theory of Transmission of Information. *Plenary Session of the Academy of Sciences of the USSR on Scientific Problems of the Automation of Industry*. (in Russian).

Kolmogorov, A. N. (1963). Theory of Transmission of Information. In *AMS Translations, Series 2*, Volume 33, 291–321. AMS.

Kolmogorov, A. N. and Y. A. Rozanov (1960). On Strong Mixing Conditions for Stationary Gaussian Processes. *Theory of Probability and its Applications 5*(2), 204–208.

Kotz, S. (1966). Recent Results in Information Theory. *Journal of Applied Probability 3*(1), 1–93.

Kullback, S. (1968). *Information Theory and Statistics*. Dover.

Kullback, S. and R. A. Leibler (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Loève, M. (1978). *Probability Theory* (4th ed.), Volume II. Springer.

Marton, K. (2003). Measure Concentration and Strong Mixing. *Studia Scientiarum Mathematicarum Hungarica 40*(1-2), 95–113.

Maruyama, G. (1949). The Harmonic Analysis of Stationary Stochastic Processes. *Memoirs of the Faculty of Science, Kyusyu University, Series A 4*(1), 45–106.

McKeague, I. W. (1981). On the Capacity of Channels with Gaussian and Non-Gaussian Noise. *Information and Control 51*(2), 153–173.

McMillan, B. (1953). The Basic Theorems of Information Theory. *The Annals of Mathematical Statistics 24*(2), 196–219.

Mittelbach, M. (2012). Some Contributions to the Fundamentals of Continuous-Time Information Theory. Diploma Thesis. Fachrichtung Mathematik, Institut für Mathematische Stochastik, Technische Universität Dresden.

Mittelbach, M. and E. A. Jorswieck (2013). Information Regular and ψ-Mixing Channels. In *Proceedings of the IEEE International Symposium on Information Theory, ISIT*, Istanbul, Turkey.

Nakamura, Y. (1975). Ergodicity and Capacity of Information Channels with Noise Sources. *Journal of the Mathematical Society of Japan 27*(2), 213–221.

Neuhoff, D. L. and P. C. Shields (1979). Channels with Almost Finite Memory. *IEEE Transactions on Information Theory 25*(4), 440–447.

Neuhoff, D. L. and P. C. Shields (1982a). Channel Distances and Representation. *Information and Control 55*, 238–264.

Neuhoff, D. L. and P. C. Shields (1982b). Channel Entropy and Primitive Approximation. *The Annals of Probability 10*(1), 188–198.

Neuhoff, D. L. and P. C. Shields (1982c). Indecomposable Finite State Channels and Primitive Approximation. *IEEE Transactions on Information Theory 28*(1), 11–18.

Olver, F. W. J., D. W. Lozier, R. F. Boisvert, and C. W. Clark (Eds.) (2010). *NIST Handbook of Mathematical Functions*. Cambridge University Press.

Peligrad, M. and W. B. Wu (2010). Central Limit Theorem for Fourier Transforms of Stationary Processes. *The Annals of Probability 38*(5), 2009–2022.

Petersen, K. (1983). *Ergodic Theory*. Cambridge University Press.

Pfaffelhuber, E. (1971). Channels with Asymptotically Decreasing Memory and Anticipation. *IEEE Transactions on Information Theory 17*(4), 379–385.

Philipp, W. (1969). The Central Limit Problem for Mixing Sequences of Random Variables. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 12*, 155–171.

Pinsker, M. S. (1960). Information and Information Stability of Random Variables and Processes. *Problemy Peredachi Informatsii, Akad. Nauk SSSR 7*. (in Russian).

Pinsker, M. S. (1963). *Arbeiten zur Informationstheorie V: Information und Informationsstabilität zufälliger Größen und Prozesse*. Deutscher Verlag der Wissenschaften.

Pinsker, M. S. (1964). *Information and Information Stability of Random Variables and Processes*. Holden-Day.

Pinsker, M. S. (1966). Some Mathematical Questions of the Theory of Information Transmission. *Kybernetika 2*(2), 117–146. (in Russian).

Pinsker, M. S. (2007). Some Mathematical Questions of the Theory of Information Transmission. *Problems of Information Transmission 43*(4), 380–392.

Pérez, A. (1957). Notions généralisées d'incertitude, d'entropie et d'information du point de vue de la théorie de martingales. In *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (held 1956)*, 183–208. (in French).

Priestley, M. B. (1981a). *Spectral Analysis and Time Series*, Volume 1. Academic Press.

Priestley, M. B. (1981b). *Spectral Analysis and Time Series*, Volume 2. Academic Press.

Prudnikov, A. P., Y. A. Brychov, and O. I. Marichev (1986). *Integrals and Series: Elementary Functions*, Volume I. Gordon and Breach.

Pursley, M. B. (1977). Equivalence of Two Notions of Continuity for Stationary Continuous-Time Information Sources. *Journal of Multivariate Analysis 7*(2), 286–291.

Revuz, D. and M. Yor (1999). *Continuous Martingales and Brownian Motion*. Springer.

Rosenblatt, M. (1956). A Central Limit Theorem and a Strong Mixing Condition. *Proceedings of the National Academy of Sciences of the United States of America 42*, 43–47.

Rosenblatt, M. (1985). *Stationary Sequences and Random Fields*. Birkhäuser.

Rosenblatt-Roth, M. (1964). The Concept of Entropy in Probability Theory and its Application in the Theory of Information Transmission Through Communication Channels. *Theory of Probability and its Applications 9*(2), 212–235.

Rosenblatt-Roth, M. (1967). Approximations in Information Theory. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 545–564. University of California Press.

Rosinski, J. and T. Zak (1997). The Equivalence of Ergodic and Weak Mixing for Infinitely Divisible Processes. *Journal of Theoretical Probability 10*(1), 73–86.

Rudin, W. (1987). *Real and Complex Analysis* (3rd ed.). McGraw-Hill.

Samson, P.-M. (2000). Concentration of Measure Inequalities for Markov Chains and ϕ-Mixing Processes. *The Annals of Probability 28*(1), 416–461.

Sasvári, Z. (2013). *Multivariate Characteristic and Correlation Functions.* Walter de Gruyter.

Schwarte, H. (1996). Approaching Capacity of a Continuous Channel by Discrete Input Distributions. *IEEE Transactions on Information Theory 42*(2), 671–675.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal 27*, 379–423, 623–656.

Shannon, C. E. and W. Weaver (1949). *The Mathematical Theory of Communication.* University of Illinois Press.

Shiryaev, A. N. (Ed.) (1992). *Selected Works of A. N. Kolmogorov: Information Theory and the Theory of Algorithms*, Volume III. Kluwer Academic.

Shiryaev, A. N. (1995). *Probability* (2nd ed.). Springer.

Simon, M. K. (2006). *Probability Distributions Involving Gaussian Random Variables: A Handbook for Engineers and Scientists.* Springer.

Steele, J. M. (1997). *Probability Theory and Combinatorial Optimization.* SIAM.

Takano, K. (1957). On the Basic Theorems of Information Theory. *Annals of the Institute of Statistical Mathematics 9*(1), 53–77.

Takano, S. (1974). On Some Transmissionary Invariant Properties and Information Stability. *Information and Control 4*, 329–340.

Takano, S. (1977). On a Strong Law in Information Stability. *IEEE Transactions on Information Theory 23*(5), 623–625.

Thomasian, A. J. (1961). Error Bounds for Continuous Channels. In *Fourth London Symposium on Information Theory*, 46–60. Butterworth Scientific Publications.

Vajda, I. (1967). A Synchronization Method for Totally Ergodic Channels. In *Transactions of the Fourth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (held 1965)*, 611–625.

Verdú, S. and T. S. Han (1994). A General Formula for Channel Capacity. *IEEE Transactions on Information Theory 40*(4), 1147–1157.

Volkonskii, V. A. and Y. A. Rozanov (1959). Some Limit Theorems for Random Functions I. *Theory of Probability and its Applications 4*(2), 178–197.

Volkonskii, V. A. and Y. A. Rozanov (1961). Some Limit Theorems for Random Functions II. *Theory of Probability and its Applications 6*, 186–198.

Wagner, T. J. (1968). A Coding Theorem for Abstract Memoryless Channels. *Information and Control 12*(5), 489–498.

Walters, P. (1982). *An Introduction to Ergodic Theory*. Springer.

Wolfowitz, J. (1960). A Note on the Strong Converse of the Coding Theorem for the General Discrete Finite-Memory Channel. *Information and Control 3*(1), 89–93.

Wolfowitz, J. (1978). *Coding Theorems of Information Theory* (3rd ed.). Springer.

Wu, W. B. (2005). Fourier Transforms of Stationary Processes. *Proceedings of the American Mathematical Society 133*(1), 285–293.

Wyner, A. D. (1966). The Capacity of the Band-Limited Gaussian Channel. *Bell System Technical Journal 45*(3), 359–395.

Wyner, A. D. (1971). On the Intersymbol Interference Problem for the Gaussian Channel. *Bell System Technical Journal 50*(7), 2355–2363.

Wyner, A. D. (1978). A Definition of Conditional Mutual Information for Arbitrary Ensembles. *Information and Control 38*(1), 51–59.

Yoshihara, K. (1964). Simple Proofs for the Strong Converse Theorems in Some Channels. *Kodai Mathematical Seminar Reports 4*, 213–222.

Zhang, R. and T. Weissman (2005). Discrete Denoising for Channels with Memory. *Communications in Information & Systems 5*(2), 257–288.