On Stationarity in the Context of MIMO Fading Channels

Intermediate Report of NSC Project

"Capacity Analysis of Various Multiple-Antenna Multiple-Users Communication Channels with Joint Estimation and Detection"

Date:	31 May 2008
Project-Number:	NSC 96-2221-E-009-012-MY3
Project Duration:	1 August 2007 – 31 July 2010
Funded by:	National Science Council, Taiwan

Author:	Stefan M. Moser
Organization:	Information Theory Laboratory Department of Communication Engineering National Chiao Tung University
Address:	Engineering Building IV, Office 727 1001 Ta Hsueh Rd. Hsinchu 30010, Taiwan
E-mail:	stefan.moser@ieee.org

Abstract

Stationarity is investigated in the context of a general multiple-input–multiple-output (MIMO) fading channel with memory. It is shown that stationary processes behave nicely as one does expect.

Concretely, this reports concentrates on three topics with regard to stationarity: firstly it is proven that—under weak conditions on the channel—any stationary channel model will have a capacity-achieving input distribution that is stationary. This statement holds in general and is not restricted to MIMO fading channels.

Secondly, it is shown that *entropy rate* is well-defined also in the context of differential entropy and stationary processes. Moreover, it can be generalized to more complex forms of conditional entropy rates without problems as long as all involved processes are stationary. Again, this are general statements not specific to MIMO fading.

Thirdly, in the context of MIMO fading it is shown that the memory of a stationary process fades in a way that one expects.

Keywords: Channel capacity, fading number, flat fading, stationarity.

Contents

Acknowledgments **1**

1 Introduction **2**
1.1 General Background . . . 2
1.2 The Fading Number . . . 6
1.3 Stationarity . . . 8

2 Definitions and Notation **10**
2.1 Notation . . . 10
2.2 The Channel Model . . . 12

3 Main Results **15**
3.1 Capacity-Achieving Input Distributions and Stationarity . . . 15
3.2 Generalization of Entropy Rates . . . 16
3.3 Limited Memory of a Quasi-Stationary Process . . . 18

4 Derivations of Main Results **19**
4.1 Proof of Theorem 4 . . . 19
4.2 Derivations for Entropy Rates . . . 21
4.2.1 Proof of Theorem 6 . . . 21
4.2.2 Proof of Theorem 7 . . . 22
4.3 Proof of Theorem 8 . . . 24

5 Discussion & Conclusion **26**

Bibliography **27**

Acknowledgments

I would like to offer many thanks to Tobias Koch, Amos Lapidoth, Su Yu Ted, Chen Po-Ning, Li Ying, Rudolf Blazek, and Chen Guan-Yu for long and fruitful discussions.

Chapter 1

Introduction

1.1 General Background

The importance of mobile communication systems nowadays needs not to be emphasized. Worldwide millions of people rely daily on their mobile phone. While for the user a mobile phone looks very similar to an old-fashioned wired telephone, the engineering technique behind it is very much different. The reason for this is that in a wireless communication system several physical effects occur that change the behavior of the channel completely compared with wired communication:

- The signal may find many different paths from the sender to the receiver via various different reflections (buildings, trees, etc.). Therefore the receiver receives multiple copies of the same signal, however, since each path has different length and different attenuation, the various copies of the signal will arrive at different times and with different strength.

- Since the transmitter and/or the receiver might be in motion while transmitting, a physical phenomenon called *Doppler effect* occurs: the frequency of the transmitted signal is shifted depending on the relative movement between receiver and transmitter.

- Since receiver and transmitter are moving and because the environment is changing permanently (*e.g.*, movements by wind, passing cars, people, etc.), the different signal paths are constantly changing, too.

The first two effects lead to a channel that not only adds noise to the transmitted signal (as this is the case for the traditional wired communication channel), but also changes the amplitude of the signal (so called *fading*) and in extreme cases introduces inter-symbol interference. Both effects can be combatted using appropriate transmissions schemes and coding.

The fact of the time variant nature of the channel is more difficult to deal with. Nowadays, usually a wireless communication system uses training sequences that are regularly transmitted between real data in order to measure the channel state, and then this knowledge is used to detect the data. This approach has the advantage that the system design can be split into two parts: one part dealing with estimating the channel and one part doing the detection under the assumption that the channel state is perfectly known.

The big disadvantage of the separate estimation and detection is that it is rather inefficient because bandwidth is lost for the transmission of the training sequences.

Particularly, if the channel is fast changing, the estimates will quickly become poor and the amount of needed training data will be exuberantly large.

A more promising approach is to design a system that uses the received data carrying the information at the same time for estimating the channel state. Such a *joint estimation and detection* approach will be particularly important for future systems where the required data rates are considerably larger than the rates provided by present systems (like, *e.g.*, GSM).

A further advantage of such joint estimation and detection systems is that they allow fair and realistic approximations to the physically feasible data rates. To elaborate more on this point, we need to briefly review some basic facts from information theory: in his famous landmark paper "A Mathematical Theory of Communication" [1] Claude E. Shannon proved that for every communication channel there exists a maximal rate—denoted *capacity*—above which one cannot transmit information reliably, *i.e.*, the probability of making decoding errors tends to one. On the other hand for every rate below the capacity it is theoretically possible to design a system such that the error probability is as small as one wishes. Of course, depending on the aimed probability of error, the system design will be rather complex and one will encounter possibly very long delays between the start of the transmission until the signal can be decoded. Particularly the latter is a large obstacle in real systems because most communication systems cannot afford large delays. Nevertheless, the capacity shows the ultimate limit of communication rate of the available channel and is therefore fundamental for the understanding of the channel and also for the judgment of implemented systems regarding their efficiency.

As mentioned above, in the situation of wireless communication channels the channel capacity is limited due to two main sources of transmission errors. Firstly, the receiver introduces thermal noise that can be well modeled by an additive random noise process. Secondly, because the signals are electromagnetic waves transmitted through air, the received signals suffer from random fluctuations in the magnitude and phase. This effect, known as *fading*, can be described by a multiplicative random noise process.

While the additive noise can be well approximated by an independent and identically distributed (IID) complex Gaussian process for almost all channels of interest, the detailed properties of the multiplicative noise depends on many parameters, system-internal and -external, and should therefore be kept as general as possible. Unfortunately, the analysis of the channel capacity in such generality is very difficult so that commonly the model is simplified in certain aspects.

One possible simplification is to assume that the receiver *perfectly* knows the fading realizations. This assumption is based on the idea that the transmitter will firstly transmit some known training symbols from which the receiver learns the current state of the multiplicative noise process. The capacity is then computed without taking into account the estimation scheme. It is common to call this *the coherent capacity of fading channels*. Such an approach will definitely lead to an overly optimistic capacity value because

- even with a large amount of training data the channel knowledge will never be perfect, but only an estimate; and because
- the data rate that is wasted for the training symbols is completely ignored.

In this project we will not make this simplification, but stick with *noncoherent detection* where the receiver has no additional knowledge about the channel state.

Note that the receiver is free to do anything in its power to gain knowledge about the fading based on the received signals.

Marzetta and Hochwald [2] simplify the noncoherent channel model by assuming that during blocks consisting of several symbol periods the fading remains constant, while the fading coefficients corresponding to different blocks are assumed to be independent. This model is generally known as *block fading model*. Note that it is pessimistic to assume that the blocks are independent of each other because memory provides additional information about the current fading level which in general will increase capacity. However, it is more problematic to conjecture that the fading coefficients are *perfectly* constant during one block. This means that for high enough signal-to-noise ratios (SNR) and for long enough blocks the receiver can get an (almost) perfect estimate of the fading value within a block and use this knowledge to decode the received signal similarly to coherent detection. For larger SNR this seems to be overly optimistic. Indeed, as shown in [2] for single-input–single-output (SISO) Gaussian block fading and in [3] for multiple-input–multiple-output (MIMO) Gaussian block fading, the capacity of the block fading channel grows logarithmically in the SNR at high SNR, *i.e.*, the capacity has the same growth rate as the coherent capacity (and, as a matter of fact, as the capacity of an additive noise channel without fading, too).

In [4] Liang and Veeravalli generalize the SISO Gaussian block fading model by allowing some temporal correlation between the different fading coefficients within one block. They show that the rank of the block correlation matrix is crucial when determining the high-SNR channel capacity: if we have a rank-deficient correlation matrix, the effect of perfect predictability comes into play again similar to the situation of Marzetta and Hochwald [2]. This then again leads to a logarithmic growth of capacity. For a full-rank correlation matrix this is not true anymore. In this case the channel model reduces to a special case of the more general model described next.

The most general models only restrict the random noise processes to be stationary and ergodic, with additional variations in the exact fading law, the number of antennas, and the memory [5]–[13]. In [5] the authors investigate a memoryless SISO Rayleigh fading channel and derive some bounds. In [6] it is shown that the capacity-achieving input distribution for the memoryless SISO Rayleigh fading channel is discrete. In [7]–[9] the channel model is then generalized to MIMO and to general non-Gaussian fading distributions (possibly with memory) where the fading process is assumed to be *regular*, *i.e.*, its differential entropy rate is finite. The complementary situation of *nonregular* fading processes has been studied in [10]–[13].

It turns out that the capacity at high SNR is very sensitive to the exact assumptions of the channel model, in particular to the regularity assumption. If we assume a regular fading process, then the capacity grows only double-logarithmically in the SNR at high SNR [7, Theorem 4.2], [9, Theorem 6.10]. This means that at high power such a channel becomes extremely power-inefficient in the sense that whenever the capacity shall be increased by only one bit, the SNR needs to be squared or, on a dB-scale, the SNR needs to be doubled! So the high-SNR behavior is dramatically different from the optimistic models mentioned above.

For nonregular Gaussian fading the high-SNR behavior of capacity depends on the specific power spectral density and can be anything between the logarithmic and the double-logarithmic growth [11].

However, it is interesting to observe that for low SNR the difference between the different models is relatively small. Indeed, the capacity of regular fading channels

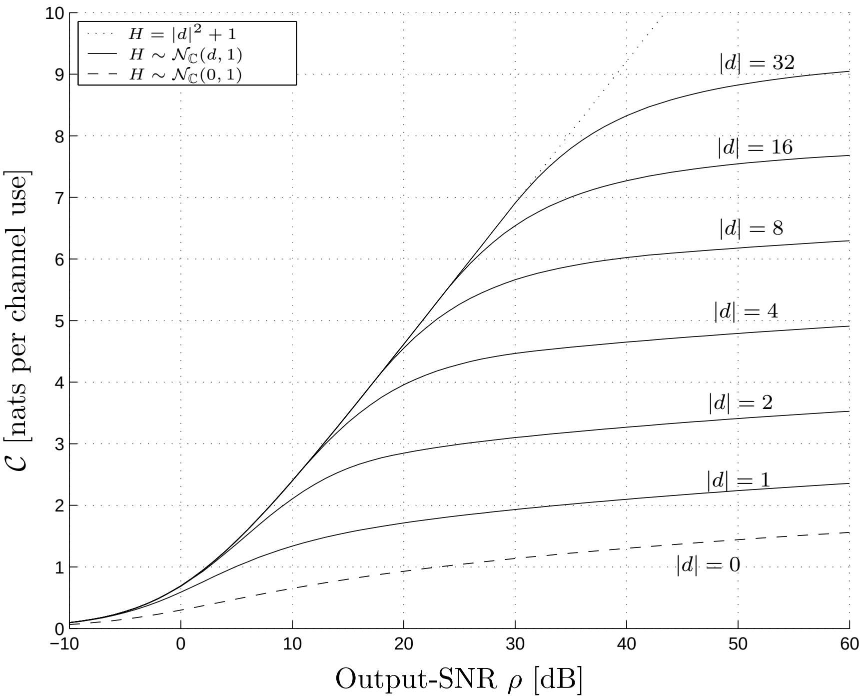

Figure 1: An upper bound on the capacity of a Rician fading channel as a function of the output-SNR $\rho = (1 + |d|^2)\text{SNR}$ for different values of the specular component d. The dashed line corresponds to the situation of a Rayleigh fading channel with a zero line-of-sight component $d = 0$. The dotted line depicts the capacity of an additive Gaussian noise channel (without fading) of equal output-SNR ρ, namely $\log(1 + \rho)$.

usually shows a very distinct turn at a certain SNR level where the growth rate changes from logarithmic to double-logarithmic. As an example Figure 1 shows the capacity of a noncoherent Rician fading channel with various values of the line-of-sight component. One clearly sees that the capacity curve, while growing logarithmically at lower SNR, suddenly has a sharp bend at a certain threshold where its growth becomes very slow. Moreover, one sees that this threshold depends strongly on the channel law, *i.e.*, on the line-of-sight component.

We conclude that at lower SNR the exact choice of the channel model has only a small impact on the capacity analysis, *i.e.*, the described simplifications (even the assumption of coherent detection) are useful in that regime. However, at high SNR many simplifications seem to lose their validity. Based on this observation we immediately ask ourselves whether we can say something about the separation between these two regimes. Particularly, in the situation of a regular fading model, we would like to know more about the threshold between the efficient low- to medium-SNR regime where the capacity grows logarithmically in the SNR and the highly inefficient high-SNR regime with a double-logarithmic growth. The dependence of this threshold on some system parameters like the number of antennas, the memory in the channel, or the availability of feedback might give valuable insight in good design criteria of wireless and mobile communication systems.

1.2 The Fading Number

In an attempt to more precisely quantify the mentioned threshold between the power-efficient and the power-inefficient regime, [7, Sec. IV.C] and [9, Sec. 6.5.2] define the *fading number* χ as the second term in the high-SNR asymptotic expansion of capacity, *i.e.*, at high SNR the channel capacity can be expressed as

$$\mathcal{C}(\text{SNR}) = \log(1 + \log(1 + \text{SNR})) + \chi + o(1). \tag{1.1}$$

Here, $o(1)$ denotes some terms that tend to zero as $\text{SNR} \uparrow \infty$.

Based on (1.1) we define the *high-SNR regime* to be the region where the $o(1)$-terms in (1.1) are negligible, *i.e.*, we say that a wireless communication system operates in the inefficient high-SNR regime if its capacity can be well approximated by

$$\mathcal{C}(\text{SNR}) \approx \log(1 + \log(1 + \text{SNR})) + \chi. \tag{1.2}$$

The important point to notice is that due to the extremely slow growth of $\log(1 + \log(1 + \text{SNR}))$ the fading number χ is usually the dominant term in the lower range of the high-SNR regime. In other words, $\log(1 + \log(1 + \text{SNR}))$ is only much larger than χ for extremely large values of SNR. An illustration of this behavior is given in Figure 2.

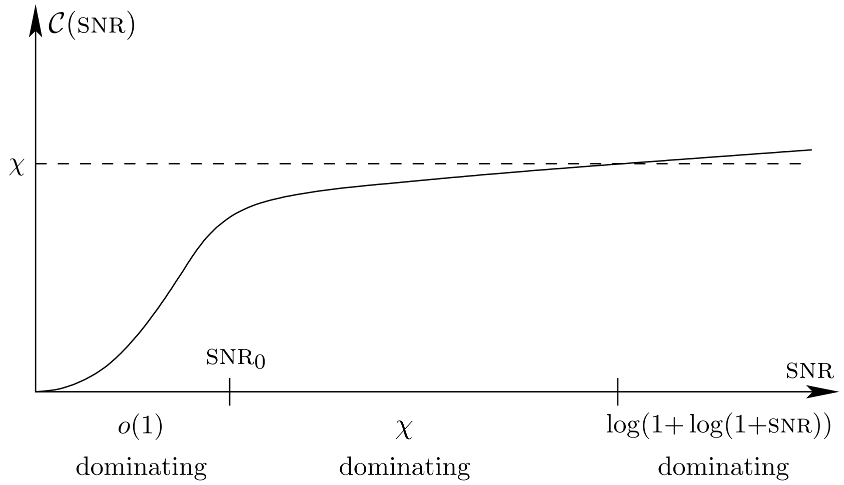

Figure 2: Illustration of the different regimes of a typical regular fading channel. At low SNR the $o(1)$ terms are dominant, in the lower range of the high-SNR regime the fading number χ is dominant, and only at very high SNR the $\log(1 + \log(1 + \text{SNR}))$ term takes the lead.

The fading number is therefore strongly connected to the point where the bend of the capacity curve occurs. As an example consider the following situation [13], [14]: assume for the moment that the threshold SNR_0 lies somewhere between 30 and 80 dB (it can be shown that this is a reasonable assumption for many channels that are encountered in practice). In this case, the threshold capacity $\mathcal{C}_0 = \mathcal{C}(\text{SNR}_0)$ must be somewhere in the following interval:

$$\log\bigl(1 + \log\bigl(1 + 30\text{ dB}\bigr)\bigr) + \chi \le \mathcal{C}_0 \le \log\bigl(1 + \log\bigl(1 + 80\text{ dB}\bigr)\bigr) + \chi \tag{1.3}$$

i.e.,

$$\chi + 2.1\text{ nats} \le \mathcal{C}_0 \le \chi + 3\text{ nats}. \tag{1.4}$$

Hence, even though we have assumed a wide range from 30 to 80 dB, the capacity changes only very little (this is because the log log-term is growing extremely slowly). Hence, we get the following *rule of thumb*.

Conjecture 1 ([13], [14])**.** *A communication system over a noncoherent regular fading channel*[1] *that operates at rates appreciably above $\chi + 2$ nats is in the high-SNR regime and therefore extremely power-inefficient.*

The fading number can therefore be regarded as quality attribute of the channel: the larger the fading number is, the higher is the maximum rate at which the channel can be used without being extremely power-inefficient. It follows from this observation that a good system design will aim at achieving a large fading number.

So far explicit expressions for the fading number are known in the special situation of general SISO fading channels with memory[2] [7, Theorem 4.41], [9, Theorem 6.41]:

$$\chi\bigl(\{H_k\}\bigr) = \log \pi + \mathsf{E}\bigl[\log |H_0|^2\bigr] - h\bigl(\{H_k\}\bigr) \tag{1.5}$$

and of general single-input–multiple-output (SIMO) fading channels with memory [8, Theorem 1], [9, Theorem 6.44]:

$$\chi\bigl(\{\mathbf{H}_k\}\bigr) = \chi_{\text{IID}}\left(\mathbf{H}_0 \,\middle|\, \mathbf{H}_{-\infty}^{-1}, \{\hat{\mathbf{H}}_\ell e^{\mathrm{i}\Theta_\ell}\}_{\ell=1}^{\infty}\right). \tag{1.6}$$

Here $\chi_{\text{IID}}\bigl(\mathbf{H} \,\big|\, \mathbf{S}\bigr)$ denotes the memoryless SIMO fading number with partial side-information $\mathbf{S}$ at the receiver [7, Note 4.31], [9, Eq. (6.194)]:

$$\chi_{\text{IID}}\bigl(\mathbf{H} \,\big|\, \mathbf{S}\bigr) = h_\lambda\bigl(\hat{\mathbf{H}} e^{\mathrm{i}\Theta} \,\big|\, \mathbf{S}\bigr) + n_{\mathrm{R}} \mathsf{E}\bigl[\log \|\mathbf{H}\|^2\bigr] - \log 2 - h\bigl(\mathbf{H} \,\big|\, \mathbf{S}\bigr). \tag{1.7}$$

The fading number of the multiple-input–single-output (MISO) fading channel has only been derived for the memoryless case [7, Theorem 4.27], [9, Theorem 6.27]:

$$\chi(\mathbf{H}^{\mathsf{T}}) = \sup_{\|\hat{\mathbf{x}}\|=1} \bigl\{\log \pi + \mathsf{E}\bigl[\log |\mathbf{H}^{\mathsf{T}}\hat{\mathbf{x}}|^2\bigr] - h(\mathbf{H}^{\mathsf{T}}\hat{\mathbf{x}})\bigr\}. \tag{1.8}$$

This fading number is achievable by inputs that can be expressed as the product of a constant unit vector in $\mathbb{C}^{n_{\mathrm{T}}}$ and a circularly symmetric, scalar, complex random variable of the same law that achieves the memoryless SISO fading number [7]. Hence, the asymptotic capacity of a memoryless MISO fading channel is achieved by beam-forming where the beam-direction is chosen not to maximize the SNR, but the fading number.

For MISO fading with memory some bounds have been found [15]–[17]:

$$\chi\bigl(\{\mathbf{H}_k^{\mathsf{T}}\}\bigr) \leq \sup_{\hat{\mathbf{x}}_{-\infty}^0} \Bigl\{ \log \pi + \mathsf{E}\bigl[\log |\mathbf{H}_0^{\mathsf{T}}\hat{\mathbf{x}}_0|^2\bigr] - h\bigl(\mathbf{H}_0^{\mathsf{T}}\hat{\mathbf{x}}_0 \,\big|\, \{\mathbf{H}_\ell^{\mathsf{T}}\hat{\mathbf{x}}_\ell\}_{\ell=-\infty}^{-1}\bigr)\Bigr\} \tag{1.9}$$

and

$$\chi\bigl(\{\mathbf{H}_k^{\mathsf{T}}\}\bigr) \geq \sup_{\hat{\mathbf{x}}} \Bigl\{ \log \pi + \mathsf{E}\bigl[\log |\mathbf{H}_0^{\mathsf{T}}\hat{\mathbf{x}}|^2\bigr] - h\bigl(\mathbf{H}_0^{\mathsf{T}}\hat{\mathbf{x}} \,\big|\, \{\mathbf{H}_\ell^{\mathsf{T}}\hat{\mathbf{x}}\}_{\ell=-\infty}^{-1}\bigr)\Bigr\}. \tag{1.10}$$

The MIMO case has been solved recently in the memoryless situation [18]:

$$\chi(\mathbb{H}) = \sup_{Q_{\hat{\mathbf{X}}}\text{circ. sym.}} \left\{ h_\lambda\left(\frac{\mathbb{H}\hat{\mathbf{X}}}{\|\mathbb{H}\hat{\mathbf{X}}\|}\right) + n_{\mathrm{R}}\mathsf{E}\Bigl[\log \|\mathbb{H}\hat{\mathbf{X}}\|^2\Bigr] - \log 2 - h\bigl(\mathbb{H}\hat{\mathbf{X}} \,\big|\, \hat{\mathbf{X}}\bigr)\right\}. \tag{1.11}$$

[1] For more details about the exact assumptions made in this report we refer to Section 2.2.

[2] For an explanation of the notation used in this report we refer to Section 2.1.

In this project we aim for the derivation of the most general case, the MIMO fading number with memory both in space and time. In this most general case we only make one main assumption: the channel model is assumed to be *stationary*, *i.e.*, all noise processes have a stationary probability distribution. This report concentrates on this aspect of the channel model and derives some useful consequences that follow from this assumption. It will turn out that in spite of the generality of our channel model that still allows for a very large family of possible fading distribution, the fact that we assume stationary processes might help us considerably in the derivations of the MIMO fading number with memory.

1.3 Stationarity

We start with a formal definition of stationarity.

Definition 2. *A discrete-time stochastic process* $\{V_k\}$ *is called* stationary *if for every positive integer* $n \in \mathbb{N}$*, for every choice of integers* $k_1, \ldots, k_n$*,* $k_i \in \mathbb{Z}$ *for* $i = 1, \ldots, n$*, and for every* $\kappa \in \mathbb{Z}$ *we have*

$$\Pr[V_{k_1+\kappa} = v_1, \ldots, V_{k_n+\kappa} = v_n] = \Pr[V_{k_1} = v_1, \ldots, V_{k_n} = v_n], \qquad \forall v_1, \ldots, v_n \tag{1.12}$$

i.e.*, the probability distribution of any subset of random variables from the process does not change when shifting over time.*

In this report we will derive some properties of stationary processes and stationary channel models. Even though these results were derived with the main intention of learning more about MIMO fading channels, most of them are very general and in no way restricted to fading channels.

More precisely, we are going to derive the following results: In Section 3.1 we will show that—under weak conditions on the channel—a stationary channel has a capacity-achieving input distribution that is stationary. This seems very intuitive, however, we are not aware of any rigorous proof in the literature. There we will also introduce the concept of *quasi-stationarity*: a finite sequence of random variables is said to be quasi-stationary if any subset of random variables from the sequence has a probability distribution that does not change when it is shifted in time (within the range of the finite sequence). Hence, a quasi-stationary block of random variables is basically a "stationary" process of finite duration.

In Section 3.2 we investigate the entropy rates of stationary processes. It is a fundamental result of (discrete) entropy that an "average" entropy, called *entropy rate*, exists for stationary processes. Similarly to the entropy that describes the uncertainty of a random variable, the entropy rate accurately describes the uncertainty in a random process.

We will generalize the definition of entropy rate for finite-alphabet processes to differential entropy rates of continuous-alphabet processes and to even more complicated forms of conditional differential entropy rates. We will show that due to the stationarity assumption all limits still exist and these definitions make sense.

Finally, in Section 3.3 we investigate one particular mutual information term that intuitively is supposed to tend to zero because from a practical point of view the memory of any process is supposed to fade away once we let the time tend to infinity. However, it turns out that this is rather difficult to prove. Again, the clue to a mathematically correct derivation is stationarity (or, actually, quasi-stationarity).

The derivations and proofs for these statements can be found in Chapter 4, and we conclude in Chapter 5.

Next, in the following chapter, we will give some more details about our notation and about the channel model used in the situation of fading channels.

Chapter 2

Definitions and Notation

2.1 Notation

As is by now fairly customary, we usually try to use upper-case letters for random quantities and lower-case letters for their realizations. This rule becomes awkward when dealing with matrices because matrices are usually written in upper case even if they are deterministic. To better differentiate between scalars, vectors, and matrices we have resorted to using different fonts for the different quantities. Upper-case letters such as X are used to denote scalar random variables taking value in the reals $\mathbb{R}$ or in the complex plane $\mathbb{C}$. Their realizations are typically written in lower-case, *e.g.*, x. Random vectors in the m-dimensional complex Euclidean space $\mathbb{C}^m$ are described by bold face capitals, *e.g.*, $\mathbf{X}$; for their realizations we use bold lower-case, *e.g.*, $\mathbf{x}$. Deterministic matrices are denoted by upper-case letters but of a special font, *e.g.*, H; and random matrices are denoted using another special upper-case font, *e.g.*, $\mathbb{H}$.

However, there will be a few exceptions to these rules. Since they are widely used in the literature, we will stick with the common customary shape of the entropy $H(\cdot)$ of a discrete random variable and of the mutual information functional $I(\cdot;\cdot)$. Moreover, we have decided to use the capital Q to denote the probability distribution of on input of a channel. In particular, Q_X and $Q_{\mathbf{X}}$ denote the probability distribution of a random variable X and random vector $\mathbf{X}$, respectively. Given an alphabet $\mathcal{A}$ we denote the set of all probability distributions over $\mathcal{A}$ by $\mathcal{P}(\mathcal{A})$.

The capacity is denoted by $\mathcal{C}$, the energy per symbol by $\mathcal{E}$, and the signal-to-noise ratio is denoted by $\textsc{snr}$.

We use the shorthand H_a^b for $(H_a, H_{a+1}, \ldots, H_b)$. For more complicated expressions, such as

$$(\mathbf{H}_a^{\mathsf{T}}\hat{\mathbf{x}}_a, \mathbf{H}_{a+1}^{\mathsf{T}}\hat{\mathbf{x}}_{a+1}, \ldots, \mathbf{H}_b^{\mathsf{T}}\hat{\mathbf{x}}_b)$$

we use the dummy variable ℓ to clarify notation: $\{\mathbf{H}_\ell^{\mathsf{T}}\hat{\mathbf{x}}_\ell\}_{\ell=a}^b$.

The subscript k is reserved to denote discrete time. Curly brackets are used to distinguish between a random process and its manifestation at time k: $\{X_k\}$ is a discrete random process over time, while X_k is the random variable of this process at time k.

Hermitian conjugation is denoted by $(\cdot)^\dagger$, and $(\cdot)^{\mathsf{T}}$ stands for the transpose (without conjugation) of a matrix or vector. We use $\|\cdot\|$ to denote the Euclidean norm of vectors or the Euclidean operator norm of matrices. That is,

$$\|\mathbf{x}\| \triangleq \sqrt{\sum_{t=1}^{m} |x^{(t)}|^2}, \quad \mathbf{x} \in \mathbb{C}^m \tag{2.1}$$

$$\|\mathsf{A}\| \triangleq \max_{\|\hat{\mathbf{w}}\|=1} \|\mathsf{A}\hat{\mathbf{w}}\|. \tag{2.2}$$

Thus, $\|\mathsf{A}\|$ is the maximal singular value of the matrix A.

The Frobenius norm of matrices is denoted by $\|\cdot\|_{\mathrm{F}}$ and is given by the square root of the sum of the squared magnitudes of the elements of the matrix, *i.e.*,

$$\|\mathsf{A}\|_{\mathrm{F}} \triangleq \sqrt{\operatorname{tr}(\mathsf{A}^\dagger \mathsf{A})} \tag{2.3}$$

where $\operatorname{tr}(\cdot)$ denotes the the trace of a matrix. Note that for every matrix A

$$\|\mathsf{A}\| \leq \|\mathsf{A}\|_{\mathrm{F}} \tag{2.4}$$

as can be verified by upper-bounding the squared magnitude of each of the components of $\mathsf{A}\hat{\mathbf{w}}$ using the Cauchy-Schwarz inequality.

We will often split a complex vector $\mathbf{v} \in \mathbb{C}^m$ up into its magnitude $\|\mathbf{v}\|$ and its *direction*

$$\hat{\mathbf{v}} \triangleq \frac{\mathbf{v}}{\|\mathbf{v}\|} \tag{2.5}$$

where we reserve this notation exclusively for unit vectors, *i.e.*, throughout this report every vector carrying a hat, $\hat{\mathbf{v}}$ or $\hat{\mathbf{V}}$, denotes a (deterministic or random, respectively) vector of unit length

$$\|\hat{\mathbf{v}}\| = \|\hat{\mathbf{V}}\| = 1. \tag{2.6}$$

To be able to work with such *direction vectors* we shall need a differential entropy-like quantity for random vectors that take value on the unit sphere in $\mathbb{C}^m$. Note that with respect to a probability distribution over $\mathbb{C}^m$, the surface of the unit sphere in $\mathbb{C}^m$ has zero measure such that the corresponding differential entropy is undefined. We therefore introduce a new probability space that only lives on the surface of the unit sphere in $\mathbb{C}^m$ and denote its measure by λ. If a random vector $\hat{\mathbf{V}}$ takes value in the unit sphere and has the density $p^{\lambda}_{\hat{\mathbf{V}}}(\hat{\mathbf{v}})$ with respect to λ, then we shall let

$$h_\lambda(\hat{\mathbf{V}}) \triangleq -\mathsf{E}\left[\log p^{\lambda}_{\hat{\mathbf{V}}}(\hat{\mathbf{V}})\right] \tag{2.7}$$

if the expectation is defined.

We note that just as ordinary differential entropy is invariant under translation, so is $h_\lambda(\hat{\mathbf{V}})$ invariant under rotation. That is, if U is a deterministic unitary matrix, then

$$h_\lambda(\mathsf{U}\hat{\mathbf{V}}) = h_\lambda(\hat{\mathbf{V}}). \tag{2.8}$$

Also note that $h_\lambda(\hat{\mathbf{V}})$ is maximized if $\hat{\mathbf{V}}$ is uniformly distributed on the unit sphere, in which case

$$h_\lambda(\hat{\mathbf{V}}) = \log c_m \tag{2.9}$$

where c_m denotes the surface area of the unit sphere in $\mathbb{C}^m$

$$c_m = \frac{2\pi^m}{\Gamma(m)}. \tag{2.10}$$

The definition (2.7) can be easily extended to conditional entropies: if $\mathbf{W}$ is some random vector, and if conditional on $\mathbf{W} = \mathbf{w}$ the random vector $\hat{\mathbf{V}}$ has density $p^{\lambda}_{\hat{\mathbf{V}}|\mathbf{W}}(\hat{\mathbf{v}}|\mathbf{w})$ then we can define

$$h_{\lambda}\big(\hat{\mathbf{V}} \,\big|\, \mathbf{W} = \mathbf{w}\big) \triangleq -\mathsf{E}\Big[\log p^{\lambda}_{\hat{\mathbf{V}}|\mathbf{W}}(\hat{\mathbf{V}}|\mathbf{W}) \,\Big|\, \mathbf{W} = \mathbf{w}\Big] \tag{2.11}$$

and we can define $h_{\lambda}\big(\hat{\mathbf{V}} \,\big|\, \mathbf{W}\big)$ as the expectation (with respect to $\mathbf{W}$) of $h_{\lambda}\big(\hat{\mathbf{V}} \,\big|\, \mathbf{W} = \mathbf{w}\big)$.

Based on these definitions we have the following lemma.

Lemma 3. *Let $\mathbf{V}$ be a complex random vector taking value in $\mathbb{C}^m$ and having differential entropy $h(\mathbf{V})$. Let $\|\mathbf{V}\|$ denote its norm and $\hat{\mathbf{V}}$ denotes its direction as in* (2.5)*. Then*

$$\begin{aligned} h(\mathbf{V}) &= h(\|\mathbf{V}\|) + h_{\lambda}\big(\hat{\mathbf{V}} \,\big|\, \|\mathbf{V}\|\big) + (2m-1)\mathsf{E}[\log \|\mathbf{V}\|] &(2.12)\\ &= h_{\lambda}\big(\hat{\mathbf{V}}\big) + h\big(\|\mathbf{V}\| \,\big|\, \hat{\mathbf{V}}\big) + (2m-1)\mathsf{E}[\log \|\mathbf{V}\|] &(2.13)\end{aligned}$$

whenever all the quantities in (2.12) *and* (2.13)*, respectively, are defined. Here $h(\|\mathbf{V}\|)$ is the differential entropy of $\|\mathbf{V}\|$ when viewed as a real (scalar) random variable.*

Proof. This follows from a change of variables. Let $\mathbf{W}$ denote the real random vector in $\mathbb{R}^{2m}$ that consists of the real and imaginary part of $\mathbf{V}$ stacked on top of each other. Then we define

$$R \triangleq \|\mathbf{W}\| \qquad \text{and} \qquad \hat{\mathbf{W}} \triangleq \frac{\mathbf{W}}{\|\mathbf{W}\|} \tag{2.14}$$

and note that the infinitesimal volume $\mathrm{d}\mathbf{w}$ in the $2m$-dimensional Euclidean space corresponds to $\mathrm{d}r \cdot r^{2m-1}\,\mathrm{d}\hat{\mathbf{w}}$ where $\mathrm{d}\hat{\mathbf{w}}$ denotes an infinitesimal area on the unit sphere in $\mathbb{R}^{2m}$. Hence, the joint probability densities can be written as

$$\begin{aligned} p_R\big(\|\mathbf{v}\|\big)p^{\lambda}_{\hat{\mathbf{V}}|R}\big(\hat{\mathbf{v}} \,\big|\, \|\mathbf{v}\|\big) &= p^{\lambda}_{\hat{\mathbf{V}}}\big(\hat{\mathbf{v}}\big)p_{R|\hat{\mathbf{V}}}\big(\|\mathbf{v}\| \,\big|\, \hat{\mathbf{v}}\big) &(2.15)\\ &= \|\mathbf{v}\|^{2m-1}p_{\mathbf{V}}(\mathbf{v}). &(2.16)\end{aligned}$$

The result now follows from $h(\mathbf{V}) = -\mathsf{E}[\log p_{\mathbf{V}}(\mathbf{V})]$. □

We shall write $\mathbf{X} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}, \mathsf{K})$ if $\mathbf{X} - \boldsymbol{\mu}$ is a circularly symmetric, zero-mean, complex Gaussian random vector of covariance matrix $\mathsf{E}\big[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\dagger}\big] = \mathsf{K}$. By $X \sim \mathcal{U}([a,b])$ we denote a random variable that is uniformly distributed on the interval $[a,b]$.

All rates specified in this report are in nats per channel use, *i.e.*, $\log(\cdot)$ denotes the natural logarithmic function. The abbreviation *RHS* stands for *right-hand side* and *LHS* stands for *left-hand side*.

2.2 The Channel Model

We consider a channel with n_{T} transmit antennas and n_{R} receive antennas whose time-k output $\mathbf{Y}_k \in \mathbb{C}^{n_{\mathrm{R}}}$ is given by

$$\mathbf{Y}_k = \mathbb{H}_k \mathbf{x}_k + \mathbf{Z}_k. \tag{2.17}$$

Here $\mathbf{x}_k \in \mathbb{C}^{n_\mathrm{T}}$ denotes the time-k channel input vector; the random matrix $\mathbb{H}_k \in \mathbb{C}^{n_\mathrm{R} \times n_\mathrm{T}}$ denotes the time-k fading matrix; and the random vector $\mathbf{Z}_k \in \mathbb{C}^{n_\mathrm{R}}$ denotes additive noise.

We assume that the fading process $\{\mathbb{H}_k\}$ and the additive noise process $\{\mathbf{Z}_k\}$ are independent and of a joint law that does not depend on the channel input $\{\mathbf{x}_k\}$.

The random vector process $\{\mathbf{Z}_k\}$ is assumed to be a spatially and temporally white, zero-mean, circularly symmetric, complex Gaussian random process, *i.e.*, $\{\mathbf{Z}_k\}$ is temporally IID $\sim \mathcal{N}_\mathbb{C}\left(\mathbf{0}, \sigma^2 \mathsf{I}_{n_\mathrm{R}}\right)$ for some $\sigma^2 > 0$. Here $\mathsf{I}_{n_\mathrm{R}}$ denotes the $n_\mathrm{R} \times n_\mathrm{R}$ identity matrix.

As for the multi-variate fading process $\{\mathbb{H}_k\}$, we shall only assume that it is stationary, ergodic, of finite second moment

$$\mathsf{E}\left[\|\mathbb{H}_k\|_\mathrm{F}^2\right] < \infty \tag{2.18}$$

and of finite differential entropy rate

$$h(\{\mathbb{H}_k\}) > -\infty \tag{2.19}$$

(the *regularity* assumption). Hence the components of $\mathbb{H}_k$ are in general correlated and depend on the past. Moreover, note that we do not necessarily assume that $\{\mathbb{H}_k\}$ is Gaussian, but allow any distribution that satisfies the above assumptions, *i.e.*, that is stationary, ergodic, regular and of finite second moment. The important special case of Gaussian fading is analyzed in more detail in a separate publication [19].

We would like to briefly comment about these assumptions. The assumption of stationarity reflects our lack of knowledge about the exact dependence of the fading law on time. Obviously we can not assume stationarity for all time as the fading law will change drastically if, *e.g.*, we move from an urban to a rural area. However, in a certain setting and for a reasonable time period, stationarity seems a natural choice. Note that the block fading model [2] is not stationary.

Ergodicity reflects our assumption that we are allowing very large blocklengths so that the channel "averages out." For systems with strong delay constraints this assumption will not be justified. Finally, by asking for a fading process that is regular we ensure that the fading process is "fully random" in the (engineering) sense that even if the past is perfectly known, the present values of the fading cannot be predicted error-free.[3] This assumption will be appropriate in certain situations and will not be in others. It seems therefore clear to us that both situations, regular and nonregular fading, should be investigated. We would like to emphasize once more that at high SNR this assumption has a dramatic effect on the capacity behavior [11].

As for the input, we consider two different constraints: a peak-power constraint or an average-power constraint. We use $\mathcal{E}$ to denote the maximal allowed instantaneous power in the former case, and to denote the allowed average power in the latter case. For both cases we set

$$\textsc{snr} \triangleq \frac{\mathcal{E}}{\sigma^2}. \tag{2.20}$$

The capacity $\mathcal{C}(\textsc{snr})$ of the channel (2.17) is given by

$$\mathcal{C}(\textsc{snr}) = \lim_{n \uparrow \infty} \frac{1}{n} \sup I\left(\mathbf{X}_1^n; \mathbf{Y}_1^n\right) \tag{2.21}$$

[3] Note that this is not a strictly mathematical explanation in general, but it is precise in the special case of a spatially independent Gaussian fading process.

where the supremum is over the set of all probability distributions on $\mathbf{X}_1^n$ satisfying the constraints, *i.e.*,

$$\|\mathbf{X}_k\|^2 \leq \mathcal{E}, \quad \text{almost surely}, \quad k = 1, 2, \ldots, n \tag{2.22}$$

for a peak-power constraint, or

$$\frac{1}{n} \sum_{k=1}^{n} \mathsf{E}\left[\|\mathbf{X}_k\|^2\right] \leq \mathcal{E} \tag{2.23}$$

for an average-power constraint.

From [7, Theorem 4.2], [9, Theorem 6.10] we have

$$\varlimsup_{\textsc{snr}\uparrow\infty} \Big\{\mathcal{C}(\textsc{snr}) - \log(1 + \log(1 + \textsc{snr}))\Big\} < \infty. \tag{2.24}$$

Note that [7, Theorem 4.2], [9, Theorem 6.10] is stated under the assumption of an average-power constraint only. However, since a peak-power constraint is more stringent than an average-power constraint, (2.24) also holds in the situation of a peak-power constraint.

The fading number χ is now defined as in [7, Definition 4.6], [9, Definition 6.13] by

$$\chi(\{\mathbb{H}_k\}) \triangleq \varlimsup_{\textsc{snr}\uparrow\infty} \Big\{\mathcal{C}(\textsc{snr}) - \log(1 + \log(1 + \textsc{snr}))\Big\}. \tag{2.25}$$

Prima facie the fading number depends on whether a peak-power constraint (2.22) or an average-power constraint (2.23) is imposed on the input. However, it will turn out that the MIMO fading number with memory is identical for both cases.

Chapter 3

Main Results

3.1 Capacity-Achieving Input Distributions and Stationarity

One of the main assumptions about our channel model is that the fading process and the additive noise are *stationary*. From an intuitive point of view it is clear that a stationary channel model should have a capacity-achieving input distribution that is also stationary. Unfortunately, we are not aware of a rigorous proof of this claim.

In [8, Lemma 5], [9, Lemma B.1] it is proven that—apart from edge effects—the optimum input distribution can be assumed to have equal marginals. Here we will extend this statement and prove that the capacity can be approached up to an $\epsilon > 0$ by a distribution that looks stationary apart from edge effects.

Theorem 4. *Consider a channel model with input $\mathbf{x}_k \in \mathbb{C}^{n_\mathrm{T}}$ and output $\mathbf{Y}_k \in \mathbb{C}^{n_\mathrm{R}}$. Assume that the channel is both* stationary *and* unaffected by zero input vectors $\mathbf{0}$ *in the following sense: for every choice of $n \in \mathbb{N}$ and $t \in \mathbb{Z}$, for some integers $\underline{n} < -|t|$ and $\overline{n} > n + |t|$, and for every distribution $Q \in \mathcal{P}(\mathbb{C}^{n_\mathrm{T}\times n})$ we have*

$$I\Big(\mathbf{0}_{\underline{n}}^{0+t}, \mathbf{X}_{1+t}^{n+t}, \mathbf{0}_{n+1+t}^{\overline{n}}\,;\, \mathbf{Y}_{\underline{n}}^{\overline{n}}\Big) = I\big(\mathbf{X}_1^n; \mathbf{Y}_1^n\big) \tag{3.1}$$

whenever both $\mathbf{X}_{1+t}^{n+t}$ on the LHS and $\mathbf{X}_1^n$ on the RHS have the same distribution Q.

Now fix some nonnegative integer κ and some power $\mathcal{E} > 0$. Then for every $\epsilon > 0$ there corresponds some positive integer $\eta = \eta(\mathcal{E}, \epsilon)$ and some distribution $Q_{\mathcal{E},\epsilon}^{\kappa+1} \in \mathcal{P}\big(\mathbb{C}^{n_\mathrm{T}\times(\kappa+1)}\big)$ such that for a blocklength n sufficiently large there exists some input $\mathbf{X}_1^n$ satisfying the following:

1. *The input $\mathbf{X}_1^n$ nearly achieves capacity in the sense that*

$$\frac{1}{n} I\big(\mathbf{X}_1^n; \mathbf{Y}_1^n\big) \geq \mathcal{C}(\mathcal{E}) - \epsilon. \tag{3.2}$$

2. *For every integer μ with $0 \leq \mu \leq \kappa$, every length-$(\mu + 1)$ block of adjacent vectors*

$$(\mathbf{X}_\ell, \dots, \mathbf{X}_{\ell+\mu}) \tag{3.3}$$

taken from within the sequence

$$\mathbf{X}_\eta, \mathbf{X}_{\eta+1}, \dots, \mathbf{X}_{n-2\eta+2} \tag{3.4}$$

has the same joint distribution $Q_{\mathcal{E},\epsilon}^{\mu+1}$, where this distribution $Q_{\mathcal{E},\epsilon}^{\mu+1}$ is given as the corresponding marginal distribution of $Q_{\mathcal{E},\epsilon}^{\kappa+1}$.

3. *In particular, all vectors in* (3.4) *have the same marginal* $Q^1_{\mathcal{E},\epsilon}$.

4. *The marginal distribution* $Q^1_{\mathcal{E},\epsilon}$ *gives rise to a second moment* $\mathcal{E}$*:*

$$\mathsf{E}\left[\|\mathbf{X}_k\|^2\right] = \mathcal{E}, \qquad k \in \{\eta, \dots, n-2\eta+2\}. \tag{3.5}$$

5. *The first* $\eta-1$ *vectors and the last* $2(\eta-1)$ *vectors satisfy the power constraint possibly strictly:*

$$\mathsf{E}\left[\|\mathbf{X}_k\|^2\right] \leq \mathcal{E}, \quad k \in \{1, \dots, \eta-1\} \cup \{n-2\eta+3, \dots, n\}. \tag{3.6}$$

Proof. The proof is based on a shift-and-mix argument using the fact that when using deterministic zeros at the input, the corresponding output yields zero information. The details are given in Section 4.1. □

Remark 5. *Neglecting the edge-effects for the moment, Theorem 4 basically says that, for every* $\mu \leq \kappa$*, every block of* $\mu+1$ adjacent *vectors has the same distribution independent of the time shift. From this it immediately follows that the distribution of every subset of (not necessarily adjacent) vectors of a* $\mu+1$ *block does not change when the vectors are shifted in time (simply marginalize those vectors out that are not member of the subset). Hence, Theorem 4* almost *proves that the capacity-achieving input distribution is stationary: the only problem are the edge effects. Note that* κ *can be chosen freely, but has to remain fixed until* n *has been loosened to infinity.* I.e.*, to get rid of the edge effects one needs to firstly let* n *tend to infinity, before one can let* κ *grow.*

In the remainder of this report we will refer to $Q^{\kappa+1}_{\mathcal{E},\epsilon}$ *and to a block of vectors* $\mathbf{X}_0^\kappa \sim Q^{\kappa+1}_{\mathcal{E},\epsilon}$ *as* quasi-stationary.

3.2 Generalization of Entropy Rates

It is generally known that (discrete) entropy of a stationary discrete-time, discrete-alphabet process $\{V_k\}$ behaves nicely in the sense that we can define the *entropy rate* as an average entropy over a block

$$H(\{V_k\}) \triangleq \lim_{n\uparrow\infty} \frac{1}{n} H(V_1, \dots, V_n), \tag{3.7}$$

i.e., that this limit always exists. Moreover, another also intuitive expression that defines an entropy of $\{V_k\}$ as uncertainty about V_0 once we know the infinite past,

$$\tilde{H}(\{V_k\}) \triangleq H(V_0|V_{-1}, V_{-2}, \dots) \tag{3.8}$$

is also always defined and, very pleasingly, actually coincides with the official definition (3.7).

We will now show that this nice behavior can be extended to differential entropy once we additionally assume that the differential entropy is finite (regularity assumption).

Theorem 6. *Let* $\{\mathbb{H}_k\}$ *be stationary, ergodic, of finite energy and regular, as given in Section 2.2. Then*

1. *the sequence* $\frac{1}{n} h(\mathbb{H}_1^n)$ *is nonincreasing in* n*;*

2. *the sequence $h(\mathbb{H}_n|\mathbb{H}_1^{n-1})$ is nonincreasing in $n\in\mathbb{N}$;*

3. *for all n we have*

$$h(\mathbb{H}_n|\mathbb{H}_1^{n-1}) \leq \frac{1}{n}h(\mathbb{H}_1^n), \qquad \forall\, n\in\mathbb{N}; \tag{3.9}$$

and

4. *the limits exist and are equal:*

$$h(\{\mathbb{H}_k\}) \triangleq \lim_{n\uparrow\infty}\frac{1}{n}h(\mathbb{H}_1^n) = \lim_{n\uparrow\infty} h(\mathbb{H}_n|\mathbb{H}_1^{n-1}). \tag{3.10}$$

Proof. See Section 4.2.1. □

This entropy rate can easily be extended to more complex forms like, *e.g.*,

$$\begin{aligned}
&h\left(\{\|\mathbb{H}_k\hat{\mathbf{X}}_k\|\}\,\middle|\,\left\{\frac{\mathbb{H}_k\hat{\mathbf{X}}_k}{\|\mathbb{H}_k\hat{\mathbf{X}}_k\|}\right\}\right)\\
&\qquad\triangleq \lim_{n\uparrow\infty}\frac{1}{n}h\left(\{\|\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\|\}_{\ell=1}^n\,\middle|\,\left\{\frac{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell}{\|\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\|}\right\}_{\ell=1}^n\right); && (3.11)\\
&h\left(\{\Psi_k^{(1)}\}\,\middle|\,\{\Psi_k^{(r)}-\Psi_k^{(1)}\}_{r=2}^{n_\mathrm{R}}, \{|\mathbf{H}_k^{(r)\mathsf{T}}\hat{\mathbf{X}}_k|\}_{r=1}^{n_\mathrm{R}}\right)\\
&\qquad\triangleq \lim_{n\uparrow\infty}\frac{1}{n}h\left(\{\Psi_\ell^{(1)}\}_{\ell=1}^n\,\middle|\,\{\Psi_\ell^{(r)}-\Psi_\ell^{(1)}\}_{\substack{r=2,\ldots,n_\mathrm{R}\\ \ell=1,\ldots,n}}, \{|\mathbf{H}_\ell^{(r)\mathsf{T}}\hat{\mathbf{X}}_\ell|\}_{\substack{r=1,\ldots,n_\mathrm{R}\\ \ell=1,\ldots,n}}\right); && (3.12)\\
&h\left(\{\mathbf{H}_k^{(1)\mathsf{T}}\hat{\mathbf{X}}_k\}\,\middle|\,\left\{\frac{\mathbf{H}_k^{(r)\mathsf{T}}\hat{\mathbf{X}}_k}{\mathbf{H}_k^{(1)\mathsf{T}}\hat{\mathbf{X}}_k}\right\}_{r=2}^{n_\mathrm{R}}, \{\hat{\mathbf{X}}_k\}\right)\\
&\qquad\triangleq \lim_{n\uparrow\infty}\frac{1}{n}h\left(\{\mathbf{H}_\ell^{(1)\mathsf{T}}\hat{\mathbf{X}}_\ell\}_{\ell=1}^n\,\middle|\,\left\{\frac{\mathbf{H}_\ell^{(r)\mathsf{T}}\hat{\mathbf{X}}_\ell}{\mathbf{H}_\ell^{(1)\mathsf{T}}\hat{\mathbf{X}}_\ell}\right\}_{\substack{r=2,\ldots,n_\mathrm{R}\\ \ell=1,\ldots,n}}, \hat{\mathbf{X}}_1^n\right). && (3.13)
\end{aligned}$$

Here $\mathbf{H}_k^{(r)}$ denotes the r-th row of $\mathbb{H}_k$, and $\Psi_k^{(r)}$ denotes the phase of $\mathbf{H}_k^{(r)\mathsf{T}}\hat{\mathbf{X}}_k$.

Also these entropy rates are all well-defined because the underlying processes are stationary. This is proven in the following theorem for only one particular case that, however, is representative for all other cases.

Theorem 7. *Let $\{\mathbb{H}_k\}$ be stationary, ergodic, of finite energy and regular, as given in Section 2.2. Let $\{\hat{\mathbf{X}}_k\}$ be a stationary unit-vector process. Then*

1. *the sequence $\frac{1}{n}h\bigl(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^n\,\big|\,\hat{\mathbf{X}}_1^n\bigr)$ is nonincreasing in n;*

2. *the sequence $h\bigl(\mathbb{H}_n\hat{\mathbf{X}}_n\,\big|\,\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n-1},\hat{\mathbf{X}}_1^n\bigr)$ is nonincreasing in n;*

3. *for all $n\in\mathbb{N}$ we have*

$$h\bigl(\mathbb{H}_n\hat{\mathbf{X}}_n\,\big|\,\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n-1},\hat{\mathbf{X}}_1^n\bigr) \leq \frac{1}{n}h\bigl(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^n\,\big|\,\hat{\mathbf{X}}_1^n\bigr), \qquad \forall\, n\in\mathbb{N}; \tag{3.14}$$

and

4. *the limits exist and are equal:*

$$\begin{aligned}
h\bigl(\{\mathbb{H}_k\hat{\mathbf{X}}_k\}\,\big|\,\{\hat{\mathbf{X}}_k\}\bigr) &\triangleq \lim_{n\uparrow\infty}\frac{1}{n}h\bigl(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^n\,\big|\,\hat{\mathbf{X}}_1^n\bigr) && (3.15)\\
&= \lim_{n\uparrow\infty} h\bigl(\mathbb{H}_n\hat{\mathbf{X}}_n\,\big|\,\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n-1},\hat{\mathbf{X}}_1^n\bigr). && (3.16)
\end{aligned}$$

Proof. See Section 4.2.2. □

3.3 Limited Memory of a Quasi-Stationary Process

In this section we want to look at the term

$$I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbf{Y}_k \,\big|\, \mathbf{X}_{k-\kappa}^k, \{\mathbb{H}_\ell \mathbf{X}_\ell\}_{\ell=k-\kappa}^{k-1}\big).$$

From an intuitive point of view this term should tend to zero as κ tends to infinity:

$$\begin{aligned}
&I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbf{Y}_k \,\big|\, \mathbf{X}_{k-\kappa}^k, \{\mathbb{H}_\ell \mathbf{X}_\ell\}_{\ell=k-\kappa}^{k-1}\big) \\
&\quad = I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbf{Y}_k, \mathbf{X}_k \,\big|\, \mathbf{X}_{k-\kappa}^{k-1}, \{\mathbb{H}_\ell \mathbf{X}_\ell\}_{\ell=k-\kappa}^{k-1}\big) \\
&\qquad - I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbf{X}_k \,\big|\, \mathbf{X}_{k-\kappa}^{k-1}, \{\mathbb{H}_\ell \mathbf{X}_\ell\}_{\ell=k-\kappa}^{k-1}\big) && (3.17)\\
&\quad \leq I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbf{Y}_k, \mathbf{X}_k \,\big|\, \mathbf{X}_{k-\kappa}^{k-1}, \{\mathbb{H}_\ell \mathbf{X}_\ell\}_{\ell=k-\kappa}^{k-1}\big) && (3.18)\\
&\quad \leq I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbf{Y}_k, \mathbf{X}_k, \mathbb{H}_k \,\big|\, \mathbf{X}_{k-\kappa}^{k-1}, \{\mathbb{H}_\ell \mathbf{X}_\ell\}_{\ell=k-\kappa}^{k-1}\big) && (3.19)\\
&\quad = I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbb{H}_k \,\big|\, \mathbf{X}_{k-\kappa}^{k-1}, \{\mathbb{H}_\ell \mathbf{X}_\ell\}_{\ell=k-\kappa}^{k-1}\big) && (3.20)\\
&\quad = I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbb{H}_k \,\big|\, \hat{\mathbf{X}}_{k-\kappa}^{k-1}, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=k-\kappa}^{k-1}, \{\|\mathbf{X}_\ell\|\}_{\ell=k-\kappa}^{k-1}\big) && (3.21)\\
&\quad = \mathsf{E}\Big[I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbb{H}_k \,\big|\, \{\hat{\mathbf{X}}_\ell = \hat{\mathbf{x}}_\ell\}_{\ell=k-\kappa}^{k-1}, \{\mathbb{H}_\ell \hat{\mathbf{x}}_\ell\}_{\ell=k-\kappa}^{k-1}\big)\Big] && (3.22)\\
&\quad \leq \sup_{\hat{\mathbf{x}}_{k-\kappa}^{k-1}} I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbb{H}_k \,\big|\, \{\mathbb{H}_\ell \hat{\mathbf{x}}_\ell\}_{\ell=k-\kappa}^{k-1}\big) && (3.23)\\
&\quad = \sup_{\hat{\mathbf{e}}_{-\kappa}^{-1}} I\big(\mathbb{H}_{-k+1}^{-\kappa-1}; \mathbb{H}_0 \,\big|\, \{\mathbb{H}_\ell \hat{\mathbf{e}}_\ell\}_{\ell=-\kappa}^{-1}\big) && (3.24)\\
&\quad \leq \sup_{\hat{\mathbf{e}}_{-\kappa}^{-1}} I\big(\mathbb{H}_{-\infty}^{-\kappa-1}; \mathbb{H}_0 \,\big|\, \{\mathbb{H}_\ell \hat{\mathbf{e}}_\ell\}_{\ell=-\kappa}^{-1}\big) && (3.25)\\
&\quad \triangleq \tilde{\delta}_1(\kappa). && (3.26)
\end{aligned}$$

Here, the first equality follows from the chain rule; (3.18) follows from the non-negativity of mutual information; in (3.19) we include additional random matrices in the mutual information; then (3.20) follows from the independence of the present input and output on the past fading when conditioned on the present fading; in (3.21) we introduce $\hat{\mathbf{X}} \triangleq \mathbf{X}/\|\mathbf{X}\|$; in (3.22) we drop $\|\mathbf{X}_\ell\|$ because conditional on $\hat{\mathbf{X}}_\ell$ they are independent of the other quantities; in (3.23) we upper bound the expectation by a supremum; (3.24) follows from stationarity; and (3.25) follows again from inclusion of additional terms into the mutual information.

Note that $\tilde{\delta}_1(\kappa)$ does neither depend on k nor on the input $\{\mathbf{X}_k\}$. Intuitively, it should tend to zero as κ tends to infinity, however, mathematically this is not clear at all because we have a supremum and because the terms in the conditioning are only n_{R}-dimensional, while the arguments of the mutual information term are $n_{\mathrm{R}} \times n_{\mathrm{T}}$ dimensional.

Luckily, we can avoid these problems when we rely on the fact that $\mathbf{X}_{k-\kappa}^k$ is quasi-stationary.

Theorem 8. *Let $\mathbf{X}_{k-\kappa}^k$ be quasi-stationary as defined in Theorem 4 and Remark 5. Then*

$$I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbf{Y}_k \,\big|\, \mathbf{X}_{k-\kappa}^k, \{\mathbb{H}_\ell \mathbf{X}_\ell\}_{\ell=k-\kappa}^{k-1}\big) \leq \delta_1(\kappa) \tag{3.27}$$

where $\delta_1(\kappa)$ does neither depend on k nor on the input $\{\mathbf{X}_k\}$ and tends to zero as κ tends to infinity.

Proof. See Section 4.3. □

Chapter 4

Derivations of Main Results

4.1 Proof of Theorem 4

The proof follows the same lines as the proofs of [8, Lemma 5] and [9, Lemma B.1]. It is based on a shift-and-mix argument.

Fix some arbitrary $\epsilon > 0$, $\mathcal{E} > 0$, and an integer $\kappa > 0$. Recalling that

$$\mathcal{C}(\mathcal{E}) = \lim_{n\uparrow\infty} \frac{1}{n} \sup I(\mathbf{X}_1, \ldots, \mathbf{X}_n; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) \tag{4.1}$$

where the supremum is over all joint distributions on $(\mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathbb{C}^{n_\mathrm{T}\times n}$ under which $\sum_{k=1}^{n} \mathsf{E}\left[\|\mathbf{X}_k\|^2\right] = n\mathcal{E}$, we conclude that there must exist some integer $\eta \geq 1$ and some joint distribution $Q^* \in \mathcal{P}(\mathbb{C}^{n_\mathrm{T}\times\eta})$ such that if $(\mathbf{X}_1, \ldots, \mathbf{X}_\eta) \sim Q^*$ then

$$\frac{1}{\eta}\sum_{k=1}^{\eta} \mathsf{E}\left[\|\mathbf{X}_k\|^2\right] = \mathcal{E} \tag{4.2}$$

and

$$\frac{1}{\eta} I(\mathbf{X}_1, \ldots, \mathbf{X}_\eta; \mathbf{Y}_1, \ldots, \mathbf{Y}_\eta) > \mathcal{C}(\mathcal{E}) - \frac{\epsilon}{2}. \tag{4.3}$$

Let $\mathbb{W}$ be a $n_\mathrm{T} \times \left(\eta \cdot \left\lceil \frac{\kappa}{\eta} + 1 \right\rceil\right)$ random matrix whose distribution consists of $\left\lceil \frac{\kappa}{\eta} + 1 \right\rceil$ independent $n_\mathrm{T} \times \eta$ blocks that are distributed according to Q^*. The distribution of $\mathbb{W}$ can then be written as the product of $\left\lceil \frac{\kappa}{\eta} + 1 \right\rceil$ distributions Q^*:

$$Q_{\mathbb{W}}\left(\mathbf{w}_1, \ldots, \mathbf{w}_{\eta\left\lceil \frac{\kappa}{\eta}+1\right\rceil}\right) = \prod_{i=0}^{\left\lceil \frac{\kappa}{\eta}+1\right\rceil-1} Q^*\left(\mathbf{w}_{i\cdot\eta+1}, \ldots, \mathbf{w}_{i\cdot\eta+\eta}\right), \qquad \mathbf{w}_\ell \in \mathbb{C}^{n_\mathrm{T}}. \tag{4.4}$$

Let us next compute the marginal distribution of $Q_{\mathbb{W}}$ for a certain block of length $\kappa+1$, $(\mathbf{W}_\ell, \ldots, \mathbf{W}_{\ell+\kappa})$. This marginal distribution depends on the particular choice of the starting point ℓ of the block, however, note that in total different choices of ℓ will result in at most η different marginal distributions. This follows from the definition of $Q_{\mathbb{W}}$ in (4.4). Let $Q_{\mathcal{E},\epsilon}^{\kappa+1}$ be the probability law on $\mathbb{C}^{n_\mathrm{T}\times(\kappa+1)}$ that is a mixture of these η different block-marginals of $Q_{\mathbb{W}}$, *i.e.*, for every Borel set $\mathcal{B} \subseteq \mathbb{C}^{n_\mathrm{T}\times(\kappa+1)}$

$$Q_{\mathcal{E},\epsilon}^{\kappa+1}(\mathcal{B}) \triangleq \frac{1}{\eta}\sum_{\ell=1}^{\eta} Q_{\mathbb{W}}\left[(\mathbf{W}_\ell, \ldots, \mathbf{W}_{\ell+\kappa}) \in \mathcal{B}\right]. \tag{4.5}$$

Note that in the situation when $\kappa < \eta$, $Q_{\mathcal{E},\epsilon}^{\kappa+1}$ can alternatively be written as

$$
\begin{aligned}
Q_{\mathcal{E},\epsilon}^{\kappa+1}(\mathcal{B}) \triangleq \frac{1}{\eta}\sum_{\ell=1}^{\eta-\kappa} & Q^*\big[(\mathbf{X}_\ell, \mathbf{X}_{\ell+1}, \dots, \mathbf{X}_{\ell+\kappa}) \in \mathcal{B}\big] \\
& + \frac{1}{\eta}\sum_{\ell=\eta-\kappa+1}^{\eta} Q^*\big[(\mathbf{X}_\ell, \dots, \mathbf{X}_\eta) \in \mathcal{B}_{1,\dots,\eta-\ell+1}\big] \\
& \qquad\qquad \cdot Q^*\big[(\mathbf{X}_1, \dots, \mathbf{X}_{\ell-1}) \in \mathcal{B}_{\eta-\ell+2,\dots,\eta}\big]
\end{aligned}
\tag{4.6}
$$

where we used $\mathcal{B}_{i,\dots,j}$ to denote the set of all corresponding $n_\mathrm{T} \times (j-i+1)$ submatrices of $\mathcal{B}$ that are created by taking only columns i to j of each matrix in $\mathcal{B}$.

Note further that from our definition it follows that $Q_{\mathcal{E},\epsilon}^{\kappa+1}$ is *quasi-stationary* in the sense that if $\mathbf{X}_0^\kappa \sim Q_{\mathcal{E},\epsilon}^{\kappa+1}$ then every length-$(\mu+1)$ subblock $\mathbf{X}_\ell^{\ell+\mu}$ has the same distribution $Q_{\mathcal{E},\epsilon}^{\mu+1}$ for all $\ell = 0, \dots, \kappa-\mu$. The distribution $Q_{\mathcal{E},\epsilon}^{\mu+1}$ can be computed from $Q_{\mathcal{E},\epsilon}^{\kappa+1}$ as marginal distribution, $0 \le \mu \le \kappa$.

In the theorem we have assumed that n is given and sufficiently large. In particular we will assume that $n \gg \max\{\eta, \kappa\}$. We shall next describe the required input distribution as follows: let

$$
\nu \triangleq \left\lfloor \frac{n-\eta+1}{\eta} \right\rfloor \tag{4.7}
$$

and let the length-$(n+\eta-1)$ sequence $\tilde{\mathbb{X}}$ of random n_T-vectors be defined by

$$
\tilde{\mathbb{X}} \triangleq \Big(\underbrace{\mathbf{0}, \dots, \mathbf{0}}_{\eta-1}, \underbrace{\mathbf{\Xi}_1^{(1)}, \dots, \mathbf{\Xi}_\eta^{(1)}}_{\eta}, \dots, \dots, \underbrace{\mathbf{\Xi}_1^{(\nu)}, \dots, \mathbf{\Xi}_\eta^{(\nu)}}_{\eta}, \mathbf{0}, \dots, \mathbf{0}\Big) \tag{4.8}
$$

so that

$$
\tilde{\mathbf{X}}_k = \begin{cases} \mathbf{0}, & \text{if } 1 \le k \le \eta-1 \\ \mathbf{\Xi}_{(k \bmod \eta)+1}^{\lfloor k/\eta \rfloor}, & \text{if } \eta \le k \le (\nu+1)\eta - 1 \\ \mathbf{0}, & \text{if } (\nu+1)\eta \le k \le n+\eta-1 \end{cases} \tag{4.9}
$$

where $\mathbf{0}$ is the zero n_T-vector and where

$$
\Big\{\big(\mathbf{\Xi}_1^{(j)}, \dots, \mathbf{\Xi}_\eta^{(j)}\big)\Big\}_{j=1}^{\nu} \quad \text{are IID} \sim Q^*. \tag{4.10}
$$

If we choose $\{\tilde{\mathbf{X}}_k\}$ as input for our channel, then it follows from the fact that zeros have no effect and from (4.3) that

$$
\frac{1}{n} I\big(\tilde{\mathbf{X}}_1^n; \mathbf{Y}_1^n\big) > \frac{1}{n} \cdot \nu\eta \left(\mathcal{C}(\mathcal{E}) - \frac{\epsilon}{2}\right). \tag{4.11}
$$

Again, since the lead-in and trailing zeros are of no consequence and since shifting does not change mutual information, this same mutual information results if we shift $\tilde{\mathbf{X}}_k$ by t (provided that $0 \le t \le \eta-1$ and n is large enough so that we do not lose any nonzero input vector):

$$
\frac{1}{n} I\big(\tilde{\mathbf{X}}_{1+t}^{n+t}; \mathbf{Y}_1^n\big) > \frac{1}{n} \cdot \nu\eta \left(\mathcal{C}(\mathcal{E}) - \frac{\epsilon}{2}\right). \tag{4.12}
$$

Consequently, if we define $\mathbf{X}_1, \dots, \mathbf{X}_n$ by the mixture of the time shift of $\tilde{\mathbb{X}}$, *i.e.*,

$$
\mathbf{X}_k \triangleq \tilde{\mathbf{X}}_{k+T}, \qquad 1 \le k \le n \tag{4.13}
$$

where

$$T \sim \mathcal{U}(\{0, \ldots, \eta - 1\}) \tag{4.14}$$

is independent of $\tilde{\mathbb{X}}$, then by the concavity of mutual information in the input distribution we obtain that

$$\begin{aligned} \frac{1}{n} I\left(\mathbf{X}_1^n; \mathbf{Y}_1^n\right) &> \frac{\eta\nu}{n}\left(\mathcal{C}(\mathcal{E}) - \frac{\epsilon}{2}\right) && (4.15) \\ &= \frac{\eta \left\lfloor \frac{n-\eta+1}{\eta} \right\rfloor}{n}\left(\mathcal{C}(\mathcal{E}) - \frac{\epsilon}{2}\right) && (4.16) \end{aligned}$$

which exceeds $\mathcal{C}(\mathcal{E}) - \epsilon$ for sufficiently large n.

Except at the edges, the above mixture guarantees that every block of $\mu + 1$ vectors has the same distribution

$$Q_{\mathcal{E},\epsilon}^{\mu+1}(\mathcal{B}) \triangleq \frac{1}{\eta} \sum_{\ell=1}^{\eta} Q_{\mathbb{W}}\big[(\mathbf{W}_\ell, \ldots, \mathbf{W}_{\ell+\mu}) \in \mathcal{B}\big] \tag{4.17}$$

for every μ, $0 \le \mu \le \kappa$ and every Borel set $\mathcal{B} \subseteq \mathbb{C}^{n_\mathrm{T} \times (\mu+1)}$, *i.e.*, $\mathbf{X}_1^n$ is (apart from the edges) quasi-stationary.

Note further that by (4.2) we have for $\mu = 0$

$$\int_{\mathbb{C}^{n_\mathrm{T}}} \|\mathbf{x}\|^2 \, \mathrm{d}Q_{\mathcal{E},\epsilon}^1(\mathbf{x}) = \mathcal{E}. \tag{4.18}$$

The power in the edges can be smaller than $\mathcal{E}$ because of the mixture with deterministic zero vectors.

4.2 Derivations for Entropy Rates

4.2.1 Proof of Theorem 6

We start with the proof of the second statement which follows directly from the fact that conditioning cannot increase differential entropy:

$$\begin{aligned} h(\mathbb{H}_{n+1} | \mathbb{H}_1^n) &\le h(\mathbb{H}_{n+1} | \mathbb{H}_2^n) && (4.19) \\ &= h\big(\mathbb{H}_n \,\big|\, \mathbb{H}_1^{n-1}\big) && (4.20) \end{aligned}$$

where the last equality follows from stationarity.

We next use the second statement to prove the third:

$$\begin{aligned} h(\mathbb{H}_1^n) &= \sum_{k=1}^{n} h\big(\mathbb{H}_k \,\big|\, \mathbb{H}_1^{k-1}\big) && (4.21) \\ &\ge \sum_{k=1}^{n} h\big(\mathbb{H}_n \,\big|\, \mathbb{H}_1^{n-1}\big) && (4.22) \\ &= n h\big(\mathbb{H}_n \,\big|\, \mathbb{H}_1^{n-1}\big) && (4.23) \end{aligned}$$

where (4.21) follows from the chain rule; and (4.22) follows from the second statement.

Next, we prove the first statement:

$$
\begin{aligned}
h\big(\mathbb{H}_1^{n+1}\big) &= h(\mathbb{H}_1^n) + h\big(\mathbb{H}_{n+1} \,\big|\, \mathbb{H}_1^n\big) && (4.24)\\
&\le h(\mathbb{H}_1^n) + h\big(\mathbb{H}_n \,\big|\, \mathbb{H}_1^{n-1}\big) && (4.25)\\
&\le h(\mathbb{H}_1^n) + \frac{1}{n} h(\mathbb{H}_1^n) && (4.26)\\
&= \frac{n+1}{n} h(\mathbb{H}_1^n) && (4.27)
\end{aligned}
$$

where (4.24) follows from the chain rule; where (4.25) follows from the second statement; and where (4.26) follows from the third statement.

Finally, to prove the fourth statement we note that for an arbitrary $m \in \mathbb{N}$

$$
\begin{aligned}
h\big(\mathbb{H}_1^{n+m}\big) &= h\big(\mathbb{H}_1^n\big) + h\big(\mathbb{H}_{n+1}^{n+m} \,\big|\, \mathbb{H}_1^n\big) && (4.28)\\
&= h\big(\mathbb{H}_1^n\big) + \sum_{k=n+1}^{n+m} h\big(\mathbb{H}_k \,\big|\, \mathbb{H}_1^{k-1}\big) && (4.29)\\
&\le h\big(\mathbb{H}_1^n\big) + \sum_{k=n+1}^{n+m} h\big(\mathbb{H}_{n+1} \,\big|\, \mathbb{H}_1^n\big) && (4.30)\\
&= h\big(\mathbb{H}_1^n\big) + m h\big(\mathbb{H}_{n+1} \,\big|\, \mathbb{H}_1^n\big). && (4.31)
\end{aligned}
$$

Here in (4.28) and (4.29) we use the chain rule; and (4.30) follows from the second statement. Hence,

$$
\underbrace{\frac{1}{n+m} h\big(\mathbb{H}_1^{n+m}\big)}_{\to \frac{1}{m} h\left(\mathbb{H}_1^m\right) \text{ as } m\uparrow\infty} \le \underbrace{\frac{1}{n+m} h(\mathbb{H}_1^n)}_{\to 0 \text{ as } m\uparrow\infty} + \underbrace{\frac{m}{n+m}}_{\to 1 \text{ as } m\uparrow\infty} h(\mathbb{H}_{n+1}|\mathbb{H}_1^n). \tag{4.32}
$$

Letting m tend to infinity we then get

$$
\lim_{m\uparrow\infty} \frac{1}{m} h\big(\mathbb{H}_1^{n+m}\big) \le h(\mathbb{H}_{n+1}|\mathbb{H}_1^n) \tag{4.33}
$$

which combined with the third statement proves the fourth statement.

4.2.2 Proof of Theorem 7

We start with the proof of the second statement which follows directly from the fact that conditioning cannot increase differential entropy:

$$
\begin{aligned}
h\big(\mathbb{H}_{n+1}\hat{\mathbf{X}}_{n+1} \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^n, \hat{\mathbf{X}}_1^{n+1}\big) &\le h\big(\mathbb{H}_{n+1}\hat{\mathbf{X}}_{n+1} \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=2}^n, \hat{\mathbf{X}}_2^{n+1}\big) && (4.34)\\
&= h\big(\mathbb{H}_n\hat{\mathbf{X}}_n \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n-1}, \hat{\mathbf{X}}_1^n\big) && (4.35)
\end{aligned}
$$

where the last equality follows from stationarity.

We next use the second statement to prove the third:

$$
\begin{aligned}
h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^n \,\big|\, \hat{\mathbf{X}}_1^n\big) &= \sum_{k=1}^n h\big(\mathbb{H}_k\hat{\mathbf{X}}_k \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{k-1}, \hat{\mathbf{X}}_1^n\big) && (4.36)\\
&= \sum_{k=1}^n h\big(\mathbb{H}_k\hat{\mathbf{X}}_k \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{k-1}, \hat{\mathbf{X}}_1^k\big) && (4.37)\\
&\ge \sum_{k=1}^n h\big(\mathbb{H}_n\hat{\mathbf{X}}_n \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n-1}, \hat{\mathbf{X}}_1^n\big) && (4.38)\\
&= n h\big(\mathbb{H}_n\hat{\mathbf{X}}_n \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n-1}, \hat{\mathbf{X}}_1^n\big) && (4.39)
\end{aligned}
$$

where (4.36) follows from the chain rule; (4.37) from the fact that conditional on $\hat{\mathbf{X}}_1^k$, $\hat{\mathbf{X}}_{k+1}^n$ is independent of all other random variables in the expression; and (4.38) follows from the second statement.

Next, we prove the first statement:

$$\begin{aligned}
&h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n+1} \,\big|\, \hat{\mathbf{X}}_1^{n+1}\big) \\
&\quad = h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n} \,\big|\, \hat{\mathbf{X}}_1^{n+1}\big) + h\big(\mathbb{H}_{n+1}\hat{\mathbf{X}}_{n+1} \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n}, \hat{\mathbf{X}}_1^{n+1}\big) && (4.40)\\
&\quad \le h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n} \,\big|\, \hat{\mathbf{X}}_1^{n}\big) + h\big(\mathbb{H}_{n}\hat{\mathbf{X}}_{n} \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n-1}, \hat{\mathbf{X}}_1^{n}\big) && (4.41)\\
&\quad \le h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n} \,\big|\, \hat{\mathbf{X}}_1^{n}\big) + \frac{1}{n} h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n} \,\big|\, \hat{\mathbf{X}}_1^{n}\big) && (4.42)\\
&\quad = \frac{n+1}{n} h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n} \,\big|\, \hat{\mathbf{X}}_1^{n}\big) && (4.43)
\end{aligned}$$

where (4.40) follows from the chain rule; where (4.41) follows from the second statement and from the fact that given $\hat{\mathbf{X}}_1^n$ the random vector $\hat{\mathbf{X}}_{n+1}$ is independent of $\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^n$, and where (4.42) follows from the third statement.

Finally, to prove the fourth statement we note that for an arbitrary $m \in \mathbb{N}$

$$\begin{aligned}
&h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n+m} \,\big|\, \hat{\mathbf{X}}_1^{n+m}\big) \\
&\quad = h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n} \,\big|\, \hat{\mathbf{X}}_1^{n+m}\big) + h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=n+1}^{n+m} \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n}, \hat{\mathbf{X}}_1^{n+m}\big) && (4.44)\\
&\quad = h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n} \,\big|\, \hat{\mathbf{X}}_1^{n}\big) + \sum_{k=n+1}^{n+m} h\big(\mathbb{H}_k\hat{\mathbf{X}}_k \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{k-1}, \hat{\mathbf{X}}_1^{n+m}\big) && (4.45)\\
&\quad = h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n} \,\big|\, \hat{\mathbf{X}}_1^{n}\big) + \sum_{k=n+1}^{n+m} h\big(\mathbb{H}_k\hat{\mathbf{X}}_k \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{k-1}, \hat{\mathbf{X}}_1^{k}\big) && (4.46)\\
&\quad \le h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n} \,\big|\, \hat{\mathbf{X}}_1^{n}\big) + \sum_{k=n+1}^{n+m} h\big(\mathbb{H}_{n+1}\hat{\mathbf{X}}_{n+1} \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n}, \hat{\mathbf{X}}_1^{n+1}\big) && (4.47)\\
&\quad = h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n} \,\big|\, \hat{\mathbf{X}}_1^{n}\big) + m h\big(\mathbb{H}_{n+1}\hat{\mathbf{X}}_{n+1} \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n}, \hat{\mathbf{X}}_1^{n+1}\big). && (4.48)
\end{aligned}$$

Here in (4.44) and (4.45) we use the chain rule; (4.46) follows from the fact that conditional on $\hat{\mathbf{X}}_1^k$, $\hat{\mathbf{X}}_{k+1}^{n+m}$ are independent of all other random variables in the expression; and (4.47) follows from the second statement. Hence,

$$\begin{aligned}
&\underbrace{\frac{1}{n+m} h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n+m} \,\big|\, \hat{\mathbf{X}}_1^{n+m}\big)}_{\to \frac{1}{m} h\left(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{m} \,\middle|\, \hat{\mathbf{X}}_1^{m}\right) \text{ as } m\uparrow\infty} \\
&\quad \le \underbrace{\frac{1}{n+m} h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n} \,\big|\, \hat{\mathbf{X}}_1^{n}\big)}_{\to 0 \text{ as } m\uparrow\infty} + \underbrace{\frac{m}{n+m}}_{\to 1 \text{ as } m\uparrow\infty} h\big(\mathbb{H}_{n+1}\hat{\mathbf{X}}_{n+1} \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n}, \hat{\mathbf{X}}_1^{n+1}\big). && (4.49)
\end{aligned}$$

Letting m tend to infinity we then get

$$\lim_{m\uparrow\infty} \frac{1}{m} h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{m} \,\big|\, \hat{\mathbf{X}}_1^{m}\big) \le h\big(\mathbb{H}_{n+1}\hat{\mathbf{X}}_{n+1} \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=1}^{n}, \hat{\mathbf{X}}_1^{n+1}\big) \qquad (4.50)$$

which combined with the third statement proves the fourth statement.

4.3 Proof of Theorem 8

We assume that $\mathbf{X}_{k-\kappa}^{k}$ is distributed according to the quasi-stationary distribution $Q_{\mathcal{E},\epsilon}^{\kappa+1}$. We bound as follows:

$$\begin{aligned}
&I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbf{Y}_k \,\big|\, \mathbf{X}_{k-\kappa}^{k}, \{\mathbb{H}_\ell \mathbf{X}_\ell\}_{\ell=k-\kappa}^{k-1}\big) \\
&\quad = I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbf{Y}_k \,\big|\, \hat{\mathbf{X}}_{k-\kappa}^{k}, \{\|\mathbf{X}_\ell\|\}_{\ell=k-\kappa}^{k}, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=k-\kappa}^{k-1}\big) && (4.51)\\
&\quad \le I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbf{Y}_k, \mathbb{H}_k \hat{\mathbf{X}}_k \,\big|\, \hat{\mathbf{X}}_{k-\kappa}^{k}, \{\|\mathbf{X}_\ell\|\}_{\ell=k-\kappa}^{k}, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=k-\kappa}^{k-1}\big) && (4.52)\\
&\quad = I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbb{H}_k \hat{\mathbf{X}}_k \,\big|\, \hat{\mathbf{X}}_{k-\kappa}^{k}, \{\|\mathbf{X}_\ell\|\}_{\ell=k-\kappa}^{k}, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=k-\kappa}^{k-1}\big) && (4.53)\\
&\quad = I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbb{H}_k \hat{\mathbf{X}}_k \,\big|\, \hat{\mathbf{X}}_{k-\kappa}^{k}, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=k-\kappa}^{k-1}\big) && (4.54)\\
&\quad = I\big(\mathbb{H}_{-k+1}^{-\kappa-1}; \mathbb{H}_0 \hat{\mathbf{X}}_k \,\big|\, \hat{\mathbf{X}}_{k-\kappa}^{k}, \{\mathbb{H}_\ell \hat{\mathbf{X}}_{k+\ell}\}_{\ell=-\kappa}^{-1}\big). && (4.55)
\end{aligned}$$

Here, in (4.51) we split the vectors $\mathbf{X}_\ell$ up into magnitude and direction; in (4.52) we add the additional term $\mathbb{H}_k \hat{\mathbf{X}}_k$ to the argument of mutual information; in (4.53) we drop $\mathbf{Y}_k$ because given $\mathbb{H}_k \hat{\mathbf{X}}_k$ it is independent of the other random quantities; then in (4.54) we remove the conditioning on $\|\mathbf{X}_\ell\|$ because it does not provide any useful information; and in the last step (4.55) we made use of the stationarity of $\{\mathbb{H}_k\}$.

Similar to the derivation of the upper bound, in the following we will again introduce a shorthand and rename $\hat{\mathbf{X}}_{k-\kappa}^{k}$ by $\hat{\mathbf{X}}_{-\kappa}^{0}$. Note that since the upper bound that is derived in this appendix will not depend on $\{\hat{\mathbf{X}}_k\}$, we lose the dependence on k in the end anyway.

Hence, letting $\hat{\mathbf{X}}_{-\kappa}^{0} \sim Q_{\mathcal{E},\epsilon}^{\kappa+1}$, we rewrite (4.55) as follows:

$$\begin{aligned}
&I\big(\mathbb{H}_{-k+1}^{-\kappa-1}; \mathbb{H}_0 \hat{\mathbf{X}}_k \,\big|\, \hat{\mathbf{X}}_{k-\kappa}^{k}, \{\mathbb{H}_\ell \hat{\mathbf{X}}_{k+\ell}\}_{\ell=-\kappa}^{-1}\big) \\
&\quad = I\big(\mathbb{H}_{-k+1}^{-\kappa-1}; \mathbb{H}_0 \hat{\mathbf{X}}_0 \,\big|\, \hat{\mathbf{X}}_{-\kappa}^{0}, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{-1}\big) && (4.56)\\
&\quad \le I\big(\mathbb{H}_{-\infty}^{-\kappa-1}; \mathbb{H}_0 \hat{\mathbf{X}}_0 \,\big|\, \hat{\mathbf{X}}_{-\kappa}^{0}, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{-1}\big) && (4.57)\\
&\quad = h\big(\mathbb{H}_0 \hat{\mathbf{X}}_0 \,\big|\, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{-1}, \hat{\mathbf{X}}_{-\kappa}^{0}\big) - h\big(\mathbb{H}_0 \hat{\mathbf{X}}_0 \,\big|\, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{-1}, \hat{\mathbf{X}}_{-\kappa}^{0}, \mathbb{H}_{-\infty}^{-\kappa-1}\big) && (4.58)\\
&\quad \le \frac{1}{\kappa+1} h\big(\{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{0} \,\big|\, \hat{\mathbf{X}}_{-\kappa}^{0}\big) - h\big(\mathbb{H}_0 \hat{\mathbf{X}}_0 \,\big|\, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{-1}, \hat{\mathbf{X}}_{-\kappa}^{0}, \mathbb{H}_{-\infty}^{-\kappa-1}\big). && (4.59)
\end{aligned}$$

Here in (4.57) we add more terms to the argument of mutual information; and (4.59) follows from the third statement of Theorem 7.

Now note that for $\hat{\mathbf{X}}_{-\kappa}^{0}$ being quasi-stationary and for all $i \in \{-\kappa, \ldots, -1\}$ we have

$$\begin{aligned}
&h\big(\mathbb{H}_i \hat{\mathbf{X}}_i \,\big|\, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{i-1}, \hat{\mathbf{X}}_{-\kappa}^{i}, \mathbb{H}_{-\infty}^{-\kappa-1}\big) \\
&\quad = h\big(\mathbb{H}_{i+1} \hat{\mathbf{X}}_{i+1} \,\big|\, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=-\kappa+1}^{i}, \hat{\mathbf{X}}_{-\kappa+1}^{i+1}, \mathbb{H}_{-\infty}^{-\kappa}\big) && (4.60)\\
&\quad = h\big(\mathbb{H}_{i+1} \hat{\mathbf{X}}_{i+1} \,\big|\, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=-\kappa+1}^{i}, \hat{\mathbf{X}}_{-\kappa+1}^{i+1}, \hat{\mathbf{X}}_{-\kappa}, \mathbb{H}_{-\infty}^{-\kappa}\big) && (4.61)\\
&\quad = h\big(\mathbb{H}_{i+1} \hat{\mathbf{X}}_{i+1} \,\big|\, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=-\kappa+1}^{i}, \hat{\mathbf{X}}_{-\kappa}^{i+1}, \mathbb{H}_{-\infty}^{-\kappa}, \mathbb{H}_{-\kappa} \hat{\mathbf{X}}_{-\kappa}\big) && (4.62)\\
&\quad \le h\big(\mathbb{H}_{i+1} \hat{\mathbf{X}}_{i+1} \,\big|\, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=-\kappa+1}^{i}, \hat{\mathbf{X}}_{-\kappa}^{i+1}, \mathbb{H}_{-\infty}^{-\kappa-1}, \mathbb{H}_{-\kappa} \hat{\mathbf{X}}_{-\kappa}\big) && (4.63)\\
&\quad = h\big(\mathbb{H}_{i+1} \hat{\mathbf{X}}_{i+1} \,\big|\, \{\mathbb{H}_\ell \hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{i}, \hat{\mathbf{X}}_{-\kappa}^{i+1}, \mathbb{H}_{-\infty}^{-\kappa-1}\big) && (4.64)
\end{aligned}$$

where (4.60) follows from the stationarity of $\{\mathbb{H}_k\}$ and the quasi-stationarity of $\hat{\mathbf{X}}_{-\kappa}^{0}$ (note that $i < 0$ so that $i + 1 \le 0$); in (4.61) we add $\hat{\mathbf{X}}_{-\kappa}$ which, conditional on $\hat{\mathbf{X}}_{-\kappa+1}^{i+1}$, is independent of the other random quantities; then in (4.62) we add $\mathbb{H}_{-\kappa}\hat{\mathbf{X}}_{-\kappa}$ to the conditioning which does not change anything as it is a function of the given terms $\mathbb{H}_{-\kappa}$ and $\hat{\mathbf{X}}_{-\kappa}$; and the inequality (4.63) then follows by dropping $\mathbb{H}_{-\kappa}$ which cannot reduce entropy.

Therefore,

$$\begin{aligned}
& h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{0} \,\big|\, \hat{\mathbf{X}}_{-\kappa}^{0}, \mathbb{H}_{-\infty}^{-\kappa-1}\big) \\
&\quad = \sum_{i=-\kappa}^{0} h\big(\mathbb{H}_i\hat{\mathbf{X}}_i \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{i-1}, \hat{\mathbf{X}}_{-\kappa}^{0}, \mathbb{H}_{-\infty}^{-\kappa-1}\big) && (4.65)\\
&\quad = \sum_{i=-\kappa}^{0} h\big(\mathbb{H}_i\hat{\mathbf{X}}_i \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{i-1}, \hat{\mathbf{X}}_{-\kappa}^{i}, \mathbb{H}_{-\infty}^{-\kappa-1}\big) && (4.66)\\
&\quad \le \sum_{i=-\kappa}^{0} h\big(\mathbb{H}_0\hat{\mathbf{X}}_0 \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{-1}, \hat{\mathbf{X}}_{-\kappa}^{0}, \mathbb{H}_{-\infty}^{-\kappa-1}\big) && (4.67)\\
&\quad = (\kappa+1) h\big(\mathbb{H}_0\hat{\mathbf{X}}_0 \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{-1}, \hat{\mathbf{X}}_{-\kappa}^{0}, \mathbb{H}_{-\infty}^{-\kappa-1}\big). && (4.68)
\end{aligned}$$

Here, (4.65) follows from the chain rule; in (4.66) we drop $\hat{\mathbf{X}}_{i+1}^{0}$ because conditioned on $\hat{\mathbf{X}}_{-\kappa}^{i}$ they are independent of the other random quantities; and in (4.67) we apply (4.64) several times to each term of the sum.

Hence we have

$$h\big(\mathbb{H}_0\hat{\mathbf{X}}_0 \,\big|\, \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{-1}, \hat{\mathbf{X}}_{-\kappa}^{0}, \mathbb{H}_{-\infty}^{-\kappa-1}\big) \ge \frac{1}{\kappa+1} h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{0} \,\big|\, \hat{\mathbf{X}}_{-\kappa}^{0}, \mathbb{H}_{-\infty}^{-\kappa-1}\big). \quad (4.69)$$

Using this in (4.59) and (4.55) we finally get

$$\begin{aligned}
& I\big(\mathbb{H}_1^{k-\kappa-1}; \mathbf{Y}_k \,\big|\, \mathbf{X}_{k-\kappa}^{k}, \{\mathbb{H}_\ell\mathbf{X}_\ell\}_{\ell=k-\kappa}^{k-1}\big) \\
&\quad \le \frac{1}{\kappa+1} h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{0} \,\big|\, \hat{\mathbf{X}}_{-\kappa}^{0}\big) - \frac{1}{\kappa+1} h\big(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{0} \,\big|\, \hat{\mathbf{X}}_{-\kappa}^{0}, \mathbb{H}_{-\infty}^{-\kappa-1}\big) && (4.70)\\
&\quad = \frac{1}{\kappa+1} I\big(\mathbb{H}_{-\infty}^{-\kappa-1}; \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{0} \,\big|\, \hat{\mathbf{X}}_{-\kappa}^{0}\big) && (4.71)\\
&\quad = \frac{1}{\kappa+1} I\big(\mathbb{H}_{-\infty}^{-\kappa-1}; \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{0}, \hat{\mathbf{X}}_{-\kappa}^{0}\big) && (4.72)\\
&\quad \le \frac{1}{\kappa+1} I\big(\mathbb{H}_{-\infty}^{-\kappa-1}; \{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{0}, \hat{\mathbf{X}}_{-\kappa}^{0}, \mathbb{H}_{-\kappa}^{0}\big) && (4.73)\\
&\quad = \frac{1}{\kappa+1} I\big(\mathbb{H}_{-\infty}^{-\kappa-1}; \mathbb{H}_{-\kappa}^{0}\big) && (4.74)\\
&\quad = \frac{1}{\kappa+1} h(\mathbb{H}_{-\kappa}^{0}) - \frac{1}{\kappa+1} h(\mathbb{H}_{-\kappa}^{0}|\mathbb{H}_{-\infty}^{-\kappa-1}) && (4.75)\\
&\quad = \frac{1}{\kappa+1} h(\mathbb{H}_{-\kappa}^{0}) - \frac{1}{\kappa+1} \sum_{i=-\kappa}^{0} h(\mathbb{H}_i|\mathbb{H}_{-\infty}^{-i-1}) && (4.76)\\
&\quad = \frac{1}{\kappa+1} h(\mathbb{H}_{-\kappa}^{0}) - \frac{1}{\kappa+1} \sum_{i=-\kappa}^{0} h(\mathbb{H}_0|\mathbb{H}_{-\infty}^{-1}) && (4.77)\\
&\quad = \frac{1}{\kappa+1} h(\mathbb{H}_{-\kappa}^{0}) - h(\mathbb{H}_0|\mathbb{H}_{-\infty}^{-1}) && (4.78)\\
&\quad \triangleq \delta_1(\kappa). && (4.79)
\end{aligned}$$

Here, (4.72) follows because $\hat{\mathbf{X}}_{-\kappa}^{0}$ is independent of $\mathbb{H}_{-\infty}^{-\kappa-1}$; in (4.73) we add $\mathbb{H}_{-\kappa}^{0}$ to the argument of mutual information; in (4.74) we drop $(\{\mathbb{H}_\ell\hat{\mathbf{X}}_\ell\}_{\ell=-\kappa}^{0}, \hat{\mathbf{X}}_{-\kappa}^{0})$ because conditional on $\mathbb{H}_{-\kappa}^{0}$ it is independent of $\mathbb{H}_{-\infty}^{-\kappa-1}$; and (4.77) follows from stationarity.

Note that $\delta_1(\kappa)$ does neither depend on k nor on the input $\{\mathbf{X}_k\}$ and that by Theorem 6 it monotonically tends to zero as κ tends to infinity.

Chapter 5

Discussion & Conclusion

We have proven three results about stationary fading channels: the first two are fundamental results that are true in a much more general context and are in no way restricted to the situation of MIMO fading channels: the first results states that a stationary channel has a capacity-achieving input distribution that is stationary. This is intuitively very pleasing and confirms our belief that stationary processes in general behave "the way engineers expect them to behave."

The second result concerns entropy rates of various forms. We have shown that entropy rates are also well-defined for processes with a continuous alphabet as long as the processes are stationary (and the entropy rate is finite). Moreover, we also generalize this definition to include various sorts of conditional entropy rates and show that they are also well-defined. These conditional entropy rates are strongly matched to the situation of MIMO fading channels, however, this is not necessary and similar type of results can be derived in many other situations, too. Even though we have shown the proof only for two cases, it is clearly seen that the fundament of the proof is the stationarity assumption only.

As third result we prove that a certain mutual information expression of the form

$$I\big(X_{-\infty}^{-\kappa}; X_0 \,\big|\, X_{-\kappa+1}^{-1}\big)$$

tends to zero as κ tends to infinity. This corresponds to the intuition that a process is supposed to forget about things that happened an infinite amount of time before. This result is more strongly bound to the context of MIMO fading channel. Nevertheless it contains some general message: it is a further example for the nice behavior of stationary processes. Note, however,that in order to prove this mathematically, it is crucial not to leave the scope of stationarity, *i.e.*, steps introducing a supremum or similar might open the door to unrealistic, but mathematically dangerous processes, which will destroy the nice properties of the family of stationary distributions.

We believe that all three results are of interests in many different situations.

Bibliography

[1] C. E. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, July and October 1948.

[2] T. L. Marzetta and B. M. Hochwald, “Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading,” *IEEE Transactions on Information Theory*, vol. 45, no. 1, pp. 139–157, January 1999.

[3] L. Zheng and D. N. C. Tse, “Communicating on the Grassmann manifold: a geometric approach to the non coherent multiple-antenna channel,” *IEEE Transactions on Information Theory*, vol. 48, no. 2, pp. 359–383, February 2002.

[4] Y. Liang and V. V. Veeravalli, “Capacity of noncoherent time-selective Rayleigh-fading channels,” *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3095–3110, December 2004.

[5] G. Taricco and M. Elia, “Capacity of fading channel with no side information,” *Electronics Letters*, vol. 33, no. 16, pp. 1368–1370, July 31, 1997.

[6] I. C. Abou-Faycal, M. D. Trott, and S. Shamai (Shitz), “The capacity of discrete-time memoryless Rayleigh fading channels,” *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1290–1301, May 2001.

[7] A. Lapidoth and S. M. Moser, “Capacity bounds via duality with applications to multiple-antenna systems on flat fading channels,” *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2426–2467, October 2003.

[8] ——, “The fading number of single-input multiple-output fading channels with memory,” *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 437–453, February 2006.

[9] S. M. Moser, “Duality-based bounds on channel capacity,” Ph.D. dissertation, Swiss Federal Institute of Technology (ETH), Zurich, October 2004, Diss. ETH No. 15769. [Online]. Available: http://moser.cm.nctu.edu.tw/

[10] T. Koch, “On the asymptotic capacity of multiple-input single-output fading channels with memory,” Master’s thesis, Signal and Information Processing Laboratory, ETH Zurich, Switzerland, April 2004, supervised by Prof. Dr. Amos Lapidoth.

[11] A. Lapidoth, “On the asymptotic capacity of stationary Gaussian fading channels,” *IEEE Transactions on Information Theory*, vol. 51, no. 2, pp. 437–446, February 2005.

[12] T. Koch and A. Lapidoth, "Degrees of freedom in non-coherent stationary MIMO fading channels," in *Proceedings Winter School on Coding and Information Theory*, Bratislava, Slovakia, February 20–25, 2005, pp. 91–97.

[13] ——, "The fading number and degrees of freedom in non-coherent MIMO fading channels: a peace pipe," in *Proceedings IEEE International Symposium on Information Theory (ISIT)*, Adelaide, Australia, September 4–9, 2005, pp. 661–665.

[14] A. Lapidoth, "The fading number and degrees of freedom: A peace pipe," presentation, Shushan Purim, Israel, March 27, 2005.

[15] S. M. Moser, "Bounds on the fading number of multiple-input single-output fading channels with memory," in *Proceedings International Wireless Communications and Mobile Computing Conference (IWCMC)*, Vancouver, Canada, July 3–6, 2006, pp. 3–8.

[16] ——, "On the fading number of multiple-input single-output fading channels with memory," in *Proceedings IEEE International Symposium on Information Theory (ISIT)*, Seattle, WA, USA, July 9–14, 2006, pp. 1056–1060.

[17] ——, "Bounds on the fading number of multiple-input single-output fading channels with memory," Department of Communication Engineering, National Chiao Tung University (NCTU), final report of National Science Council (NSC) project NSC 94-2218-E-009-037, from December 1, 2005, to July 31, 2006.

[18] ——, "The fading number of memoryless multiple-input multiple-output fading channels," *IEEE Transactions on Information Theory*, vol. 53, no. 7, pp. 2652–2666, July 2007.

[19] T. Koch, A. Lapidoth, and S. M. Moser, "Fading number and degrees of freedom," 2008, in preparation.