Master Thesis

Dynamical characterization of Markov processes with varying order

Michael Bauer

Chemnitz, July 1, 2008

Tutor: Prof. Dr. Günter Radons
Bauer, Michael
Dynamical characterization of Markov processes with varying order
Master Thesis, Faculty of Natural Sciences
University of Technology Chemnitz, July 2008
Acknowledgement

I would like to thank Prof. Dr. Günter Radons for his assistance contributing numerous ideas and valuable reference hints. Furthermore, I appreciate the interest of Holger Fiedler, Ines Hartwig, Mario Heidernätsch, and Roland Lange resulting in annotations and fruitful discussions. My particular thanks are dedicated to my whole family as well as to Marlen Benduhn and her family, who all provided support to me studying at Chemnitz University of Technology during the last years.
Abstract

Time-delayed actions appear as an essential component of numerous systems especially in evolution processes, natural phenomena, and particular technical applications and are associated with the existence of a memory. Under common conditions, external forces or state dependent parameters modify the length of the delay with time. Consequently, an altered dynamical behavior emerges, whose characterization is compulsory for a deeper understanding of these processes. In this thesis, the well-investigated class of time-homogeneous finite-state Markov processes is utilized to establish a variation of memory length by combining a first-order Markov chain with a memoryless Markov chain of order zero. The fluctuations induce a non-stationary process, which is accomplished for two special cases: a periodic and a random selection of the available Markov chains. For both cases, the Kolmogorov-Sinai entropy as a characteristic property is deduced analytically and compared to numerical approximations to the entropy rate of related symbolic dynamics. The convergences of per-symbol and conditional entropies are examined in order to recognize their behavior when identifying unknown processes. Additionally, the connection from Markov processes with varying memory length to hidden Markov models is illustrated enabling further analysis. Hence, the Kolmogorov-Sinai entropy of hidden Markov chains is calculated by means of Blackwell’s entropy rate involving Blackwell’s measure. These results are used to verify the previous computations.
## Contents

### List of Figures
- iii

### List of Tables
- v

### 1 Introduction
- 1

### 2 Theoretical backgrounds
- 3
  - 2.1 Entropies
    - 2.1.1 Shannon entropy
    - 2.1.2 Block entropy
    - 2.1.3 Per-symbol entropy
    - 2.1.4 Conditional entropy
    - 2.1.5 Kolmogorov-Sinai entropy
    - 2.1.6 Graphs of entropies and their characteristics
  - 2.2 Markov chains
    - 2.2.1 Definitions
    - 2.2.2 State diagrams
    - 2.2.3 Probability distributions on the state space
    - 2.2.4 Transformations between different orders
    - 2.2.5 Properties
    - 2.2.6 Stationary distribution
    - 2.2.7 Kolmogorov-Sinai entropy
  - 2.3 Hidden Markov chains
    - 2.3.1 Definitions
    - 2.3.2 Probability distributions and stationary distribution
    - 2.3.3 Kolmogorov-Sinai entropy

### 3 Examples of stationary Markov processes
- 19
  - 3.1 First-order Markov chain
  - 3.2 Markov chain without memory

### 4 Numerical calculations
- 24
  - 4.1 Implementation
  - 4.2 Examples
    - 4.2.1 First-order Markov chain
## CONTENTS

4.2.2 Markov chain without memory ........................................... 27

5 Variations of order ................................................................. 31
5.1 Analytical cases ................................................................. 31

5.2 Periodic variations .............................................................. 32
5.2.1 Preliminaries ................................................................. 33
5.2.2 Derivation of the Kolmogorov-Sinai entropy ......................... 34
5.2.3 Summary ................................................................. 40
5.2.4 Example ................................................................. 41
5.2.5 Investigation of the convergence ........................................ 44

5.3 Random variations ............................................................ 47
5.3.1 Preliminaries ................................................................. 47
5.3.2 Derivation of the Kolmogorov-Sinai entropy ......................... 49
5.3.3 Alternative derivation and hidden Markov chains .................. 52
5.3.4 Summary ................................................................. 56
5.3.5 Example ................................................................. 57
5.3.6 Dependence on probability of occurrence ......................... 60

6 Blackwell’s formula ............................................................... 64
6.1 Theory .......................................................................... 64
6.2 Kolmogorov-Sinai entropy for random variations ..................... 68
6.2.1 Hidden Markov chain ......................................................... 68
6.2.2 Derivation ................................................................. 70
6.2.3 Example ................................................................. 72

6.3 Kolmogorov-Sinai entropy for periodic variations ..................... 75
6.3.1 Hidden Markov chain ......................................................... 76
6.3.2 Derivation ................................................................. 79
6.3.3 Example ................................................................. 81

7 Summary ........................................................................... 87
7.1 Results and conclusions ..................................................... 87
7.2 Outlook ......................................................................... 88

Bibliography ................................................................. 91
## List of Figures

2.1 Characteristic evolution of the conditional entropies of various processes . 8
2.2 State diagram of a first-order Markov chain with two states . . . . . . . . 11

4.1 Extraction of subsequences with a moving window . . . . . . . . . . . . 25
4.2 Entropy versus block length for a first-order Markov chain . . . . . . . 28
4.3 Entropy versus block length for a memoryless Markov chain . . . . . . . 29

5.1 Entropy versus block length for periodic variation . . . . . . . . . . . . 43
5.2 Logarithmic differences between numerical and analytical entropy rates . 45
5.3 Fitted and extrapolated entropy versus block length for periodic variation 46
5.4 Two autonomous first-order Markov chains with binary output . . . . . . 54
5.5 Two coupled first-order binary Markov chains for random variation . . . 55
5.6 Entropy versus block length for random variation . . . . . . . . . . . . 59
5.7 Conditional entropies dependent on $n$ and $p(M(1))$ for random variation 61
5.8 Entropy rate and Kolmogorov-Sinai entropy versus $p(M(1))$ . . . . . . 63

6.1 Mapping of $\Phi$ from the hidden states to the output symbols . . . . . . 70
6.2 Distribution of the probability mass after observation . . . . . . . . . . . 73
6.3 Distributions after one observation for random variation . . . . . . . . . 75
6.4 Two coupled first-order binary Markov chains for periodic variation . . . 78
6.5 Swapping of the probability mass between the machines . . . . . . . . . 79
6.6 Distributions after one observation for periodic variation . . . . . . . . . 83
6.7 Distributions after two observations for periodic variation . . . . . . . . . 85
6.8 Blackwell’s entropy rate versus block length for periodic variation . . . . 86
List of Tables

3.1 Transition probabilities of a first-order Markov chain . . . . . . . . . . . 20
3.2 Transition probabilities of a memoryless Markov chain . . . . . . . . . . . 22
5.1 Comparison of numerical and analytical Kolmogorov-Sinai entropies . . 62
6.1 Transition probabilities of a hidden Markov chain with random variation . 69
6.2 Assignment of the output symbols to the hidden states via Φ . . . . . . . 69
6.3 Transition probabilities of a hidden Markov chain with periodic variation 77
1 Introduction

In numerous natural and technical processes, time-delayed effects contribute a significant impact on the dynamical behavior of the system. In particular, machines like lathes for woodturning and metal spinning excite time-delayed dependencies caused by the work piece [Wie97]. The unintended effects are difficult to predict and lead to a destruction of the tool. Furthermore, the usage of time-delayed feedback control in chaotic systems [JBO+97] establishes stability and time-delayed neural networks are utilized to improve speech recognition methods [WHH+89].

In recent investigations, a variation of the time delay has been taken into consideration [KCR+04]. This resembles the reality of time-delayed processes more accurately as the length of the delay changes with time. These phenomena are encountered due to system parameters or external forces depending on time. The evolution of time-delayed processes is described by delay differential equations, which can be solved numerically [BZ03]. An enhancement to fluctuating delays exists.

The solution of time-delayed equations raises the question of stability of their trajectories as well as their sensitivity to initial conditions. This can be measured by Lyapunov exponents investigating the divergence of nearby trajectories. The Lyapunov exponents provide estimation at the amount of chaos contained in the system and are closely related to the Kolmogorov-Sinai entropy [BGS76, LPT97] according to Pesin’s theorem. An extensive overview for the theoretical backgrounds is given by [Cho05] introducing measure and ergodic theory explained with discrete dynamical systems and their computer simulations.

Markov processes belong to a class of stochastic processes, which are well investigated. They are equipped with a memory and thus represent a process with time delay. If two or more Markov processes with different memory lengths are combined and selected arbitrarily, it will result in a process with varying time delay. Furthermore, Markov processes are often related to time-series analysis [Tjo90] and interpreted as the generator of such a series. For this purpose, the entropy rate of the source is determined providing a variety of practical applications, such as DNA analysis [Alm83] in bioinformatics and data compression [Fan96] in computer science.

As an extension to Markov processes with further practical usage, hidden Markov processes are introduced. In contrast to ordinary Markov processes, they do not reveal their
inner state, when creating an output symbol. Thus, they are identified with an observation in noise and gained a broad interest in communication theory. Nevertheless, an analytical calculation of the Kolmogorov-Sinai entropy cannot be deduced in general up to today [ZKD05]. The widespread applications of hidden Markov processes are connected to pattern classification [DHS01] being utilized in speech recognition [Rab89], as well as machine learning [Bis06, Mac03] and analysis and modeling of protein molecules [KBM+94].

By means of finite-state Markov processes with fluctuating memory length, some properties can be adopted to delay differential equations with varying delay length. Hereinafter, the investigations comprise the numerical computation of entropy rates as well as explicit analytical expressions for the Kolmogorov-Sinai entropies of Markov processes with special fluctuation of memory length. Especially, a periodic as well as a random variation between two Markov processes of different memory lengths is particularized. The derived analytical results are verified and compared to entropy calculations of hidden Markov processes accomplished by Blackwell’s formula [Bla57].
2 Theoretical backgrounds

2.1 Entropies

The entropy is an important property of a stochastic process. It is a measure specifying the uncertainty associated with its random variable. Hence, it describes the amount of information contained in a sequence of states generated by the stochastic process. Equivalently, the entropy quantifies the average amount of information, which is missed, if the state of the random variable is not known [SW63].

Furthermore, a system being sensitive to its initial conditions produces information. The explanation is, if two different initial conditions are as close together that they cannot be distinguished experimentally, they will develop into two states, which are distinguishable after a finite time [ER85]. This behavior leads to the definition of Lyapunov exponents [EFS98, KS04] and illustrates their connection to information and therefore to entropy [McM53].

A usage of entropies can be found in a variety of applications. In general, it is used to characterize stochastic processes, determining their sensitivity to initial conditions and consequently their chaotic behavior [GP83]. Moreover, by calculating the entropy of a text, its information content can be extracted. As a result, a limit for the best lossless compression of a text is defined by the entropy [Fan96]. Thus, entropy and its calculation gained several applications in modern communication technology and computer science.

2.1.1 Shannon entropy

In order to measure the uncertainty associated with a random variable, the calculation of the Shannon entropy is a crucial basis.

A finite alphabet \( \mathcal{A} \), called state space, consists of \( \lambda \) different states depicted by symbols \( c^l \) where \( l = 1, \ldots, \lambda \). For each symbol \( c^l \) the probability \( p_l \) is given, describing its relative occurrence in a sequence of symbols, and thus, \( 0 \leq p_l \leq 1 \) and \( \sum_{l=1}^{\lambda} p_l = 1 \).
As Shannon defined in [Sha48], the entropy is given by

$$H = -k \sum_{l=1}^{\lambda} p_l \ln(p_l) \quad \text{with } k = \text{const}, \quad (2.1)$$

where all terms of the sum involving $p_l = 0$ can be neglected according to L'Hôpital’s rule.

If the symbols $c_l$ are equally likely with $p_l = \frac{1}{\lambda}$ $\forall l$, the entropy will reach its maximum. If the constant $k = 1$, the maximum value will be $\ln(k)$. By using $k = \frac{1}{\ln(k)}$ instead, hence the entropy is normalized to the range of zero to one. This normalization allows the comparison of entropies of alphabets with different cardinalities. Dividing the sum from (2.1) by $\ln(k)$ is equal to changing the base of the logarithm, which leads to

$$H = -\sum_{l=1}^{\lambda} p_l \log_\lambda(p_l) \quad (2.2)$$

the normalized Shannon entropy.

To summarize, the Shannon entropy represents the mean uncertainty when predicting symbols. It can also be interpreted as the gain of information when extending the sequence by one symbol.

### 2.1.2 Block entropy

The idea of the Shannon entropy is enhanced by grouping symbols together. In order to obtain these groups, subsequences of length $n \in \mathbb{N}$ are cut out of the sequence. These $n$-words $(c_1, \ldots, c_n)$ consist of $n$ distinct symbols $c_i \in \mathcal{A} \equiv \{c^1, \ldots, c^\lambda\}$, where $i$ is the index for the symbols within the sequence. The entropy of the subsequences is called entropy of blocks [SG96] and can be regarded as joint entropy in a system of $n$ random variables.

By means of the probabilities for the occurrence of the $n$-words, the entropy of blocks of length $n$

$$H_n := - \sum_{c_1,\ldots,c_n \in \mathcal{A}^n} p(c_1,\ldots,c_n) \log_\lambda(p(c_1,\ldots,c_n)) \quad (2.3)$$

can be calculated. This is equivalent to (2.2), whereby the $n$-words are regarded as symbols $c'$ of another finite alphabet $\mathcal{A}' = \mathcal{A}^n$ with a larger cardinality. This new state space $\mathcal{A}'$ contains $\lambda^n$ states depicted by the symbols $c'$.

Furthermore, it should be noted that the sum (2.3) has to be over all realizations of possible subsequences of length $n$. Thus, the number of terms of the sum increases exponentially with the length $n$ of the subsequence complicating the numerical calculation of block
entropies. To overcome this drawback, [KS96] proposes to rewrite differences of block entropies as averages of decay rates and to adapt the sampling.

Khinchin [Khi57a] has shown, that $H_m \geq H_n$ holds for all $m > n$. Therefore, the block entropy is monotonically nondecreasing due to an increase in the length of the subsequences. The equality will hold if an additional symbol does not result in a higher uncertainty. This is caused by correlations existing between the symbols $c_i$, which lead to structural regularities in the sequence. The effect is a much smaller uncertainty for predicting the next symbol in the sequence, the entropy of blocks grows more slowly.

### 2.1.3 Per-symbol entropy

So far, the joint entropy has an extensive character because a longer subsequence leads to a monotonically nondecreasing uncertainty. In order to remove the influence of the word length from entropy, (2.3) is divided by the length $n$ of the sequence resulting in

$$h_n := \frac{H_n}{n} \quad (2.4)$$

the mean uncertainty per symbol. With the occurrence of correlations, mentioned in 2.1.2, the entropy $h_n$ decreases with increasing length $n$ of the words. Thus, the uncertainty for predicting a symbol will shrink if the correlations are strong. Vice versa, the gain of information will become smaller if the subsequence is enlarged.

By calculating the mean uncertainty per symbol of infinitely long subsequences, the limit

$$h := h_\infty = \lim_{n \to \infty} \frac{H_n}{n} \quad (2.5)$$

represents the minimal mean uncertainty per symbol, which is referred to as source entropy [Khi57a]. It specifies the unavoidable amount of randomness, which is produced by the source. With this property, it is the entropy production rate and can be used as a measure for chaos.

### 2.1.4 Conditional entropy

Another approach to eliminate the extensive character of the block entropy is the usage of conditional entropies. The conditional entropy

$$\Delta H_n := H_{n+1} - H_n = H_{n+1|n} \quad \text{with } n = 1, 2, \ldots \quad (2.6)$$

describes the change of the block entropy when enlarging the $n$-word by one symbol. Moreover, definition (2.6) is completed by

$$\Delta H_0 := H_1 \quad (2.7)$$
By applying (2.6) to infinitely long subsequences, the limit

\[ h' := \Delta H_\infty = \lim_{n \to \infty} \Delta H_n \]  

(2.8)
again represents the unavoidable uncertainty for the prediction of the next symbol in the sequence. Thus, it is the entropy rate of the source, which generated the sequence. Alternatively, it can be interpreted as the gain of information from the \( n + 1 \)-word when knowing the \( n \)-word.

By rewriting the conditional entropy from (2.6) with sums,

\[
\Delta H_n = - \sum_{c_1, \ldots, c_{n+1}} p(c_1, \ldots, c_{n+1}) \log \lambda \left( p(c_1, \ldots, c_{n+1}) \right)
+ \sum_{c_1, \ldots, c_n} p(c_1, \ldots, c_n) \log \lambda \left( p(c_1, \ldots, c_n) \right)
\]

(2.9)
is obtained, which can be transformed by using the definition of conditional probabilities

\[ p(c_1, \ldots, c_{n+1}) = p(c_{n+1}|c_1, \ldots, c_n) p(c_1, \ldots, c_n) \]

(2.10)
to

\[
\Delta H_n = - \left( \sum_{c_1, \ldots, c_{n+1}} p(c_1, \ldots, c_{n+1}) \log \lambda \left( p(c_{n+1}|c_1, \ldots, c_n) \right) 
+ \sum_{c_1, \ldots, c_n} \sum_{c_{n+1}} p(c_{n+1}|c_1, \ldots, c_n) p(c_1, \ldots, c_n) \log \lambda \left( p(c_1, \ldots, c_n) \right) \right)
+ \sum_{c_1, \ldots, c_n} p(c_1, \ldots, c_n) \log \lambda \left( p(c_1, \ldots, c_n) \right)
\]

(2.11)
and simplified in consequence of

\[
\sum_{c_{n+1}} p(c_{n+1}|c_1, \ldots, c_n) = 1,
\]

(2.12)
due to opposite algebraic signs, to

\[
\Delta H_n = - \sum_{c_1, \ldots, c_{n+1}} p(c_1, \ldots, c_{n+1}) \log \lambda \left( p(c_{n+1}|c_1, \ldots, c_n) \right).
\]

(2.13)
Equation (2.13) is a good basis for an analytical derivation of conditional entropies, and, by applying the limit from (2.8), the entropy rate of the source can be calculated exactly.
2.1.5 Kolmogorov-Sinai entropy

Both per-symbol entropy from [2.1.3] and conditional entropy from [2.1.4] represent some kind of entropy production rate or equivalently a gain of information by enlarging the sequence by one symbol.

In [CT91], it is shown that for every stationary stochastic process both limits in (2.5) and (2.8) exist and even result in equal values, whereby

\[ h_{KS} := h' = h \]  (2.14)

becomes the entropy rate of the associated process.

This entropy rate is called the metric entropy or Kolmogorov-Sinai entropy \( h_{KS} \) in the case of a dynamical system [ER85]. This transition is comprehensible because any stochastic process has a correspondence to a measure-preserving dynamical system and vice versa [Pet83].

However, the speed of convergence of (2.5) and (2.8) is different and depends strongly on the structure of the stochastic process. In general, the conditional entropy (2.8) converges faster, mostly by an exponential law as stated in [Gra86] and [SG86].

Furthermore, different correlations of the symbols within the sequences have a significant impact on the entropy production rate of the source. Due to the normalization of the entropy in (2.2) two borderline cases exist. The first case is a source which generates symbols periodically. The resulting sequence will contain maximum redundancy, thus the conditional entropy will be zero if all symbols of the period are known. In the second case of independent and identically distributed symbols produced by the source, there is no redundancy at all. Hence, no information can be gained by knowing longer subsequences, the entropy rate reaches its maximum value of one.

2.1.6 Graphs of entropies and their characteristics

For several types of processes, the evolution of the entropies shows a characteristic behavior. Especially, the conditional entropy is considerable for a characterization and detection of a process. The values in figure [2.1] have been created numerically with the program described in [4.1].

Figure [2.1] illustrates the evolutions of various processes and enables their characterization depending on the behavior. For this purpose, three typical types of processes are depicted including the entropies of a Bernoulli source, a periodic source, and two Markov sources with different orders.
Figure 2.1: Characteristic evolution of the conditional entropies of various processes. A Bernoulli source does not contain any correlations, thus the uncertainty cannot be reduced leading to the conditional entropies staying at the same level. A periodic source generates correlations up to the period length, and then, the sequence can be predicted resulting in no uncertainty. This is expressed by the conditional entropies being equal to zero. Markov chains of order $m$ possess a memory of length $m$ and hence create sequences with such correlations. The uncertainty can be improved up to the length of the memory and then stays constant at the level of the entropy rate originating from the source.

The Bernoulli source does not contain any correlations between the generated symbols. Thus, not any information will be gained if a sequence is enlarged by one symbol. This results in a constant series of conditional entropies being equal to the Kolmogorov-Sinai entropy.

On the contrary, the periodic sequence depicted in figure 2.1 generating the symbols AABA has a period length of four. Thus, starting with subsequences of block length $n = 4$, all correlations which exist between the symbols are included, and a prediction
without uncertainty is possible. Moreover, all contained subsequences whose lengths are equal to the period length minus one or larger occur equally often. Consequently, they induce the same block entropies, and hence, the conditional entropies are equal to zero from $\Delta H_3$ on due to the definition (2.6). In conclusion, any periodic process possesses a Kolmogorov-Sinai entropy equal to zero, which can be interpreted as a non-existence of chaos.

The third type is a Markov process, which is particularized in 2.2. In short, Markov chains contain a memory corresponding to their order. Hence, a generated sequence involves only correlations with the length of this memory. As a result, the block entropies for $n$ larger than the memory length increase by the same amount induced by the randomness originating from the source. Thus, the conditional entropies are constant and equal to the Kolmogorov-Sinai entropy starting from the $n$ which coincides with the order of the Markov process. This leads to a characteristic bend in the evolution of the conditional probability.

2.2 Markov chains

Markov processes are a specific, well-known subclass of stochastic processes, which have their own properties to generate future evolutions. Markov processes gained a particular importance for describing processes in chemistry and physics [vK07] because they reveal a close similarity to processes found in nature. The most popular example of a Markov process is the Brownian motion, which describes the random movement of particles.

Inevitably, it is distinguished between the continuous-time and the discrete-time case. In particular, the latter is called a Markov chain and comprises a sequence of random variables known as a time series. Hereinafter, discrete-time finite-state Markov chains are investigated.

Markov processes provide a large field of application. Their usage is not constrained to creating time series or sequences of states. It can also be applied to investigate processes, time series, and symbol sequences. Furthermore, Markov chains are the basis for an extension to hidden Markov chains and several related applications.

2.2.1 Definitions

Markov chains are discrete-time stochastic processes and hence inherit properties from them. A discrete-time stochastic process consists of a collection of random variables $\{C_n\}$. The index $n \in \mathbb{Z}^+$ indicates the sequence, in which the random variables appear and can be identified as a kind of discrete-time parameter indicating the number of time
THEORETICAL BACKGROUNDS

steps. Furthermore, the random variables \( C_n \) are mapped to the realizations \( c_n \), which belong to a state space. The sequence of realizations \( c_n \) builds up a Markov chain. In a finite-state process, the aforementioned state space is a countable set \( \mathcal{A} \), which consists of \( \lambda \) symbols \( e^l \), where \( l = 1, \ldots, \lambda \).

A Markov chain is characterized by the feature that future states only depend on its most recent predecessors. This is called the Markov property and is often restricted to exactly one previous state. This causes the next state to be independent of all states except the current one. The transitions between the states are described by conditional probabilities.

However, the Markov property is defined more generally. If the generation of a future state is based on \( m \) preceding states, it will be called a Markov chain of order \( m \) [Par62]. Consequently, the Markov chain is said to have a memory of length \( m \) and a Markov chain of order zero is called memoryless.

For Markov chains of order \( m \) the conditional probability

\[
p(C_{n+1} = c_{n+1} | C_{n-m+1} = c_{n-m+1}, \ldots, C_n = c_n) = p(c_{n+1} | c_{n-m+1}, \ldots, c_n) \tag{2.15}
\]

depends on exactly \( m \) preceding states.

The states can be merged to vectors forming a new state space, where

\[
i \equiv \{c_{n-m+1}, \ldots, c_n\} \tag{2.16a}
\]
\[
j \equiv \{c_{n-m+2}, \ldots, c_{n+1}\} \tag{2.16b}
\]

are the compound states. This new state space is extended to \( \lambda^m \) different states.

Using the extended state space, the conditional probability (2.15) can be simplified to

\[
p(c_{n+1} | c_{n-m+1}, \ldots, c_n) = p(j | i, n) = p_{i-j}^{(n)} = p_{ij}^{(n)} \tag{2.17}
\]

and if the states \( i \) and \( j \) are identified by some integer values \( i, j = 1, \ldots \lambda^m \), a matrix representation will be given by

\[
(M_{ij}^{(n)}) = (p_{ij}^{(n)}) = \begin{pmatrix}
p_{11}^{(n)} & \cdots & p_{1\lambda^m}^{(n)} \\
\vdots & \ddots & \vdots \\
p_{\lambda^m 1}^{(n)} & \cdots & p_{\lambda^m \lambda^m}^{(n)}
\end{pmatrix} \tag{2.18}
\]

named Markov matrix \( M \). The conditional probabilities as well as the Markov matrix remain dependent on the time step parameter \( n \). It should be noticed, that the time step parameter \( n \) is written in parentheses to avoid confusion with exponents.

The Markov matrix \( M \) turns out to be a transition probability matrix, which describes the changes between states and their related probabilities. The probability of the transition
from state $i$ to state $j$ is located in the $i^{th}$ row and the $j^{th}$ column of $M$. The elements of $M$ are nonnegative real numbers and thus $M$ has to be a stochastic matrix at all times $n$, which can be recognized by

$$\sum_j M_{ij}^{(n)} = 1 \quad \forall i$$

(2.19)

the row sum being equal to one in each row. The explanation is that any current state performs a transition with certainty expressed by a probability vector in each row.

A Markov chain with a finite number of states, transitions between them, and actions on the states is also called a finite-state machine. An example for such an action is the generation of an output symbol. Thus, symbol sequence generators based on Markov chains are often referred to as machines.

### 2.2.2 State diagrams

It is common to illustrate Markov chains by directed graphs. These graphs are often called state diagrams since they include the states of the Markov chain as well as arrows with directions to define transitions between the states. To each transition, a transition probability is assigned, denoted by a conditional probability. From this graph, the transition matrix can be extracted in an effortless manner.

Figure 2.2: State diagram of a first-order Markov chain consisting of two states A and B. The arrows illustrate the transition between the states and are provided with appropriate conditional probabilities.

Figure 2.2 depicts a state diagram of a first-order Markov chain consisting of two states. The two states A and B are connected by arrows indicating the possible transitions between the states. At the arrows, conditional probabilities declare the probability of transition between the associated states. For this purpose, the current state is written as the condition of the probability to change to the next state.


2.2.3 Probability distributions on the state space

On the state space of a Markov chain, a probability distribution can be defined. It represents the probability of occurrence of the according states at a time step $n$ and is denoted by a row vector

$$\mathbf{\pi}(n).$$

For each $n$, the elements of $\mathbf{\pi}(n)$, which are nonnegative real numbers, sum up to one. Hence, these are stochastic vectors, being also called probability vectors.

The initial state vector is referred to as $\mathbf{\pi}(0)$ and describes the starting distribution on the state space. When regarding symbol sequences, it corresponds to the first symbol $c_1$ of the sequence. In particular, it assigns a probability to all symbols of the alphabet $\mathcal{A}$, with which they occur as the first symbol. Any Markov chain of order $m$ requires $m$ predefined distributions on the state space, which represent the initial memory of the Markov chain.

Because a change of the states is described and accomplished by the transition matrix $M$, a probability distribution on the states can be iterated likewise. Applying matrix $M$ to a state vector redistributes the probability mass to the states forming a new state vector. The vector-matrix product

$$\mathbf{\pi}(n)M(n) = \mathbf{\pi}(n+1)$$

(2.21)

generates the next probability distribution on the states from the current one. To realize further transitions, the Markov matrix appropriate to the time step $n$ has to be multiplied.

A Markov matrix, being time-homogeneous, as defined in 2.2.5, and therefore independent of the time step $n$, enables a $k$-step transition to consist solely of $k$ equal transition matrices $M$. Thus, the change of the probability distribution on the states can be written as

$$\mathbf{\pi}(n)M^k = \mathbf{\pi}(n+k)$$

(2.22)

a vector-matrix product involving a matrix power.

2.2.4 Transformations between different orders

When constructing the transition matrix of a $m^{th}$-order Markov chain, the compound states in (2.16a) and (2.16b) have to be compatible. This implies the matching of the symbols $c_{n-m+2}, \ldots, c_n$ in both $i$ and $j$. If no match is accomplished the corresponding matrix element will contain a structural zero. This causes only few elements to differ from zero, which represent the available transitions.
The method of redefining the state space from [2.2.1], as seen in (2.18), can be used to transform any Markov chain of order \( m \) to a first-order Markov chain [EFS98]. Due to extending the state space, the dimensions of the matrix grow as seen in (2.18). Likewise, the state vector contains the states of the new state space and therefore is extended to the same dimension.

If processes of different orders are combined, the transition matrices will have to be adjusted. In order to enable the evolution of the state vector by calculating the vector-matrix product in (2.21), the dimensions of all matrices and the state vector have to match. This is achieved by filling smaller matrices of lower order Markov chains with redundant rows.

For a Markov chain of order \( m' \) with \( m' < m \), all rows which represent \( m' \) identical last states consist of equal elements. This is obvious because for this chain only \( m' \) previous states have to be taken into account, ignoring the preceding \( m - m' \) ones. In other words, the next symbol \( c_{n+1} \) is independent of the symbols \( c_{n-m+1}, \ldots, c_{n-m'} \).

However, an adjustment of the matrices affects their product as well. By inserting redundant rows, the rank of the matrix decreases and is no full rank anymore. The reason is the definition of the rank coinciding with the maximal number of linearly independent rows of a matrix. Furthermore, the rank of the product of the matrices \( M \) and \( M' \)

\[
\text{rank}(MM') \leq \min(\text{rank}(M), \text{rank}(M'))
\]

(2.23)
cannot exceed the minimum rank of the two involved matrices [Gan77]. This inequality is not influenced by the commutativity of the matrices. Thus, if multiplying the matrices of a memoryless Markov chain and a first-order Markov chain, the resulting matrix will always have a rank corresponding to the memoryless system.

### 2.2.5 Properties

In addition to standard stochastic processes, Markov chains possess several specific properties. Some of them are important, in order to investigate other properties or derive analytical expressions.

Time-homogeneous Markov chains are characterized by

\[
p(C_{n+1} = x|C_n = y) = p(C_n = x|C_{n-1} = y) \quad \forall n \quad (x, y \in \mathcal{A}),
\]

(2.24)
claiming equality for the transition between identical states for all times \( n \). Thus, such a Markov chain can be described by a single time-independent transition matrix \( M \). However, it should be noted, that a variation between several time-homogeneous Markov chains with different matrices reintroduces a time-dependence to the process again.

According to [CT91], a Markov chain will be called irreducible if each state can be reached from any other state of the Markov chain within a finite number of transitions. This property becomes necessary for the existence of stationary distributions.
Furthermore, the aperiodicity of a Markov chain will occur if the largest common factor of the lengths of different transition paths from one state to itself is equal to one \cite{CT91}. This means, that a return of one state is not only encountered after numbers of transitions, which are a multiple of a natural number larger than one. The latter would be the case if the process were periodic.

These two properties suffice to prove ergodicity of a Markov chain. A Markov chain is said to be ergodic if it is both irreducible and aperiodic \cite{Cra68, Pak69}. Thus, a stationary distribution on the states can be found.

### 2.2.6 Stationary distribution

In order to obtain a stationary distribution on the states of the Markov chain, it has to comply with several requirements. An irreducible, aperiodic time-homogeneous Markov chain possesses exactly one stationary distribution, where all its components are positive \cite{vK07}. It is also called the steady state vector \cite{HL94, KGB87} or equilibrium distribution \cite{GH90, HS90} of the Markov chain.

Time-homogeneous Markov chains are described by one single time-independent transition matrix by applying it to the state vector according to (2.21). If vector $\mathbf{\pi}^*$ denotes the stationary distribution, it will be invariant to the multiplication with matrix $M$, leading to

$$\mathbf{\pi}^* = \mathbf{\pi}^* M$$

a stationary equation. Hence, vector $\mathbf{\pi}^*$ is a normalized left eigenvector of the Markov matrix $M$. The required eigenvector corresponds to the eigenvalue one. For stochastic matrices, the eigenvalue one always exists and is the largest of the system \cite{Dei91}.

Another approach to obtain the stationary distribution is the repeated multiplication with the matrix $M$ itself expressed as the matrix power with $k$ in (2.22). In other words, the system evolves to a stationary state over time, thus the long-term probabilities become independent of an initial distribution. If $k$ tends to infinity, the initial distribution vector will be mapped to a fixed point induced by the linear transformation associated to the matrix $M$. For any irreducible, aperiodic Markov process, this fixed point is a unique stationary distribution and denoted by $\mathbf{\pi}^*$. Simultaneously, by tending $k$ to infinity,

$$\lim_{k \to \infty} (M^k)_{ij} = (M^*)_{ij} = \mathbf{\pi}^*_j \quad \forall i$$

$M^k$ converges to a rank-one matrix $M^*$, which consists solely of identical rows. In $M^*$, each row is equal to the steady state vector $\mathbf{\pi}^*$, and thus, the Markov chain is independent of its initial distribution. These deductions are stated by the Perron-Frobenius theorem and are particularized and proofed in \cite{HJ85} and \cite{Sen06}.
2.2.7 Kolmogorov-Sinai entropy

For a deduction of the Kolmogorov-Sinai entropy for Markov chains, the two approaches via per-symbol entropy in \[2.1.3\] and conditional entropy in \[2.1.4\] can be pursued. Nevertheless, the definition of the block entropies \[2.3\] is necessary for both.

For a Markov chain of order \(m\), all block entropies \(H_n\) for block lengths greater or equal to the order of the Markov chain can directly be calculated by

\[
H_n = H_m + (n - m)(H_{m+1} - H_m) \quad (n \geq m)
\]  

\[2.27\]

according to [Gat72]. The interpretation is, if all correlations are included in the sub-sequence, the block entropy will always increase by the same amount \(H_{m+1} - H_m\) with increasing block length \(n\).

From \[2.5\] and \[2.27\], the source entropy

\[
h = \lim_{n \to \infty} \frac{H_n}{n} = \lim_{n \to \infty} \left( \frac{H_m}{n} + \frac{n}{n}(H_{m+1} - H_m) - \frac{m}{n}(H_{m+1} - H_m) \right)
\]

\[2.28\]

\[
= H_{m+1} - H_m
\]

can be deduced by means of per-symbol entropy and reveals a close correspondence to the conditional entropy \[2.6\]. This is in accordance to [Khi57a], that for any stationary and ergodic process, the per-symbol entropy will converge to the entropy of source \[2.5\] if \(n\) tends to infinity. However, in \[2.28\] no limit \(n \to \infty\) is necessary to get the entropy rate because \(h\) of a Markov source only depends on two specific block entropies. The block entropy of a block length equal to the order \(m\) of the Markov chain and the one, whose length exceeds the order \(m\) by one, determine the source entropy directly.

The dependence on \(H_{m+1}\) and \(H_m\) is obvious due to length of the memory of the Markov chain. If subsequences longer than the memory length of the process are cut out, not any gain of knowledge will be possible. This is because in those subsequences only correlations with the length of the memory exist and longer subsequences will not contain any additional information. In other words, more information has not been put into the symbol sequence by its generator.

In order to determine an analytical expression for the Kolmogorov-Sinai entropy of a Markov chain, the general conditional entropy \(\Delta H_n\) from \[2.13\] is used. The limit from \[2.8\] tending \(n\) to infinity can be neglected. The reason is \[2.14\], claiming the equality of both per-symbol and conditional entropy, together with \[2.28\], stating the per-symbol entropy’s independence of \(n\).
For a Markov chain of order \(m\), the entropy rate involves only the last \(m\) symbols generated by the source. Thus, (2.14) can be written as

\[
\Delta H_n = - \sum_{c_{n-m+1}, \ldots, c_{n+1}} p(c_{n-m+1}, \ldots, c_{n+1}) \log \lambda \ p(c_{n+1} | c_{n-m+1}, \ldots, c_n) \tag{2.29}
\]

depending on the symbols \(c_{n-m+1}\) to \(c_{n+1}\). The joint probability in (2.29) is expressed by

\[
p(c_{n-m+1}, \ldots, c_{n+1}) = p(c_{n-m+1}, \ldots, c_n) \cdot p(c_{n+1} | c_{n-m+1}, \ldots, c_n) \tag{2.30}
\]

the conditional probability generating the symbol \(c_{n+1}\) given a specific sequence \(c_{n-m+1}\) to \(c_n\).

In equivalence to (2.16), a new state space is formed by grouping the symbols together and identifying them by \(i\) and \(j\) \((i, j \in \mathbb{Z}^+)\). Moreover, the conditional probabilities \(p_{i \rightarrow j}\) are replaced by their corresponding element \(M_{ij}\) of the transition matrix \(M\). The joint probability \(p(c_{n-m+1}, \ldots, c_n)\) is represented by the steady state vector on the extended state space and hence is substituted by \(\vec{\pi}^*\).

Then (2.29) is simplified to

\[
h_{KS} = - \sum_{i, j} \vec{\pi}^*_i M_{ij} \log \lambda (M_{ij}) \tag{2.31}
\]

an analytical expression for the Kolmogorov-Sinai entropy of a Markov chain.

The formula for the Kolmogorov-Sinai entropy (2.31) is reasonable because it considers the transition probabilities between all states and weights them with the probability of occurrence of each state. The close similarity to the Shannon entropy (2.2) becomes apparent.

### 2.3 Hidden Markov chains

An extension to Markov chains is introduced by hidden Markov processes. In general, they consist of two stochastic processes, basically of a Markov chain with states and transition probabilities between the states. However, these states are not observable and remain hidden. Thus, an additional stochastic process is defined on the states, generating an observable output symbol in accordance to a state-dependent probability distribution. Due to this definition, these processes are called hidden Markov chains or functions of finite Markov chains [BK57].

In information and communication theory, hidden Markov chains are sometimes referred to as Markov chains observed in noise. Especially the reconstruction of the original signal from the reception of such a perturbed signal is of great importance. This led to the
2.3 Hidden Markov Chains

theory of hidden Markov models, which has been established by [BP66] and analyzes the output symbols to estimate the transition probabilities between the hidden states. A comprehensive introduction to hidden Markov models is given by [RJ86]. Besides, an overview of statistical and information-theoretic aspects of hidden Markov processes is presented in [EM02].

Based on the analysis of hidden Markov processes and their close relation to various phenomena in nature and technology further practical applications have been developed. In particular, hidden Markov models are applied in speech recognition [Rab89] and pattern recognition [DHS01] as well as in bioinformatics to analyze biological sequences [DEKM98, Edd96].

2.3.1 Definitions

Hidden Markov chains are bivariate discrete-time stochastic processes consisting of two random variables, which are connected. The unobservable part of a hidden Markov chain is defined analogously to an ordinary Markov chain in 2.2.1. This finite-state Markov chain $X$ generates a collection of random variables $\{X_n\}$, where $n \in \mathbb{Z}^+$ indicates the sequence of their occurrence. The random variables $X_n = x_n$ are elements of a countable state space comprising $I$ states $x_i = i$ with $i = 1, \ldots, I$. The transitions between these states are time-homogeneous and specified by a transition matrix $M$, which contains conditional probabilities in order to describe the dependence on previous states.

As a characteristic property, the states which belong to the ordinary Markov chain remain hidden and cannot be observed. Instead, these states are mapped to an output space $C$ consisting of $\lambda$ different symbols $c_l$ where $l = 1, \ldots, \lambda$. The state space $C$ is also denoted by a finite alphabet $\mathcal{A}$. The mapping is accomplished by a function $\Phi$, which in general is a stochastic process and generates an output symbol according to a state-dependent probability distribution. With

$$C_n = \Phi(X_n),$$

$\{C_n\}$ represents a sequence of independent random variables without direct dependence on its history but is instead conditioned solely on $X_n$.

The sequence of output symbols is used to gain some information about the distribution on the hidden states. Specifically, hidden Markov models for speech or pattern recognition apply this idea and estimate the original state. Depending on the given parameters, three canonical problems are associated with hidden Markov models, whose solution is accomplished by the forward-backward algorithm, the Viterbi algorithm, and the Baum-Welch algorithm, respectively, explained in [Rab89].
2.3.2 Probability distributions and stationary distribution

On the state space of the Markov chain contained in the hidden Markov chain, a probability distribution denoted by vector $\vec{\pi}^{(n)}$ is defined equivalently to \[2.2.3\]. In short, it represents the probability of occurrence of each state at a time step $n$. In addition to ordinary Markov chains, this probability distribution is mapped to the output symbols according to $\Phi$.

The underlying time-homogeneous Markov chain will have a stationary distribution if it is both irreducible and aperiodic, as mentioned in \[2.2.6\]. Consequently, the output generated by $\Phi$ obtains a stationary distribution as well. Thus, $C$ is a stationary process itself, which causes the entropy rate to be well defined.

2.3.3 Kolmogorov-Sinai entropy

The Kolmogorov-Sinai entropy of the hidden Markov chain is influenced by the underlying Markov chain as well as the mapping function $\Phi$. Actually, the sequence of the output symbols $\{C_n\}$ establishes a stationary process but does not form a Markov chain in general. Hence, an analytical deduction of the entropy rate is complicated, and its expression, like the one proposed by Blackwell \[Bl57\], remains difficult to evaluate. Quite recently, the analyticity of hidden Markov chains has been shown under positivity assumptions \[HM06\]. However, various approaches exist to calculate lower and upper bounds \[OW04\] of the entropy rate as well as solutions to specific hidden Markov chains such as binary hidden Markov processes \[JS04, ZK05\].

Due to the output sequence being a stationary process, both per-symbol entropy and conditional entropy tend to the same limit, resulting in the entropy rate of the associated process \[CT91\]. In particular, this becomes relevant for the numerical calculation of the Kolmogorov-Sinai entropy. However, even a calculation based on the conditional entropy $H(C_n|C_{n-1}, \ldots, C_1)$ of the output sequence can converge arbitrarily slow and not any knowledge is available, how close to the limit the computed value is.
3 Examples of stationary Markov processes

As an illustration, the following examples utilize the theoretical preliminaries from [2] concerning Markov processes and their Kolmogorov-Sinai entropy. For this purpose, two Markov chains with different orders are investigated. The Markov chain with order zero demonstrates a Bernoulli process and its connection to higher-order Markov chains. In contrast to this memoryless Markov chain, a first-order Markov chain is established. However, the confinement to those two Markov chains is without loss of generality because, as shown in [2.2.4], any higher-order Markov chain can be reduced to a first-order Markov process.

The two Markov chains will be reused in chapter [5] in order to generate a varying memory length. Hence, to ensure their compatibility, they are defined on the same state space \( \mathcal{A} \) comprising the symbols A and B causing \( \lambda = 2 \). Markov chains whose state space consists of only two symbols are often referred to as binary Markov processes [MFG93, Vas07].

Furthermore, both processes are irreducible and aperiodic, guaranteeing the existence of a stationary distribution. These properties are crucial to calculate the Kolmogorov-Sinai entropy.

3.1 First-order Markov chain

First-order Markov chains are the prevailing subclass of Markov processes. Their popularity originates from the fact that many common processes are of order one and any Markov chain of higher order can be transformed to a Markov process of order one. This is achieved by extending its state space as outlined in [2.2.4].

However, many processes in nature are often modeled by first-order Markov processes because their future state only depends on the current state. Thus, a long-term memory is not necessary and a transition matrix can directly be determined by observing the transitions between the states and their probability. A practical realization is given by a random walk on a finite one-dimensional lattice.

For a binary first-order Markov chain with states A and B, there exist four possible transitions. They are tabulated in table [3.1] and associated with their corresponding conditional probabilities describing the transitions.
Table 3.1: Overview of the possible transitions from any current state to a next state and their corresponding conditional probabilities for a first-order Markov chain. The current state influences the transition probability according to the next state.

Table 3.1 illustrates the different transitions from a current state to the next state. The row sum for each row yields one because a transition is performed with certainty from any state. If symbol A is the current state, the memory of the Markov chain will become relatively strong. Thus, the probability to regenerate A is higher than to change to state B. On the contrary, being on state B causes a weak but still apparent memory. In state B, it is more likely to switch to state A than to rest there. In conclusion, it seems to be obvious, that state A is preferred and occurs more likely in the long term.

The conditional probabilities from table 3.1 are transferred into the matrix

\[
M = \begin{pmatrix}
0.8 & 0.2 \\
0.6 & 0.4 \\
\end{pmatrix}
\]

which describes the first-order binary Markov chain completely. The matrix rank equaling two can be comprehended without effort. Due to the row sums being equal to one, \(M\) is called a stochastic matrix. By applying \(M\) to a current state vector \(\vec{\pi}(n)\), which comprises the probabilities of the states at time step \(n\), the vector-matrix product generates a future state vector \(\vec{\pi}(n+1)\).

In table 3.1 it is clearly recognizable that any state can be reached by any other one in a finite number of transitions. As described in 2.2.5, this property is called irreducibility. Analogously, the aperiodicity of \(M\) can be identified because the different transition paths from one state to itself can be of arbitrary length due to the existence of solely non-zero probabilities. As a direct consequence of these two properties, the Markov chain is said to be ergodic and there exists exactly one stationary distribution on its states.

The steady state vector \(\vec{\pi}^*\) corresponding to this first-order Markov chain is obtained by finding the left eigenvector of \(M\), which belongs to the eigenvalue one. This results in

\[
\vec{\pi}^* = \begin{pmatrix}
0.75 \\
0.25 \\
\end{pmatrix}
\]

and provides information on the long-term probabilities, the states are occurring with.
The same result will be found if matrix $M$ is repeatedly applied to itself. For sufficiently large powers of $M$, the numerically resulting matrix product

$$\lim_{k \to \infty} M^k = \begin{pmatrix} 0.75 & 0.25 \\ 0.75 & 0.25 \end{pmatrix}$$

(3.3)

consists solely of identical rows, which are equal to the stationary state vector, as formulated in (2.26).

For an interpretation of the steady state vector (3.2), the stationary transition matrix (3.3) offers a descriptive notation. By writing down the transition probabilities contained in (3.3) following table 3.1, the conditional probabilities

$$p(A|A) = p(A|B) = 0.75$$

(3.4a)

$$p(B|A) = p(B|B) = 0.25$$

(3.4b)

are detected to be in pairs identical. Obviously, they do not depend on the condition anymore. This drops the relevance of the current state and only the probabilities of the occurrence of symbol A and symbol B, respectively, remain. Thus, the long-term distribution of the probability mass is 0.75 on state A and 0.25 on state B. When using this first-order Markov chain as a generator of a symbol sequence, an infinitely long sequence will contain a portion of 0.75 of symbols A and 0.25 of Bs. This supports the conjecture, that symbol A is more likely generated than B.

With both transition matrix and stationary distribution given, the Kolmogorov-Sinai entropy of this first-order Markov process can be calculated. The necessary values are introduced into (2.31), resulting in

$$h_{KS} = -(0.75 \cdot 0.8 \log_2(0.8) + 0.75 \cdot 0.2 \log_2(0.2) + 0.25 \cdot 0.6 \log_2(0.6) + 0.25 \cdot 0.4 \log_2(0.4))'$$

(3.5)

which can be computed and finally rounded to a value of

$$h_{KS} = 0.78418$$

(3.6)

for the Kolmogorov-Sinai entropy.

### 3.2 Markov chain without memory

Discrete-time finite-state stochastic processes, which do not involve any previous states for generating the next state, can be regarded as Markov chains of order zero. Those Markov processes are said to be memoryless because they do not include any knowledge about their history.
The most popular representative of binary Markov chains without memory is the Bernoulli process. It consists of a sequence of independent random variables taking states from a state space which contains only two states. For instance, a coin tossing experiment provides a practical application for Bernoulli processes.

A Bernoulli process with states A and B only has two probabilities \( p(A) \) and \( p(B) \) for the occurrence of each state. In contrast to first-order Markov chains, native transitions do not exist because future states of Bernoulli processes lack their dependence on the current state. However, such dependence can be introduced artificially by establishing conditional probabilities, whose values in fact do not depend on the condition. Hence, the conditional probabilities for the same event but different conditions

\[
\begin{align*}
  p(A|A) &= p(A|B) = p(A) \quad (3.7a) \\
  p(B|A) &= p(B|B) = p(B) \quad (3.7b)
\end{align*}
\]

are equal and correspond to the probability of occurrence of each event. Subsequently, the transitions are stipulated in table 3.2 with their corresponding probability.

<table>
<thead>
<tr>
<th>current state</th>
<th>next state</th>
<th>( A )</th>
<th>( B )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>( p(A</td>
<td>A) = 0.4 )</td>
<td>( p(A</td>
</tr>
<tr>
<td>B</td>
<td>( p(B</td>
<td>A) = 0.4 )</td>
<td>( p(B</td>
</tr>
</tbody>
</table>

Table 3.2: Overview of the possible transitions from any current state to a next state and their corresponding conditional probabilities for a Markov chain without memory. The current state does not influence the transition probability according to the next state.

Table 3.2 exemplifies the different transitions from a current state to the next state. Again, the probabilities in each row sum up to one, claiming a transition from any state to occur certainly. Even the independence of the current state is obvious because each column contains the same values. This underlines the memorylessness of the Bernoulli process. Furthermore, in this example, it is slightly more likely to obtain state B than state A in the long term.

The conditional probabilities from table 3.2 are used to create the matrix

\[
M = \begin{pmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{pmatrix}
\]  

(3.8)

describing the Bernoulli process completely. With its row sum of one, it is a memoryless stochastic matrix containing solely redundant rows as depicted in 2.2.4. As a consequence, the matrix does not possess a full rank, but rather a rank of one. If \( M \) is applied...
to a current state vector $\vec{\pi}^{(n)}$ comprising the probabilities of the states, the vector-matrix product will create the next state vector $\vec{\pi}^{(n+1)}$.

Again, the irreducibility and the aperiodicity of the memoryless Markov chain can be verified based on table 3.2. Thus, this Bernoulli process is ergodic and possesses exactly one stationary distribution on its states.

The steady state vector $\vec{\pi}^*$ belonging to this Markov chain of order zero can be determined by the eigenvector, which corresponds to eigenvalue one. The resulting stationary distribution is given by

$$\vec{\pi}^* = \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix}$$ (3.9)

and represents the long-term probabilities of the states.

Besides, the stationary distribution can directly be deduced from the transition matrix without any effort because $M$ consists only of identical rows. This is the criterion for the stationary transition matrix, which otherwise has to be calculated by taking the matrix $M$ to the power of $k$ and tending $k$ to infinity. Caused by the memorylessness of the Bernoulli process, this is not necessary. In particular, any state vector is directly mapped to the steady state vector because a previous distribution on the states is not of any relevance for the transitions.

In summary, the stationary distribution on the states reveals a probability of 0.4 to be in state A and 0.6 to be in B. If this Bernoulli process is used to generate a symbol sequence, an infinitely long sequence will consist of a portion of 0.4 of symbols A and 0.6 of symbols B.

Then, the Kolmogorov-Sinai entropy of this Bernoulli process, resembling a Markov process of order zero, can be calculated analytically with the help of (2.31). The insertion of the stationary state vector and the transition probabilities leads to

$$h_{KS} = -(0.4 \cdot 0.4 \log_2(0.4) + 0.4 \cdot 0.6 \log_2(0.6)) + (0.6 \cdot 0.4 \log_2(0.4) + 0.6 \cdot 0.6 \log_2(0.6))$$ (3.10)

and can be computed and finally rounded to a value of

$$h_{KS} = 0.97095$$ (3.11)

for the Kolmogorov-Sinai entropy.

The resulting entropy rate of the Bernoulli process is close to one, thus its prediction is difficult. The reason is the structure of the matrix, involving the two probabilities $p(A)$ and $p(B)$ being very close to each other. This is similar to an unbiased coin toss, where the outcome of heads and tails is equally likely. For such a coin tossing experiment, the entropy rate amounts one, which can be computed easily. A Kolmogorov-Sinai entropy of one represents the highest possible entropy. Hence, a reasonable prediction is not possible for such a process.
4 Numerical calculations

In computer science and information theory, the estimation of entropy gained a crucial importance to reduce the size of data transmission. A reduction is achieved with various data compression methods. In particular, lossless compression methods are of broad interest, which are implemented by entropy encoding techniques, like Huffman coding [Huf52] and arithmetic coding [Ris76].

The lossless compression algorithms are based on the entropy of the data stream, which is represented by a sequence of symbols. Moreover, the best realizable compression efficiency is limited by the source entropy, which quantifies the unavoidable amount of randomness originating from the data source. Hence, a lower entropy rate of a symbol source allows a better compression and vice versa. Furthermore, the entropy can be used to measure the amount of similarity between data streams.

In order to estimate the entropy rate of a data stream, its symbol sequences have to be analyzed. According to 2.1, the block entropies (2.3) are determined by counting the occurrence of the different $n$-words. These absolute frequencies are converted to relative frequencies, with which the entropy for blocks of length $n$ is obtained. Then, the entropy rate can be calculated using both per-symbol entropy (2.4) and conditional entropy (2.6) by tending $n$ to infinity.

4.1 Implementation

The development of the program to analyze the block entropies and the entropy rate of arbitrary symbol sequences is split into two modules. The first part contains the symbol source and generates a sequence consisting of a predefined number of symbols selected from a specified alphabet. The characteristics of the source are implemented by adding desired correlations to the symbols. Thus, an uncorrelated Bernoulli source can be established as well as a strongly correlated periodic process. In general, each type of symbol generator is conceivable, for instance a Markov source with arbitrary memory length as well as even a variation between different sources. In addition, the text of a file can be put in and converted to a compatible source.

The second module analyzes the symbol sequence and calculates the entropies. In order to calculate the block entropies for length $n$, the analyzing algorithm cuts out subsequences
of this length $n$. For this purpose, the index of the starting symbol is increased beginning at the first position within the sequence. Hence, the next $n$-word is investigated. This continues until the subsequence reaches the end of the generated symbol sequence. In other words, a window of width $n$ is moved over the sequence extracting all $n$-words in succession as illustrated in figure 4.1.

\[ \begin{array}{ccccccccc}
\cdots & M^{(k)} & M^{(k+1)} & M^{(k+2)} & M^{(k+3)} & M^{(k+4)} & M^{(k+5)} & M^{(k+6)} & M^{(k+7)} & \cdots \\
\cdots & A & A & B & A & B & B & A & B & \cdots \\
\end{array} \]

(a)

\[ \begin{array}{ccccccccc}
\cdots & M^{(k)} & M^{(k+1)} & M^{(k+2)} & M^{(k+3)} & M^{(k+4)} & M^{(k+5)} & M^{(k+6)} & M^{(k+7)} & \cdots \\
\cdots & A & A & B & A & B & B & A & B & \cdots \\
\end{array} \]

(b)

\[ \begin{array}{ccccccccc}
\cdots & M^{(k)} & M^{(k+1)} & M^{(k+2)} & M^{(k+3)} & M^{(k+4)} & M^{(k+5)} & M^{(k+6)} & M^{(k+7)} & \cdots \\
\cdots & A & A & B & A & B & B & A & B & \cdots \\
\end{array} \]

(c)

Figure 4.1: Extraction of subsequences with a moving window. A window cutting out four symbols is moved over the sequence extracting all subsequences in succession. Then, the relative frequencies are determined and the desired entropy for blocks of length four can be calculated. Additionally, the figure shows the association of the symbols to the machine, from which they originated.

Next, the program counts the occurrences of the extracted $n$-words. The length $n$ of the longest word, which should be extracted, is supplied to the program. Then, the $n$ symbols of each word are used to construct a tree of depth $n$ by creating child nodes recursively corresponding to the symbols. Thus, each path to a leaf node represents a specific $n$-word, whereas shorter paths are identified by shorter words. In each involved node, a counter is incremented, counting the occurrence of the subsequences up to a length $n$.

The implementation of the trees is implemented quite memory efficiently. In general, each node contains as many child nodes as different symbols in the alphabet exist. Nevertheless, the creation of child nodes is restricted to those, which are induced by the
subsequence and therefore required. In summary, each node consists of a counter of data type long, and additionally, each non-leaf node has at least one pointer to a child node.

The probabilities of the words are estimated by their relative frequencies of occurrence within the symbol sequence. For this purpose, the tree is parsed recursively, and the absolute frequency of a subsequence is divided by the total number of extracted subsequences with the same length yielding their relative frequency. This relative frequency approximates the probability of occurrence and is used to calculate the contribution to the sum of the block entropy. The sums of the different block entropies are held in an array of length $n$ and all contributions are summed in their appropriate element. After the parsing is finished, the array contains the block entropies for block lengths from one to $n$.

According to (2.4), the block entropies are divided by the block length and yield the per-symbol entropy. Equivalently, the necessary block entropies are inserted into (2.6) in order to obtain the conditional entropies in dependence on $n$.

It should be noticed, that the determined entropies are independent of any initial symbol. Initial symbols have to be defined for various types of symbol sources, for instance Markov chains. Specifically, the memory of a Markov chain of order $m$ has to be filled with $m$ symbols initially. By moving the cutting window over the sequence, each symbol of the sequence serves as an initial condition. Hence, the observation results in a time average of the initial conditions and thus loses this influence.

Furthermore, the requirement of fairly long sequences should be remarked. The relative frequencies will only approximate the probability of occurrence well enough if a large amount of $n$-words is contained within the sequence. Consequently, the sequence necessitates a great number of symbols or alternatively the subsequences are not allowed to exceed a certain length. The insufficient approximation of the probabilities will appear in the block entropies noticeably.

4.2 Examples

The behavior of the entropies is determined numerically. For this purpose, the two classes of processes, represented by the first-order Markov chain from 3.1 and the memoryless Markov chain, denoted Bernoulli process, from 3.2 are investigated. Firstly, they are transformed into symbol generators, and later on, the various entropies are calculated. The resulting figures provide a detection of useful characteristics belonging to Bernoulli processes and Markov chains.
4.2 Examples

4.2.1 First-order Markov chain

To analyze the behavior of the entropies with respect to the word length $n$, a symbol generator is constructed, which creates a sequence corresponding to the first-order Markov chain from [3.1]. The transition probabilities in matrix (3.1) describe the selection process of the source to output a symbol based on the previously generated symbol.

With the Markov source, a sequence comprising one billion symbols is generated. In this sequence, blocks up to length $n = 24$ are cut out and the block entropies are computed. Subsequently, the per-symbol entropy $h_n$ as well as the conditional entropy $\Delta H_n$ with respect to $n$ is calculated.

Figure 4.2 depicts the behavior of both per-symbol entropy $h_n$ and conditional entropy $\Delta H_n$ with respect to $n$. The Kolmogorov-Sinai entropy $h_{KS} = 0.78418$, as determined in [3.1] is plotted as a straight line independent of $n$. The convergence of both numerically calculated entropies is obvious; however, there is a qualitative difference. The per-symbol entropy converges gradually to the limit with increasing $n$. On the contrary, the conditional entropy reaches the value of the entropy rate immediately at $n = 1$. In fact, this agrees with the explanations in [2.2.7] once if all correlations are included in the subsequence, the block entropy will increase by identical amounts. In other words, the uncertainty for the prediction of the next symbol can be reduced as long as not all correlations are contained in the subsequences. Hence, the convergence of the conditional entropy represents a characteristic property of Markov chains and the order of the Markov chain can be deduced by the position of the bend in the behavior of the conditional entropy.

Moreover, the problem concerning the insufficient approximation of the probabilities via relative frequencies becomes apparent at $n = 20$ in figure 4.2. The conditional entropy is at a constant level from $n = 1$ on and starts to decrease below the Kolmogorov-Sinai entropy at $n = 20$. This is caused by the length of the subsequences involving 20 symbols. Thus, more than one million ($2^{20}$) different variants of those words exist, being counted in a sequence which comprises solely one billion symbols. This relation gets even worse with increasing $n$, causes the probabilities to be inaccurate, and destroys the reliability of the numerical computation.

As a consequence of the asymmetrical difference in (2.6), it should be noted that the values of $h_1$ and $\Delta H_0$ are equal but are located at different positions $n$ due to their definition.

4.2.2 Markov chain without memory

The analysis of the entropies and their behaviors in dependence on the word length $n$ is accomplished by transforming the transition matrix (3.8) of the Bernoulli process into a
symbol generator. Then, the produced sequence possesses the characteristics of a memoryless Markov chain because the selection of the symbols occurs independently of previously generated symbols.

With the Bernoulli source, a sequence including one billion symbols is produced. Subsequences of block lengths up to \( n = 24 \) are extracted from the sequence computing their block entropies. Moreover, both per-symbol entropy \( h_n \) and conditional entropy \( \Delta H_n \) are deduced in dependence on \( n \).

Figure 4.3 illustrates the behavior of both per-symbol entropy \( h_n \) and conditional entropy...
4.2 EXAMPLES

Figure 4.3: Numerically computed per-symbol entropy $h_n$ and conditional entropy $\Delta H_n$ versus block length $n$ compared to analytically calculated Kolmogorov-Sinai entropy for a memoryless Markov chain. Both per-symbol and conditional entropy converge instantaneously to the analytical entropy rate. For word lengths $n \geq 20$, the conditional entropies become inaccurate due to insufficient approximation of the probabilities and differ from the Kolmogorov-Sinai entropy.

$\Delta H_n$ with respect to $n$. The Kolmogorov-Sinai entropy $h_{KS} = 0.97095$ calculated in 3.2 is plotted as a straight line independent of $n$. Both numerically computed entropies converge immediately to the analytical value of the entropy rate and depict a constant series. An explanation is given in 2.1.4 for Markov chains of order $m$, which can be applied to Bernoulli process with $m = 0$ as well. According to this, the block entropies increase by the same amount as soon as all correlations between the symbols are included in the subsequences. For a memoryless Markov process, this is accomplished with the first symbol because not any history is involved in the symbol creation process. In other words, the uncertainty for the prediction of the next symbol of a Bernoulli process cannot be lowered by knowing more than one symbol. Hence, the immediate convergence of both per-symbol
and conditional entropy is a typical property of Markov chains without memory called Bernoulli processes.

Besides, figure [4.3] reveals the expected problems regarding an inaccurate approximation of the probabilities via relative frequencies. Beginning with $n = 20$, a significant deviation from the Kolmogorov-Sinai entropy is obvious as the conditional entropy drops down the limiting entropy rate. Analogously to [4.2.1], exactly $2^{20}$ combinations of symbols forming words of length 20 exist, which is too much compared to the total length of the sequence comprising only one billion symbols. With increasing $n$, the error grows and no reliable numerical computation of the entropy can be achieved.
5 Variations of order

5.1 Analytical cases

A variation of the order within the Markov chain is achieved by consecutively selecting one of the transition matrices and multiplying it to the previous state vector. This selection, resulting in a specific matrix multiplication, can be done in a variety of ways. It has to be considered that the matrix multiplication is not commutative in general. Thus, beside the frequency of the matrices their sequence is significant for the resulting matrix product, which is applied to the initial state vector.

For the selection of the sequence, infinite many possibilities exist between the two special cases of being purely arbitrary without any correlations and being a strongly correlated periodic process. Furthermore, a Markov chain of any order is imaginable as well as even a Markov chain with fluctuating order. By all means, those special cases can be investigated analytically. This gives the opportunity to calculate dynamical characteristics directly by use of the transition probabilities from the applied matrices and other known parameters.

For an analytical purpose, the limit (2.8) is applied to (2.13) implying

\[ h_{KS} = - \lim_{n \to \infty} \sum_{c_1, \ldots, c_{n+1}} p(c_1, \ldots, c_{n+1}) \log \lambda (p(c_{n+1}|c_1, \ldots, c_n)) \]  

(5.1)

the Kolmogorov-Sinai entropy. The Kolmogorov-Sinai entropy represents the entropy rate of the associated process and thus is a measure for the amount of chaos emitted by the source of the sequence.

Due to the variation between several Markov chains, (5.1) is not sufficient. It is required to include the connection of the probabilities to the machines. Hence, a derivation analogously to (2.1.4) is necessary considering the \( n \) machines \( M^{(1)} \) to \( M^{(n)} \), which create the symbols \( c_2 \) to \( c_{n+1} \) in succession based on the initial symbol \( c_1 \). For this purpose, the joint probabilities are extended and contain the machines as well, establishing

\[ p(c_1, \ldots, c_{n+1}, M^{(1)}, \ldots, M^{(n)}) \]  

(5.2)

the probability to obtain the sequence \( c_1, \ldots, c_{n+1} \) in conjunction with the machine sequence \( M^{(1)}, \ldots, M^{(n)} \). Consequently, (2.9) is extended by replacing the joint probabilities
with \(5.2\) yielding

\[
\Delta H_n = - \sum_{c_1, \ldots, c_{n+1}} p(c_1, \ldots, c_{n+1}, M^{(1)}, \ldots, M^{(n)}) \log \lambda(p(c_1, \ldots, c_{n+1}, M^{(1)}, \ldots, M^{(n)}))
\]

\[
+ \sum_{c_1, \ldots, c_n} p(c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)}) \log \lambda(p(c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)}))
\]

\[(5.3)\]

the conditional entropy as a difference of the two block entropies \(H_{n+1}\) and \(H_n\). With the formal definition of the conditional probability

\[
p(c_1, \ldots, c_{n+1}, M^{(1)}, \ldots, M^{(n)}) = \frac{p(c_{n+1}, M^{(n)}|c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)}) p(c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)})}{p(c_1, \ldots, c_{n+1})}
\]

\[(5.4)\]

a transformation of (5.3) is achieved analogously to (2.11). Moreover, the simplification

\[
\sum_{c_{n+1}} p(c_{n+1}, M^{(n)}|c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)}) = 1
\]

\[(5.5)\]

is introduced, and due to opposite algebraic signs, the derivation yields the analytical expression

\[
h_{KS} = - \lim_{n \to \infty} \sum_{c_1, \ldots, c_{n+1}} p(c_1, \ldots, c_{n+1}, M^{(1)}, \ldots, M^{(n)}) \log \lambda(p(c_1, \ldots, c_{n+1}, M^{(1)}, \ldots, M^{(n-1)}))
\]

\[(5.6)\]

for the Kolmogorov-Sinai entropy with respect to the applied machine sequence.

In (5.6) the necessary probabilities are substituted by the appropriate transition probabilities of the specific process. Depending on the correlation properties, the sum can be simplified resulting in an explicit expression for calculating the entropy rate by means of the transition matrices. Consequently, there will not be any requirement to generate a long sequence of symbols from the source and calculate the entropy rate based on the relative frequencies of occurrence as described in 2.1.2. For specific generators, this will evade the problem of slow convergence of both source entropy and conditional entropy.

### 5.2 Periodic variations

One special case for varying the order of a Markov chain is given by a periodic variation. A periodic variation can be done in a variety of ways and with arbitrary many Markov chains of different orders.
Hereinafter, a periodic variation between two Markov chains of order one is investigated. The restriction to Markov chains with memory length one is not of any relevance because as mentioned in 2.2.4, the transition matrices of an arbitrary order can be transformed to order one. Furthermore, an extension to three or more underlying Markov chains can be deduced analogously to the derivation.

The two involved Markov chains are described by their corresponding transition matrices \( M(0) \) and \( M(1) \). It is often referred to these transition matrices as machines, which underlines the character of different sources generating a sequence of symbols.

A periodic variation embodies a strong correlation for the selection of the machines. Hence, not any random influence exists for determining the sequence of machines. The sequence of machines is completely predictable in contrast to the generated symbols, which contain the different randomness of the symbol generation on the machines.

### 5.2.1 Preliminaries

The periodic variation between both Markov chains leads to

\[
\hat{\pi}^{(0)} M(0) M(1) M(0) M(1) M(0) M(1) M(0) M(1) \cdots \tag{5.7}
\]

an infinite product of periodically alternating matrices applied to an initial state vector \( \hat{\pi}^{(0)} \). Caused by the non-commutative character of the matrix multiplication, a rearrangement of the matrix product for simplification is not possible.

However, the infinite periodic matrix product can be rewritten by grouping two matrices in pairs together. The product \( M(0) M(1) \) results in a new stochastic matrix \( M(01) \) representing a two-step transition matrix. This new machine \( M(01) \) evolves a state vector to a new state vector in the same way the application of matrix \( M(0) \) followed by \( M(1) \) does.

This grouping changes (5.7) to

\[
\begin{align*}
\hat{\pi}^{(0)} & M(0) M(1) \underbrace{M(0)}_{\hat{\pi}^{(0)}} M(1) \cdots \\
& \Rightarrow \hat{\pi}^{(0)} M(01) M(01) M(01) \cdots \tag{5.8}
\end{align*}
\]

an infinite product of matrix \( M(01) \), which is applied to the initial state vector \( \hat{\pi}^{(0)} \).

Moreover, there exists another way of grouping the matrices in (5.7) to form two-step transition matrices. The reason is the noncommutativity of matrix products, which gives different results depending on the sequence of the matrices the product is performed with. The product \( M(1) M(0) \) occurs likewise in the sequence of matrices resulting in a second
Variations of order

stochastic matrix $M(10)$. This new machine $M(10)$ generates a new state vector by applying $M(1)$ followed by $M(0)$ to the current state vector and hence accomplishes a two-step transition.

In rare cases, when $M(0)$ and $M(1)$ commutate, both matrices $M(01)$ and $M(10)$ are identical. In general, this is not the case and (5.7) can be written as

$$\tilde{\pi}(0) M(0) M(1) M(0) \cdots \Rightarrow \tilde{\pi}(0) M(10) M(10) \cdots \quad (5.9)$$

an infinite product of matrix $M(10)$, which is applied to a modified initial state vector $\tilde{\pi}(0)$.

Both machines $M(01)$ and $M(10)$ can be considered as independent Markov chains with altered transition matrices. Thus, the impact on an initial state vector by applying only one matrix in succession is the same as explained in [2.2.6] If matrix $M(01)$ and $M(10)$ are aperiodic and irreducible, there will exist a stationary distribution within the state vector for each machine. However, these stationary distributions are different between the two machines in general. This leads to a dualism between the two stationary distributions.

The dualism between the steady states is obvious when cutting subsequences of the same length out of a long sequence. The subsequence starting at the current index position involved basically a product of matrices $M(01)$ in the symbol generator, whereby a subsequence beginning at the next index position involved a product of matrices $M(10)$ and vice versa. This also holds for subsequences of odd lengths by using one matrix to modify the state vector and grouping the remaining matrices in pairs.

Due to the continuous alternation between the steady states, an analytical calculation of the Kolmogorov-Sinai entropy with (2.31) is not possible. The reason is that (2.31) is based on a stationary distribution on the available states, which cannot be satisfied by a periodic variation of different transition matrices. Particularly, there exists a fluctuation between the steady state vectors $\tilde{\pi}^\ast(M(01))$ and $\tilde{\pi}^\ast(M(10))$, which has to be examined.

5.2.2 Derivation of the Kolmogorov-Sinai entropy

The Kolmogorov-Sinai entropy of a process with periodic variation between two Markov chains can be deduced from (5.6). The conditional probability therein is transformed to

$$p(c_{n+1}, M^{(n)}|c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)}) = \frac{p(c_1, \ldots, c_{n+1}, M^{(1)}, \ldots, M^{(n)})}{p(c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)})}$$

$$= \frac{p(c_{n+1}|c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n)}) \cdot p(c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n)})}{p(c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)})} \quad (5.10)$$
an expression, which can further be simplified because of the deterministic machine selection. This is accomplished by regarding the deterministic machine selection as a first-order Markov process, whose transition probabilities equal either one

\[ p(M(0)|M(1)) = p(M(1)|M(0)) = 1 \] (5.11)

for changing the current machine or zero

\[ p(M(0)|M(0)) = p(M(1)|M(1)) = 0 \] (5.12)

for remaining on the machine.

As a result, the joint probability

\[ p(c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n)}) = p(c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)}) \cdot p(M^{(n)}|M^{(n-1)}) \] (5.13)

involves a conditional probability describing the selection of machine \( M^{(n)} \) after machine \( M^{(n-1)} \) was used. Then, the fraction in (5.10) is eliminated leading to

\[ p(c_{n+1}, M^{(n)}|c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)}) = p(c_{n+1}|c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n)}) \cdot p(M^{(n)}|M^{(n-1)}) \] (5.14)

being substituted into the argument of the logarithm of (5.6).

Without \( p(M^{(n)}|M^{(n-1)}) \), (5.14) represents the probability to generate a symbol \( c_{n+1} \) assuming that it has been preceded by a specific sequence of symbols \( c_1 \) to \( c_n \) being created on a given sequence of machines \( M^{(1)} \) to \( M^{(n)} \). In the investigated system, the generating of symbols is realized by first-order Markov chains reducing the dependence on the history to only one symbol. Thus, the conditional probability in (5.14)

\[ p(c_{n+1}|c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n)}) = p(c_{n+1}|c_n, M^{(n)}) = p_{M^{(n)}|c_n=c_{n+1} = M^{(n)}_{c_n=c_{n+1}}} \] (5.15)

is replaced by the transition probability on the current machine, which only requires one preceding symbol \( c_n \), instead of the whole subsequence.

Moreover, in (5.6) the joint probability of the symbols, which belong to the subsequence, in conjunction with machine sequence can be rewritten to

\[ p(c_1, \ldots, c_{n+1}, M^{(1)}, \ldots, M^{(n)}) = p(M^{(1)}, \ldots, M^{(n)}) \cdot p(c_1, \ldots, c_{n+1}|M^{(1)}, \ldots, M^{(n)}) \] (5.16)

based on the definition of the conditional probability.

The second factor on the right hand side of (5.16) represents the probability of the observation \( c_1 \) to \( c_{n+1} \) conditioned by the usage of machines \( M^{(1)} \) to \( M^{(n)} \). Because each
symbol only depends on the current machine,
\[ p(c_1, \ldots, c_{n+1} | M^{(1)}, \ldots, M^{(n)}) = p(c_1) \cdot p(c_2 | c_1, M^{(1)}) \cdot p(c_3 | c_2, M^{(2)}) \cdots p(c_{n+1} | c_n, M^{(n)}) \]
\[ = p(c_1) \cdot p_{M^{(1)}}(c_1 \rightarrow c_2) \cdot p_{M^{(2)}}(c_2 \rightarrow c_3) \cdots p_{M^{(n)}}(c_n \rightarrow c_{n+1}) \]
\[ = p(c_1) \cdot M^{(1)}_{c_1 \rightarrow c_2} \cdot M^{(2)}_{c_2 \rightarrow c_3} \cdots M^{(n)}_{c_n \rightarrow c_{n+1}} \]
(5.17)
is replaced by a product of conditional probabilities, which describe the dependence of the symbols on each other. Furthermore, these conditional probabilities can be expressed by the elements of the transition current matrix. Because of the characteristics of a first-order Markov chain, the probability of a symbol depends only on one preceding symbol.

The joint probability of the machines on the right hand side of (5.16) can be expressed by
\[
\begin{align*}
p(M^{(1)}, \ldots, M^{(n)}) &= p(M^{(1)}) \cdot p(M^{(2)} | M^{(1)}) \cdots p(M^{(n)} | M^{(n-1)}) \\
&= p(M^{(n)}) \cdot p(M^{(n-1)} | M^{(n)}) \cdots p(M^{(1)} | M^{(2)})
\end{align*}
\]
(5.18)
specifying the probability of the first machine multiplied by the conditional probabilities which describe the selection. Reversely, the probability of the last machine can be predefined and the conditional probabilities describe the transition to the preceding machine. As deduced in (5.11) and (5.12), the conditional probabilities for the machine selection are either one or zero, and hence, the sum over the machine sequences in (5.6) chooses only the ones compatible to a periodic selection. Any non-periodic machine selection involves at least one conditional probability equal to zero in the products of (5.18) and does not contribute to the sum. On the contrary, for a periodic selection each conditional probability is equal to one. In this case, the probability of the machine sequence
\[ p(M^{(1)}, \ldots, M^{(n)}) = p(M^{(1)}) = p(M^{(n)}) \]
(5.19)
depends only on the machine, at which the process is started or, alternatively, finished. In other words, (5.18) is simplified for the reason of the periodic variation being a deterministic process.

When generating a long sequence of symbols by periodic variation of two machines, naturally every second symbol has been created on the same machine. Similarly, if subsequences of same length \(n+1\) are cut out of this long sequence by moving the cutting window forward by one symbol as explained in 4.1, the first symbol will be generated on the same machine for every second subsequence. This fact is also applicable to the last symbol of the subsequence. Hence, the extraction of the subsequences out of a long sequence resembles a time average of the initial conditions, as explained in 4.2.1, causing the subsequences to start on each machine equally often. Equivalently, this holds for the last machine of the subsequence. This results in
\[ p(M^{(n)} = M(0)) = p(M^{(n)} = M(1)) = 0.5 \]
(5.20)
the initial condition for the machine selection.

The aforementioned conditional probabilities in (5.17) are represented by the appropriate transition probabilities on the machines, which generated the specific subsequence. The joint probability resolves to

\[
p(c_1, \ldots, c_{n+1} | M(0), \ldots, M^{(n)}) = p(c_1) \cdot M(0)_{c_1 \rightarrow c_2} \cdot M(1)_{c_2 \rightarrow c_3} \cdot M(0)_{c_3 \rightarrow c_4} \cdots \cdot M(0/1)_{c_n \rightarrow c_{n+1}},
\]

which will be the case if the generating of the symbols starts on machine \( M(0) \). The periodic variation of the Markov chains can be identified in (5.21) by the alternating usage of machines \( M(0) \) and \( M(1) \).

In addition to (5.21), when starting on machine \( M(1) \), the joint probability

\[
p(c_1, \ldots, c_{n+1} | M(1), \ldots, M^{(n)}) = p(c_1) \cdot M(1)_{c_1 \rightarrow c_2} \cdot M(0)_{c_2 \rightarrow c_3} \cdot M(1)_{c_3 \rightarrow c_4} \cdots \cdot M(1/0)_{c_n \rightarrow c_{n+1}}
\]

uses the transition probabilities on the other machines, respectively. Again, the alternation between the machines \( M(1) \) and \( M(0) \) is obvious.

In both (5.21) and (5.22), it depends on the length \( n + 1 \) of the subsequence, in order to determine the correct machine for the transition probabilities of the last symbol. This is illustrated by \( M(0/1) \) and \( M(1/0) \) and has to be considered in the next steps of the derivation.

Due to (5.19), the sum over the machine sequence in (5.6) can be reduced to a sum over the last machine \( M^{(n)} \). The reason is that only compatible machine sequences contribute to the sum, incompatible ones yield zero and are neglected. Hence, it is sufficient to sum over the last machine and simultaneously define all conditional probabilities to be equal to one in order to form a compatible sequence.

Because only two machines are involved, the sum over the last machine can be expanded and results in a sum of two sums over the symbol observation \( c_1 \) to \( c_{n+1} \). The first sum \( S(n,M(0)) \) deals with all subsequences of length \( n + 1 \), whose last symbol was generated on machine \( M(0) \), and the second one \( S(n,M(1)) \) with those generated on \( M(1) \), respectively. Each of the sum still contains the probability of the machine sequence from (5.16), which was simplified through (5.19) because the periodic machine selection is a deterministic process.

The expression for the Kolmogorov-Sinai entropy

\[
h_{KS} = - \lim_{n \to \infty} \left( p(M(0))S(n,M(0)) + p(M(1))S(n,M(1)) \right)
\]

(5.23)
gets the initial condition for the machine selection \((5.20)\) inserted and contains the sums

\[
S(n, M(0)) = \sum_{c_1, \ldots, c_{n+1}} p(c_1) \cdots M(1)_{c_1 \rightarrow c_n} \cdot M(0)_{c_n \rightarrow c_{n+1}} \log_2 (M(0)_{c_{n} \rightarrow c_{n+1}}) \quad (5.24a)
\]

\[
S(n, M(1)) = \sum_{c_1, \ldots, c_{n+1}} p(c_1) \cdots M(0)_{c_1 \rightarrow c_n} \cdot M(1)_{c_n \rightarrow c_{n+1}} \log_2 (M(1)_{c_{n} \rightarrow c_{n+1}}) \quad (5.24b)
\]

representing the two possible patterns of symbol creation in the subsequences. The term \(p(c_1) \cdots M(1)_{c_1 \rightarrow c_n}\) within \((5.24a)\) describes the joint probability of the symbols \(c_1\) to \(c_n\) taking account of the \(n^{th}\) symbol being generated by machine \(M(1)\). In \((5.24b)\) \(c_n\) is generated by machine \(M(0)\), respectively.

To determine the suitable sequence of machines ending in machine \(M(1)\), the length \(n\) of the sequence is significant. From this sequence, especially the machine, which generated the first symbol of the \(n\)-word is of interest. For this purpose, the transition part of the joint probability for \(n\) symbols is separated into \(n - 1\) conditional probabilities according to \((5.17)\). Consequently, a differentiation between odd and even word lengths has to be introduced and

\[
\cdots M(1)_{c_1 \rightarrow c_n} = \begin{cases} 
  M(0)_{c_1 \rightarrow c_2} \cdots M(0)_{c_{n-2} \rightarrow c_{n-1}} \cdot M(1)_{c_{n-1} \rightarrow c_n} & : n \text{ odd} \\
  M(1)_{c_1 \rightarrow c_2} \cdots M(0)_{c_{n-2} \rightarrow c_{n-1}} \cdot M(1)_{c_{n-1} \rightarrow c_n} & : n \text{ even}
\end{cases} \quad (5.25)
\]

consists of two cases. In the exact same manner, two cases exist

\[
\cdots M(0)_{c_1 \rightarrow c_n} = \begin{cases} 
  M(1)_{c_1 \rightarrow c_2} \cdots M(0)_{c_{n-2} \rightarrow c_{n-1}} \cdot M(0)_{c_{n-1} \rightarrow c_n} & : n \text{ odd} \\
  M(0)_{c_1 \rightarrow c_2} \cdots M(1)_{c_{n-2} \rightarrow c_{n-1}} \cdot M(0)_{c_{n-1} \rightarrow c_n} & : n \text{ even},
\end{cases} \quad (5.26)
\]

in which the last symbol of the \(n\)-word is produced by machine \(M(0)\).

In both \((5.25)\) and \((5.26)\) the distinction between the cases for even and odd length \(n\) can clearly be seen. While the last symbols are created on the same sequence of machines, the difference has to be made with the starting machine. In other words, the first machine of the periodic variation has to be chosen properly in order to evolve a sequence of machines ending with \(M(1)\) or \(M(0)\), respectively. Thus, in two words of length \(n\) and \(n + 1\) the sequence of machines, which generated the last \(n\) symbols, is identical.

The equality of the sequence of machines is a basis for further simplification. The machines are grouped together in pairs in backward direction beginning with the last machine. This can be done as a result of

\[
\sum_{c_k} M(0)_{c_{k-1} \rightarrow c_k} \cdot M(1)_{c_k \rightarrow c_{k+1}} = M(01)_{c_{k-1} \rightarrow c_{k+1}} \quad (5.27)
\]

summing over all intermediate symbols \(c_k\). Hence, \(M(01)_{c_{k-1} \rightarrow c_{k+1}}\) describes a two-step transition on matrix \(M(01)\). For this purpose, the product of matrix \(M(0)\) with \(M(1)\) is
combined to a new stochastic matrix $M(01)$ analogously to 5.2.1. Respectively, the compound matrix $M(10)$ is formed by the product of matrix $M(1)$ with $M(0)$ simplifying

$$
\sum_{c_k} M(1)_{c_k-1 \rightarrow c_k} \cdot M(0)_{c_k \rightarrow c_k+1} = M(10)_{c_k-1 \rightarrow c_k+1}
$$

(5.28)

for sequences terminated by machine $M(0)$.

Reapplying (5.27) to (5.25) and (5.28) to (5.26) and summing over the intermediate symbols $c_2$ to $c_{n-1}$ leads to

$$
(\ldots M(1))_{c_1 \rightarrow c_n} = \begin{cases} 
(M(01) \cdots M(01))_{c_1 \rightarrow c_n} & : n \text{ odd} \\
(M(1) \cdot M(01) \cdots M(01))_{c_1 \rightarrow c_n} & : n \text{ even}
\end{cases}
$$

(5.29)

and

$$
(\ldots M(0))_{c_1 \rightarrow c_n} = \begin{cases} 
(M(10) \cdots M(10))_{c_1 \rightarrow c_n} & : n \text{ odd} \\
(M(0) \cdot M(10) \cdots M(10))_{c_1 \rightarrow c_n} & : n \text{ even}
\end{cases}
$$

(5.30)

involving a product of several identical matrices. For odd $n$ the compound matrix is taken to the power of $n - \frac{1}{2}$ and for even $n$ the exponent is $n - \frac{2}{2}$ resulting in

$$
(\ldots M(1))_{c_1 \rightarrow c_n} = \begin{cases} 
(M(01)^{n-\frac{1}{2}})_{c_1 \rightarrow c_n} & : n \text{ odd} \\
(M(1) \cdot M(01)^{n-\frac{2}{2}})_{c_1 \rightarrow c_n} & : n \text{ even}
\end{cases}
$$

(5.31)

and

$$
(\ldots M(0))_{c_1 \rightarrow c_n} = \begin{cases} 
(M(10)^{n-\frac{1}{2}})_{c_1 \rightarrow c_n} & : n \text{ odd} \\
(M(0) \cdot M(10)^{n-\frac{2}{2}})_{c_1 \rightarrow c_n} & : n \text{ even}
\end{cases}
$$

(5.32)

as simplified expressions.

Both (5.31) and (5.32) are substituted back into the sums (5.24a) and (5.24b), to which the limit from (5.23) has to be applied. In the course of tending $n$ to infinity, the power of the matrix will yield its stationary transition matrix if the matrix is irreducible andaperiodic. This also affects the cases for even $n$ eliminating the first matrix due to the stationary transition matrix. It can be understood by the fact that a stationary transition matrix only consists of equal rows. This means that not any distinction is made, which the last state of the system was. Thus, when multiplying an arbitrary stochastic matrix with a matrix of equal rows the result will be the same matrix with equal rows again.

In the limit to infinity, the cases for even and odd $n$ in (5.31) and (5.32) become redundant and can be expressed by one stationary transition matrix each. If the last symbol of the $n$-word is generated on machine $M(1)$, the stationary transition matrix of $M(01)$ will be necessary and $M(10)$ will be used in case of machine $M(0)$.

Furthermore, the symbol $c_1$, which is the first symbol in the word, drops it relevance because a steady state is independent of initial conditions. This is obvious when regarding
the stationary transition matrix. As mentioned above, it only consists of equal rows and therefore has no dependence on the previous symbol. The matrix can be reduced to the stationary distribution on the states.

Additionally, the limit eliminates \( n \), thus \( c_n \) and \( c_{n+1} \) become two symbols from \( A \) named \( c \) and \( c' \). The steady state vectors are denoted by \( \pi^*(M(10)) \) for matrix \( M(10) \) and \( \pi^*(M(01)) \) for \( M(01) \). They are introduced into the sums (5.24a) and (5.24b), which only have to sum over \( c \) and \( c' \) due to the aforementioned simplifications of the limit. The sums can be reinserted into (5.23) resulting in an analytical expression of the Kolmogorov-Sinai entropy.

The Kolmogorov-Sinai entropy of a Markov chain with a periodic variation of two transition probability matrices can be calculated by

\[
\begin{align*}
    h_{KS} &= - \left( \frac{1}{2} \sum_{c,c'} \pi^*_c(M(10)) \cdot M(1)_{c \rightarrow c'} \log \lambda (M(1)_{c \rightarrow c'}) \\
           &\quad + \frac{1}{2} \sum_{c,c'} \pi^*_c(M(01)) \cdot M(0)_{c \rightarrow c'} \log \lambda (M(0)_{c \rightarrow c'}) \right) \\
\end{align*}
\]

(5.33)

analytically. It differentiates between the two possible patterns of machine usage and involves the stationary state of the compound matrices \( M(10) \) and \( M(01) \) as well as the transition probabilities of both machines \( M(0) \) and \( M(1) \).

### 5.2.3 Summary

An interpretation of (5.33) reveals the concepts behind this analytical solution. Firstly, the stationary distribution of the symbols \( c \) is taken for a sequence of machines ending with machine \( M(0) \). This is achieved by grouping the machines together to a two-step machine \( M(10) \). Thereafter, machine \( M(1) \) is the next one in this sequence and describes the transition from the current symbol \( c \) to the next symbol \( c' \). This has to be summed over all combinations of symbols \( c \) and \( c' \), which can occur. With this, a close connection to the Kolmogorov-Sinai entropy of first-order Markov chains (2.31) is apparent. In the same way, the other sequence of machines terminated by \( M(1) \) is taken into account. Finally, both sums are weighted according to their appearance, which is equal due to periodic variation.

The analytical derivation leading to solution (5.33) provides some aspects, which coincide with the reflections in 5.2.1. The approach of having two stationary distributions from \( M(10) \) and \( M(01) \) only had to be completed by the next step transition on the corresponding matrix and the weighting between the two steady states. Furthermore, simplifications enabled by the limit to infinity reduced the distinction of cases for odd and even lengths resulting in expressions similar to (2.31).
5.2.4 Example

The correctness of the analytical expressions derived in [5.2.2] can be examined by applying it to an example case of a Markov chain with periodic variation of memory. At first, the Kolmogorov-Sinai entropy can be determined by inserting the products of the transition matrices and their stationary distributions into the analytical formula (5.33). Then, a symbol sequence generator based on the given Markov chain with periodically varying order is implemented. Thus, the block entropies can be calculated numerically for various block lengths and an approximate computation of the entropy rate via both per-symbol entropy (2.4) and conditional entropy (2.6) can be done. Both ought to converge to one value, which can be compared to the analytically calculated Kolmogorov-Sinai entropy.

To exemplify a Markov chain with periodic variation of order, the first-order Markov chain from [3.1] is combined with the memoryless Markov chain from [3.2]. The periodic variation is achieved by alternating between those two processes as described in [5.2]. Thus, a product of periodically changing transition matrices is multiplied to an initial state vector.

To distinguish between the two processes, the matrix corresponding to the Bernoulli process, which is a Markov chain of order zero, is denoted by

\[ M(0) = M = \begin{pmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{pmatrix} \] (5.34)

and the matrix for the first-order Markov chain by

\[ M(1) = M' = \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} \] , (5.35)

respectively.

In order to calculate the Kolmogorov-Sinai entropy analytically based on (5.33), the matrix product comprising the two given transition matrices has to be determined for all combinations. The first combination results in the two-step transition matrix

\[ M(10) = M(1)M(0) = \begin{pmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{pmatrix} \] (5.36)

being equal to the involved matrix of the memoryless Markov chain. The reason for that fact is obvious. As mentioned in [3.2], any state vector \( \tilde{x} \) is mapped directly to the steady state vector \( \tilde{x}^* \) of the Bernoulli process. It ignores the previous state and, by implication, its creation with another matrix too.

The two-step transition matrix of the second available combination yields

\[ M(01) = M(0)M(1) = \begin{pmatrix} 0.68 & 0.32 \\ 0.68 & 0.32 \end{pmatrix} \] (5.37)
a different matrix compared to $M(10)$. Hence, the noncommutativity of the two matrices $M(0)$ and $M(1)$ is apparent.

Furthermore, the two-step transition matrix $M(01)$ consists solely of identical rows. This originates from the different ranks of the matrices as remarked in 2.2.4. Thus, a two-step transition on a Markov chain with variation of memory involving a memoryless Markov chain is always independent of the previous state vector.

For both matrices $M(10)$ and $M(01)$ exactly one stationary state vector exists because they represent irreducible, aperiodic processes as shown in 3.2. Due to the ranks of $M(10)$ and $M(01)$ being one, there is not any requirement to calculate the eigenvectors. The steady state vector $\pi^*$ of each two-step transition matrix can directly be extracted. This leads to the stationary state vectors

$$\pi^*(M(10)) = \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix} \quad (5.38)$$

corresponding to matrix $M(10)$ and

$$\pi^*(M(01)) = \begin{pmatrix} 0.68 \\ 0.32 \end{pmatrix} \quad (5.39)$$

for matrix $M(01)$ respectively.

The steady state vectors (5.38) and (5.39) and the transition matrix suffice to calculate the Kolmogorov-Sinai entropy. They are inserted into (5.33) resulting in

$$h_{KS} = -\left( \frac{1}{2} \right) \left( 0.4 \cdot 0.8 \log_2(0.8) + 0.4 \cdot 0.2 \log_2(0.2) \right. + 0.6 \cdot 0.6 \log_2(0.6) + 0.6 \cdot 0.4 \log_2(0.4)) \\
+ \left. \frac{1}{2} \left( 0.68 \cdot 0.4 \log_2(0.4) + 0.68 \cdot 0.6 \log_2(0.6) + 0.32 \cdot 0.4 \log_2(0.4) + 0.32 \cdot 0.6 \log_2(0.6) \right) \right), \quad (5.40)$$

which can be computed and finally rounded to a value of

$$h_{KS} = 0.92115 \quad (5.41)$$

for the Kolmogorov-Sinai entropy.

The behavior of both per-symbol entropy $h_n$ and conditional entropy $\Delta H_n$ with growing block length $n$ is illustrated in figure 5.1. Additionally, the analytically computed value of the Kolmogorov-Sinai entropy is drawn as a straight line being independent of $n$ because it is expected to be the limit.
5.2 Periodic variations

Figure 5.1: Numerically computed per-symbol entropy $h_n$ and conditional entropy $\Delta H_n$ versus block length $n$ compared to analytically calculated Kolmogorov-Sinai entropy for a periodic variation between a first-order Markov chain and a Bernoulli process. A different but slow convergence of each numerical value towards the analytical entropy rate is detected, which shows a noticeable difference even for subsequences consisting of 20 symbols.

Figure 5.1 was generated by creating a symbol sequence of one billion symbols. Afterwards, all blocks of symbols consisting of the same block length $n$ were cut out of the sequence. Their relative frequencies were determined, and hence, the block entropy for the specific block length was calculated. By means of the block entropies, both the per-symbol and the conditional entropy are computed and then plotted according to $n$.

The converging of both per-symbol entropy and conditional entropy towards the analytically calculated Kolmogorov-Sinai entropy can be anticipated as seen in figure 5.1. However, a quantitative difference is obvious regarding the speed of convergence. As mentioned in 2.1.3, the conditional entropy converges much faster compared to the per-symbol entropy. Nevertheless, even for subsequences of block length 20, the gap to the
Kolmogorov-Sinai entropy is noticeable. On the other hand, using larger \( n \)-words requires a far longer sequence in order to approximate the probabilities by relative frequencies, which are measured by the occurrence of the words. For \( n = 20 \), there exist \( 2^{20} \) distinct words already, which is more than one million.

The periodic variation between the two Markov chains influences the entropy rate of the new process. The Kolmogorov-Sinai entropy \( h_{KS} = 0.92115 \) of the Markov chain with periodic variation lies in between the values of those corresponding to the standalone Markov chains (0.78418 and 0.97095). However, it is not the mean value of the entropy rates.

### 5.2.5 Investigation of the convergence

The slow convergence of both per-symbol and conditional entropy discovered in figure 5.1 raises the question if their limit is indeed given by the analytical calculated value (5.41). An obvious approach is available by examining the convergence for its exponential decline.

For this purpose, the difference of the numerically computed entropies to the analytical Kolmogorov-Sinai entropy is determined and depicted in a semilogarithmic plot. In particular, the differences are plotted logarithmically against the block length \( n \) on a linear scale as illustrated in figure 5.2. A linear decrease in a semilogarithmic plot is identified by an exponential decline on an ordinary linear scale.

In figure 5.2 from \( n = 3 \) on, both differences apparently show a proper linear decrease confirming the exponential behavior. The individual slopes of both entropies reveal the speed of the exponential decline. Specifically, a larger absolute value of the slope implies a faster convergence of the exponential function and thus of the entropy rate. Hence, the conditional entropy reduces the gap to the Kolmogorov-Sinai faster, which was discussed in 5.2.4 and can obviously be seen in figure 5.1.

Figure 5.2 enables the extraction of the slope and the crossing with the \( y \)-axis corresponding to the linear equation

\[
g(x) = ax + b
\]  

in slope-intercept form. The fitting is accomplished in the range of \( 2 \leq n \leq 20 \) and yields the parameters

\[
a = -0.0360 \quad \text{and} \quad b = -2.521
\]  

for the per-symbol entropy and

\[
a = -0.0847 \quad \text{and} \quad b = -2.539
\]  

for the conditional entropy. Hence, the different slope is validated being steeper for the conditional entropy.
5.2 Periodic variations

Figure 5.2: Behavior of the differences from numerically computed per-symbol entropy \( h_n \) and conditional entropy \( \Delta H_n \) to the analytical Kolmogorov-Sinai entropy in a semilogarithmic plot versus linear block length \( n \) for a periodic variation between a first-order Markov chain and a Bernoulli process. Both entropies can be identified by linear functions confirming their exponential decline to the analytical value. The slope corresponding to the conditional entropy is steeper than for the per-symbol entropy, which illustrates the faster convergence of the conditional entropy.

The linear equation (5.42) has been established in the semilogarithmic plot and has to be transformed to a linear scale in order to be depicted into figure 5.1. For this purpose, (5.42) becomes the argument of the exponential function, which is shifted by the Kolmogorov-Sinai entropy, resulting in

\[
f(x) = e^{ax+b} + 0.92115 \quad (5.45)
\]
a function converging to the analytical entropy rate of the process.

The parameters in (5.45) are replaced by (5.43) and (5.44), respectively, and the exponential functions are plotted in conjunction with the evolutions of the numerical entropies as
Figure 5.3: Numerically computed per-symbol entropy $h_n$ and conditional entropy $\Delta H_n$ and their exponential fits versus block length $n$ compared to analytically calculated Kolmogorov-Sinai entropy for a periodic variation between a first-order Markov chain and a Bernoulli process. The fitted graphs agree well with the numerical values and confirm the exponential decline. The exponential functions are extrapolated up to a block length $n = 50$ and illustrate the convergence to the analytically determined Kolmogorov-Sinai entropy. For block length $n \geq 21$, the conditional entropy abandons the fitted function because the relative frequencies become inaccurate due to the growing number of subsequence combinations and the finite length of the sequence.

Furthermore, the numerical values are depicted for block lengths up to $n = 25$, and the exponential approximations extrapolate the entropies up to $n = 50$.

Figure 5.3 verifies the exponential decline of the numerical entropy rate with respect to $n$. The fitted exponential functions agree well with the computed values and simultaneously, by means of the extrapolation the convergence to the analytical Kolmogorov-Sinai entropy is confirmed.
Nevertheless, starting at \( n = 21 \), the computed conditional entropy leaves the exponential fit taking smaller values. The reason for this is discussed in [4.1] and originates from the relative frequencies, which do not approximate the probabilities of occurrence well enough due to the exponentially growing number of combinations for the subsequences. It issues from the numerical computations based on a sequence comprising only one billion symbols.

### 5.3 Random variations

Another special case for varying the order of a Markov chain is achieved by a random variation. Again, a random variation can be established in a variety of ways and with arbitrary many Markov chains of different orders.

As done in [5.2], a random variation between two Markov chains of order one is investigated. However, this limitation is not significant because any Markov chain of arbitrary order can be transformed to a memory length of one. This transformation by defining new states and grouping states together is described in [2.2.4] and yields transition matrices compatible to first order. Besides, the first derivation is done for a system with arbitrary many randomly fluctuating Markov chains and later on adjusted to two. Even the second deduction via hidden Markov chains can easily be extended to systems of three or more Markov chains with random selection.

During the derivation, the two contributing Markov chains are denoted by their related matrices \( M(0) \) and \( M(1) \) or by \( M(i) \) with \( i \in \mathbb{Z}^+ \) in a more general case with more than two matrices. The character of different sources generating a sequence of symbols is emphasized by referring to these transition matrices as machines.

The random variation does not include any correlations, the selection of the machines is a Bernoulli process. Thus, the sequence of machines is purely random and cannot be predicted. In contrast to periodic variation, knowledge about previously used machines does not reduce the uncertainty to predict future machines. The generated symbols contain this random in addition to the random of the symbol generation on the machines.

#### 5.3.1 Preliminaries

A random variation between the two Markov chains \( M(0) \) and \( M(1) \) leads exemplarily to

\[
\bar{\pi}(0)M(1)M(0)M(1)M(0)M(1)M(0)M(0)M(1)M(0)M(0)M(1)M(1)M(1)\cdots
\]  

(5.46)
an infinite product of randomly alternating matrices applied to the initial state vector $\vec{\pi}^{(0)}$. This random product

$$T = \prod_{k=1}^{n} M^{(k)}.$$  \quad (5.47)

requires the matrices to be multiplied on the right hand side. The index of the machine within the sequence is denoted by $k \in \mathbb{Z}^+$. Hence, the $k^{th}$ matrix, which is applied to the initial state vector $\vec{\pi}^{(0)}$, is called $M^{(k)}$. Machine $M^{(k)} \in \{M(0), M(1)\}$ uses an existing $k$-word $(c_1, \ldots, c_k)$ to generate symbol $c_{k+1}$.

There exist infinite possibilities forming such a product $T$ of matrices. The number of combinations grows exponentially with the number $n$ of applied matrices according to $2^n$. Because this value $n$ is in correspondence to the word length, an $n+1$-word can be created by $2^n$ combinations of machines. In general, a simplification rearranging the matrices in the product is not possible because the matrices do not commutate. Thus, beside the probability of their occurrence, the sequence of the matrices is significant.

Rewriting the matrices in groups like discussed in 5.2.1 is not possible for random variation. The reason is the expected existence of all combinations, whose number grows exponentially with the size of the group. Hence, the number of groups is indefinite and does not support an analytical solution based on grouped matrices.

Furthermore, the machines within the sequence can occur with different probabilities. These probabilities are denoted $p(M(i))$, where $i \in \mathbb{Z}^+$ is the index of the machine. In general, for the probability of occurrence, the constraint

$$\sum_i p(M(i)) = 1$$  \quad (5.48)

is given. Specifically, the examined process only consists of two machines reducing $i$ to $i \in \{0, 1\}$. The involved machines $M(0)$ and $M(1)$ are denoted by $M$ and $M'$, respectively. However, the analytical derivation permits an arbitrary number of machines.

The states which are obtained during the process are dependent on the sequence and frequency of the matrices. Each matrix lets the state vector converge exponentially to the stationary distribution of this matrix. The more often the same matrix is applied, the closer the state will tend to the stationary distribution of this matrix. Nevertheless, as soon as the other matrix is used, the state vector starts tending towards stationary distribution of this applied matrix.

If one of the Markov chains is a Markov chain of order zero, its matrix will have a rank of one. A matrix of rank one only consists of identical rows, the current state does not have influence on the next one. Thus, the state vector will change instantaneously instead of exponentially to the steady state if the matrix without memory is applied. For this reason, the system loses its dependence on the initial state once the Markov chain of order zero has been used.
5.3.2 Derivation of the Kolmogorov-Sinai entropy

The derivation of the Kolmogorov-Sinai entropy can be accomplished based on (5.6). The conditional probability therein is transformed to

\[ p(c_{n+1}, M^{(n)} | c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)}) = \frac{p(c_1, \ldots, c_{n+1}, M^{(1)}, \ldots, M^{(n)})}{p(c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)})} \]

an expression, which can further be simplified because of the uncorrelated machine selection. As a result, the joint probability

\[ p(c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n)}) = p(c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)}) \cdot p(M^{(n)}) \]

is a product of independent events, and the fraction in (5.49) is eliminated leading to

\[ p(c_{n+1} | c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)}) = p(c_{n+1} | c_n, M^{(n)}) \cdot p(M^{(n)}) \]

being substituted into the argument of the logarithm of (5.6).

Without \( p(M^{(n)}) \), (5.51) represents the probability to generate a symbol assuming that it has been preceded by a specific sequence of symbols being created on a given sequence of machines. In the examined system, the generating of symbols is realized by first-order Markov chains reducing the dependence on the history to only one symbol. Thus, the conditional probability in (5.51)

\[ p(c_{n+1} | c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n-1)}) = p(c_{n+1} | c_n, M^{(n)}) \]

is replaced by the transition probability on the current machine, which only requires one preceding symbol \( c_n \), instead of the whole subsequence.

Furthermore in (5.6), the joint probability of the symbols, which belong to the subsequence, in conjunction with machine sequence can be rewritten to

\[ p(c_1, \ldots, c_{n+1}, M^{(1)}, \ldots, M^{(n)}) = p(M^{(1)}, \ldots, M^{(n)}) \cdot p(c_1, \ldots, c_{n+1} | M^{(1)}, \ldots, M^{(n)}) \]

based on the definition of the conditional probability.

The second factor on the right hand side of (5.53) represents the probability of the observation \( c_1 \) to \( c_{n+1} \) conditioned by the usage of machines \( M^{(1)} \) to \( M^{(n)} \). Because each
symbol only depends on the current machine, the conditional probability in (5.53) is substituted by a product of transition probabilities

\[ p(c_1, \ldots, c_{n+1} | M^{(1)}, \ldots, M^{(n)}) = p(c_1) \cdot p(c_2 | c_1, M^{(1)}) \cdot p(c_3 | c_2, M^{(2)}) \cdots p(c_{n+1} | c_n, M^{(n)}) \]

\[ = p(c_1) \cdot M^{(1)}_{c_1 \rightarrow c_2} \cdot M^{(2)}_{c_2 \rightarrow c_3} \cdots M^{(n)}_{c_n \rightarrow c_{n+1}} \]

(5.54)

expressed by the elements of the transition matrices \( M \). The product in (5.54) describes the dependence of the symbols on each other forming a word of length \( n + 1 \). Because of the random variation between different machines \( M(i) \in \mathcal{M} \), (5.54) considers for all \( 1 \leq k \leq n \), on which machine \( M^{(k)} \) the \( k \)th symbol has been created. Thus, it represents the desired random product of matrices.

Naturally, due to all combinations of \( n \) machines for a \( n + 1 \)-word, the joint probabilities have to be summed up. Moreover, the joint probabilities have to be weighted by the probability, with which the given sequence of machines \( M^{(1)}, \ldots, M^{(n)} \) occurs. These obvious facts are already contained in (5.6)

\[ \sum_{M^{(1)}, \ldots, M^{(n)}} p(c_1, \ldots, c_{n+1}, M^{(1)}, \ldots, M^{(n)}) \log_{\lambda} (p(M^{(n)}) \cdot M^{(n)}_{c_n \rightarrow c_{n+1}}) \]

\[ = \sum_{M^{(1)}, \ldots, M^{(n)}} p(M^{(1)}, \ldots, M^{(n)}) \cdot p(c_1) \cdot M^{(1)}_{c_1 \rightarrow c_2} \cdots M^{(n)}_{c_n \rightarrow c_{n+1}} \log_{\lambda} (p(M^{(n)}) \cdot M^{(n)}_{c_n \rightarrow c_{n+1}}), \]

(5.55)

after separating the sums and inserting (5.53) and (5.54).

Due to an uncorrelated selection of the machines, the joint probability of the machine sequence in (5.53) and (5.55) consists solely of independent events and can be expressed as

\[ p(M^{(1)}, \ldots, M^{(n)}) = p(M^{(1)}) \cdot p(M^{(2)}) \cdots p(M^{(n)}) \]

(5.56)

the product of the probabilities belonging to each event.

This enables a rewriting of the elements on the right hand side of the sum (5.55). The probabilities of the occurrence of a machine are grouped to the transition probability on this machine. Additionally, the sums can be separated and yield

\[ \sum_{M^{(1)}, \ldots, M^{(n)}} p(c_1, \ldots, c_{n+1}, M^{(1)}, \ldots, M^{(n)}) = \]

\[ p(c_1) \cdot \left( \sum_{M^{(1)}} p(M^{(1)}) \cdot M^{(1)}_{c_1 \rightarrow c_2} \right) \cdots \left( \sum_{M^{(n)}} p(M^{(n)}) \cdot M^{(n)}_{c_n \rightarrow c_{n+1}} \right), \]

(5.57)
5.3 Random variations

A product of sums. In each sum, \( M^{(k)} \) can be any machine of \( \mathcal{M} \), and hence, all sums are of the same structure.

Induced by the similar sums over all machines in each part of (5.57), it can be simplified by means of

\[
\sum_{M(i)} p(M(i)) \cdot M(i)_{c_k \rightarrow c_{k+1}} = E_{c_k \rightarrow c_{k+1}} \tag{5.58}
\]

describing a transition from symbol \( c_k \) to \( c_{k+1} \) on a new machine \( E \). The transition matrix can be deduced from (5.58) by taking all elements \( c_k \) and \( c_{k+1} \) into account. This leads to

\[
E = \sum_{M(i)} p(M(i)) \cdot M(i) \tag{5.59}
\]

a simple weighting of the matrices, which are summed up in order to get matrix \( E \). Due to this construction, \( E \) can be called the effective matrix of the process, set up by random variation of first-order Markov chains. The effective matrix \( E \) itself is a stochastic matrix with Markov property and represents an effective first-order Markov chain.

Finally, the transitions (5.58) on the effective matrix are substituted back into (5.57) and therefore are used to transform (5.6) to

\[
h_{KS} = - \lim_{n \to \infty} \sum_{c_1, \ldots, c_{n+1}} p(c_1) \cdot E_{c_1 \rightarrow c_2} \cdot E_{c_2 \rightarrow c_3} \cdots E_{c_{n-1} \rightarrow c_n} \cdot E_{c_n \rightarrow c_{n+1}} \cdot \log \lambda (E_{c_n \rightarrow c_{n+1}}), \tag{5.60}
\]

which leaves the summation over all combinations of subsequences of length \( n + 1 \).

Due to the transitions being on the effective matrix \( E \) without any exception, the \( n-1 \)-step transition from symbol \( c_1 \) to \( c_n \) can be calculated by taking matrix \( E \) to the power of \( n-1 \). This is achieved by summing over the intermediate symbols \( c_2 \) to \( c_{n-1} \), which eliminates them. In the resulting expression

\[
h_{KS} = - \lim_{n \to \infty} \sum_{c_1, \ldots, c_{n+1}} p(c_1) \cdot (E^{n-1})_{c_1 \rightarrow c_n} \cdot E_{c_n \rightarrow c_{n+1}} \cdot \log \lambda (E_{c_n \rightarrow c_{n+1}}), \tag{5.61}
\]

the limit tends \( n \) to infinity. As a consequence, \( E^{n-1} \) will become the stationary transition matrix if \( E \) is aperiodic and irreducible.

Furthermore, the first symbol \( c_1 \) of the subsequence loses its influence because every steady state is independent of initial conditions. This is comprehensible by means of the stationary transition matrix. It only contains identical rows, which reflects the independence of the previous symbol. Then, the matrix can be expressed by the stationary distribution on the states.

The limit changes \( c_n \) and \( c_{n+1} \) to two symbols from \( \mathcal{A} \) denoted by \( c \) and \( c' \). The steady state vector of matrix \( E \) is named \( \vec{\pi}^*(E) \). The sum has to include \( c \) and \( c' \) to ensure the correct transitions to the last symbol.
The Kolmogorov-Sinai entropy of a Markov chain with a random variation of transition probability matrices can be calculated by

\[ h_{KS} = - \sum_{c,c'} \pi^*_c(E) \cdot E_{c \rightarrow c'} \log \lambda(E_{c \rightarrow c'}) \] (5.62)

analytically. The various involved transition matrices and their probabilities of occurrence are responsible for the effective transition matrix. As expected, (5.62) is in correspondence to (2.31). This supports the conclusion that \( E \) is a first-order Markov chain itself.

Actually, the examined system involves only two matrices \( M \) and \( M' \) occurring with \( p(M) \) and \( p(M') = 1 - p(M) \). In this special case, (5.59) yields

\[ E = p(M) \cdot M + (1 - p(M)) \cdot M' \] (5.63)

for the effective transition matrix. With this effective matrix, the Kolmogorov-Sinai entropy is calculated by using (5.62) as stated above.

### 5.3.3 Alternative derivation of the Kolmogorov-Sinai entropy and correspondence to hidden Markov chains

The derivation of the Kolmogorov-Sinai entropy based on (5.6), as it had been done for periodic variation of the machines, raised difficulties for randomly varying machines in 5.3.2. The limit in (5.6), which tends \( n \) to infinity, caused infinite combinations of machines and thus complicated a simplification of the sum. In the periodic case, only two combinations are contained even in infinitely long sequences of machines. This enabled the sums (5.24a) and (5.24b) in (5.23) to be substituted by (5.31) and (5.32).

Another approach with a redefined state space will avoid to sum over infinite combinations of machines. Therefore, the probabilities are rewritten to consist of two distinct state spaces. The first state space contains the symbols that are generated, which is in accordance to 5.2.2. The second state space comprises the machines generating the symbols. A joint transition probability combines the transitions of both state spaces forming two state transition probabilities.

The two state transition probabilities

\[ p(c_{n+1}, M^{(n+1)}|c_1, \ldots, c_n, M^{(1)}, \ldots, M^{(n)}) \] (5.64)

accomplish two actions, one on every state space. (5.64) represents the probability to generate symbol \( c_{n+1} \) preceded by the word \( c_1, \ldots, c_n \) and to change simultaneously to machine \( M^{(n+1)} \) after machine sequence \( M^{(1)}, \ldots, M^{(n)} \) occurred. It should be annotated that the creation of symbol \( c_{n+1} \) is accomplished on machine \( M^{(n)} \) as discussed in 5.3.1.
5.3 Random variations

The machine $M^{(n+1)} \in \{M, M'\}$ is to be selected independently of previous machines according to a Bernoulli process. Thus, the conditional probabilities for the transition of the machines can be replaced by the appropriate probabilities of occurrence of the machines. However, machine $M^{(n)}$ cannot be neglected because symbol $c_{n+1}$ is created on this machine. For this purpose, machine $M^{(n)}$ remains in the conditional probability.

Furthermore, each machine is identified by a first-order Markov chain. The creation of the next symbol only depends on the current symbol in the sequence. Hence, the $n$ preceding symbols become irrelevant and only the last symbol of the sequence is important.

When collecting these facts, (5.64) can be reduced to

$$p(c_{n+1}, M^{(n+1)} | c_n, M^{(n)}) (5.65)$$

and simplified to

$$p(c_{n+1} | c_n, M^{(n)}) \cdot p(M^{(n+1)}) = M_{c_n \rightarrow c_{n+1}}^{(n)} \cdot p(M^{(n+1)}), (5.66)$$

which takes the independence of the machine selection into consideration. The term $M_{c_n \rightarrow c_{n+1}}^{(n)}$ symbolizes the generation of symbol $c_{n+1}$ on machine $M^{(n)}$ when the current symbol is $c_n$.

The introduction of the redefined state space and the simplification of the transition probabilities reveals a strong similarity to hidden Markov chains. Figure 5.4 illustrates two unconnected first-order Markov chains with binary output A and B denoted by the machines $M$ and $M'$. On each machine, the transition probabilities are only related to the current symbol. The machines occur with their associated probabilities $p(M)$ and $p(M')$. The confinement to two output symbols does not restrict an analytical derivation but can be generalized to arbitrary many output symbols.

The Bernoulli process for the selection of the next machine provides the possibility to connect the machines among each other. In order to join both machines, at least one transition from a symbol on one machine to a symbol on the other machine and vice versa has to be present. Because these transitions do not depend on previous machines, they can be introduced as ordinary independent transitions. Thus, the conditional probabilities continue to be solely determined by the current symbol. The transitions between the machines are expressed by multiplying their probability of occurrence to these conditional probabilities.

The connection of the machines with its transitions between the symbols and the two first-order Markov chains is exemplified in figure 5.5. Furthermore, the adjustment of the transition probabilities can be recognized, and the switching between the machines during the process is obvious.
Variations of order

Figure 5.4: Two autonomous first-order Markov chains with binary output. The transition probabilities on each machine are solely related to the current symbol. The machines are selected with probability \( p(M) \) and \( p(M') \).

The connected system in figure 5.5 produces the same symbols on different machines. If only these output symbols can be observed, the applied machines will be unidentified. Thus, the machines resemble an internal state of the system, which is, however, significant for the output. This establishes a hidden Markov chain with twice as many internal states as output symbols. Accordingly, four internal states and two output symbols exist in figure 5.5.

Internal states consist of both machine and output symbol \( (M^{(k)}c_k) \), which confirms a factor of two between the number of internal states and output symbols for the examined system. This is consistent to the reflections that led to (5.64). In figure 5.5 the internal states are \( MA, MB, M'A, \) and \( M'B \), the output symbols are A and B. In principle, transitions between all internal states are possible.

In this hidden Markov chain, the internal states are mapped to the output symbols in a straightforward way. Each internal state is assigned to only one output symbol. The internal states \( MA \) and \( M'A \) generate symbol A and the other internal states \( MB \) and \( M'B \) output symbol B. Hence, it is a special case of a hidden Markov chain but does not
Figure 5.5: Two coupled first-order Markov chains with binary output for random variation. Due to the random variation, transitions between all states are possible. The transition probabilities on each machine are related to the current symbol \(A\) or \(B\) as well as to the current machine \(M\) or \(M'\). The conditional probabilities are multiplied by the probabilities of machine occurrence \(p(M)\) and \(p(M')\), respectively.

introduce restrictions to the analytical derivation.

A stationary distribution among the machines will be established if the hidden Markov chain generates \(n\)-words and \(n\) tends to infinity. However, on the symbols within each machine, there will not be gained such a stationary distribution. The reason is the exponential convergence to this distribution, which requires the continued applying of one machine. As a matter of fact, the distribution among the machines suffices to deduce the necessary transitions \(M_{c_n}^{(n)} \rightarrow c_{n+1}\) in (5.66). In particular, the stationary distribution among the machines is equal to the corresponding probabilities of the machine selection \(p(M)\) and \(p(M')\). This originates from the selection being a Bernoulli process.

Based on figure 5.5, the transitions between the output symbols can be written down by using the knowledge of the transition between the internal states. For the transition from
symbol A to symbol A, the conditional probability

\[ p_{A \rightarrow A} = p(M) \cdot (M_{A \rightarrow A} \cdot p(M) + M_{A \rightarrow A} \cdot p(M')) + p(M') \cdot (M'_{A \rightarrow A} \cdot p(M) + M'_{A \rightarrow A} \cdot p(M')) \]  

is determined. It is calculated by the probability to be on machine \( M \) and multiplies the transition on machine \( M \) from A to A with the probability to stay on machine \( M \) as well as with the probability to switch to machine \( M' \), respectively. Then the result for the case to be on machine \( M' \) is added.

Analogously to (5.67), the transitions between the other output symbols can be deduced. However, this is not necessary because the transitions between the internal states involve only those states whose output symbols match, two as seen in (5.67). Hence, (5.67) can be generalized to

\[ p_{c_k \rightarrow c_{k+1}} = p(M) \cdot (M_{c_k \rightarrow c_{k+1}} \cdot p(M) + M_{c_k \rightarrow c_{k+1}} \cdot p(M')) + p(M') \cdot (M'_{c_k \rightarrow c_{k+1}} \cdot p(M) + M'_{c_k \rightarrow c_{k+1}} \cdot p(M')) . \]  

The parenthesized expressions can be factorized, yielding the sum

\[ p(M) + p(M') = 1, \]  

which simplifies (5.68) to

\[ p_{c_k \rightarrow c_{k+1}} = p(M) \cdot M_{c_k \rightarrow c_{k+1}} + p(M') \cdot M'_{c_k \rightarrow c_{k+1}}. \]  

In summary, the elements of the matrix for each transition are weighted with the probability of occurrence of the matrix resulting in a new matrix with modified transition probabilities. Hence, considering \( p(M') = 1 - p(M) \), the new effective matrix is

\[ E = p(M) \cdot M + (1 - p(M)) \cdot M' \]  

representing an ordinary first-order Markov chain. In order to calculate the Kolmogorov-Sinai entropy for the effective machine \( E \), the methods from 2.2.6 and 2.2.7 can be reused.

5.3.4 Summary

The analytical deduction of the Kolmogorov-Sinai entropy involved a sum over all possible combinations for the sequence of the machines. Due to their independent selection via a Bernoulli process, a factorization could be done, and the sum was split into individual terms. The similarity of these terms provided a simplification, which led to (5.62) in the end.
The derived expression for the Kolmogorov-Sinai entropy is identical to the one of a conventional first-order Markov chain. The identity is obvious when determining an effective first-order Markov chain from the transition matrices of the varying process and their probabilities of occurrence according to (5.59).

The second approach to an analytical derivation did not include the combinations of machines but instead combined the two existing state spaces of symbols and machines. After connecting the states, the system could be identified by a hidden Markov chain with internal states. The transitions between the output symbols could be written down based on the internal transitions. The result confirmed the first approach.

To summarize, the transformation into a hidden Markov chain simplified the derivation and revealed the close relation to Markov chains with random variation of the transition matrix. Thus, this special case of a hidden Markov chain, where each internal state is mapped to only one output state, can be translated into an effective first-order Markov chain. Moreover, the effective first-order Markov chain can be transformed into a process with random variation between other first-order Markov chains.

### 5.3.5 Example

An evaluation of the analytical expressions derived in [5.3.2](#) can be accomplished by an example case using a Markov chain with random variation of order. For this purpose, the Kolmogorov-Sinai entropy can be calculated by the help of (5.62) with an effective first-order Markov chain and its stationary distribution. Besides, the parameters of the Markov chain with randomly fluctuating memory length are utilized to implement a generator for symbol sequences. Within the sequence, symbol blocks with various block lengths are analyzed in order to approximate the entropy rate via both per-symbol entropy (2.4) and conditional entropy (2.6). Both entropies are expected to converge to the analytically computed Kolmogorov-Sinai entropy.

The realization of a Markov chain with random variation of memory is achieved by connecting the first-order Markov chain from [3.1](#) and the memoryless Markov chain from [3.2](#). Alternating both processes arbitrarily without any interdependence accomplishes the random variation as explained in [5.3](#). Hence, an unpredictable product of matrices is multiplied to an initial state vector.

The matrix of the involved Bernoulli process, which resembles a Markov chain of order zero, is identified by

\[
M(0) = M = \begin{pmatrix}
0.4 & 0.6 \\
0.4 & 0.6
\end{pmatrix},
\]  

(5.72)
and the matrix for the first-order Markov chain by

\[ M(1) = M' = \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} \]  

respectively, in order to differentiate between both processes.

Equally important, the random selection of the transition matrices is influenced by the frequencies, they are occurring with. Thus, a probability for each matrix is introduced taking into account that they have to sum up to one. In this example, matrix \( M(0) \) is selected with a probability of

\[ p(M(0)) = 0.3 \]  

and \( M(1) \) with

\[ p(M(1)) = 0.7, \]

respectively complying with the constraint of the sum being equal to one.

With the probabilities of the matrices and the matrices itself, it is possible to determine the effective Markov chain according to (5.59). The matrix of this effective Markov chain is denoted by \( E \) and yields

\[ E = 0.3 \cdot \begin{pmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{pmatrix} + 0.7 \cdot \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} = \begin{pmatrix} 0.68 & 0.32 \\ 0.54 & 0.46 \end{pmatrix} \]  

in this example.

The effective Markov matrix \( E \) is indeed a stochastic matrix and possesses a full rank of two. Hence, it is comparable to matrices of first-order Markov chains. Moreover, the properties like irreducibility and aperiodicity pertain, and thus, exactly one stationary distribution vector exists. The exact calculation of the left eigenvector corresponding to the eigenvalue one results in

\[ \vec{\pi}^*(E) = \begin{pmatrix} \frac{27}{43} \\ \frac{16}{43} \end{pmatrix} \]  

and is a measure for the long-term probabilities of the states’ occurrences.

With the effective transition matrix and its steady state vector, all necessary parameters are available in order to calculate the Kolmogorov-Sinai entropy with (5.62). The insertion of the values leads to

\[ h_{KS} = -\left( \frac{27}{43} \cdot 0.68 \log_2(0.68) + \frac{27}{43} \cdot 0.32 \log_2(0.32) \\
+ \frac{16}{43} \cdot 0.54 \log_2(0.54) + \frac{16}{43} \cdot 0.46 \log_2(0.46) \right) \]  

and is computed as well as finally rounded to a value of

\[ h_{KS} = 0.93824 \]
for the Kolmogorov-Sinai entropy.

The behavior of both per-symbol entropy $h_n$ and conditional entropy $\Delta H_n$ depending on the block length $n$ is illustrated in figure [5.6]. In addition, the Kolmogorov-Sinai, which is determined analytically, is plotted as a straight line independent of $n$ and indicates the limit of the entropy rate.

Figure [5.6] was produced by generating a symbol sequence of three billion symbols. Thereafter, all blocks involving the same number $n$ of symbols were cut out of the sequence. The program determines their relative frequencies and calculates the block entropy for the specific block length. Both per-symbol entropy and conditional entropy are deduced calculationally from the block entropies and depicted in accordance to $n$. 

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5.6.png}
\caption{Numerically computed per-symbol entropy $h_n$ and conditional entropy $\Delta H_n$ versus block length $n$ compared to analytically calculated Kolmogorov-Sinai entropy for a random variation between a first-order Markov chain and a Bernoulli process. The per-symbol entropy converges confidently towards the analytical entropy rate, which is reached by the computed conditional entropy immediately due to the effective first-order Markov character.}
\end{figure}
For the convergence of the numerical entropies, a significant difference can clearly be seen. The per-symbol entropy aims at the limit given by the Kolmogorov-Sinai entropy slowly but steadily. For a block length of \( n = 20 \), only a small deviation remains. On the contrary, the conditional entropy converges immediately to the Kolmogorov-Sinai entropy and verifies the analytical result exactly. The numerical value extracted from the conditional entropy amounts 0.93825, and thus, its difference to the analytical value can be neglected.

The implementation of the Markov chain with randomly fluctuating memory is based on two independent Markov chains as described above. It does not use the effective Markov chain. Furthermore, the plot of the conditional entropy in figure 5.6 shows its behavior analyzed from the generated sequence of symbols. This behavior coincides with the ones found for first-order Markov chains because all \( \Delta H_n \) for \( n \) greater-than or equal to the order of the Markov chain are at one level. Thus, this compound process is identified as a first-order Markov chain. This agrees well with the effective matrix emerging during the derivation in 5.3.2.

The random variation between the two Markov chains has a great impact on the entropy rate of the resulting process. The Kolmogorov-Sinai entropy \( h_{\text{KS}} = 0.93824 \) of the Markov chain with random variation of memory lies in between the values of those corresponding to the autonomous Markov chains (0.78418 and 0.97095). In particular, there seems to be a strong dependence on the probability of selection of each Markov chain.

5.3.6 Dependence on probability of occurrence

The random fluctuation of the memory length is influenced considerably by the probability of occurrence of the two involved Markov chains. This has been signified in 5.3.5 and appears obviously in (5.59) and (5.63), respectively.

In order to investigate the dependence on the Kolmogorov-Sinai entropy, various Markov chains with random fluctuation of order are implemented. Therefore, the transition matrices of the first-order Markov chain \( M(1) \) and the Bernoulli process \( M(0) \) from 5.3.5 are reused. However, the probabilities of their occurrence \( p(M(1)) \) and \( p(M(0)) = 1 - p(M(1)) \), respectively, are modified.

The specified parameters implement the corresponding Markov chain with random fluctuation of order, which generates a symbol sequence. The block entropies for different block lengths of subsequences are determined and used to calculate the conditional entropies. According to their limit of convergence for large \( n \), the numerical entropy rate \( \Delta H_\infty \) is extracted, which can be compared to the analytical Kolmogorov-Sinai entropy.
Figure 5.7: Numerically computed conditional entropy $\Delta H_n$ versus block length $n$ in comparison with different $p(M(1))$ for a random variation between a first-order Markov chain and a Bernoulli process. The behavior of the conditional entropies resembles the entropy rate of a first-order Markov chain except for $p(M(1)) = 0$, which is a Markov process without memory. The numerical estimation of the Kolmogorov-Sinai entropy $\Delta H_\infty$ emphasizes the dependence on $p(M(1))$.

Figure 5.7 illustrates the behavior of the conditional entropy dependent on the block length $n$ for various processes with different $p(M(1))$. Due to modifying the probability of occurrence of the first-order Markov chain, quantitative differences in the development of the conditional entropies are detected. However, they cannot be distinguished qualitatively except for $p(M(1)) = 0$. The behavior for $p(M(1)) = 0$ involves only matrix $M(0)$ and hence resembles a Bernoulli process without any memory. As a result, the conditional entropy consists of a constant series of $\Delta H_n$ as mentioned in 4.2.2. On the contrary, the conditional entropy of any process with $p(M(1)) > 0$ shows the typical behavior of a first-order Markov chain described in 4.2.1. This agrees well with matrix $E$ from (5.63) representing an effective first-order Markov chain.
The calculation of the Kolmogorov-Sinai entropy of the effective matrix $E$ can be done via \[\text{(5.62)}\] analytically. Thus, a dependence on $p(M(1))$ establishes a plot, which compares the numerical entropy rate to the analytical Kolmogorov-Sinai entropy. For this purpose, various entropy rates are computed and combined with the analytical Kolmogorov-Sinai entropy in one plot.

<table>
<thead>
<tr>
<th>$p(M(1))$</th>
<th>analytical $h_{KS}$</th>
<th>numerical $\Delta H_{\infty}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.97095</td>
<td>0.97097</td>
</tr>
<tr>
<td>0.1</td>
<td>0.98494</td>
<td>0.98497</td>
</tr>
<tr>
<td>0.2</td>
<td>0.99383</td>
<td>0.99382</td>
</tr>
<tr>
<td>0.3</td>
<td>0.99708</td>
<td>0.99706</td>
</tr>
<tr>
<td>0.4</td>
<td>0.99402</td>
<td>0.99401</td>
</tr>
<tr>
<td>0.5</td>
<td>0.98386</td>
<td>0.98386</td>
</tr>
<tr>
<td>0.6</td>
<td>0.96566</td>
<td>0.96569</td>
</tr>
<tr>
<td>0.7</td>
<td>0.93824</td>
<td>0.93814</td>
</tr>
<tr>
<td>0.8</td>
<td>0.90016</td>
<td>0.90016</td>
</tr>
<tr>
<td>0.9</td>
<td>0.84960</td>
<td>0.84960</td>
</tr>
<tr>
<td>1.0</td>
<td>0.78418</td>
<td>0.78419</td>
</tr>
</tbody>
</table>

Table 5.1: Comparison of numerically and analytically determined Kolmogorov-Sinai entropy in dependence on $p(M(1))$ for a random variation between a first-order Markov chain and a Bernoulli process. The values agree well and show a non-linear evolution between the entropy rates of $M(0)$ at $p(M(1)) = 0$ and $M(1)$ at $p(M(1)) = 1$.

The results of both numerical and analytical calculation of the Kolmogorov-Sinai entropy $h_{KS}$ and $\Delta H_{\infty}$, respectively, are listed in table 5.1. The numerical analysis was realized on a sequence consisting of 100 million symbols. The values are consistent with each other, although some minor deviations are recognizable. Obviously, the entropy rate does not evolve linearly between the Kolmogorov-Sinai entropy of $M(0)$ at $p(M(1)) = 0$ and the one of $M(1)$ at $p(M(1)) = 1$ and hence is depicted in dependence on $p(M(1))$.

Table 5.1 is visualized in figure 5.8 and shows a good agreement between numerical and analytical values. Despite that, it is noteworthy to find a maximum at around $p(M(1)) = 0.3$ because it is somehow unexpected. The entropy rate of the autonomous Bernoulli process can be found at $p(M(1)) = 0$, and analogously, the entropy rate of the first-order Markov chain is plotted at $p(M(1)) = 1$. As expected, a mixing of these two processes leads to a modified Kolmogorov-Sinai entropy. Astonishingly, even a higher entropy rate is reached than the two involved processes possess when applied separately. The reason originates from combining the two transition matrices to an effective transition matrix according to \[\text{(5.59)}\]. If the probability $p(M(1))$ is chosen properly, the resulting
Figure 5.8: Comparison of analytical and numerical Kolmogorov-Sinai entropy versus probability of occurrence $p(M(1))$ of a first-order Markov matrix $M(1)$ in a random variation with a Bernoulli process. Both calculations agree well and depict an increase of the entropy rate due to a mixing of both processes with a suitable probability $p(M(1))$.

effective matrix will contain elements which are approximately close together. The closer these transition probabilities are, the more the entropy rate tends to one. For instance, when examining an unbiased coin toss with its matrix elements being equal to 0.5, it is comprehensible, that the Kolmogorov-Sinai entropy reaches one. Besides, the structures of the involved matrices are significant.

In summary, the calculation shows a reasonable correspondence between the Kolmogorov-Sinai entropies from both analytical and numerical solution. In addition, two independent first-order Markov chains can be combined by random selection with a specific probability in order to establish a new first-order Markov chain with higher Kolmogorov-Sinai entropy.
6 Blackwell’s formula

The character of hidden Markov chains consisting of an ordinary Markov chain and a stochastic output process raises difficulties in calculating its Kolmogorov-Sinai entropy as described in 2.3. A decent approach to determine the entropy rate of a hidden Markov chain is given by Blackwell [Bla57]. However, it involves an integral of a certain function with respect to a measure and therefore remains hard to evaluate in general cases.

However, an investigation of some specific hidden Markov chains, as they emerge from Markov chains with variation of memory length, can be accomplished. The close relation between Markov chains with varying order and hidden Markov chains was explained in 5.3.3. They represent a special case because the mapping function $\Phi$ is not stochastic but deterministic. Furthermore, several states are mapped to solely one output symbol, and thus, those processes are called aggregated Markov processes.

As an illustration, Blackwell’s formula is used to calculate the Kolmogorov-Sinai entropy for the two cases investigated in 5. At first, the parameters for both types of processes are identified in a general manner to show the universality. Then, the two specific examples are examined enabling a comparison to the Kolmogorov-Sinai entropy determined in 5.2.4 and 5.3.5, respectively.

6.1 Theory

The theory of Blackwell’s entropy is based on hidden Markov chains as introduced in 2.3. The underlying finite-state Markov chain has to be stationary and is represented by a transition matrix $\tilde{M}$ containing conditional probabilities in order to describe the transitions between the states. The $I$ hidden states of $X$ are mapped to the output space $C$, also called a finite alphabet $\mathcal{A}$, by a stochastic function $\Phi$ according to

$$C_n = \Phi(X_n),$$

(6.1)

where $n$ denotes the time step parameter. The cardinality of $C$ equals $\lambda$ and does not exceed the number of hidden states, hence, $\lambda \leq I$. The mapping via $\Phi$ and the stationarity of $\tilde{M}$ cause $\{C_n\}$ to be a stationary process as well.

On the state space of the hidden Markov chain, a probability distribution $\tilde{\pi}^{(n)}$ is defined and additionally another probability distribution $\tilde{\pi}^{(n)}$ is introduced. This vector $\tilde{\pi}^{(n)}$ can
be distinguished from \( \tilde{\pi}^{(n)} \) by its conditional dependence on the observed output symbols. Hence, the \( i \)th element of vector \( \tilde{\pi} \)

\[
\tilde{\pi}^{(n)}_i (c_1, \ldots, c_n) = p(x_n = i|c_1, \ldots, c_n)
\]

represents the probability to be in state \( i \), after the \( n \)-word consisting of the symbols \( c_1 \) to \( c_n \) has been detected. In other words, a probability distribution on the hidden states is given by (6.2) when the output sequence is known. The definition (6.2) is completed by an initial distribution

\[
\tilde{\pi}^{(0)} = \tilde{\pi}^*
\]

being the stationary distribution of the Markov chain.

Instead of \( \tilde{\pi}^* \) for the initial distribution (6.3), any other distribution on \( X \) is appropriate but suffers from a modified convergence behavior. The reason is obvious, that the probability mass has to be distributed among the states. In case of the stationary distribution, this has already been achieved and thus represents the long-term behavior of the distribution for \( n \) tending to infinity.

The probability distribution \( \tilde{\pi}^{(n)} \) depends on \( n \) symbols, hence for any combination of these symbols forming a \( n \)-word a separate vector \( \tilde{\pi}^{(n)} \) exists. This manifests the existence of \( \lambda^n \) different \( \tilde{\pi}^{(n)} \) and each occurs with the probability \( p(c_1, \ldots, c_n) \) of the \( n \)-word. Thus, a stationary distribution of probability distributions is established. Equivalently, Blackwell shows, that \( \{ \tilde{\pi}^{(n)} \} \) is a stationary Markov process and denotes its limiting stationary probability distribution, as \( n \) tends to infinity, by the distribution function \( Q(\tilde{\pi}) \), known as Blackwell’s measure.

Blackwell’s measure consists of several Heaviside step functions \( \Theta \) starting to be equal to one when its argument becomes zero. For a fixed \( n \), each \( \tilde{\pi}^{(n)} \) is the argument of a step function multiplied by its probability of occurrence. Hence, at each \( \tilde{\pi}^{(n)} \), a step with the height of its probability is performed, in order to induce

\[
Q(\tilde{\pi}) = \sum_{c_1, \ldots, c_n} p(\tilde{\pi}^{(n)}(c_1, \ldots, c_n)) \cdot \Theta(\tilde{\pi} - \tilde{\pi}^{(n)}(c_1, \ldots, c_n))
\]

the distribution function.

Furthermore, from the transition matrix \( \tilde{M} \) special matrices \( \tilde{M}_c \) with \( c \in \mathcal{A} \) are extracted, whose elements

\[
(\tilde{M}_c)_{ij} = \begin{cases} 
\tilde{M}_{ij} : & \text{if } \Phi(j) = c \\
0 : & \text{else}
\end{cases}
\]

depend on the mapping function \( \Phi \). Thus, each column \( j \) of \( \tilde{M}_c \) will be equal to the \( j \)th column of matrix \( \tilde{M} \), if the state \( j \) is mapped to the output symbol \( c \) and otherwise filled with zeros. In short, these matrices \( \tilde{M}_c \) distribute the probability mass to those states
which produce symbol $c$. However, it should be noted that $\tilde{M}_c$ is not a stochastic matrix anymore because it lacks the row sums of one.

By means of the matrices $\tilde{M}_c$, a probability distribution $\tilde{\pi}$ is spread to the states which create symbol $c$. By implication, the vector elements of the vector-matrix product

$$(\tilde{\pi}\tilde{M}_c)_i = p(c,i) \quad (6.6)$$

yields the joint probability to be in state $i$ and generate symbol $c$. Summing over all states in these joint probabilities for a given $c$ results in

$$p(c) = \sum_i p(c,i) = \tilde{\pi}\tilde{M}_c\tilde{\eta} \quad (6.7)$$

the probability to receive symbol $c$, where $\tilde{\eta}$ is a vector consisting solely of elements equal to one. Dividing the joint probabilities in (6.6) by the probability (6.7) for the specific symbol, conditional probabilities

$$p(i|c) = \frac{p(c,i)}{p(c)} = \frac{(\tilde{\pi}\tilde{M}_c)_i}{\tilde{\pi}\tilde{M}_c\tilde{\eta}} \quad (6.8)$$

are obtained for each state $i$, given the output symbol $c$. Hence, summing over all states $i$ yields

$$\sum_i p(i|c) = 1 \quad (6.9)$$

because the system has to be in some state after $c$ was observed. Blackwell denotes (6.8) by a vector function

$$\vec{f}_c(\tilde{\pi}) = \frac{\tilde{\pi}\tilde{M}_c}{\tilde{\pi}\tilde{M}_c\tilde{\eta}} \quad (6.10)$$

defined in the same manner, which is applied to the probability distribution $\tilde{\pi}$ and generates a new probability distribution.

By means of $\vec{f}_c$, the current probability mass on the states is distributed to those states which generate symbol $c$ and thus describes the development of $\tilde{\pi}$ depending on the time step $n$. This is achieved by applying $c = c_n$ at time step $n$ in order to generate the probability distribution $\tilde{\pi}^{(n)}$. Consequently, it results in

$$\tilde{\pi}^{(n+1)} = f_{c_{n+1}}(\tilde{\pi}^{(n)}), \quad (6.11)$$

called the random dynamical iteration.

Finally, based on the given definitions, Blackwell establishes the analytical expression

$$h_{KS} = H(C) = -\int \sum_c \tilde{\pi}\tilde{M}_c\eta \log \left( \frac{\tilde{\pi}\tilde{M}_c\eta}{\tilde{\pi}\tilde{M}_c\tilde{\eta}} \right) dQ(\tilde{\pi}) \quad (6.12)$$
for the entropy rate of \( \{ c_n \} \) with respect to a given \( n \). It reveals to be the integral over a function involving a sum of logarithms with respect to a measure defined by Blackwell. It should be remarked, that the Kolmogorov-Sinai entropy in (6.12) is normalized to the cardinality \( \lambda \) of the alphabet \( \mathcal{A} \), as done in 2.1.1 in contrast to Blackwell’s normalization to a binary alphabet.

The sum contained in the analytical expression for the entropy rate of a hidden Markov chain (6.12) unveils a certain similarity to the Kolmogorov-Sinai entropy of an ordinary Markov chain. Additionally to the sum, the integral with respect to the specific measure is incorporated for hidden Markov chains.

Moreover, the integral with respect to the distribution function \( Q(\tilde{\pi}) \) can be rewritten as an integral over its probability density function \( \rho(\tilde{\pi}) \) with respect to \( \tilde{\pi} \). This is expressed by

\[
dQ(\tilde{\pi}) = \rho(\tilde{\pi}) \, d\tilde{\pi},
\]

and thus, \( \rho(\tilde{\pi}) \) is the derivative of \( Q(\tilde{\pi}) \). The derivative of a Heaviside step function \( \Theta(x) \) results in the Dirac delta distribution \( \delta(x) \), which has an infinitely large peak at \( x = 0 \), is equal to zero elsewhere, and satisfies the constraint

\[
\int_{-\infty}^{\infty} \delta(x) \, dx = 1 \quad x \in \mathbb{R}.
\]

Consequently, \( dQ(\tilde{\pi}) \) with \( Q(\tilde{\pi}) \) from (6.4) can be replaced by

\[
dQ(\tilde{\pi}) = \sum_{c_1, \ldots, c_n} p(\tilde{\pi}(n)(c_1, \ldots, c_n)) \cdot \delta(\tilde{\pi} - \tilde{\pi}(n)(c_1, \ldots, c_n)) \, d\tilde{\pi}
\]

a sum of weighted delta peaks for a fixed \( n \) considering all combinations of \( n \)-words.

In order to evaluate (6.12), the integral involving the delta distributions has to be determined. In particular, the integral over a function \( g(x) \) multiplied by the shifted delta distribution results in

\[
\int_{-\infty}^{\infty} g(x) \delta(x - x_0) \, dx = g(x_0) \quad x, x_0 \in \mathbb{R}
\]

the selection of the value of function \( g(x = x_0) \) at the shifted position \( x_0 \). Hence, the integral in (6.12) is eliminated and introduces a sum over all \( \tilde{\pi}(n) \) and substitutes \( \tilde{\pi} \) by

\[
p(\tilde{\pi}(n)(c_1, \ldots, c_n)) \cdot \tilde{\pi}(n)(c_1, \ldots, c_n).
\]

In summary, Blackwell’s formula offers a structured analytical expression to determine the Kolmogorov-Sinai entropy of any hidden Markov chain complying with the requirements. However, for most cases of the underlying Markov chain and the involved mapping function, the measure is complicated, and the integral becomes difficult to evaluate. Nevertheless, even some special cases exist enabling an investigation via Blackwell’s formula.
6.2 Kolmogorov-Sinai entropy for random variations

A random variation between two Markov chains is achieved as explained in 5.3.1. Again, only two Markov chains are involved, which are also called machines and denoted by $M$ and $M'$. The selection of the machines is controlled by taking the probabilities of occurrence $p(M)$ and $p(M') = 1 - p(M)$ into consideration. The selection process has to be uncorrelated forming a Bernoulli process.

Moreover, the state space of each Markov chain consists solely of two states represented by a binary alphabet $\mathcal{A} \equiv \{A,B\}$, thus $\lambda = 2$. This simplifies the comprehensibility of Blackwell’s entropy rate but can easily be enhanced to a larger state space. Even an extension to a variation between three or more Markov chains is feasible.

6.2.1 Hidden Markov chain

The hidden Markov chain originating from the random variation between two binary Markov chains is established by grouping the two machines and the two output symbols in pairs together. This leads to a Markov chain with four unobservable states

$$x_n \in \{MA, MB, M'A, M'B\} \quad (6.17)$$

as depicted in 5.3.3, which can alternatively be numbered by integers.

For a random variation of memory, any transition between the hidden states is available. As explained explicitly in 5.3.3, the conditional probabilities are modified and include a transition between the machines as well. The current machine is necessary to determine the correct transition probabilities between the symbols. On the contrary, the selection of the next machine is independent of the current machine. This simplifies the conditional probability

$$p(c_{n+1}, M^{(n+1)}|c_n, M^{(n)}) = M_{c_n \rightarrow c_{n+1}} \cdot p(M^{(n+1)}) \quad (6.18)$$

to a product of the transition probability on the current machine and the probability to select the next machine. These transition probabilities between the hidden states are summarized in table 6.1 resembling the transition matrix $\tilde{M}$ of the hidden Markov chain.

Additionally, a hidden Markov chain possesses a mapping function $\Phi$ which transforms the realizations $x_n$ of the hidden states to observable output symbols $c_n$. Due to the construction of this hidden Markov chain, the output symbol can be extracted from the notation of the hidden state by omitting the current machine identifier. The entire definition of $\Phi$ is given in table 6.2 and illustrated in figure 6.1. The enumeration of the states denoted by $i$ should be noted and is utilized to address the states in vectors like $\tilde{\pi}$ more intuitively.
6.2 Kolmogorov-Sinai entropy for random variations

<table>
<thead>
<tr>
<th>current state</th>
<th>next state</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA</td>
<td>MA → A · p(M)</td>
</tr>
<tr>
<td>MB</td>
<td>M′B → A · p(M)</td>
</tr>
</tbody>
</table>

Table 6.1: Overview of the possible transitions from any current state to a next state and their corresponding conditional probabilities of the hidden Markov chain representing the random variation between two Markov chains. The transition probability between the symbols on the current machine is multiplied by the probability to select the next machine.

<table>
<thead>
<tr>
<th>i</th>
<th>x_n</th>
<th>c_n = Φ(x_n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MA</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>MB</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>M′A</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>M′B</td>
<td>B</td>
</tr>
</tbody>
</table>

Table 6.2: Assignment of the two output symbols to the four hidden states of the underlying Markov chain by means of the mapping function Φ. Caused by the construction of the hidden Markov chain and the notation of the internal states, Φ neglects the correspondence to the current machine in order to generate the appropriate output symbol. The states are enumerated to identify them in vectors.

Inevitably, the properties of the underlying Markov chain represented by its transition matrix have to be investigated. It is necessary that the matrix is both irreducible and aperiodic. Then, a stationary distribution \( \tilde{\pi}^* \) on the hidden states exists. By means of Φ, it is mapped to a stationary distribution on the output symbols simultaneously. Due to this, the hidden Markov chain represents a stationary process.

The stationary distribution \( \tilde{\pi}^* \) is used as the initial distribution \( \tilde{\pi}^{(0)} \). The random dynamical iteration (6.11) generates the successive distributions on the hidden states depending on the symbols that have been observed. By taking the probability of the observation into account, Blackwell’s measure is obtained.
6.2.2 Derivation

The general expressions from Blackwell have to be determined for the special hidden Markov chain resembling a Markov chain with random variation of memory length. To construct the random dynamical iteration (6.11), the transition matrices $\tilde{M}_A$ and $\tilde{M}_B$ have to be prepared.

The matrix

$$\tilde{M}_A = \begin{pmatrix} M_{A\rightarrow A} \cdot p(M) & 0 & M_{A\rightarrow A} \cdot p(M') & 0 \\ M_{B\rightarrow A} \cdot p(M) & 0 & M_{B\rightarrow A} \cdot p(M') & 0 \\ M'_{A\rightarrow A} \cdot p(M) & 0 & M'_{A\rightarrow A} \cdot p(M') & 0 \\ M'_{B\rightarrow A} \cdot p(M) & 0 & M'_{B\rightarrow A} \cdot p(M') & 0 \end{pmatrix}, \quad (6.19)$$

which distributes the probability mass to those hidden states that produce output A, consists of two zero-filled columns and two being extracted from the transition matrix of the
6.2 Kolmogorov-Sinai Entropy for Random Variations

hidden Markov chain. Likewise, $\tilde{M}_B$ is structured, but in

$$\tilde{M}_B = \begin{pmatrix} 0 & M_{A\rightarrow B} \cdot p(M) & 0 & M_{A\rightarrow B} \cdot p(M') \\ 0 & M_{B\rightarrow B} \cdot p(M) & 0 & M_{B\rightarrow B} \cdot p(M') \\ 0 & M'_{A\rightarrow B} \cdot p(M) & 0 & M'_{A\rightarrow B} \cdot p(M') \\ 0 & M'_{B\rightarrow B} \cdot p(M) & 0 & M'_{B\rightarrow B} \cdot p(M') \end{pmatrix},$$ (6.20)

the zero-filled columns are swapped compared to $\tilde{M}_A$. This is reasonable because the columns filled with zeros belong to the states which produce another output symbol than the desired one. Hence, the probability mass is not distributed to those states.

Furthermore, it should be remarked that the sum $\tilde{M}_A + \tilde{M}_B$ returns the transition matrix of the underlying Markov chain. This is evident because to any state solely one output symbol is assigned.

Subsequently, the matrices are used to establish the function $f_c(\tilde{\pi})$ for the random dynamical iteration according to (6.10). Firstly, the denominator of $f_c(\tilde{\pi})$

$$\tilde{\pi} M_c \tilde{\eta} = \tilde{\pi}_1 \cdot M_{A\rightarrow c} \cdot p(M) + \tilde{\pi}_2 \cdot M_{B\rightarrow c} \cdot p(M) + \tilde{\pi}_3 \cdot M'_{A\rightarrow c} \cdot p(M) + \tilde{\pi}_4 \cdot M'_{B\rightarrow c} \cdot p(M) + \tilde{\pi}_1 \cdot M'_{A\rightarrow c} \cdot p(M') + \tilde{\pi}_2 \cdot M'_{B\rightarrow c} \cdot p(M')$$ (6.21)

is determined and can be significantly simplified by factorizing similar terms and utilizing $p(M) + p(M') = 1$. Then, the simplification yields

$$\tilde{\pi} M_c \tilde{\eta} = \tilde{\pi}_1 \cdot M_{A\rightarrow c} + \tilde{\pi}_2 \cdot M_{B\rightarrow c} + \tilde{\pi}_3 \cdot M'_{A\rightarrow c} + \tilde{\pi}_4 \cdot M'_{B\rightarrow c}$$ (6.22)

an expression which is independent of $p(M)$ and $p(M')$.

Afterwards, the two functions $\tilde{f}_A$ and $\tilde{f}_B$ are considered separately. In the numerator of $\tilde{f}_A$ a factorization is possible and hence results in the vector

$$\tilde{\pi} M_A = \begin{pmatrix} \left( \tilde{\pi}_1 \cdot M_{A\rightarrow A} + \tilde{\pi}_2 \cdot M_{B\rightarrow A} + \tilde{\pi}_3 \cdot M'_{A\rightarrow A} + \tilde{\pi}_4 \cdot M'_{B\rightarrow A} \right) \cdot p(M) \\ 0 \\ \left( \tilde{\pi}_1 \cdot M_{A\rightarrow A} + \tilde{\pi}_2 \cdot M_{B\rightarrow A} + \tilde{\pi}_3 \cdot M'_{A\rightarrow A} + \tilde{\pi}_4 \cdot M'_{B\rightarrow A} \right) \cdot p(M') \\ 0 \end{pmatrix},$$ (6.23)

having two elements equal to zero. Again, these elements represent the states which do not produce symbol A. In order to obtain $\tilde{f}_A(\tilde{\pi})$, (6.23) has to be divided by (6.22) and $c$ is inserted appropriately. This reduces the expression to

$$\tilde{f}_A(\tilde{\pi}) = \begin{pmatrix} p(M) \\ 0 \\ p(M') \end{pmatrix}. $$ (6.24)
In correspondence to (6.23), the numerator of $\tilde{f}_B$ is formed by

$$\tilde{\pi} \tilde{M}_B = \begin{pmatrix} 0 \\ \left( \tilde{\pi}_1 \cdot M_{A\rightarrow B} + \tilde{\pi}_2 \cdot M_{B\rightarrow B} + \tilde{\pi}_3 \cdot M'_{A\rightarrow B} + \tilde{\pi}_4 \cdot M'_{B\rightarrow B} \right) \cdot p(M) \\ 0 \end{pmatrix} \cdot p(M')$$

(6.25)

and contains two zeros, which are located in the elements related to output symbol A. Moreover, $\tilde{f}_B(\tilde{\pi})$ is deduced likewise by dividing (6.25) by (6.22). Hence, the second vector for the random dynamical iteration

$$\tilde{f}_B(\tilde{\pi}) = \begin{pmatrix} 0 \\ p(M) \\ 0 \end{pmatrix}$$

(6.26)

is prepared.

When examining (6.24) and (6.26) carefully, their independence of $\tilde{\pi}$ becomes apparent, thus $\tilde{f}_c(\tilde{\pi}) = \tilde{f}_c$. As a matter of fact, any distribution on the hidden Markov chain is spread to those states that can produce the observed symbol. Simultaneously, the state related to machine $M$ gets a portion $p(M)$ of the probability mass and $M'$ gets $p(M')$, respectively. Hence, the random dynamical iteration always maps the probability distribution to one of the two states (6.24) and (6.26) with respect to the current symbol observation. As a consequence, Blackwell’s measure only consists of the two distribution vectors corresponding to the output symbols which are occupied with the probability of the symbol’s occurrence.

The random dynamical iteration induced by $\tilde{f}_A(\tilde{\pi})$ is illustrated in figure 6.2(a). The probability mass is delivered to the states which produce output A and is split between the two machines. Caused by the independence of the current distribution on the states, the iteration will always lead to the same distribution vector if symbol A is observed. Moreover, it is reasonable that the states generating symbol B do not receive any probability mass. In figure 6.2(b) the analogous case is depicted for $\tilde{f}_B(\tilde{\pi})$.

### 6.2.3 Example

As mentioned above, the example from [5.3.5] is continued, and its Kolmogorov-Sinai entropy should be verified via Blackwell’s formula.

The Markov chain with random variation of order is constructed from the matrix of the Bernoulli process (5.72) and the matrix (5.73) of a first-order Markov chain. The probabilities for performing the uncorrelated selection of these two matrices are $p(M) = 0.3$ and $p(M') = 0.7$ as stipulated in [5.74] and [5.75], respectively.
As defined in table 6.1, the conjunction of both matrices and their corresponding probabilities of occurrence establish the transition matrix

$$\tilde{M} = \begin{pmatrix}
0.12 & 0.18 & 0.28 & 0.42 \\
0.12 & 0.18 & 0.28 & 0.42 \\
0.24 & 0.06 & 0.56 & 0.14 \\
0.18 & 0.12 & 0.42 & 0.28 \\
\end{pmatrix} \quad (6.27)$$

between the unobservable states of the hidden Markov chain. The irreducibility and aperiodicity of $\tilde{M}$ can be confirmed, and thus, a stationary distribution exists on its hidden states. The steady state vector

$$\tilde{\pi}^* (\tilde{M}) = \begin{pmatrix}
0.188372 \\
0.11628 \\
0.439535 \\
0.260465 \\
\end{pmatrix} \quad (6.28)$$

can be evaluated. It expresses that the hidden state $MA$ is accepted with a probability of approximately 0.188, $MB$ with 0.112 and the other states respectively. Furthermore, the states are mapped to observable symbols by means of $\Phi$ causing a stationary distribution on the output as well. It can be calculated by summing the stationary probabilities of
those states which map to the same symbol. Thus, the long-term probabilities for the
observation of $A$
\[ p(A) = 0.627907 \quad (6.29) \]
and
\[ p(B) = 0.372093 \quad (6.30) \]
for symbol $B$ are given and sum up to one perfectly.

Due to the independence of $\vec{\pi}$, $\vec{f}_A$ and $\vec{f}_B$ are given without any effort. By inserting $p(M)$ and $p(M')$ into (6.24), $\vec{f}_A$ yields
\[ \vec{f}_A = \begin{pmatrix} 0.3 \\ 0 \\ 0.7 \\ 0 \end{pmatrix}, \quad (6.31) \]
and analogously,
\[ \vec{f}_B = \begin{pmatrix} 0 \\ 0.3 \\ 0 \\ 0.7 \end{pmatrix} \quad (6.32) \]
is obtained. Hence and in consequence of (6.11), each probability distribution at iteration $n$ depends solely on the current output symbol and is either equal to $\vec{f}_A$ or $\vec{f}_B$. Consequently, based on their definition, these distributions on the states occur with the probabilities of the observed symbols from (6.29) and (6.30).

Figure 6.3 illustrates the possible distributions on the states after observation of a certain symbol as well as their relative frequencies of occurrence. For each observed symbol, only one distribution on the states exists. Moreover, the distributions on the states after observation of the current symbol are independent of the previous distribution, and therefore, they are not influenced by previously observed symbols. Hence, these two distributions are the only ones that occur, regardless how many symbols have been observed.

Finally, the two distributions and their relative frequencies of occurrence are assembled and establish
\[ Q(\vec{\pi}) = 0.627907 \cdot \Theta(\vec{\pi} - \vec{f}_A) + 0.372093 \cdot \Theta(\vec{\pi} - \vec{f}_B) \quad (6.33) \]
Blackwell’s measure.

According to (6.1), the derivative of (6.33) yields
\[ dQ(\vec{\pi}) = 0.627907 \cdot \delta(\vec{\pi} - \vec{f}_A) d\vec{\pi} + 0.372093 \cdot \delta(\vec{\pi} - \vec{f}_B) d\vec{\pi} \quad (6.34) \]
6.3 Kolmogorov-Sinai entropy for periodic variations

A periodic variation between two Markov chains is achieved as explained in 5.2.1. According to this, only two Markov chains are involved, which are also called machines and de-
noted by $M$ and $M'$. The machines are selected purely deterministic, alternating between $M$ and $M'$. Hence, the selection process is periodic and therefore strongly correlated.

Furthermore, the state space of each Markov chain contains two states depicted by symbols A and B from a binary alphabet $\mathcal{A}$. This causes $\lambda = 2$ and provides an assessable derivation of Blackwell’s entropy rate. An extension to a larger state space or a periodic variation between three or more Markov chains is realizable.

### 6.3.1 Hidden Markov chain

The periodic variation of the memory length can be modeled on a hidden Markov chain as well. For this purpose, the two involved machines and the two output symbols are grouped together in pairs. The resulting Markov chain consists of the four states

$$x_n \in \{MA, MB, M' A, M' B\}, \quad (6.37)$$

which are unobservable and can alternatively be numbered by integers.

For a periodic varying memory length, not every transition between the hidden states is available. A switch between the machines is necessary in any iteration prohibiting the inter-symbol transitions on each machine. Thus, only eight of the sixteen transitions are allowed introducing zeros to the transition matrix. Simultaneously, the other conditional probabilities remain unmodified. In other words, there exist four conditional probabilities to describe the machine selection

$$p(M|M) = 0,$$
$$p(M'|M) = 1,$$
$$p(M|M') = 1,$$
$$p(M'|M') = 0,$$ \quad (6.38)

which are either zero or one and therefore select the machines deterministically. These values have to be multiplied to the transition probabilities of $M$ and $M'$, respectively, reducing them to zero or leaving them unaffected. The transitions between the hidden states and their conditional probabilities are tabulated in table 6.3 resembling the transition matrix $\tilde{M}$ of the hidden Markov chain.

In figure 6.4, the transitions from table 6.3 are depicted showing the absence of several transitions whose probabilities are zero. This is caused by the periodic variation requiring a change of the machine in each step. Thus, only transitions between machines are available, and transitions on the same machine are prohibited. Additionally, in figure 6.4, it is obvious that the transition probabilities on the hidden states are retained equal to those of the uncoupled Markov chains.
In addition to the transition matrix, a hidden Markov chain defines a mapping function $\Phi$ transforming the realizations $x_n$ of the hidden states to output symbols $c_n$, which can be observed. Induced by the special structure of the hidden Markov chain, the output symbol which is produced by $\Phi$ can be extracted from the notation of the hidden state by omitting the identifier of the current machine. The complete definition of $\Phi$ is given in table 6.1, where the enumeration of the states denoted by $i$ should be remarked. This is utilized to address the states in vectors like $\tilde{\pi}$ more intuitively.

Moreover, an investigation of the properties possessed by the underlying Markov chain is unavoidable. Obviously, the transition matrix $\tilde{M}$ is indeed irreducible because each state can be reached by any other one within a finite number of transitions. On the contrary, the second necessary property does not hold for this transition matrix. The aperiodicity is not fulfilled because always a multiple of two steps is necessary to return from a state to itself. Hence, the process is periodic agreeing with the construction of the hidden Markov chain from a periodic variation between two Markov chains.

As a result, the requirements of ergodicity are not met and a stationary distribution does not exist causing a permanent dependence on the initial conditions. This is comprehensible because the probability mass which was initially assigned to the machines is swapped between the machines and returns after every even number of transitions. Nevertheless, a spreading between the symbols on each machine occurs and consequently two stationary-like distributions are formed, which are obtained alternately and depend on the initial probability mass distribution between the machines. The swapping process is illustrated in figure 6.5 showing the sums of the probability mass on each machine and their exchange between the machines in every step.

Taking the aforementioned considerations into account, the probability mass has to be dis-
Figure 6.4: Two coupled first-order Markov chains with binary output for periodic variation. Due to the periodic variation, only transitions between some specific states are possible, involving a jump to the other machine in each step. The transition probabilities on the machines remain unchanged, but instead a change of the machine is required defining a kind of schedule.

distributed uniformly between the machines for a stationary distribution. Hence, the swapping still appears but cannot be recognized anymore. Then, the probability mass between the symbols adjusts to the stationary distribution resulting in a steady state vector on the states. This vector is chosen as initial distribution to smoothen the random dynamical iteration, which otherwise would swap the probability mass in each iteration.

The stationary distribution on the states of the hidden Markov chain can be determined by the left eigenvector corresponding to eigenvalue one as described in 2.2.6. Due to $\tilde{M}$ being a stochastic matrix, the eigenvalue one indeed exists. Thus, the left eigenvector is the desired steady state vector because it has to be invariant to applying matrix $\tilde{M}$ and is used as the initial distribution $\tilde{\pi}^{(0)}$. Then, the random dynamical iteration (6.11) creates the successive distributions on the hidden states depending on the symbols which have been observed. By taking the probability of the observation into account, Blackwell’s measure is obtained.
6.3.2 Derivation

The general expressions from Blackwell have to be adapted to the special hidden Markov chain resembling a Markov chain with periodic variation of memory length. To construct the random dynamical iteration (6.11), the transition matrices $\tilde{M}_A$ and $\tilde{M}_B$ have to be determined.

The matrix

$$
\tilde{M}_A = \begin{pmatrix}
0 & 0 & 0 & M_{A\to A} & 0 \\
0 & 0 & 0 & M_{B\to A} & 0 \\
M'_{A\to A} & 0 & 0 & 0 \\
M'_{B\to A} & 0 & 0 & 0
\end{pmatrix},
$$

(6.39)

which distributes the probability mass to those states that produce output A, consists of two zero-filled columns and two others being extracted from the transition matrix $\tilde{M}$ of the hidden Markov chain. Likewise, $\tilde{M}_B$ is structured, but in

$$
\tilde{M}_B = \begin{pmatrix}
0 & 0 & 0 & M_{A\to B} \\
0 & 0 & 0 & M_{B\to B} \\
0 & M'_{A\to B} & 0 & 0 \\
0 & M'_{B\to B} & 0 & 0
\end{pmatrix},
$$

(6.40)
the zero-filled columns are swapped compared to \( \tilde{M}_A \). This is comprehensible because the zero-filled columns belong to the states which produce another output symbol than the desired one. Hence, the probability mass is not distributed to those states.

It should be annotated that the sum \( \tilde{M}_A + \tilde{M}_B \) yields the transition matrix of the hidden Markov chain. This is plausible because each state outputs only one symbol, and thus, the combination of the two cases has to resemble the complete matrix.

Furthermore, the matrices \( \tilde{M}_A \) and \( \tilde{M}_B \) are required to derive \( \tilde{f}_c(\tilde{\pi}) \) for the random dynamical iteration according to (6.10). At first, the denominator of \( \tilde{f}_A(\tilde{\pi}) \)

\[
\tilde{\pi} \tilde{M}_c \tilde{\eta} = \tilde{\pi}_1 \cdot M_{A\rightarrow c} + \tilde{\pi}_2 \cdot M_{B\rightarrow c} + \tilde{\pi}_3 \cdot M'_{A\rightarrow c} + \tilde{\pi}_4 \cdot M'_{B\rightarrow c} 
\]  

(6.41)

is determined but cannot be simplified any further.

Thereafter, the two functions are deduced separately caused by their different numerators. The numerator of \( \tilde{f}_A \) yields the vector

\[
\tilde{\pi} \tilde{M}_A = \begin{pmatrix} \tilde{\pi}_3 \cdot M'_{A\rightarrow A} + \tilde{\pi}_4 \cdot M'_{B\rightarrow A} \\ 0 \\ \tilde{\pi}_1 \cdot M_{A\rightarrow A} + \tilde{\pi}_2 \cdot M_{B\rightarrow A} \\ 0 \end{pmatrix} 
\]  

(6.42)

as well as

\[
\tilde{\pi} \tilde{M}_B = \begin{pmatrix} 0 \\ \tilde{\pi}_3 \cdot M'_{A\rightarrow B} + \tilde{\pi}_4 \cdot M'_{B\rightarrow B} \\ \tilde{\pi}_1 \cdot M_{A\rightarrow B} + \tilde{\pi}_2 \cdot M_{B\rightarrow B} \end{pmatrix} 
\]  

(6.43)

does for the numerator of \( \tilde{f}_B \), respectively. These results are joined by dividing (6.42) by (6.41) and replacing \( c \) by \( A \) appropriately. It evaluates to elongated expressions and is simplified by reciprocal fractions to

\[
\tilde{f}_A(\tilde{\pi}) = \begin{pmatrix} (1 + \frac{\tilde{\pi}_1 \cdot M_{A\rightarrow A} + \tilde{\pi}_2 \cdot M_{B\rightarrow A}}{\tilde{\pi}_3 \cdot M'_{A\rightarrow A} + \tilde{\pi}_4 \cdot M'_{B\rightarrow A}})^{-1} \\ 0 \\ (1 + \frac{\tilde{\pi}_1 \cdot M_{A\rightarrow A} + \tilde{\pi}_2 \cdot M_{B\rightarrow A}}{\tilde{\pi}_3 \cdot M'_{A\rightarrow A} + \tilde{\pi}_4 \cdot M'_{B\rightarrow A}})^{-1} \end{pmatrix},
\]  

(6.44)

and the division of (6.43) by (6.41) and inserting \( c = B \) leads to

\[
\tilde{f}_B(\tilde{\pi}) = \begin{pmatrix} 0 \\ (1 + \frac{\tilde{\pi}_1 \cdot M_{A\rightarrow B} + \tilde{\pi}_2 \cdot M_{B\rightarrow B}}{\tilde{\pi}_3 \cdot M'_{A\rightarrow B} + \tilde{\pi}_4 \cdot M'_{B\rightarrow B}})^{-1} \\ (1 + \frac{\tilde{\pi}_1 \cdot M_{A\rightarrow B} + \tilde{\pi}_2 \cdot M_{B\rightarrow B}}{\tilde{\pi}_3 \cdot M'_{A\rightarrow B} + \tilde{\pi}_4 \cdot M'_{B\rightarrow B}})^{-1} \end{pmatrix},
\]  

(6.45)
respectively.

To summarize, the random dynamical iteration of the hidden Markov chain for periodic variation of memory length keeps its dependence on the current state. Thus, the probability mass is redistributed in each iteration to the states corresponding to the observed symbol. Consequently, the number of distribution vectors establishing Blackwell’s measure increases with each iteration. These distribution vectors are occupied with the probability of the observed symbol sequences, which can be determined by a product of conditional probabilities.

### 6.3.3 Example

Blackwell’s formula continues the example in 5.2.4 and the Kolmogorov-Sinai entropy should be verified.

The Markov chain with periodic variation of memory length is constructed from the matrix of the Bernoulli process (5.34) and the matrix (5.35) of a first-order Markov chain. The switching between the machines takes place purely deterministic and does not involve any other probabilities.

As prepared in table 6.3, the connection of both matrices and their periodic variation establish the transition matrix

\[
\tilde{M} = \begin{pmatrix}
0 & 0 & 0.4 & 0.6 \\
0 & 0 & 0.4 & 0.6 \\
0.8 & 0.2 & 0 & 0 \\
0.6 & 0.4 & 0 & 0
\end{pmatrix}
\] (6.46)

between the unobservable states of the hidden Markov chain. Due to the absence of ergodicity, arbitrary stationary-like distributions exist as discussed in 6.3.1 However, the left eigenvector corresponding to eigenvalue one yields a stationary distribution, where not any swapping between the machines is detected. This eigenvector

\[
\tilde{\pi}^{(0)} = \tilde{\pi}^* (\tilde{M}) = \begin{pmatrix}
0.34 \\
0.16 \\
0.2 \\
0.3
\end{pmatrix}
\] (6.47)

resembles the steady state vector and is completely invariant to applying it to matrix \(\tilde{M}\). This is obvious and can be understood by summing up the probabilities which belong to machine \(M\), \(0.34 + 0.16 = 0.5\), and the ones of machine \(M'\), \(0.2 + 0.3 = 0.5\). In fact, both are equal to 0.5, and thus, the swapping of the probability mass between the machines cannot be detected.
Moreover, a stationary distribution on the output symbols exists. By summing up those states which map to the same output symbol, the long-term probabilities for the observation of symbol A

\[ p(A) = 0.54 \] (6.48)

and

\[ p(B) = 0.46 \] (6.49)

for symbol B are given and sum up to one perfectly.

The functions \( \vec{f}_A \) (6.44) and \( \vec{f}_B \) (6.45) for the random dynamical iteration are filled with the appropriate transition probabilities from matrix \( \tilde{M} \). Subsequently, the random dynamical iteration is applied to the initial distribution (6.47) evaluating to

\[ \tilde{\vec{f}}^{(1)} = f_{c_1}(\tilde{\vec{\pi}}(0)) = \begin{cases} \vec{f}_A(\tilde{\vec{\pi}}(0)) = \begin{pmatrix} 0.6296 \\ 0 \\ 0.3704 \end{pmatrix} & : \text{observed A with } p(A) = 0.54 \\ \vec{f}_B(\tilde{\vec{\pi}}(0)) = \begin{pmatrix} 0.3478 \\ 0 \\ 0.6522 \end{pmatrix} & : \text{observed B with } p(B) = 0.46 \end{cases} \] (6.50)

a distribution depending on the observation of the first symbol \( c_1 \). Figure 6.6 illustrates the occurring distributions on the states with their relative frequencies under consideration of the first observed symbol.

Finally, this induces the probability distribution function

\[ Q(\tilde{\vec{\pi}}) = 0.54 \cdot \Theta(\tilde{\vec{\pi}} - \tilde{\vec{f}}_A(\tilde{\vec{\pi}}(0))) + 0.46 \cdot \Theta(\tilde{\vec{\pi}} - \tilde{\vec{f}}_B(\tilde{\vec{\pi}}(0))) \] (6.51)

referred to as Blackwell’s measure, which is related to one observed symbol.

According to (6.1), the derivative of (6.51) results to

\[ dQ(\tilde{\vec{\pi}}) = 0.54 \cdot \delta(\tilde{\vec{\pi}} - \tilde{\vec{f}}_A(\tilde{\vec{\pi}}(0))) \, d\tilde{\vec{\pi}} + 0.46 \cdot \delta(\tilde{\vec{\pi}} - \tilde{\vec{f}}_B(\tilde{\vec{\pi}}(0))) \, d\tilde{\vec{\pi}} \] (6.52)

the probability density function with its delta peaks, which select the two probability distributions \( \tilde{\vec{f}}_A(\tilde{\vec{\pi}}(0)) \) and \( \tilde{\vec{f}}_B(\tilde{\vec{\pi}}(0)) \) in the integration of (6.12). After simplification of the integral,

\[ h_{KS} = -0.54 \cdot \sum_c \tilde{f}_A(\tilde{\vec{\pi}}(0)) \tilde{M}_c \eta \log_\lambda(\tilde{f}_A(\tilde{\vec{\pi}}(0))) \tilde{M}_c \eta \]

\[ -0.46 \cdot \sum_c \tilde{f}_B(\tilde{\vec{\pi}}(0)) \tilde{M}_c \eta \log_\lambda(\tilde{f}_B(\tilde{\vec{\pi}}(0))) \tilde{M}_c \eta \] (6.53)
6.3 KOLMOGOROV-SINAI ENTROPY FOR PERIODIC VARIATIONS

Figure 6.6: Probability distributions on the states corresponding to the first observed symbol, being A in (a) or symbol B in (b) for a periodic variation of memory. For each output symbol only one distribution on the states exists resulting from the initial distribution. The relative frequencies of each distribution depend on the probability of occurrence of the observed symbol.

is obtained and represents a sum of two expressions originating from each observed symbol, which is weighted by its probability of observation. Within these terms, another sum takes the two possible transitions into account producing the next symbol under consideration of the current probability distribution on the states.

The result of the entropy rate for \( n = 1 \) evaluates to

\[
h_{KS} = 0.99515 \tag{6.54}
\]

based on the observation of solely one output symbol. Compared to the analytically deduced entropy rate (5.41) with its value of 0.92115, it differs tremendously. Indeed, as explained in 6.1, Blackwell’s measure is defined as the limiting stationary probability distribution as \( n \) tends to infinity. Thus, the next random dynamical iteration has to be performed.

The second iteration is achieved by applying the random dynamical iteration

\[
\tilde{\pi}^{(2)} = f_{c_2}(\tilde{\pi}^{(1)}) = f_{c_2}(f_{c_1}(\tilde{\pi}^{(0)})) \tag{6.55}
\]

twice to the initial distribution. The functions \( f_c \) have to be selected in correspondence to the first observed symbol \( c_1 \) and the second one \( c_2 \). Then, the joint probability \( p(c_1,c_2) \) of the succession of the two symbols is determined by

\[
p(c_1,c_2) = p(Mc_1,M'c_2) + p(M'c_1,Mc_2)
= p(Mc_1) \cdot p(M'c_2|Mc_1) + p(M'c_1) \cdot p(Mc_2|M'c_1) \tag{6.56}
\]
the involved internal states, their stationary probability, and the transition probabilities between the states. It multiplies the probability to be on a certain state by the conditional probability to change to a compatible one taking all states into account which generate the symbols $c_1$ and $c_2$.

As depicted in figure 6.7, this leads to the four available distributions

\[
\bar{\pi}^{(2)} = \begin{cases}
\bar{f}_A(f_A(\bar{\pi}^{(0)})) = \begin{pmatrix}
0.5405 \\
0 \\
0.4595 \\
0
\end{pmatrix} : \text{observed AA with } p(A, A) = 0.296 \\
\bar{f}_A(f_B(\bar{\pi}^{(0)})) = \begin{pmatrix}
0.7377 \\
0 \\
0.2623 \\
0
\end{pmatrix} : \text{observed BA with } p(B, A) = 0.244 \\
\bar{f}_B(f_A(\bar{\pi}^{(0)})) = \begin{pmatrix}
0.1639 \\
0 \\
0.8361
\end{pmatrix} : \text{observed AB with } p(A, B) = 0.244 \\
\bar{f}_B(f_B(\bar{\pi}^{(0)})) = \begin{pmatrix}
0.5556 \\
0 \\
0.4444
\end{pmatrix} : \text{observed BB with } p(B, B) = 0.216
\end{cases}
\]  

based on two observed symbols establishing Blackwell’s measure.

The distributions on the states and their relative frequencies after an observation of two symbols are illustrated in figure 6.7. For this purpose, all sequence observations ending in the same symbol were grouped together because the probability mass is only distributed among the states generating the last observed symbol. The relative frequencies of the distributions correspond to the probability of occurrence of the observed symbol sequences. Nevertheless, the sum of relative frequencies for the last symbol has to be constant throughout all observations. This is obvious because the long-term probabilities for the observation of the symbols are given in (6.48) and (6.49).

After transforming Blackwell’s measure into a probability density function and simplifying the integral, the entropy rate for $n = 2$ results in

\[ h_{KS} = 0.9907 \]  

(6.58)

and still differs from (5.41) significantly.
6.3 Kolmogorov-Sinai entropy for periodic variations

Figure 6.7: Probability distributions on the states corresponding to two observed symbols, the last one being A in (a) or symbol B in (b) for a periodic variation of memory. Because the final probability mass is located on those states which output the last symbol, all symbol observations ending in the same symbol can be put together in one graph. The relative frequencies of each distribution depend on the probability of occurrence of the observed symbol sequence.

In order to approximate the Kolmogorov-Sinai entropy, \( n \) has to be enlarged further. This is accomplished by applying the random dynamical iteration recursively computing it in a small program. Simultaneously, the probabilities for the observation of the symbol sequence are calculated using the transition probabilities of matrix \( \tilde{M} \) that generate the appropriate symbol. These values are used to form Blackwell’s measure and deduce the entropy rate for the specific \( n \). Due to the recursion creating \( 2^n \) different probability distribution vectors, a computation can be done up to \( n = 22 \) in a reasonable amount of time.

The behavior of Blackwell’s entropy rate is illustrated in figure 6.8. The entropy rate is plotted against \( n \) and compared to the conditional entropy, which is computed numerically from a sequence comprising one billion symbols as described in 5.2.4. Additionally, the analytical value (5.41) of the Kolmogorov-Sinai entropy is inserted as a straight line showing the limit to which the entropy rates should converge.

In figure 6.8, Blackwell’s entropy rate agrees very well with the numerical estimation via conditional entropies from 5.2.4 and shows an identical convergence. However, starting from \( n = 20 \) both plots differ, revealing a drop of the conditional entropy. The reason is obvious that the numerical computation is based on a finite symbol sequence and estimates the necessary probabilities of occurrence by relative frequencies. If the subsequences reach a certain length in relation to the overall length, this estimation will become unrea-
Figure 6.8: Blackwell’s entropy rate and numerically computed conditional entropy $\Delta H_n$ versus block length $n$ compared to analytically calculated Kolmogorov-Sinai entropy for a periodic variation between a first-order Markov chain and a Bernoulli process. Blackwell’s entropy rate and conditional entropy agree well and show an identical convergence towards the analytically calculated Kolmogorov-Sinai entropy. From $n = 20$ on, the conditional entropy decreases below Blackwell’s entropy rate due to inexact estimation of probabilities via relative frequencies.

liable. On the contrary, Blackwell’s measure depends on predefined probability distributions on the states and distributes them among the states induced by the iterations. Thus, these are real probabilities and can only become incorrect due to numerical errors.
7 Summary

7.1 Results and conclusions

The investigation of Markov chains with varying order revealed aspects contributing to an improved understanding of the effects on the behavior of a fluctuating memory length. Due to the close connection to delay differential equations with time-dependent delay length, the obtained results can be assigned to their properties enabling further research and comprehension.

The fluctuation of memory length was established by an appropriate selection process between two Markov chains with different order. The analyzed processes with a random variation and a periodic variation represent borderline cases between uncorrelated and strongly correlated selection, respectively.

For a Markov chain of fixed order, the entropy rate, known as Kolmogorov-Sinai entropy, is well known and was calculated analytically by means of its transition matrix. Additionally, such analytical expressions were deduced for the presented special cases varying the order of the Markov chain. For the purpose of verification, a numerical estimation of the entropy rate of any desired symbol source was implemented. The numerical computation of entropies was realized by a counting algorithm which determined the relative frequencies of the existing subsequences within a generated symbol sequence. In order to approximate the Kolmogorov-Sinai entropy, two approaches based on block entropies were pursued calculating per-symbol and conditional entropies. However, both have to be considered for block lengths tending to infinity, which introduces computational barriers due to the exponentially growing number of combinations. As a result, even ordinary Markov chains showed a slow convergence of the per-symbol entropy to the limit of the analytical value. On the other hand, the conditional entropy approached the Kolmogorov-Sinai as soon as all correlations produced by the source are included. This characteristic behavior identifies Markov chains and their respective order. Besides, the approximation of the probabilities by relative frequencies required for the calculation of block entropies became inaccurate for long subsequences due to the exponential growth of the number of combinations and the finite size of the symbol sequence. Hence, the conditional entropy differed from the analytical value again.

Two first-order Markov chains were used to construct a Markov chain with random variation of memory length. The evolution of the numerical entropies revealed a coincidence
with the characteristic behavior of an ordinary first-order Markov chain. Consequently, the existence of an effective first-order Markov chain was deduced possessing an identical long-term behavior and thus the same Kolmogorov-Sinai entropy. The numerical value of the conditional entropy and the analytical result agreed well. Moreover, the random variation between the two Markov matrices was modified by changing the probability of occurrence belonging to each matrix. As a result, the Kolmogorov-Sinai entropy reached higher values for certain mixing probabilities than the entropy rates of the involved processes achieved separately. This behavior was encouraged by the structure of the matrices. Reversely, it can be used to combine Markov chains with low entropy rate randomly, establishing a new Markov process with higher Kolmogorov-Sinai entropy, which could be advantageous in compression theory.

A periodic fluctuating memory of a Markov chain was established by an alternating selection of two first-order Markov chains. The deduced analytical expression for the entropy rate involved the two possible products of the Markov matrices originating from grouping in pairs. Both per-symbol and conditional entropy converged slowly but confidently towards the analytically calculated Kolmogorov-Sinai entropy. Further investigations detected an exponential decline to the entropy rate confirming the limit of convergence.

Another verification dealt with Blackwell’s measure as a facility to calculate the entropy rate of hidden Markov chains. For this purpose, the connection of the Markov chains with varying memory to special types of hidden Markov chains was shown, enabling the results to be adapted to them. However, even Blackwell’s entropy rate suffered from the weak convergence for Markov chains with periodic varying memory length and simultaneously agreed well with the numerical results. For a random variation between two Markov chains, Blackwell’s entropy rate yielded the analytical determined value. By applying the findings reversely, an inference to the internal structure of a hidden Markov chain can be drawn by analyzing the behavior of the entropies.

In summary, the analytical expressions of the Kolmogorov-Sinai entropy for the special cases of memory variation contained matrices related to the matrices of the elementary Markov chains, and the values of the entropy rates were supported by the numerical results and Blackwell’s approach.

### 7.2 Outlook

According to the investigations, several enhancements can be considered advancing the analysis of Markov chains with varying order.

The problem related to finite-length symbol sequences can be reduced by generating the symbols continuously without predefining a length. However, this modification in the
implementation does not eliminate the problem of the exponentially growing number of combinations, which have to be counted and consume a large amount of memory.

Besides, additional special cases for the machine selection can be investigated. For instance, a machine selection depending on the last output symbol can cause new effects and another characteristic behavior of the entropy rates. In general, the more various cases have been investigated, the more characteristics can be used to detect internal structures of hidden Markov chains.

Equally interesting is the connection to Lyapunov exponents related to random products of matrices [FK60]. By transforming the Markov chains with varying memory into Markov maps, the Lyapunov exponents can easily be determined. This enables verification and furthermore gives a correspondence to Kolmogorov-Sinai entropies incorporating Pesin’s theorem.

Moreover, in time-series and text analysis with Markov models, numerous transition probabilities equal zero because they cannot be found in the training sequence due to its finite size. The mixing of Markov chains with different orders can be utilized to overcome problems with large numbers of combinations for higher order Markov models as proposed in [SPCW97]. The above investigations provide approaches to construct such combinations with different orders keeping the entropy rate of the original process unmodified.
Bibliography


Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig angefertigt, nicht anderweitig zu Prüfungszwecken vorgelegt und keine anderen als die angegebenen Hilfsmittel verwendet habe. Sämtliche wissentlich verwendete Textausschnitte, Zitate oder Inhalte anderer Verfasser wurden ausdrücklich als solche gekennzeichnet.

Chemnitz, den 1. Juli 2008

__________________________
Michael Bauer