Generalized Markov approximations from information theory and consequences for prediction

Dissertation

zur Erlangung des Doktorgrades
der Fakultät der Naturwissenschaften
Doctor rerum naturalium (Dr. rer. nat.)
der Bergischen Universität Wuppertal

WUB-DIS 2007-12

vorgelegt von

Detlef Holstein
geboren am 11.06.1973 in Bremen

Dezember 2007

erstellt am
MAX-PLANCK-INSTITUT FÜR PHYSIK KOMPLEXER SYSTEME
DRESDEN

Diese Dissertation kann wie folgt zitiert werden:

urn:nbn:de:hbz:468-20080255
[http://nbn-resolving.de/urn/resolver.pl?urn=urn%3Anbn%3Ade%3Ahbz%3A468-20080255]

Date of submission:	December 13, 2007

Supervisor:	Prof. Dr. H. Kantz
Second evaluator and examiner:	Prof. Dr. A. Klümper
Examiner:	Prof. Dr. R. Koppmann

Abstract

(English)

For the justification of Markov approximations novel criteria based on information theory and statistics are developed. Ignored memory under conditioning and the statistical error of redundancy are related, in order to obtain optimal conditionings in the sense of generalized Markov approximations. The introduced criteria are checked successfully by various example dynamics as autoregressive processes, generalized Hénon dynamics and Mackey-Glass dynamics. The new kind of model selection results in improved predictions, which is verified via resolution-dependent prediction errors. As an application the prediction of wind speeds is carried out and the arising problems are analyzed. In the framework of generalized Markov approximations a new method for improved estimation of the Kolmogorov-Sinai entropy is suggested.

The introduction of a new notation for entropies in time series analysis, which enables access to the generalized criterion, allows for a unified description of various situations of modelling and prediction as, e.g., arbitrary omission of time steps in conditionings, downsampling and variable future prediction times. A non-widespread generalization of the quantity 'mutual information' from two to several random variables is made known in the community of time series analysts. With this quantity redistributions of information corresponding to future random variables under arbitrary omissions of conditioning time steps in the past are calculable analytically in prediction situations from basic constituents.

A time-continuous version of the theory of entropies and entropy rates is introduced, which offers a new distinction of chaos and stochasticity. In the case of deterministic dynamics upper thresholds of uncertainties in the limit of infinitely high sampling rates can be determined. The dimension and the Kolmogorov-Sinai entropy of a dynamics can be read out from one single 3D-graph.

The relevant foundations for this work concerning stochastic processes, information theory, prediction and hydrodynamics are presented.

Abstract

(German)

Auf der Grundlage von Informationstheorie und Statistik werden neue Kriterien zur Rechtfertigung von Markov Approximationen entwickelt. Bei Konditionierung ignorierter Memory und der statistische Fehler der Redundanz werden miteinander in Beziehung gesetzt, um optimale Konditionierungen im Sinne verallgemeinerter Markov Approximationen zu erhalten. Die eingeführten Kriterien werden mit verschiedenen Beispieldynamiken wie autoregressiven Prozesssen, verallgemeinerter Hénon Dynamik und Mackey-Glass Dynamik erfolgreich überprüft. Mit auflösungsabhängigen Vorhersagefehlern wird verifiziert, dass die neue Art der Modellselektion verbesserte Vorhersagen liefert. Als Anwendung wird die Vorhersage von Windgeschwindigkeiten durchgeführt und die dabei auftretenden Probleme werden analysiert. Im Rahmen der verallgemeinerten Markov Approximationen wird eine neue Methode zur verbesserten Schätzung der Kolmogorov-Sinai Entropie vorgeschlagen.

Die Einführung einer neuen Notation für Entropien in der Zeitreihenanalyse, welche den Zugang zum verallgemeinerten Kriterium erst ermöglicht, erlaubt eine vereinheitlichte Beschreibung von verschiedenen Situationen der Modellierung und Vorhersage wie z.B. willkürliche Auslassungen von Zeitschritten in Konditionierungen, Downsampling und variable Zukunftsvorhersagezeiten. Eine nicht weit verbreitete Verallgemeinerung der Grösse 'Mutual Information' von zwei auf mehrere Zufallsvariablen wird in der Gemeinschaft der Zeitreihenanalytiker bekanntgemacht. Mit dieser Grösse werden in Vorhersagesituationen Umverteilungen von Informationen zur Zukunft korrespondierender Zufallsvariabler unter willkürlicher Auslassung konditionierender Zeitschritte der Vergangenheit aus elementaren Konstituenten analytisch berechenbar.

Eine zeitkontinuierliche Version der Theorie der Entropien und Entropie-Raten wird vorgestellt, mit der sich eine neue Unterscheidungsmöglichkeit von Chaos und Stochastizität eröffnet. Im Fall deterministischer Dynamik können obere Schranken von Unsicherheiten für den Limes unendlich hoher Sampling-Raten bestimmt werden. Die Dimension und die Kolmogorov-Sinai Entropie einer Dynamik können in einem einzigen 3D-Graphen abgelesen werden.

Es werden die für diese Arbeit relevanten Grundlagen der stochastischen Prozesse, Informationstheorie, Vorhersage und Hydrodynamik präsentiert.

Contents

1 Introduction **9**

2 Stochastic processes **13**
- 2.1 Introductory definitions of stochastics 13
- 2.2 Stochastic processes 14
  - 2.2.1 Properties of stochastic processes 14
  - 2.2.2 Gaussian processes 16
  - 2.2.3 Properties restricted to the time-discrete case 20
- 2.3 Time-discrete linear stochastic processes 21
  - 2.3.1 Autoregressive AR(p) processes 21
  - 2.3.2 Related process classes 24
- 2.4 Nonlinear stochastic processes 26

3 Information theory and geometry **27**
- 3.1 General information theory 27
  - 3.1.1 Elementary definition 27
  - 3.1.2 Generalizations of elementary definition 28
  - 3.1.3 Conditional entropy 29
  - 3.1.4 Kullback-Leibler distance 31
  - 3.1.5 Mutual information 32
  - 3.1.6 Multiple mutual information 33
- 3.2 Connection with the entropies of statistical physics 35
- 3.3 Information theory in time series analysis for the block-case 36
  - 3.3.1 Joint entropies 36
  - 3.3.2 Conditional entropies 38
  - 3.3.3 Entropy rates 41
  - 3.3.4 Jointly conditioned joint entropies 42
  - 3.3.5 Redundancy 44
- 3.4 Entropies of Markov processes 44
- 3.5 Dynamical entropies and dimensions 46
  - 3.5.1 Topological and metric entropy 46
  - 3.5.2 A few comments on dimensions 47
  - 3.5.3 Systematics of entropies and dimensions for different Renyi order 47
- 3.6 Estimation from data 47
  - 3.6.1 Fixed size methods 48

3.6.2 Further classification of estimation methods for entropies and dimensions 50
3.6.3 Statistical error of entropy estimates via correlation sum 51
3.7 Interrelation of the correlation integral, entropies and dimensions 53
3.8 Entropies in case of continuous time 55
3.8.1 Usual procedure 55
3.8.2 Discrepancies from usual differentiation and integration structure of entropies in discrete time 56
3.8.3 Time-continuous information theory 57
3.8.4 Example calculations for time-continuous entropies 61
3.8.5 Conclusion and outlook 73

4 Entropies on perforated structures 75
4.1 Motivation and aim 75
4.2 General definitions of entropies and derived quantities on perforated structures 76
4.3 Estimation of entropies in perforated case 78
4.4 Special cases of generalized prediction and modelling situations 79
4.4.1 Variable future prediction times 79
4.4.2 Arbitrary joint prediction 81
4.4.3 Perforation in conditioning of entropies 82
4.4.4 Connection of multiple mutual information and perforation 87
4.4.5 Changed sampling: Downsampling and upsampling 89
4.4.6 Example calculation of AR(2) for downsampling 92
4.5 Comment on dimensions in perforated case 94

5 Origin and description of memory 95
5.1 The problem 95
5.2 Origin and nature of memory 95
5.2.1 Resolution, effective degrees of freedom and Markovianity 95
5.2.2 Methods of complexity reduction 96
5.2.3 A priori memory-including modelling 97
5.2.4 Conclusion 97
5.3 Entropic assessment of memory 97
5.3.1 Quantities in block-case 97
5.3.2 Generalization to perforated case 98

6 Optimal Markov approximations 99
6.1 Introduction 99
6.2 Optimal usual Markov approximation 100
6.2.1 A novel criterion 100
6.2.2 Examples 100
6.2.3 Concluding remarks 103
6.3 Optimal perforated Markov approximation 103
6.3.1 A novel criterion 103
6.3.2 Simplified approximative representation of the criterion 104
6.3.3 Implementational remarks 105
6.4 Application to different given dynamics in the perforated case 106

6.4.1 AR processes . . . 106
6.4.2 Generalized Hénon map . . . 108
6.4.3 Mackey-Glass dynamics . . . 111
6.5 Application for different parameter values in the perforated case . . . 113
6.5.1 Variable future lead time . . . 113
6.5.2 Dataset length dependence of optimal perforated Markov model . . . 114
6.5.3 Optimal perforated Markov model in hypothetical case of infinite length of the dataset . . . 115
6.6 Iterative procedure for the construction of the optimal conditioning and improved estimation of the KS entropy . . . 116
6.6.1 Presentation of the method . . . 116
6.6.2 Examples . . . 118
6.6.3 Discussion of the method . . . 119
6.6.4 Delay determination and search for feedback time . . . 120
6.7 Comparison with the literature . . . 120
6.7.1 Information criteria . . . 120
6.7.2 Comparison of AIC and MDL with the newly introduced criterion . . . 121
6.7.3 Comparison of methods based on functional independence with the newly introduced criterion . . . 122
6.8 Conclusion . . . 123

7 Prediction **125**
7.1 Point prediction and prediction error . . . 125
7.2 Locally constant prediction with generalized delay vectors . . . 125
7.2.1 Specialized point prediction . . . 125
7.2.2 Empirical minimization of the prediction error . . . 126
7.2.3 Discussion on the balance of the introduced criterion . . . 129
7.2.4 Prediction from optimal perforated Markov model . . . 130
7.2.5 Comparison of empirical prediction error optimization with entropic access to prediction error optimization . . . 133
7.2.6 Detection of suitable resolution interval or even of optimal resolution for prediction . . . 133
7.3 An overview of prediction methods . . . 134
7.3.1 Classificational aspects . . . 134
7.3.2 Arrangement of the prediction method based on perforated Markov models into the general classification of prediction methods . . . 136

8 Fluids and wind **139**
8.1 Theoretical basis of hydrodynamics . . . 139
8.1.1 Navier-Stokes equation . . . 139
8.1.2 Turbulence . . . 140
8.1.3 Kolmogorov theory . . . 142
8.1.4 Boundary layer . . . 143
8.2 Theoretical basis of wind dynamics . . . 145
8.2.1 Origin of wind . . . 145
8.2.2 Atmospheric boundary layer . . . 145

8.2.3 Aspects of the physics of gusts and extreme wind events 148
8.3 Conclusion 148

9 Application of discussed concepts and proposed methods on wind data 149
9.1 Wind data 149
9.2 Stochastic modelling of wind data 150
9.3 Entropies of wind data in nonperforated case 150
9.4 Entropies of wind data in perforated case 151
9.4.1 Variable future lead time 151
9.4.2 Regular perforation by downsampling 151
9.5 Finding optimal Markov approximation 153
9.5.1 Finding optimal usual Markov approximation for wind data 153
9.5.2 Finding optimal perforated Markov approximation for wind data . . . 154
9.6 Wind speed prediction 156
9.6.1 General remarks 156
9.6.2 Prediction from optimal perforated Markov model 156
9.7 Discussion of the results for prediction with generalized delay vectors from the criterion of information theory and statistics in the context of wind data . . 159

10 Conclusion 163

11 Outlook 165

A Determinism versus stochasticity 167

B Prediction via precursors: ROC plot from clustering analysis 171
B.1 Introductory remarks 171
B.2 Central tool: ROC plot 171
B.3 Cluster analysis 172
B.4 Resulting prediction scheme 173
B.5 Results 175
B.5.1 In-sample ROC as consistency check 175
B.5.2 Out-of-sample prediction from clustered pre-structures 175
B.6 Problems with this approach 176
B.7 Research on precursors in general 176

C Is information theory a suitable starting point for predicting extremes, e.g., extreme wind events? 177

Bibliography 181

Acknowledgements 187

Chapter 1

Introduction

Reliable prediction was and will probably forever be a very important task, e.g., the prediction of natural hazards (floods or eruptions of volcanos) to save lifes, prediction in economy (crashes of stock market) to save property, or simply a weather prediction for the plan and equipment of the next day. Central behind the ability of predicting is the fact that in some way available knowledge at a certain time can be used for reducing the uncertainty about the future.

In general, dynamics can contain memory, i.e., a dependence of the future on the further past exceeding the presence exists. This is known, e.g., from the framework of delay differential equations or from maps with multiple dependences. Thus a possible dependence of the future on the further past has to be taken seriously into account especially with respect to forecasting in an optimal sense.

In the context of tasks for prediction the dynamical laws of a complex system are typically not known a priori, and hence usually time series from measurements of possibly several physical quantities at equidistant times are the starting point. Measurements can be understood as projections of the system states. Having performed a projection of arbitrary complex stochastic dynamics into one physical quantity even allows in principle for infinite memory time hidden in the time series.

For practical purposes, on the other hand, memory in infinitely many time steps in the past is not usable, because of the finiteness of computational resources. E.g., with respect to prediction, neighbor search methods in a time series, i.e., searches for similar states represented by similar time series segments, are common. Those methods find their justification in the assumption of continuity in the time evolution of states, meaning that similar states in the past are assumed to lead to similar states in the future. The longer a time series segment of conditioning as consequence of memory is chosen, the fewer similar states can be found, and statistical problems start to become crucial.

Henceforth there exists the need for the development of a criterion for a theoretically justified Markov approximation for suitable truncation of memory such that the problem of prediction is optimally manageable. This is one of the central tasks of this work. Since memory is not necessarily continuously equidistributed over the past, but can instead concentrate in certain time differences from the presence, it is imaginable that Markov approximations with omissions could be advantageous, where the truncation is more subtle than only removing all weak dependences in the further past.

The central theoretical framework for the evaluation of memory and uncertainties of random variables in general can be obtained from the entropies of information theory and derived quantities. Thus a broad space of this work is dedicated to information theory. The question concerning redistribution of information among random variables under change of the number of participating random variables has to be understood in general, in order to be able to calculate changes of information for prediction under the conditions of advanced truncations of memory.

From the quantification of losses of information and under statistical demands it is then possible to formulate a quantitative optimization procedure for obtaining optimal generalized Markov approximations. The selected generalized Markov model can serve as an input for a suitable prediction method being optimal from the point of view of information theory and statistics for a broad class of nonlinear stochastic processes. The optimization of predictions is obtained from a targeted loss of information.

An application for prediction can be found in the study of time series of wind speeds in the atmospheric boundary layer, simply measured with cup anemometers. Nowadays there exists an increased interest in extracting energy from renewable resources, especially from wind rotors, because of the global climate consequences of emission rates of, e.g., coal power plants and the existing strong risks connected with the energy from nuclear power stations. As a consequence, in the last 20 years the number of windfarms in Germany increased a lot. With this new technology also new kinds of problems have to be solved as for example the problems of prediction of power supply or the prevention of damage of wind engines due to strong winds. The solution of such problems is even more important with regard to much larger and more expensive offshore wind power farms, which are nowadays still a technological challenge.

The present work is divided into the following parts:

In chapter 2 some basics of the theory of stochastic processes are presented. Especially the Ornstein-Uhlenbeck process, autoregressive processes and subtleties of the Markov property are discussed.

Subsequently the fundamentals of information theory are presented in chapter 3. Entropic quantities are introduced in general and in the narrower framework of time series. Concerning information theory for time series a ’flow’-picture instead of a ’map’-picture is chosen consequently, what explains differences of the formulas compared to [46]. Differences in changes of the Renyi order are discussed. A non-widespread generalization of the formula of mutual information for several random variables instead of two is supported for distribution among the scientists. Various estimation methods for entropies and dimensions are classified, especially a method involving the correlation sum. A generalization of formulas of information theory from discrete to continuous time is introduced and deterministic and stochastic examples are calculated numerically. This excursus extends the framework given in the work [31] and opens a fruitful ground for in the author’s opinion powerful results.

In chapter 4 omission of time steps in conditioning of Markov models, a concept in this work called ’perforation’, and consequences for entropies are described. It is not only the aim of this chapter to be able to treat variable future prediction times (lead time), joint prediction, arbitrary perforation and changed sampling rates all for their own in the sense of information theory, but to provide a new unified notational framework, in which all such problems can be treated as special cases. This introduced framework from information theory is a suitable basis for the search of optimal generalized Markov approximations. The generalization of mutual information is used for analytically calculating the redistribution of information for prediction under omission of time steps.

An explanation of the origin of memory is given in chapter 5 and the relationship between memory and complexity-reduction is illuminated. It is tried to relate the principle of locality and memory. For the purposes of this work memory is assessed with the quantities of information theory.

In chapter 6 novel criteria based on information theory and statistics for the optimal resolution-dependent usual as well as perforated Markov approximation are introduced and applied to various known dynamics, e.g. autoregressive processes, the Hénon map and the Mackey-Glass dynamics. Different realizations of the algorithms for selecting the optimal perforated Markov model are suggested. A resolution-integrating suggestion leads to an improved method of the estimation of the Kolmogorov-Sinai(KS) entropy rate from information theory. Furthermore the algorithm serves as a delay finder for arbitrary dynamics. A comparison with usual model selection criteria and with methods based on functional independence is carried out.

In chapter 7 the local perforated point prediction is evaluated by a resolution-dependent root mean squared (rms) prediction error. For stochastic and deterministic example dynamics the resolution-dependent empirically optimal conditioning is obtained. On the basis of the selected resolution-dependent optimal perforated Markov model the resolution-dependent rms prediction error is compared to those from the conditioning in the sense of standard delay vectors. The relationship of the empirical and entropic method is explained. The variability of general prediction methods is partly exposed and the introduced method is embedded into the framework of prediction methods.

Some basics of fluid dynamics and turbulence are worked out in the first part of chapter 8. Afterwards the physics behind wind phenomena in the atmospheric boundary layer is illuminated and the used example for the application of the prediction scheme is prepared theoretically.

Chapter 9 starts from data sets of wind speed measurements in the lower atmospheric boundary layer. Estimates of entropies of the wind data are investigated under various conditions and the suggested model selecting criteria are applied to wind data. Instead of explicitly adressing the above formulated complex predictive aims concerning the wind, the consequences of the criteria from information theory and statistics for point prediction of the wind speeds are shown and the arising problems are analyzed.

In chapter 10 the essential results of this work are concluded and in chapter 11 an outlook is given.

Appendix A serves for collecting the different results of this work concerning determinism and stochasticity. In comparison with chapters 7 and 9 a completely different prediction scheme rather suitable for the prediction of extreme events is qualitatively outlined in appendix B. Here typical structures before gusts are clustered after suitable rescaling of the relevant data segments. This kind of precursor is used in a prediction scheme evaluated by the receiver operator characteristic (ROC). The use of information theory for the prediction of extremes is critically discussed in appendix C.

Chapter 2

Stochastic processes

2.1 Introductory definitions of stochastics

Axiomatically introduced probability theory is based on the notion of the so-called 'probability space' $(\Omega, \mathcal{A}, P)$. Ω is the space in which the random experiment can take its values, $\mathcal{A}$ is a suitable set of subsets of Ω and the probability measure P is a function assigning a value in $[0;1]$ to every element of $\mathcal{A}$ such that $P(\Omega) = 1, P(\phi) = 0$ (ϕ: empty set) and a suitable additivity property is fulfilled[1].

In this work random variables[2] X are assumed to be maps from Ω to $\mathbb{R}$. They do not contain randomness in their mapping rule, but transport probabilities from the basic probability space into an induced probability space. Probability densities p are given by the property

$$p(x)\,dx = P(\{X \in [x, x+dx]\}) \ . \tag{2.1}$$

Given a set of random variables $X_1, ..., X_n$, a joint probability density fulfils

$$p(x_1, ..., x_n)dx_1...dx_n = P(\{X_1 \in [x_1, x_1 + dx_1], ..., X_n \in [x_n, x_n + dx_n]\}) \ . \tag{2.2}$$

Conditional probability densities are defined from joint probability densities as

$$p(x_1, ..., x_k | y_1, ..., y_l) = \frac{p(x_1, ..., x_k, y_1, ..., y_l)}{p(y_1, ..., y_l)} \ . \tag{2.3}$$

Using this definition joint probability densities are always factorisable according to

$$\begin{aligned} p(x_{k-n}, ..., x_k) &= p(x_{k-n}, ..., x_{k-1}) \cdot p(x_k | x_{k-n}, ..., x_{k-1}) \qquad (2.4) \\ &= p(x_{k-n}) \cdot p(x_{k-n+1}|x_{k-n}) \cdot p(x_{k-n+2}|x_{k-n}, x_{k-n+1}) \cdot ... \cdot p(x_k|x_{k-n}, ..., x_{k-1}) \ . \end{aligned}$$

The expectation value of a random variable X is obtained from

$$EX = \int_{\mathbb{R}} x\, p(x)\, dx \ . \tag{2.5}$$

---

[1] A more thorough formulation in the sense of mathematics cannot be justified in the framework of this thesis, even though it is possible in principle. Further details can be found in [5] or [7].

[2] The capital letter X is used for random variables, whereas the small letter x is used for realizations of stochastic experiments. On the other hand bold letters $\mathbf{x}$ are used to indicate vectors whereas 'usual' letters indicate scalar variables.

The covariance of two random variables is defined as

$$\mathrm{Cov}(X,Y) := E[(X-EX)(Y-EY)] = E(XY) - EX \cdot EY \tag{2.6}$$

and the variance

$$\mathrm{Var}(X) := \mathrm{Cov}(X,X) \tag{2.7}$$

is included as a special case. Correlations are obtained from normalized covariances ([7], p.283 ff) by

$$\tilde{\rho}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}} \ . \tag{2.8}$$

Two random variables X, Y: $\Omega \to \mathbb{R}$ are defined to be stochastically independent with respect to the probability measure P iff (if and only if)

$$P(\{X \in A\} \cap \{Y \in B\}) = P(\{X \in A\}) \cdot P(\{Y \in B\}) \qquad \forall\, A \times B \subseteq \mathbb{R}^2. \tag{2.9}$$

Very often the shortcut 'iid' appears, which means 'independent identically distributed' refering to random variables. Stochastic independence of two random variables X and Y implies uncorrelated X and Y, i.e.,

$$P(\{X \in A\}\cap\{Y \in B\}) = P(\{X \in A\})\cdot P(\{Y \in B\}) \ \forall\, A\times B \subseteq \mathbb{R}^2 \ \Rightarrow \ \rho(X,Y) = 0. \tag{2.10}$$

The opposite implication is not valid in general.

2.2 Stochastic processes

A stochastic process is defined as a family of random variables $(X_t)_{t\in I}$ parametrized by an index set I such that [56] $\quad X : I \times \Omega \to \mathbb{R}\ ; \quad (t,\omega) \mapsto X_t(\omega)$.

Usually the elements of the index set are interpreted as times. If $I = \mathbb{R}$, the stochastic process is called time-continuous, while if $I = \mathbb{Z}$ the stochastic process is called time-discrete. Corresponding joint probability densities have to be positive, normalized, symmetric in their arguments and complete in the sense of

$$\int_{\mathbb{R}} p_n(x_1,t_1;x_2,t_2;...;x_n,t_n)dx_n = p_{n-1}(x_1,t_1;x_2,t_2;...;x_{n-1},t_{n-1}) \ . \tag{2.11}$$

2.2.1 Properties of stochastic processes

2.2.1.1 Markovianity

A stochastic process is called Markovian if for $t_k > t_{k-1} > ... > t_{k-n}$ the conditional probability densities fulfil

$$p(x_k,t_k|x_{k-n},t_{k-n};...;x_{k-1},t_{k-1}) = p(x_k,t_k|x_{k-1},t_{k-1}) \quad \forall n>1\ , t_i, x_i\ . \tag{2.12}$$

The right hand side of eq. (2.12) is called transition probability. For a Markov process, the stochastic process - pendant of eq. (2.4) simplifies to

$$\begin{aligned} p(x_{k-n}, t_{k-n}; ...; x_k, t_k) = p(x_{k-n}, t_{k-n}) \cdot p(x_{k-n+1}, t_{k-n+1} | x_{k-n}, t_{k-n}) \\ \cdot\, p(x_{k-n+2}, t_{k-n+2} | x_{k-n+1}, t_{k-n+1}) \cdot ... \cdot p(x_k, t_k | x_{k-1}, t_{k-1}) \; . \end{aligned} \tag{2.13}$$

It has to be pointed out that in a time-dependent context, factorisation in the sense of eq. (2.4) still holds, even for non-Markovian processes.

2.2.1.2 Stationarity

A stochastic process is called stationary if all joint probabilities are invariant under time translations, i.e.,

$$p_n(x_1, t_1; ...; x_n, t_n) = p_n(x_1, t_1 - \tau; ...; x_n, t_n - \tau) \quad \forall\ x_i, t_i, \tau, n \; . \tag{2.14}$$

Stationarity is assumed throughout this work, unless otherwise stated.

2.2.1.3 Homogeneity

Nonstationary Markov processes, whose transition probabilities depend on the time difference alone, are called homogeneous processes. Those processes are also called 'Markov processes with stationary transition probability' ([75], p.87).

2.2.1.4 Autocorrelation and autocovariance

In the framework of *stationary* stochastic processes, the correlation function in eq. (2.8) leads to the autocorrelation function given by

$$\rho(\tau) = \tilde{\rho}(X_t, X_{t+\tau}) = \frac{E(X_t X_{t+\tau}) - EX_t \cdot EX_{t+\tau}}{\sqrt{\mathrm{Var}(X_t) \cdot \mathrm{Var}(X_{t+\tau})}} \; . \tag{2.15}$$

The autocorrelation function without normalization by the variances is called autocovariance function. The non-centered autocovariance function is also called the *reduced* autocorrelation function[3]

$$\kappa(\tau) = E(X_t X_{t+\tau}) \tag{2.16}$$

From stationarity it is inferred that the time t is sufficiently large that transient effects can be neglected. Thus for the functions ρ and κ an explicit dependence on the time t does not appear.

The Fourier transform of the autocorrelation function is the power spectral density. This relation is known as the Wiener-Khinchin theorem. Since this work deals almost exclusively with dependence, Markovianity and describing entropies, but not explicitly with correlations, Fourier space treatments are missing completely.

---

[3] The use of merely the notion 'autocorrelation function' is not unusual in this context, but not supported by this work.

2.2.2 Gaussian processes

A stochastic process $(X_t)_{t\in I}$ on $(\Omega, \mathcal{A}, P)$ is called Gaussian, iff all finite dimensional joint probability densities are multivariate Gaussians

$$p_{n,\boldsymbol{\mu},\boldsymbol{\Theta}}(x_1, ..., x_n) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Theta})}} \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Theta}^{-1}(\mathbf{x}-\boldsymbol{\mu})] \; . \tag{2.17}$$

Here

$$\boldsymbol{\mu} = E\mathbf{X} \quad \text{and} \qquad \boldsymbol{\Theta} = E((\mathbf{X}-E\mathbf{X})^T(\mathbf{X}-E\mathbf{X})) \tag{2.18}$$

are the expectation value and the covariance matrix, respectively.

2.2.2.1 Gaussian white noise

A Gaussian process X_t with $I = \mathbb{R}$ is called Gaussian white noise η_t if as well $E(\eta_t) = 0$ and

$$E(\eta_t\eta_s) = c \cdot \delta(t-s) \; , \tag{2.19}$$

i.e., it is uncorrelated. The infinite-dimensional covariance matrix $\boldsymbol{\Theta}$ is diagonal and Gaussian white noise is stationary. Values $c = 1$ and $c = 2$ are common and for this work $c = 1$ is chosen. Since $\mathrm{Var}(\eta_t) = \infty$, Gaussian white noise is a mathematically idealized formal construction rather than a true physical process.

2.2.2.2 Standard Wiener process

A continuous-time stochastic process $\{W_t, t \geq 0\}$ with state space $(-\infty, \infty)$ is called standard (or unit) Wiener process, if

i) $W_0 = 0$
ii) $\{W_t, t \geq 0\}$ has stationary, independent increments
iii) W_t is normally distributed with $E(W_t) = 0$ and $\mathrm{Var}(W_t) = t,\ t > 0$;
i.e., $W_t \sim \mathcal{N}(0, t)$.

The standard Wiener process is a nonstationary Markov process describing Brownian motion. Almost all sample paths of the standard Wiener process are everywhere continuous, but nowhere differentiable. The non-differentiability of the standard Wiener process is a consequence of the scaling of the time increment of the standard Wiener process $dW_t \propto (dt)^{\frac{1}{2}}$, which is related to eq. (2.23). The independence of the increments alone would not be enough for inferring the non-differentiability, because independence of increments in principle would allow for other scaling relations from which non-differentiability would not follow. The formal time-derivative of the standard Wiener process $\dot{W}_t = \frac{dW_t}{dt} \equiv \eta_t$ is the Gaussian white noise process. The derivative at any point is almost certainly infinite ([29], p.68). From $W_t = \int_0^t dW_{t'} = \int_0^t \dot{W}_{t'} dt'$ and eq. (2.19) it is calculated

$$E(W_tW_s) = \int_0^t dt' \int_0^s ds' \; \delta(t'-s') = \min\{t, s\} \; . \tag{2.20}$$

For the function ρ, which because of nonstationarity of the process cannot be called 'autocorrelation function',

$$\lim_{\min\{t,s\}\to\infty} \rho(W_t, W_s) = \lim_{\min\{t,s\}\to\infty} \frac{\min\{t,s\}}{\sqrt{t}\sqrt{s}} = 1 \qquad \text{if} \quad |t-s| < \infty \tag{2.21}$$

is obtained. The nonstationarity is the reason, why this result should not be interpreted as the loss of stochasticity with increasing time, as it should be in the stationary case. Because of

$$E[(W_s - W_t)^2] = E[W_s^2 + W_t^2 - 2W_sW_t] = s + t - 2\cdot\min\{s,t\} = s - t \quad \text{for} \quad s > t\ , \tag{2.22}$$

the special case of $s = t + dt$ leads for the differential of the standard Wiener process to

$$E[(dW_t)^2] = E[(W_{t+dt} - W_t)^2] = dt\ . \tag{2.23}$$

From $E\,dW_t = (E\,\eta_t)dt = 0$ the variance

$$\mathrm{Var}(dW_t) = dt \tag{2.24}$$

and thus also the autocorrelation function

$$\rho(dW_s, dW_t) = \frac{E(dW_s dW_t)}{\sqrt{\mathrm{Var}(dW_s)}\sqrt{\mathrm{Var}(dW_t)}} = \frac{E(\eta_s\eta_t)ds\,dt}{\sqrt{ds}\sqrt{dt}} = \delta(s-t)\sqrt{ds}\sqrt{dt} \tag{2.25}$$

are obtained. Hence dW_t is uncorrelated, i.e., white Gaussian noise with distribution

$$dW_t \sim \mathcal{N}(0, dt)\ . \tag{2.26}$$

2.2.2.3 Ornstein-Uhlenbeck processes

An Ornstein-Uhlenbeck process can be defined as the stationary solution of the stochastic differential equation

$$dX_t = -\alpha X_t\,dt + \sqrt{D}\,dW_t\ . \tag{2.27}$$

dW_t is the differential of the standard Wiener process and D is the diffusion constant. Mathematically less correct, but for physicists not unusual, is

$$\dot{X}_t = -\alpha X_t + \sqrt{D}\,\dot{W}_t\ , \tag{2.28}$$

where $\dot{W}_t$ is Gaussian white noise. A successful ansatz is

$$X_t = X_0 e^{-\alpha t} + \sqrt{D}\int_0^t e^{-\alpha(t-t')} dW_{t'}\ . \tag{2.29}$$

The time evolution of the expectation value is derived as

$$EX_t = EX_0 \cdot e^{-\alpha t} \overset{t\to\infty}{\to} 0\ . \tag{2.30}$$

Here

$$E[\int_0^t e^{-\alpha(t-t')} dW_{t'}] = E[\int_0^t e^{-\alpha(t-t')}\dot{W}_{t'}dt'] = \int_0^t e^{-\alpha(t-t')}[E\dot{W}_{t'}]dt' = 0 \tag{2.31}$$

is used. For the variance it is calculated

$$\begin{aligned}\sigma_{x_t}^2 = \mathrm{Var}(X_t) &= E((X_t - EX_t)^2) \\ &= E(X_t^2) - (EX_t)^2 \\ &= [\mathrm{Var}(X_0) - \frac{D}{2\alpha}]e^{-2\alpha t} + \frac{D}{2\alpha} \\ &\overset{t\to\infty}{\to} \frac{D}{2\alpha} =: \sigma_x^2 \ . \end{aligned} \tag{2.32}$$

The time-dependence of the variance seems to suggest that the Ornstein-Uhlenbeck process is nonstationary. The truth is that the time-dependence of the variance is the change of the initial distribution reaching asymptotically the stationary equilibrium distribution. It holds that

$$E(X_t X_s) = [EX_0^2 - \frac{D}{2\alpha}]e^{-\alpha(t+s)} + \frac{D}{2\alpha}e^{-\alpha|s-t|} \ . \tag{2.33}$$

For sufficiently large time $\min\{s,t\}$ such that transient effects are avoided, the reduced autocorrelation function is calculated as

$$\kappa(s-t) = \frac{D}{2\alpha}e^{-\alpha|s-t|} \ . \tag{2.34}$$

Assuming $\tau \geq 0$

$$E(X_t X_{t+\tau}) = (EX_0^2 - \frac{D}{2\alpha})e^{-\alpha(2t+\tau)} + \frac{D}{2\alpha}e^{-\alpha\tau} \ . \tag{2.35}$$

For $\alpha > 0$ and sufficiently large t ($t \gg \frac{1}{\alpha}$)

$$\kappa(\tau) = \frac{D}{2\alpha}e^{-\alpha\tau} = \sigma_x^2 e^{-\alpha\tau} \tag{2.36}$$

results and hence also the autocorrelation function

$$\rho(\tau) = e^{-\alpha\tau} \tag{2.37}$$

is obtained as exponentially decreasing. For the time derivatives of the random variables

$$\begin{aligned}E(\frac{dX_t}{dt}\frac{dX_s}{ds}) &= E((-\alpha X_t + \sqrt{D}\eta_t)(-\alpha X_s + \sqrt{D}\eta_s)) \\ &= \alpha^2 E(X_t X_s) + D\ E(\eta_t\eta_s) \\ &= \alpha^2 E(X_t X_s) + D\delta(t-s)\end{aligned} \tag{2.38}$$

can be derived. From the stochastic differential equation it is seen that

$$D^{(1)}(x) = -\alpha x \quad \text{and} \quad D^{(2)}(x) = D = \text{const} \ , \tag{2.39}$$

i.e., the Fokker-Planck equation

$$\partial_t p(x,t) = -\partial_x D^{(1)}(x,t)p(x,t) + \partial_x^2 D^{(2)}(x,t)p(x,t) \ , \tag{2.40}$$

which corresponds to the Ornstein-Uhlenbeck process, becomes [29]

$$\partial_t p = \partial_x(\alpha x p) + D\partial_x^2 p \ . \tag{2.41}$$

It contains a linear drift coefficient and a constant diffusion coefficient. A solution can be given as

$$p(x,t|x_0,t_0) = \frac{1}{[2\pi(1-e^{-2\tau})]^{\frac{1}{2}}} e^{\frac{-(x-x_0e^{-\tau})^2}{2(1-e^{-2\tau})}} \quad \text{with} \quad \tau = t - t_0 > 0 \; , \tag{2.42}$$

where the explicit τ-dependence accounts for the transient behaviour.

As the standard Wiener process, the Ornstein-Uhlenbeck process is continuous and non-differentiable. The Ornstein-Uhlenbeck process is essentially the only process which is stationary, Gaussian and Markovian ([75], p.84).

2.2.2.4 Wiener processes from Ornstein-Uhlenbeck processes

In the limit of $\alpha \to 0$ and for initial condition $X_0 = W_0 = 0$, the (non-standard) Wiener process with constant diffusion D typically different from one is obtained from the Ornstein-Uhlenbeck process. The drift $D^{(1)}(x)$ vanishes. The stochastic differential equation in eq. (2.27) becomes

$$dX_t = \sqrt{D}\, dW_t \; . \tag{2.43}$$

For the variance from eq. (2.32) it holds

$$\begin{aligned} \lim_{\alpha\to 0} \mathrm{Var}(X_t) &= \lim_{\alpha\to 0}[(-\frac{D}{2\alpha})\cdot(1-2\alpha t) + \frac{D}{2\alpha}] \\ &= Dt \; . \end{aligned} \tag{2.44}$$

From $EX_0 = 0$

$$X_t \sim \mathcal{N}(0, Dt) \tag{2.45}$$

follows. It is possible to see that, whereas the Ornstein-Uhlenbeck process is stationary, the Wiener process is nonstationary. For the reduced autocorrelation function it holds that

$$\begin{aligned} \lim_{\alpha\to 0} E(X_t X_{t+\tau}) &= \lim_{\alpha\to 0}[\frac{D}{2\alpha}(-1+\alpha(2t+\tau)+1-\alpha\tau)] \\ &= Dt \; . \end{aligned} \tag{2.46}$$

For the autocorrelation function from eq. (2.44) and eq. (2.46), but of course also from eq. (2.37)

$$\lim_{\alpha\to 0} \rho(\tau) = 1 \tag{2.47}$$

is obtained. The result is independent of the diffusion constant D and the time difference τ. From

$$\zeta_t := \lim_{\alpha\to 0} \frac{dX_t}{dt} \tag{2.48}$$

and from eq. (2.38)

$$\lim_{\alpha\to 0} E(\frac{dX_t}{dt}\frac{dX_s}{ds}) = E(\zeta_t\zeta_s) = D\delta(t-s) \tag{2.49}$$

is inferred, i.e., it is derived that the diffusion constant plays the role of the strength of the white Gaussian noise corresponding to the general Wiener process. With eq. (2.43) it is consistently reobtained eq. (2.19)

$$E(\frac{dW_t}{dt}\frac{dW_s}{ds}) = E(\eta_t\eta_s) = \delta(t-s) \; . \tag{2.50}$$

For the increments it holds by linearity that

$$\lim_{\alpha\to 0} E(X_t - X_{t+\tau}) = 0\ . \tag{2.51}$$

This result is much stronger than, but nevertheless includes, the stationarity of the increments in the definition of the standard Wiener process.

For the covariance of increments it is calculated, starting with the Ornstein-Uhlenbeck process:

$$\begin{aligned}
E((X_t - X_{t-\tau})(X_s - X_{s-\tau})) &= E(X_tX_s - X_tX_{s-\tau} - X_sX_{t-\tau} + X_{t-\tau}X_{s-\tau}) \\
&\overset{(2.34)}{\approx} \frac{D}{2\alpha}[2e^{-\alpha|s-t|} - e^{-\alpha|s-\tau-t|} - e^{-\alpha|s-t+\tau|}] \\
&\overset{\alpha\to 0}{\approx} \frac{D}{2\alpha}[2(1-\alpha|s-t|) - 1 + \alpha|s-\tau-t| - 1 + \alpha|s-t+\tau|] \\
&\overset{D=1}{=} \frac{1}{2}[-2|s-t| + |s-\tau-t| + |s-t+\tau|] \\
&\overset{s=t+n\tau}{=} \frac{1}{2}[-2|n\tau| + |(n-1)\tau| + |(n+1)\tau|] \\
&= \begin{cases} \tau & \text{if} \quad n = 0 \\ 0 & \text{if} \quad n \neq 0 \end{cases} \ .
\end{aligned} \tag{2.52}$$

Whereas the Ornstein-Uhlenbeck process has correlated increments, the Wiener process does not. From the uncorrelatedness of the increments the stochastic independence is inferred as demanded in the definition of the standard Wiener process. In particular it follows

$$E(dW_t dW_s) = \begin{cases} dt & \text{if} \quad t = s \\ 0 & \text{if} \quad t \neq s \end{cases} \ , \tag{2.53}$$

to be compared with calculations in eq. (2.25).

With the former arguments it becomes clear, why the Wiener process, even though being a special representant of the class of Ornstein-Uhlenbeck processes, has qualitatively different properties.

2.2.3 Properties restricted to the time-discrete case

2.2.3.1 Markov processes of higher order

The definition of a Markov process is extended in discrete time by defining a *Markov process of order m* [34] to have the property

$$p(x_k|x_{k-n}, ..., x_{k-1}) = p(x_k|x_{k-m}, ..., x_{k-1}) \quad \forall\, n \geq m \geq 1\ . \tag{2.54}$$

It is necessary to remark that this definition is not generally supported by every author. Processes with $m \geq 2$ are very often already called non-Markovian. Having accepted the definition of a Markov process of order m, non-Markovianity is restricted to the case of conditioning in infinite time.

For a Markov process of order m eq. (2.4) takes the form

$$\begin{aligned} p(x_{k-n},...,x_k) &= p(x_{k-n},...,x_{k-n+m-1}) \cdot p(x_{k-n+m}|x_{k-n},...,x_{k-n+m-1}) \cdot ... \cdot p(x_k|x_{k-m},...,x_{k-1}) \\ &= p(x_{k-n},...,x_{k-n+m-1}) \prod_{j=k-n+m}^{k} p(x_j|x_{j-m},...,x_{j-1}) \; . \end{aligned} \tag{2.55}$$

2.2.3.2 Markov chains

A Markov process is called a *Markov chain* if the state space is discrete, time is discrete and the process is stationary or at least homogeneous [75].

It is useful to stress that the notion 'chain' in Markov chain has nothing to do with a Markov order larger than one. A generalization of the notion Markov chain to the case of Markov processes of higher order might be allowed, but then it is necessary to take care that the notion 'chain' is not implicitly associated with something seemingly chained in the higher order, but instead still sticks to the properties of discreteness in time and state and the stationarity property.

2.3 Time-discrete linear stochastic processes

The processes of the following classes are fully characterized by their two-point correlation-functions. Because of the Wiener-Khinchin-theorem, which connects the autocorrelation function with the power spectrum, all statistics of the process is thus contained in the power spectrum.

2.3.1 Autoregressive AR(p) processes

The stochastic process given by the dynamical law

$$X_{n+1} = \sum_{k=0}^{p-1} a_k X_{n-k} + \xi_{n+1} \;\; , \quad \xi_{n+1} \text{ Gaussian iid} \tag{2.56}$$

is called autoregressive process of order p (AR(p)). It is a particular Markov process of order p. Autoregressive processes are stationary if the roots of the characteristic polynomial lie inside the unit circle ([46], p.238). Defining the time shift operator

$$BX_n = X_{n-1} \; , \tag{2.57}$$

it is possible to express eq. (2.56) as

$$\pi(B)X_{n+1} = \xi_{n+1} \tag{2.58}$$

with a suitable polynomial π. For the autocorrelation function a recursion relation (Yule-Walker equation)

$$\rho_k = a_1\rho_{k-1} + a_2\rho_{k-2} + ... + a_p\rho_{k-p} \;\; , \quad k > 0 \tag{2.59}$$

holds. In general, the autocorrelation function of a stationary autoregressive process will consist of a mixture of damped exponentials and harmonic waves. In the following the special cases of low order are discussed.

2.3.1.1 AR(1)

AR(1) processes are given by ($a_0 \equiv a$)

$$X_{n+1} = aX_n + \xi_{n+1} \quad , \quad \xi_{n+1} \sim \mathcal{N}(0, \sigma_\xi^2) \quad \text{iid} \; . \tag{2.60}$$

For the second moment it holds that

$$E(X_{n+1}^2) = a^2 \cdot E(X_n^2) + 2a \, EX_n \cdot \underbrace{E\xi_{n+1}}_{=\,0} + E(\xi_{n+1}^2) \; . \tag{2.61}$$

From stationarity with $E(X_{n+1}^2) = E(X_n^2)$ it is derived

$$E(X_n^2) = \frac{\sigma_\xi^2}{1 - a^2} \quad \forall n \; , \tag{2.62}$$

demanding

$$|a| < 1 \; . \tag{2.63}$$

Thus the AR(1) process is stable for $|a| < 1$ and becomes unstable for $|a| \geq 1$. Stationarity with eq. (2.63) leads to:

$$EX_{n+1} = aEX_n \quad \Rightarrow \quad EX_n = 0 \;\; \forall n \quad \Rightarrow \quad E(X_n^2) = \sigma_x^2 \;\; \forall n \; . \tag{2.64}$$

Iteration of eq. (2.60) yields

$$E(X_n X_{n+k}) = a^k \cdot E(X_n^2) \; . \tag{2.65}$$

The autocorrelation function of the AR(1) shows exponential decay in the stationary case:

$$\rho_k = \frac{\kappa_k}{\sigma_x^2} = \frac{E(X_n X_{n+k})}{\sigma_x^2} = a^k \quad , \quad k \geq 0 \; . \tag{2.66}$$

Discretizing eq. (2.28) for the Ornstein-Uhlenbeck process it is derived

$$X_{n+1} = (1 - \alpha \Delta t) X_n + \sqrt{D} \sqrt{\Delta t} \, \eta_{n+1} \quad , \quad \eta_{n+1} \sim \mathcal{N}(0,1) \text{ iid} \; . \tag{2.67}$$

By identifying

$$a = e^{-\alpha \Delta t} \approx 1 - \alpha \Delta t \qquad \text{and} \qquad \xi_{n+1} = \sqrt{D} \sqrt{\Delta t} \, \eta_{n+1} \quad (= \sqrt{\Delta t} \zeta_{n+1}) \tag{2.68}$$

(Δt time between discrete steps) the AR(1) process can be regarded as the time-discrete version of an Ornstein-Uhlenbeck process. From $E(\xi_{n+1}^2) = D \Delta t \, E(\eta_{n+1}^2)$ the variance of the AR(1) process

$$\sigma_x^2 = \frac{\sigma_\xi^2}{1 - a^2} = \frac{D \Delta t}{1 - e^{-2\alpha \Delta t}} \approx \frac{D}{2\alpha} \tag{2.69}$$

is obtained as the variance of the Ornstein-Uhlenbeck process of eq. (2.32). From $\tau = k \Delta t$ the autocorrelation function of the AR(1) process

$$\rho_k = a^k = e^{-k \alpha \Delta t} = e^{-\alpha \tau} = \rho(\tau) \tag{2.70}$$

is transformed into the autocorrelation function eq. (2.37) of the Ornstein-Uhlenbeck process. As discussed before, the discrete version of the Wiener process is obtained from $\alpha \to 0$, i.e. $a = 1$, and the dynamical laws read

$$X_{n+1} = X_n + \sqrt{D}\sqrt{\Delta t}\,\eta_{n+1} \quad \text{and} \quad W_{n+1} = W_n + \sqrt{\Delta t}\,\eta_{n+1} \ . \tag{2.71}$$

With $X_0 = 0$ and $t = n\Delta t$ it is obtained

$$\mathrm{Var}(X_n) = \mathrm{Var}(X_{n-1}) + D\Delta t = \mathrm{Var}(X_0) + nD\Delta t = Dt = \mathrm{Var}(X_t) \tag{2.72}$$

the variance of the Wiener process. For $1 = \alpha\Delta t$, from eq. (2.67) the discrete version of the Gaussian white noise process results as

$$X_{n+1} = \sqrt{D}\sqrt{\Delta t}\,\eta_{n+1} \ . \tag{2.73}$$

2.3.1.2 AR(2)

The dynamical law for the AR(2) process reads

$$X_{n+1} = a_0 X_n + a_1 X_{n-1} + \xi_{n+1} \quad , \quad \xi_{n+1} \text{ Gaussian iid} \ . \tag{2.74}$$

The autocorrelation function of the AR(2) decays exponentially as for the AR(1), but a periodic modulation is possible.

The AR(2) process can be approximated by the discretization of the stochastically perturbed damped harmonic oscillator. The corresponding equation of motion reads

$$\ddot{X}(t) = -\gamma\dot{X}(t) - \omega_0^2 X(t) + b\eta(t) \ . \tag{2.75}$$

Here the parameters are the friction constant γ and the eigenfrequency ω_0. $\eta(t)$ is assumed to be δ-correlated noise:

$$\langle \eta(t)\eta(t') \rangle \;=\; \delta(t-t') \ . \tag{2.76}$$

Discretising eq. (2.75) it is obtained [49]

$$\frac{X_{n+1} - 2X_n + X_{n-1}}{\Delta t^2} = -\gamma\frac{X_{n+1} - X_{n-1}}{2\Delta t} - \omega_0^2 X_n + \frac{b}{\sqrt{\Delta t}}\eta_n \quad , \quad \eta_n \sim \mathcal{N}(0,1) \ . \tag{2.77}$$

A reformulation yields

$$X_{n+1} = \frac{4 - 2\omega_0^2\Delta t^2}{2 + \gamma\Delta t}X_n + \frac{-2 + \gamma\Delta t}{2 + \gamma\Delta t}X_{n-1} + \frac{2\Delta t^{\frac{3}{2}} b}{2 + \gamma\Delta t}\eta_n \ . \tag{2.78}$$

As the discretization of the Ornstein-Uhlenbeck process in sec. 2.3.1.1 led to an AR(1) process, the discretization of the stochastically perturbed damped harmonic oscillator leads to an *AR(2) process* with coefficients

$$a_0(\Delta t) = \frac{4 - 2\omega_0^2\Delta t^2}{2 + \gamma\Delta t} \quad , \qquad a_1(\Delta t) = \frac{-2 + \gamma\Delta t}{2 + \gamma\Delta t} \tag{2.79}$$

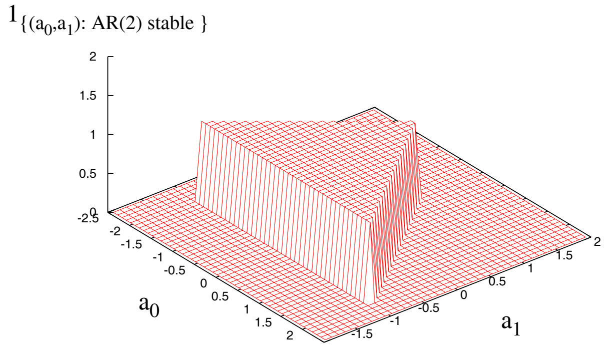


Figure 2.1: Domain of stability of the AR(2) process.

and uncorrelated Gaussian noise

$$\xi_n := \frac{2\Delta t^{\frac{3}{2}} b}{2+\gamma\Delta t}\eta_n \ . \tag{2.80}$$

The connection of the AR(2) process with the stochastically perturbed damped harmonic oscillator will be used in sec. 4.4.6. The characteristic function of the stability area of the AR(2) is shown in fig. 2.1.

In the limit $\Delta t \to 0$ the coefficients become $a_0 = 2$ and $a_1 = -1$, i.e., the top left corner of the stability triangle area is reached.

2.3.1.3 AR(3)

In chap. 7 the stability area of the two process classes shown in fig. 2.2 is needed. Even though AR processes are linear and seemingly simple, the form of domains of stability turns out to be not immediately obvious, e.g., the borders of the stability areas are not straight.

2.3.2 Related process classes

It is preferable to remark already here that the following argumentation is purely representation theory! The introduced process classes do not contain new stochastic processes, but yield only a different representation of the same processes as in the AR class.

2.3.2.1 Moving Average MA(q) processes

A moving average process is given by

$$X_{n+1} = \sum_{l=0}^{q} b_l \xi_{n+1-l} \ , \quad \xi_i \text{ Gaussian iid} \ . \tag{2.81}$$

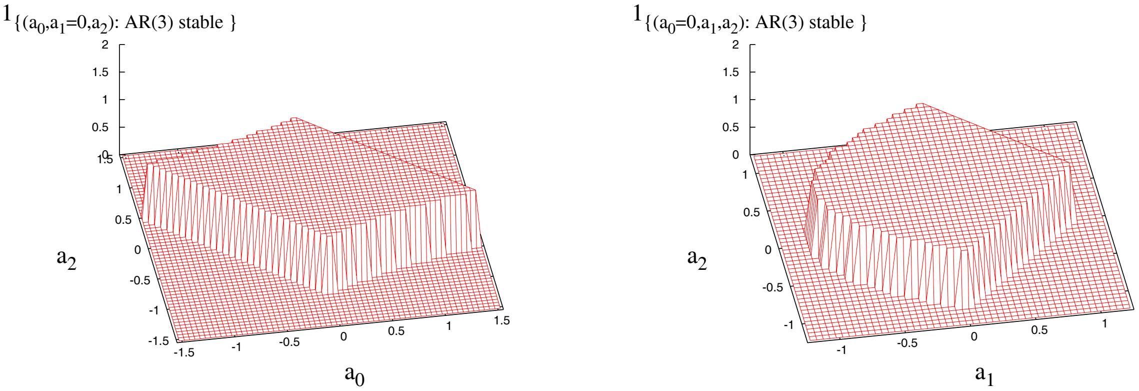


Figure 2.2: Stability area of special AR(3) processes. Left panel: Domain of stability of the AR(3) process if $a_1 = 0$ is fixed. Right panel: Domain of stability of the AR(3) process if $a_0 = 0$ is fixed.

It is called the Wold representation of time-discrete linear stochastic processes. Using the definition of the time shift operator (2.57), an MA process is also representable as

$$X_{n+1} = \psi(B)\,\xi_{n+1} \tag{2.82}$$

for a suitable polynomial ψ. Every finite order MA process is stationary for all parameter values ([16], p.37). Every AR process with finitely many coefficients not equal to zero is converted to a MA process with infinitely many coefficients not equal to zero. Hence Markov processes (of arbitrarily small order) have infinitely many coefficients in the MA representation. The order of MA processes gives the maximal length of autocorrelation different from zero. Autoregressive and moving-average processes behave in some sense complementary concerning invertibility, stationarity, finiteness of number of terms in the sum and spectral properties [11].

2.3.2.2 ARMA(p,q) processes

The ARMA representation of above introduced processes is

$$X_{n+1} = \sum_{k=0}^{p-1} a_k X_{n-k} + \sum_{l=0}^{q-1} b_l \xi_{n-l} \quad , \qquad \xi_i \text{ Gaussian iid .} \tag{2.83}$$

Again using (2.57) it is possible to write an ARMA process as

$$\phi(B)\,X_{n+1} = \theta(B)\,\xi_{n+1} \tag{2.84}$$

with suitable polynomials ϕ and θ of finite order p and q. In contrast to the classes of AR processes and MA processes, which contain a unique representation of every process in the class, the class of ARMA processes comprises arbitrarily many different representations of every process in the class, without extending the class of AR processes or MA processes. Under demand of minimizing the number of involved parameters for describing a given process, a representation can be reasonable, which does not belong to the class of AR or MA representations.

In this work, as examples of the time-discrete linear stochastic processes only models in AR representation are used, because only in this case the order of the Markovianity is immediately visible.

2.4 Nonlinear stochastic processes

In the class of stochastic processes (in general given by families of joint probability densities), which can be traced back to dynamical equations, nonlinear stochastic processes are understood as intrinsic nonlinear dynamics with imposed stochasticity from interaction of the considered subsystem with the environment, i.e., they are the unifying generalization of as well nonlinear deterministic dynamics as linear stochastic dynamics.

The space of time-discrete nonlinear stochastic processes contains all nonlinear deterministic maps perturbed by any kind of noise. The central problem of this class is that there are no general statements on the stability of the iteration procedure. From the instability (divergence of typical observables) the less but still disadvantageous property of nonstationarity and all its usual consequences (no invariant density, no autocorrelation) can be inferred.

An example in the time-continuous case are the Fokker-Planck equations obtained from a nonlinear Langevin equation [62]

$$\dot{X}_t = h(X_t, t) + g(X_t, t)\eta(t) \tag{2.85}$$

with either h or g nonlinear in X_t, not to confuse with nonlinear Fokker-Planck equations (FP-operator nonlinear in the density p). In the one-dimensional case, where a potential V with

$$-\partial_x V(x) = D^{(1)}(x) = h(x) \tag{2.86}$$

always exists, stability can be inferred if the potential walls are infinitely high. With $D^{(2)}(x) = g^2(x)$ the stationary solution reads [62]

$$p_{stat}(x) = \frac{N_0}{D^{(2)}(x)} \exp\Big(\int^x \frac{D^{(1)}(x')}{D^{(2)}(x')} dx'\Big) \ . \tag{2.87}$$

If for higher (two or three) -dimensional Fokker-Planck equations the driftfield $D^{(1)}(\mathbf{x})$ is not irrotational, then there does not exist a potential. Under such circumstances there again do not exist general statements on the stability of the dynamics.

The class of nonlinear stochastic processes is a very broad subject, which is very multifacetted and truely not overlooked by the author. A unique theory as for example in the case of the ARMA processes or for the Ornstein-Uhlenbeck processes is not available. Hence research is rather example-driven (e.g. neuronal spiking) and application-focussed (e.g. ratchets) and restricted to the framework of selected rather special models.

Yet as the central message of this section it is pointed out that the methods, which will be developed in this work, are usable also for that subclass of nonlinear stochastic processes, for which stability of the process can be assumed.

Chapter 3

Information theory and geometry

3.1 General information theory

3.1.1 Elementary definition

Entropies are functionals of probability distributions, but abbreviating the notation it should also be allowed to write entropies as functions of the corresponding random variables. The quantity

$$\boxed{H(X) = -\sum_{i=1}^{N} p(x_i)\log(p(x_i))} \qquad (3.1)$$

is defined to be the **Shannon entropy** of a discrete probability distribution. Here it is understood that

$$0\log 0 \equiv 0\ . \qquad (3.2)$$

The formula (3.1) follows uniquely from the following four axioms of Khinchin [48]:

1. Maximum for equidistribution: $H(p(x_1), ..., p(x_N)) \leq H(\frac{1}{N}, ..., \frac{1}{N}) \quad \forall\, p$
2. Continuity of $H(p(x_1), ..., p(x_N))$ in probabilities
3. Extendability: $H(p(x_1), ..., p(x_N)) = H(p(x_1), ..., p(x_N), p(x_{N+1}) = 0)$
4. Composition of random experiments[1]:
   $H(p(x_1, y_1), p(x_1, y_2), ..., p(x_N, y_M)) = H(p(x_1), ..., p(x_N))$
   $+ \sum_{x=x_1}^{x_N} p(x) H(p(y_1|x), ..., p(y_M|x))$

---

[1] The statement of the 4th Khinchin axiom is a demand on a special shape the conditional entropy with conditioning on a random variable defined in eq. (3.9) should have. The special form of the conditional entropy with conditioning on a realization $H(p(y_1|x), ..., p(y_M|x)) = -\sum_i p(y_i|x)\log p(y_i|x)$ is a consequence of the set of axioms and <u>not</u> part of the fourth axiom as claimed in ([6], p.48).

3.1.1.1 Interpretation of the entropy

Shannon entropies are interpreted as a measure for the mean uncertainty of the outcome of a random experiment. They characterize a stationary stochastic process by assessing the process-generated probability distributions. Entropies are also called measures of randomness [18]. Since the reduction of uncertainty always comes along with obtaining information, Shannon entropies are also interpreted as the mean information, which is necessary to know the outcome of a random experiment on the basis of the underlying probability distribution. Formula (3.1) can be considered as weighting the contributions to uncertainty '$-\log p(x_i)$' by the probability distribution itself.

3.1.1.2 Example: Discrete probability distributions with finite support

Given a support consisting of N (even number assumed) possible realizations x_i

- Equidistribution: $p(x_i) = \frac{1}{N} \quad \forall\, i \quad \Rightarrow \quad H = \log N$
- $p(x_i) = \frac{2}{N}$ for $i \in \{1, ..., \frac{N}{2}\}$ and $p(x_i) = 0$ for $i \in \{\frac{N}{2} + 1, ..., N\} \quad \Rightarrow \quad H = \log \frac{N}{2}$
- Certain outcome: $p(x_i) = 1$ for $i = 1$ and $p(x_i) = 0$ for $i \in \{2, ..., N\} \;\Rightarrow\; H = \log 1 = 0$

3.1.1.3 Units of entropies

The unit of entropies depends on the base of the logarithm. Given the base of the logarithm is two, information is measured via entropies in the unit of 'bits' and given the base of the logarithm is e, i.e., given the natural logarithm, one speaks of the unit 'nats'. For base of logarithm being ten, the unit is digits. With the formula for the change of the base of the logarithm

$$\log_a b = \log_c b \cdot \log_a c \;, \tag{3.3}$$

one can switch from one unit to another. For this work it is chosen to continue with the use of the natural logarithm according to ([19], p.638). Entropies do not have a unit of $\frac{1}{\text{time}}$.

3.1.2 Generalizations of elementary definition

Replacing the fourth Khinchin axiom for the Shannon entropy by additivity of the entropy for independent systems leads to the generalized definition of the Renyi entropies:

$$H^{(q)}(X) := \frac{1}{1-q} \ln \sum_{x \in \mathcal{X}} p(x)^q \;. \tag{3.4}$$

A freedom in one parameter q arises, which is called the Renyi order of the entropy. The Shannon entropies are contained as the special case for $q = 1$. Because inserting $q = 1$ leads to an undefined expression, this statement is obtained by applying the rule of l' Hospital:

$$\begin{aligned} H^{(q=1)}(X) &= \lim_{q\to 1} \frac{\frac{d}{dq}\ln(\sum_x p(x)^q)}{\frac{d}{dq}(1-q)} \\ &= \lim_{q\to 1} \frac{\frac{d}{dq}\ln(\sum_x \exp(q\,\ln p(x)))}{-1} \\ &= -\lim_{q\to 1} \frac{\sum_x \exp(q\,\ln p(x))\cdot \ln p(x)}{\sum_x \exp(q \ln p(x))} \\ &= -\sum_x p(x)\,\ln p(x)\ . \end{aligned} \tag{3.5}$$

For $q = 0$ one gets the so-called 'Hartley entropies'

$$H^{(q=0)}(X) = \ln(|\mathcal{X}|)\ . \tag{3.6}$$

Here $|\mathcal{X}|$ counts the number of elements with *nonvanishing* probability in the sense of 'box counting'. The generalized entropies are a monotonously decreasing function of the Renyi order q, i.e.,

$$H^{(q_1)}(X) \geq H^{(q_2)}(X) \qquad \text{for} \quad q_1 \leq q_2\ . \tag{3.7}$$

The interpretation of entropies given in sec. 3.1.1.1 still remains valid also for the Renyi entropies. A next generalizing step consists in defining a joint entropy for m random variables by

$$H^{(q)}(X_1,...,X_m) = \frac{1}{1-q}\ln \sum_{x_1\in\mathcal{X}_1,...,x_m\in\mathcal{X}_m} p(x_1,...,x_m)^q. \tag{3.8}$$

3.1.3 Conditional entropy

The conditional entropy of random variable Y conditioned on random variable X is defined as

$$\begin{aligned} H^{(q)}(Y|X) &:= H^{(q)}(X,Y) - H^{(q)}(X) \\ &= \frac{1}{1-q}(\ln \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)^q - \ln \sum_{x\in\mathcal{X}} p(x)^q)\ . \end{aligned} \tag{3.9}$$

$H^{(q)}(Y|X)$ is the part of the mean uncertainty in Y, which still remains after having used the information of the mean conditioning on X.

In the Shannon-case ($q = 1$) further calculations are possible by

$$\begin{aligned} H^{(q=1)}(Y|X) &= -\sum_{x,y} p(x,y)\ln p(x,y) + \sum_{x} p(x)\ln p(x) \\ &= -\sum_{x,y} p(x,y)\ln p(x,y) + \sum_{x,y} p(x,y)\ln p(x) \\ &= -\sum_{x,y} p(x,y)\ln \frac{p(x,y)}{p(x)} \\ &= -\sum_{x,y} p(x,y)\ln p(y|x) \\ &= -\sum_{x} p(x)\sum_{y} p(y|x)\ln p(y|x)\ . \end{aligned} \tag{3.10}$$

It has to be noticed that the difference of entropic terms can be merged by the usage of conditional probabilities for $q = 1$. Conditioning on a certain realization x of the random variable X allows for

$$\begin{aligned} H^{(q=1)}(Y|X=x) &= -\sum_{x'} \delta(x'-x)\sum_{y} p(y|x')\ln p(y|x') \\ &= -\sum_{y} p(y|x)\ \ln\ p(y|x)\ . \end{aligned} \tag{3.11}$$

With the given formulas a treatment of as well uncertainties in case of conditioning on random variables, what usually is done, as also uncertainies in case of conditioning on certain realizations of random variables (equivalent to conditioning on trivial random variables) or even a mixture of both cases is possible. It can be derived that

$$\begin{aligned} H^{(q=1)}(Y|X) &= \sum_{x\in\mathcal{X}} p(x) H^{(q=1)}(Y|X=x) \qquad (3.12) \\ [\quad &= \sum_{x\in\mathcal{X}} p(x)(H^{(q=1)}(X=x,Y) - \underbrace{H^{(q=1)}(X=x)}_{=0})\quad]\ . \end{aligned}$$

This allows for the interpretation that $H^{(q=1)}(Y|X)$ is the average over all available conditions according to their weight of the mean uncertainty of Y under conditioning on realizations of the random variable X. It is exactly the statement of the fourth axiom of Khinchin for Shannon entropies.

For arbitrary Renyi order $q \neq 1$ a calculation analogous to eq. (3.10) is impossible. The differences of entropies do not lead to simplified expressions with quotients of probabilities. That's the pity with generalized entropies. The conditional entropy for conditioning on a realization of the random variable X can be calculated straight from eq. (3.9)[2] as

$$H^{(q)}(Y|X=x) = \frac{1}{1-q}(\ln \sum_{x'\in\mathcal{X},y\in\mathcal{Y}} (p(y|x')\cdot\delta(x'-x))^q - \underbrace{\ln \sum_{x'\in\mathcal{X}} (\delta(x'-x))^q)}_{=0}$$

$$= \frac{1}{1-q} \ln \sum_{y\in\mathcal{Y}} (p(y|x))^q \ . \tag{3.13}$$

An equation analogous to eq. (3.12) is not valid, i.e., in general for $q \neq 1$ it holds

$$H^{(q)}(Y|X) \neq \sum_{x\in\mathcal{X}} p(x) H^{(q)}(Y|X=x) \ . \tag{3.14}$$

This is seen from inserting eq. (3.13) into (3.14) and the impossibility to retrieve eq. (3.9). Hence the interpretation that $H^{(q)}(Y|X)$ would be the remaining uncertainty averaged over all available conditions according to their weight fails for generalized entropies. The replacement of the fourth Khinchin axiom for Shannon entropies by the weaker demand of additivity of entropies for independent random variables for Renyi entropies is the immediately visible reason of this inequality for $q \neq 1$. Ineq. (3.14) is a hint for strange behaviour of conditional Renyi entropies with $q \neq 1$, because with it a very natural demand does not hold.

Though there is no simplified representation with conditional probabilities for the generalized conditional entropies as for Shannon entropies, a chain rule still holds (trivially) also for generalized entropies by

$$H^{(q)}(X,Y,Z) = H^{(q)}(X) + H^{(q)}(Y|X) + H^{(q)}(Z|X,Y) \tag{3.15}$$

with

$$H^{(q)}(Z|X,Y) = H^{(q)}(X,Y,Z) - H^{(q)}(X,Y) \ . \tag{3.16}$$

If

$$H^{(q)}(X,Y) \leq H^{(q)}(X) + H^{(q)}(Y) \tag{3.17}$$

could generally be proven, then

$$H^{(q)}(Y|X) \leq H^{(q)}(Y) \tag{3.18}$$

would follow, i.e., conditioning would in general reduce the generalized entropy interpreted as uncertainty (but compare also sec. 3.8.4.1).

3.1.4 Kullback-Leibler distance

If

$$p(x) \neq 0 \quad \Rightarrow \quad \rho(x) \neq 0 \qquad \forall x \ , \tag{3.19}$$

then the Kullback-Leibler distance for discrete distributions is defined as

$$D(p\|\rho) \equiv \sum_{x\in\mathcal{X}} p(x) \ln\left(\frac{p(x)}{\rho(x)}\right) \ . \tag{3.20}$$

---

[2] Because of the power in q, care concerning a generalization to the space-continuous case is necessary.

$D(p\|\rho)$ measures the distance between two probability distributions p and ρ ([67], p.23) in the sense of positive semidefinitness, i.e., $D \geq 0$ and $D = 0$ only for $p = \rho$. It is not a true distance between the distributions, because it is not symmetric in its arguments and also the axiom of a metric concerning the triangle inequality is in general not fulfilled. The Kullback-Leibler distance belongs to the framework of Shannon entropies. For continuous distributions a differential form holds:

$$D(f\|g) \equiv \int_{R^n} f(z) \ln\left(\frac{f(z)}{g(z)}\right) dz \, . \tag{3.21}$$

The Kullback-Leibler distance is also called Kullback-Leibler entropy, relative entropy or information gain between two distributions [18] and it appears in the literature often also with the symbols R and K.

In order to prepare the connection with mutual information of two random variables defined next, the Kullback-Leibler distance for joint distributions in two variables

$$D(p\|\rho) = \sum_{x,y} p(x,y) \ln\left(\frac{p(x,y)}{\rho(x,y)}\right) \tag{3.22}$$

is given.

3.1.5 Mutual information

Mutual information of two random variables is defined for arbitrary Renyi order q as

$$\begin{aligned} I^{(q)}(X;Y) :=& H^{(q)}(X) + H^{(q)}(Y) - H^{(q)}(X,Y) \\ =& H^{(q)}(X) - H^{(q)}(X|Y) \\ =& H^{(q)}(Y) - H^{(q)}(Y|X) \, . \end{aligned} \tag{3.23}$$

It measures the amount of information shared between two random variables X and Y. $I^{(q)}(X;Y)$ is the part of the uncertainty of one random variable, which is reduced in average, when knowing the realization of the other. Furthermore mutual information of two random variables has the following properties:

- Symmetry: $I^{(q)}(X;Y) = I^{(q)}(Y;X)$
- Self-information: $I^{(q)}(X;X) = H^{(q)}(X)$
- Positivity: $I^{(q=1)}(X;Y) \geq 0$
- $I^{(q=1)}(X;Y) = 0 \quad \Longleftrightarrow \quad X$ is independent of Y ; (for $q \neq 1$ '$\Leftarrow$' holds)
- Visualizability by Venn-diagrams (cmp. [17] and fig. 3.1 in this work)

Mutual information is also called redundancy ([34], p.6). This becomes clear as follows: If information for reducing the full uncertainty of $H^{(q)}(X,Y)$ is needed, but the information $H^{(q)}(X)$ and $H^{(q)}(Y)$ is gained, then exactly the amount of $I^{(q)}(X;Y)$ is superfluous, i.e., redundant. In later chapters, it is preferred to use the notion redundancy for a special mutual information. The notion 'transinformation' ([50], p.6) is also used for mutual information. Concerning prediction, mutual information between past and future is the information which reduces the uncertainty concerning the future. From the definition (3.23) for Renyi order $q = 1$ it is derivable that

$$I^{(q=1)}(X;Y) \equiv I(X;Y) = \sum_{x,y} p(x,y) \ln \left(\frac{p(x,y)}{p(x)p(y)} \right) . \tag{3.24}$$

Hence in the case of *two* random variables mutual information is the Kullback-Leibler distance between the joint distribution $p(x, y)$ and the product of distributions $p(x)p(y)$. The larger the Kullback-Leibler distance of the distributions $p(x, y)$ and $p(x)p(y)$ the more information is shared between $p(x)$ and $p(y)$. In other words: The larger the deviation from stochastic independence, the larger the mutual information [66].

3.1.6 Multiple mutual information

Refering to its name, a generalized *mutual information* should be the *shared* informational contribution among *all* involved random variables. In accordance with fig. 3.1 as a generalization of definition (3.23) mutual information of three random variables can be defined as [79]

$$I(X;Y;Z) := H(X)+H(Y)+H(Z)-H(X,Y)-H(X,Z)-H(Y,Z)+H(X,Y,Z) . \tag{3.25}$$

This was also already published in [8]. With this definition, $I(X;Y;Z)$ is symmetric in its arguments. Equivalent with the property for two random variables, in case of $X = Y = Z$ it is obtained that

$$I(X;X;X) = H(X) . \tag{3.26}$$

For X, Y, Z all pairwise independent, i.e., $H(X,Y) = H(X) + H(Y)$, $H(X,Z) = H(X) + H(Z)$, etc., it holds that

$$I(X;Y;Z) = 0 , \tag{3.27}$$

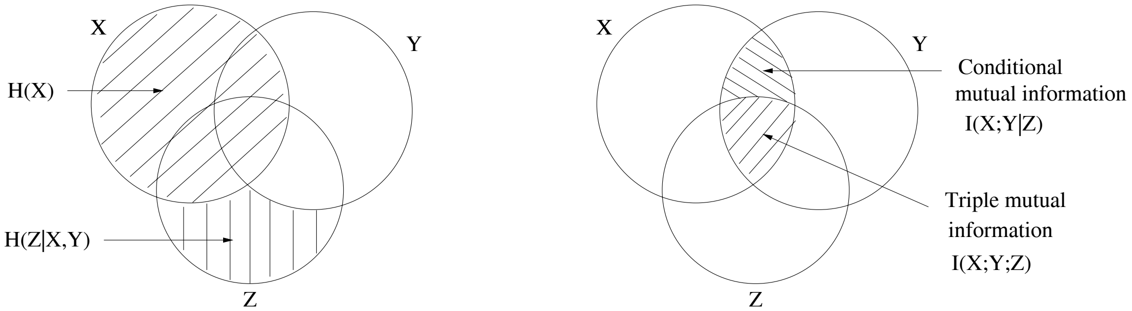


Figure 3.1: Visualization of mutual information $I(X;Y;Z)$ of three random variables by Venn diagrams and the connection to simple set theory by observing that it is simply the intersection of sets corresponding to single uncertainties. The well known conditional mutual information $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$ is also indicated.

but for *only* X and Y independent the result is

$$\begin{aligned} I(X;Y;Z) &= H(Z) - H(X,Z) - H(Y,Z) + H(X,Y,Z) \\ &= H(Z) - H(X,Z) - H(Y,Z) + H(X|Y,Z) + H(Y,Z) \\ &= H(Z) - H(X,Z) + H(X|Y,Z) \\ &= -H(X|Z) + H(X|Y,Z) \\ &\leq 0 \, , \end{aligned} \tag{3.28}$$

at least for Renyi order $q = 1$, because of monotony properties for conditional Shannon entropies. The example of $Z = X + Y$ for independently distributed random variables X and Y shows that an inequality in the last line of eq. (3.28) is the usual case, i.e., in contrast to the case of two random variables, triple mutual information can become negative. Despite this property, the term 'mutual information' is kept as already done in ([79], p.469). According to its values in the given examples, $I(X;Y;Z)$ reflects the inhomogeneities in the dependence structure of all random variables. Because of the possibility of negativity of $I(X;Y;Z)$, non-overlapping circles for X and Y in fig. 3.1 *cannot* automatically be inferred from $I(X;Y) = 0$. Furthermore the value corresponding to a set in fig. 3.1 cannot even qualitatively be associated with the size of the set by counting the basic constituents. Since the sets, which potentially get assigned negative numbers, are not isolatedly accessible by a single random experiment, the given construction does not enforce the necessity of an interpretation of random variables with negative uncertainty.

A short calculation from eq. (3.25) with the aim of reaching a form as close as possible to (3.24) yields

$$I(X;Y;Z) = \sum_{x,y,z} p(x,y,z) \ln \frac{\left[\frac{p(x,y)p(x,z)p(y,z)}{p(x)p(y)p(z)}\right]}{p(x,y,z)} \, . \tag{3.29}$$

Nevertheless, from the possibility of non-positive $I(X;Y;Z)$, an interpretation in the sense of a Kullback-Leibler distance is impossible. For the generalization of mutual information by eq. (3.25), mutual information and Kullback-Leibler distances hit each other only accidentally in the case of two random variables.

For n random variables a straightforward generalization of eq. (3.25) yields

$$\boxed{\begin{aligned} I(X_1;...;X_n) = \; & \sum_{i=1}^{n} H(X_i) \\ & - \sum_{i_1,i_2=1\,:\,i_1<i_2}^{n} H(X_{i_1}, X_{i_2}) \\ & + \sum_{i_1,i_2,i_3=1\,:\,i_1<i_2<i_3}^{n} H(X_{i_1}, X_{i_2}, X_{i_3}) \\ & \pm ... \\ & + (-1)^{n+1} H(X_1, ..., X_n) \, . \end{aligned}} \tag{3.30}$$

Formulas of this type will be needed in sec. 4.4.4, where the starting points of formulas for multiple mutual information are always pictures like in fig. 3.1, possibly higherdimensional.

Based on the Kullback-Leibler distance, in [69] another generalization of the usual mutual information of two random variables was suggested by[3]

$$I(X_1, ..., X_n) \equiv D(\mathcal{L}(X_1, ..., X_n) \| \prod_{i=1}^{n} \mathcal{L}(X_i)) \quad [\, = \sum_{i=1}^{n} H(X_i) - H(X_1, ..., X_n)\,]. \tag{3.31}$$

As an advantage, this quantity is strictly positive. However, it does not describe the shared informational contributions among all involved random variables in the sense of the intersections in fig. 3.1. The case $X_1 = ... = X_n$ yields $I(X_1; ...; X_n) = (n-1) \cdot H(X_1)$. Also for arbitrary random variables parts of the informational contributions are counted multiply. In [27]

$$I(X_0, X_1, ..., X_n) = \sum_j [H(X_j) - H(X_0, X_1, ..., X_n)] \tag{3.32}$$

is suggested. It is interpreted here as identical to eq. (3.31) with erroneously placed square brackets, because usual mutual information of two random variables is not reproduced by this formula.

The triple mutual information given by eq. (3.25) is not the same as the quantity called three-point mutual information in ([46], 2nd ed., p.298) defined in ([46], 1st ed., p.143, (9.5)), which is rather in the sense of usual (double) mutual information of a single random variable and a joint random variable composed of two single random variables.

The family of quantities obtained as the 'union of all mutual informations of certain order' according to eq. (3.30) yields a family of other candidates for a generalization of $I(X;Y)$, which fulfil the properties of symmetry in all random variables and of singly counting of informational parts. For three random variables and mutual informaton of order two this quantity would be (cmp. fig. 3.1) $I(X;Y;Z) + I(X;Y|Z) + I(Y;Z|X) + I(Z;X|Y)$. However, the idea of informational contributions shared by all involved random variables is not consequently captured by those quantities. Nevertheless, those quantities would be preferable in comparison with the definition (3.31), because no informational contributions are counted multiply.

3.2 Connection with the entropies of statistical physics

Usually the entropy in statistical physics is defined as

$$S(E) = k_B \ln N(E). \tag{3.33}$$

Here k_B is Boltzmann's constant and $N(E)$ is the number of microstates depending on the energy. $S(E)$ measures the disorder of a given physical system.

The thermodynamic entropy is just the Shannon uncertainty (up to the constant factor k_B) of the (micro-)state equidistribution, provided the macrostate is given [34]. It is reminded that the equidistribution causes the cancellation of the sum and the first probability in eq. (3.1).

---

[3] $\mathcal{L}(..)$ means 'distibution of ..'.

From the chain rule and the Kullback-Leibler distance being nonnegative one can show that the Kullback-Leibler distance of a stationary distribution and an arbitrary state distribution developing in time must decrease. Given that the stationary distribution is uniform one can show from this that the entropy of a closed system must increase [17], i.e.,

$$dS > 0 \; , \tag{3.34}$$

and verify the second law of thermodynamics in terms of information theory.

3.3 Information theory in time series analysis for the block-case

For the following stationarity is assumed. Furthermore in this section only quantities of block-type are treated, i.e., equidistancy in time is assumed. A possible dependence of the entropic quantities on a time-step τ of the time series is notationally suppressed.

3.3.1 Joint entropies

3.3.1.1 Discrete entropies (Discrete(-ized) phase space; discrete time)

The Renyi entropy of order q for block length m is given by

$$H_m^{(q)}(\epsilon) = \frac{1}{1-q} \ln \sum_{j_1,\dots,j_m} {p_{j_1,\dots,j_m}}^q \; . \tag{3.35}$$

The formula is based on m in general coupled random variables (successive in time). The probabilities $p_{j_1,\dots,j_m}$ are obtained from an integration of the underlying probability density over a certain volume, presumed the density exists. The size of this volume is related with the given resolution ϵ, which on the right side of the formula appears implicitly in the range of the j_i. Details of the integration depend on the decision if a partition of the state space or more general coverings are chosen. Of course for discrete probability distributions this formula is also valid. $H_m^{(q)}(\epsilon)$ is interpreted as a measure for the uncertainty of a realisation of the underlying joint probability distribution of m coupled random variables (at successive times) with resolution ϵ (in the state space). As a function of q, $H_m^{(q)}$ is called the spectrum of the Renyi information ([50], p.133).

$$H_{m=0}^{(q)}(\epsilon) \equiv 0 \tag{3.36}$$

reflects the fact that, where no experiment is made, there is also no uncertainty. As before the case $q = 1$ leads to the Shannon entropy given for $m = 1$ by

$$H_{m=1}^{(q=1)}(\epsilon) = -\sum_{i=1}^{N(\epsilon)} p_i \ln p_i \; . \tag{3.37}$$

For case $q = 2$ it holds that

$$H_{m=1}^{(2)}(\epsilon) = -\ln \sum_j p_j^2 \; . \tag{3.38}$$

The case with $q = 2$ is important, because the entropy estimation via the usual correlation sum (see sec. 3.6) does not contain a systematic error (no bias) for small ϵ. So the decision for use of an order q is rather estimation-guided than entropy-property guided! After decision of the order of the entropy the corresponding index q is often omitted.

Before reaching its asymptotic behaviour in m, $H_m(\epsilon)$ contains most relevant aspects of the following discussion and is analyzed by successive discrete derivatives of H_m.

3.3.1.2 Continuous entropies (Continuous phase space; discrete time)

In case of existing density the corresponding continuous entropy is defined as

$$H_m^{c,(q)}(\rho_m) := \frac{1}{1-q} \ln \int {\rho_m}^q(\mathbf{x}) d^m x \; . \tag{3.39}$$

The continuous entropies are functionals on the set of probability densities. For $q = 1$ the continuous Shannon entropy is derived (being able to commute limit procedures) again via l' Hospital as

$$H_m^c(\rho_m) \equiv H_m^{c,(q=1)}(\rho_m) = -\int \rho_m(\mathbf{x}) \ln \rho_m(\mathbf{x}) \, d^m x \; . \tag{3.40}$$

This formula is applied to the class of characteristic functions in order to discuss an elementary property of continuous entropies:

$$H_1^c(a \cdot 1_{[0,\frac{1}{a}]}(x)) = -\ln a \; . \tag{3.41}$$

From the usual behaviour of the logarithm it is known that for $a = 1$ the continuous entropy is zero, for $0 < a < 1$ it is larger than zero and for $a > 1$ it is smaller than zero. The results for different parameter value a are connected by the fact that under the coordinate transformation $y = f(x)$ densities transform as $\tilde{\rho}(y) = \frac{1}{f'(x)}\rho(x)$ and in the special case of $y = \alpha x$ the continuous entropy transforms as

$$H_1^{c,(q)}(\tilde{\rho}) = H_1^{c,(q)}(\rho) + \ln \alpha \; . \tag{3.42}$$

The main point here is that continuous entropies can become negative. Thus in contrast to the discrete entropies $H_m(\epsilon)$, the continuous entropies H_m^c are not interpretable as a measure of uncertainty.

Especially with regard to the following derivation it is useful to point out that a continuous entropy does not depend on a resolution ϵ. The continuous entropies get their use from their contribution to the calculation of the discrete entropies by

$$H_m(\epsilon) = H_m^c - m \ln \epsilon \; . \tag{3.43}$$

The argument is essentially a Taylor approximation of order zero [45]:

$$\begin{aligned}
H_m^c &= -\int_V \rho_m(\mathbf{x}) \ln \rho_m(\mathbf{x})\, d^m x \\
&= -\sum_{j_1,\dots,j_m} \int_{V_{j_1,\dots,j_m}} \rho_m(\mathbf{x}) \ln \rho_m(\mathbf{x})\, d^m x \\
&\approx -\sum_{j_1,\dots,j_m} \int_{V_{j_1,\dots,j_m}} \frac{p_{j_1,\dots,j_m}}{\epsilon^m} \ln \frac{p_{j_1,\dots,j_m}}{\epsilon^m}\, d^m x \\
&= -\sum_{j_1,\dots,j_m} \epsilon^m \frac{p_{j_1,\dots,j_m}}{\epsilon^m} \ln \frac{p_{j_1,\dots,j_m}}{\epsilon^m} \\
&= -\sum_{j_1,\dots,j_m} (p_{j_1,\dots,j_m} \ln p_{j_1,\dots,j_m} - p_{j_1,\dots,j_m} \ln \epsilon^m) \\
&= H_m(\epsilon) + m \ln \epsilon\ . \qquad (3.44)
\end{aligned}$$

For very small ϵ the '$-m\ln\epsilon$'-term in eq. (3.43) dominates the continuous entropy H_m^c, a constant with respect to ϵ. This dominance diminishes for increasing ϵ and vanishes for $\epsilon = \exp(-\frac{H_m^c}{m})$. From the reason that the uncertainty $H_m(\epsilon)$ is not allowed to become negative, a necessary condition for the validity of the Taylor approximation is that $\epsilon < \exp\frac{H_m^c}{m}$.

For order $q = 2$ also the connection between continuous entropy and the discrete entropy is deducable with similar arguments:

$$H_m^{c,(q=2)} = -\ln[\int_V {\rho_m}^2(\mathbf{x})\, d^m x] \approx -\ln[\sum_{j_1,\dots,j_m} p^2_{j_1,\dots,j_m}] - \ln\frac{1}{\epsilon^m} = H_m^{(q=2)}(\epsilon) + m\ln\epsilon\ . \quad (3.45)$$

The cases $q = 1$ and $q = 2$ allow for the empirical rule, that it is possible to get from the discrete to the continuous entropy by replacing sum by integral and probability by density. As a final remark continuous entropies are also called differential entropies [17].

3.3.2 Conditional entropies

The conditional entropy with block-type conditioning is defined as[4]

$$H_{1|m}^{(q)}(\epsilon) := H_{m+1}^{(q)}(\epsilon) - H_m^{(q)}(\epsilon)\ . \qquad (3.46)$$

$H_{1|m}^{(q)}(\epsilon)$ is interpreted as the amount of information, which (averaged over all time arguments) is needed for the prediction of the immediate future observation with an accuracy ϵ, after having used all information from the dependences of the future on the last m observations.

---

[4] With respect to the indices the notation for conditional entropies in this work is chosen to be similar to the notation for conditional probabilities in [75]. Conditional entropies show their conditioning explicitly in their dependences and not in the distinction of lower or upper case letters for the main symbol. This is not usual. The advantage of indicating the distinction of conditioned and unconditioned entropies in one symbol disappears in the transition from the nonperforated (chap. 3) to the perforated (chap. 4) case.

In the following Renyi order $q = 1$ is assumed. Quite similar to eq. (3.10), conditional Shannon entropies are expressible by conditional probabilities via

$$\begin{aligned} H_{1|m}(\epsilon) &= -\sum_{i_1,\ldots,i_{m+1}} p_{i_1,\ldots,i_{m+1}} \ln p_{i_1,\ldots,i_{m+1}} + \sum_{i_1,\ldots,i_m} p_{i_1,\ldots,i_m} \ln p_{i_1,\ldots,i_m} \\ &= -\sum_{i_1,\ldots,i_{m+1}} p_{i_1,\ldots,i_{m+1}} \ln p_{i_1,\ldots,i_{m+1}} + \sum_{i_1,\ldots,i_{m+1}} p_{i_1,\ldots,i_{m+1}} \ln p_{i_1,\ldots,i_m} \\ &= -\sum_{i_1,\ldots,i_{m+1}} p_{i_1,\ldots,i_{m+1}} \ln\, p_{i_{m+1}|i_1,\ldots,i_m} \;. \end{aligned} \tag{3.47}$$

The conditional entropies of block type can be written as

$$H_{1|m}(\epsilon) = -\sum_{i_1,..,i_{m+1}} p_{i_1,\ldots,i_{m+1}} \ln \frac{p_{i_1,\ldots,i_{m+1}}}{p_{i_1,\ldots,i_m}} \tag{3.48}$$

$$= -D(P_{m+1}\|P_m) \geq 0 \;. \tag{3.49}$$

The negativity of the function D is not in contradiction with statements in 3.1.4 because there a prerequisite for positivity is that both distributions are normalized to one ([6], p.52). However, here the normalization of the distribution P_m is violated. Hence the interpretation of the function D as a distance is not allowed, if the distributions to compare have their support in different dimensions.

It holds that

$$H_{1|0}(\epsilon) \geq H_{1|1}(\epsilon) \geq ... \geq H_{1|\infty}(\epsilon) \;. \tag{3.50}$$

This is intuitively clear, because the more conditioning is imposed, the smaller the uncertainty should be. A schematic plot can be found in ([22], p.18). From induction of the definition of the conditional entropy one gets the chain rule for entropies:

$$H_m(\epsilon) = H_{1|0}(\epsilon) + \ldots + H_{1|m-1}(\epsilon) \;. \tag{3.51}$$

This means that the total uncertainty of a joint probability distribution is representable as a sum of conditional uncertainties. From this formula and the above chain of inequalities one can see that $H_m(\epsilon)$ is a concave function of $m > 0$ for small m with finally a linear asymptote as shown in ([22], p.21).

For arranging those conditional entropies in the framework of general conditional entropies of sec. 3.1.3 one has to identify X as a random variable for the joint distribution on m successive time steps and Y as a random variable for the distribution at the immediate time step ahead. Under such conditions one gets

$$H(Y|X) = H(X,Y) - H(X) = H_{m+1} - H_m = H_{1|m} \;. \tag{3.52}$$

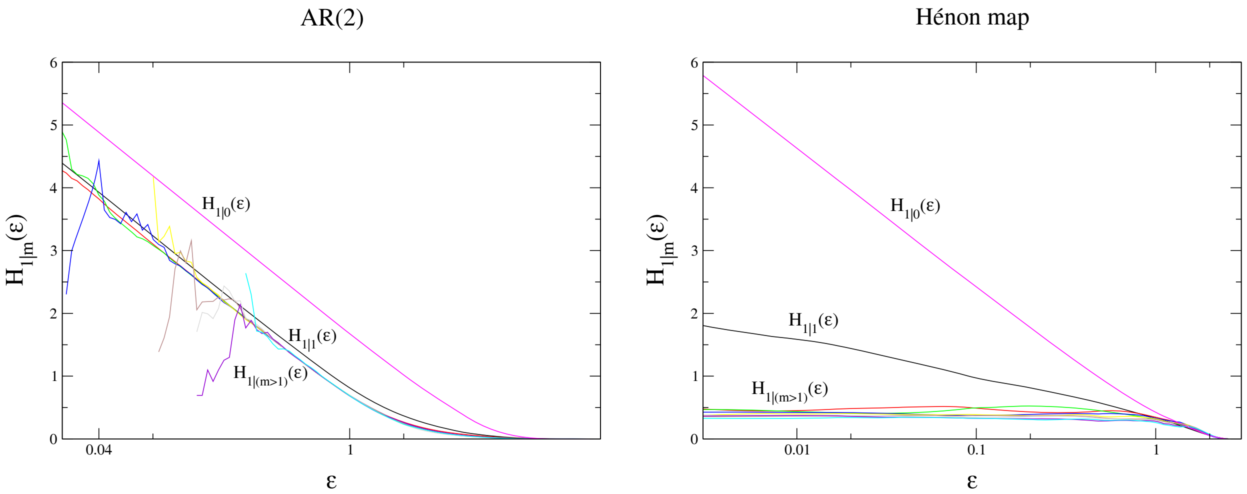


Figure 3.2: Conditional entropies. Left panel: stochastic dynamics. Right panel: chaotic dynamics.

The graphs of fig. 3.2 are obtained from the TISEAN-program d2 [40]. Numerical procedures for estimating entropies are described in sec. 3.6. On the left panel the conditional entropies for an AR(2) process ($a_0 = 0.48$; $a_1 = 0.48$), i.e. stochastic dynamics, are shown. One can see that the conditional entropies increase logarithmically with increasing resolution, i.e. decreasing ϵ. This does not change even for arbitrary high conditioning, which is seen from eq. (3.43). The difference in the curves arises from the different continuous conditional entropies

$$H^c_{1|m} = H^c_{m+1} - H^c_m = H_{1|m}(\epsilon) + \ln \epsilon \tag{3.53}$$

for variable m. Graphically they can be found as the intersection points of the virtual extension of the straight lines with the x-axis, because for $0 = H_{1|m}(\epsilon_m) = H^c_{1|m} - \ln \epsilon_m$ follows trivially $\ln \epsilon_m = H^c_{1|m}$. The ϵ_m can be interpreted as a measure of the width of the underlying distribution, which comes out exactly for the rectangle distribution and holds approximately for every other distribution. From this reasoning one can argue that increasing the width of underlying distributions leads to right-shifts of the conditional entropy graphs and higher entropy-values for the same resolution ϵ. This is shown in fig. 3.3. From

$$H_{1|m}(\epsilon) = -\ln(\epsilon) + \ln(\epsilon_m) \quad \text{and} \quad H_{1|0}(\epsilon) = -\ln(\epsilon) + \ln(\epsilon_0) \tag{3.54}$$

one gets

$$\exp(H_{1|0}(\epsilon) - H_{1|m}(\epsilon)) = \exp(H^c_{1|0} - H^c_{1|m}) = \frac{\epsilon_0}{\epsilon_m}\ . \tag{3.55}$$

A difference of conditional entropies is thus translatable into a fraction of widthes of the corresponding distributions. Especially one can see how much conditioning decreases the width of distributions.

The conditional entropies approaching zero for coarser resolution instead of reaching negative values is clear from their interpretation as uncertainty. The smoothness of this behaviour instead of a kink is a finite size effect. Eq. (3.43) loses its validity for sufficiently large ϵ, because higher orders in ϵ have to be included in the derivation in eq. (3.44). The deviation from straight lines to fluctuating behaviour for finer resolution on the other hand is known as finite sample effect.

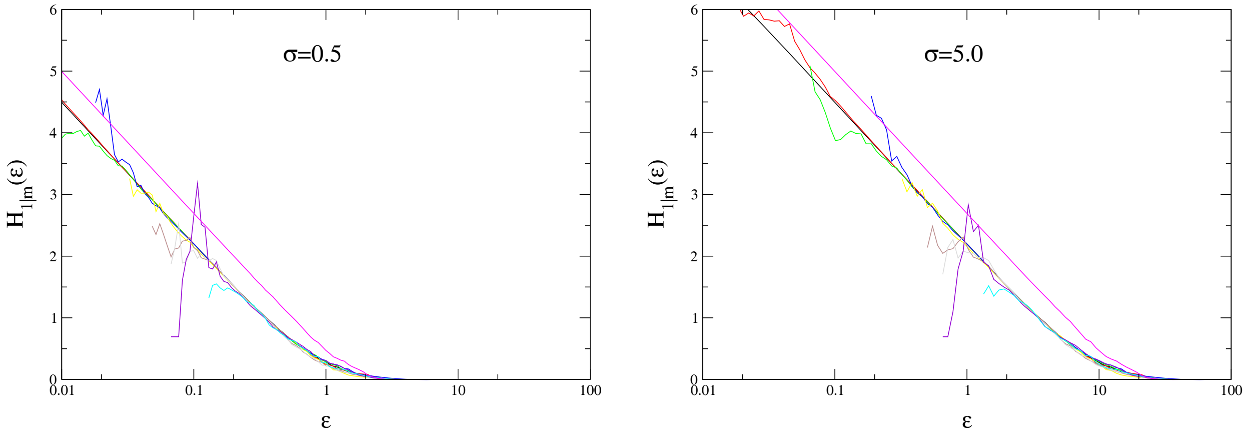


Figure 3.3: AR(1): $x_{n+1} = ax_n + \xi_{n+1}$ with $a = 0.8$; ξ_n is $\mathcal{N}(0, \sigma)$-distributed.

On the right panel of fig. 3.2 one finds the conditional entropies for the usual Hénon map

$$x_{n+1} = 1 - \alpha x_n^2 + \beta x_{n-1} \tag{3.56}$$

with standard parameters $\alpha = 1.4$ and $\beta = 0.3$, i.e., dissipative deterministic chaotic dynamics. Again the unconditioned entropy $H_1(\epsilon)$ increases logarithmically with increasing resolution. A density does not exist for the attractor of the Hénon map in dimension $m \geq 2$. Henceforth eq. (3.43) is not valid, because for all resolutions ϵ the Taylor approximation in (3.44) does not hold. Instead of $H_2(\epsilon) \sim -2 \ln \epsilon$ it is thus observed $H_2(\epsilon) \sim -D \ln \epsilon$ with $D \approx 1.22$. For sufficiently high resolution a weaker logarithmic increase in $H_{1|1}(\epsilon)$ results. The slope of approximately -0.22 instead of zero corresponds to not fully reconstructed dynamics by taking a one-dimensional embedding for the Hénon map. H_3 and H_2 only differ by a constant, what explains the behaviour of constant $H_{1|2}$ and the same is valid for all $H_{1|m}$ with $m > 2$. For coarser resolution than a threshold resolution also the conditional entropy of deterministic dynamics falls to zero.

In later chapters it will turn out that the conditional entropies described in this chapter are very special cases of more general conditioning possibilities. A further discussion about stochasticity and determinism and its consequences on entropies can be found in appendix A.

3.3.3 Entropy rates

3.3.3.1 Conditional entropy rates

In the 'flow'-picture (explicit inclusion of the time step length) the conditional entropy rate is defined as the discrete derivative of the entropy H by the time[5]

$$h_m(\epsilon, \tau) := \frac{H_{1|m}(\epsilon, \tau)}{\tau} \, . \tag{3.57}$$

---

[5] In this work the capital letter H is used for entropies and hence for information (or uncertainty), no matter if conditioned (difference of entropies) or not. The small letter h indicates the first time derivative of entropies, i.e., entropy rates. This convention is notationally close to [31] and [15].

Here τ is the elementary time interval of the underlying time series. It is the mean uncertainty per time interval of one time step ahead knowing m time steps in the past $(-(m-1), ..., 0)$. Instead of refering the entropy on a time interval, in a 'map'-picture it is usual to refer the entropy on an iteration- or time step, which then results in replacing τ by 1 and omitting it as a dependence leading to

$$h_m(\epsilon) = H_{1|m}(\epsilon) \ . \tag{3.58}$$

Thus in the 'map'-picture the conditional entropy rate can be also interpreted as conditional entropy, with suppression of the notion 'iteration step', as needed. It is of course trivial to see, but absolutely necessary to be aware of the fact that $h_m(\epsilon, \tau)$ in the 'flow'-picture and $h_m(\epsilon)$ in the 'map'-picture have different properties especially for small τ. For this reason it is not allowed to mix up both pictures. It is possible to see that in the 'map'-picture in written form entropies and entropy rates usually are exchanged vice versa. At this point the author would like to apply some pedantism in demanding that in the 'flow'-picture such an exchange should not be supported and furthermore pointing out again that entropies do not carry the unit $\frac{1}{\text{time}}$. This unusual pedantism in this area will be the key for the formulations of sec. 3.8.

3.3.3.2 Limit cases

The resolution-dependent mean uncertainty rate for the immediate future random variable, if the whole past is known, is defined as

$$h_\infty(\epsilon, \tau) := \lim_{m \to \infty} h_m(\epsilon, \tau) \ . \tag{3.59}$$

It holds that

$$h_\infty(\epsilon, \tau) = \lim_{m \to \infty} \frac{H_m(\epsilon, \tau)}{m\tau} \ . \tag{3.60}$$

Even though the limit is the same in both previous cases, the speed of convergence of the second case is in general much slower. It is defined

$$h_\infty := \lim_{\epsilon \to 0} \lim_{\tau \to 0} h_\infty(\epsilon, \tau) \ . \tag{3.61}$$

Depending on the Renyi order q, in sec. 3.5 h_∞ will be identified with the topological or metric, i.e., Kolmogorov-Sinai entropy rates.

3.3.4 Jointly conditioned joint entropies

In the following the resolution dependence is neglected.

3.3.4.1 Finite number of time steps

Supposed, l successive time steps in the immediate past are known and the outcome of k successive time steps in the immediate future is collectively requested, then the remaining uncertainty after conditioning in all k time steps is defined by

$$H_{k|l} := H_{l+k} - H_l \ . \tag{3.62}$$

If as an example there are two future time steps and l time steps in past are known, then the remaining uncertainty results as

$$H_{2|l} = H_{1|l+1} + H_{1|l} \ . \tag{3.63}$$

The available uncertainty reduction concerning k time steps in the immediate future, if realizations for l time steps in the past are known, is calculated from the mutual information as

$$\begin{aligned} I(l\,\text{past steps}; k\,\text{future steps}) &= H_l + H_k - H_{k+l} \\ &= H_1 + H_{1|1} + \cdots + H_{1|l-1} \\ &\quad + H_1 + H_{1|1} + \cdots + H_{1|k-1} \\ &\quad - H_1 - H_{1|1} - \cdots - H_{1|k+l-1} \\ &= \sum_{j=0}^{k-1} (H_{1|j} - H_{1|l+j}) \\ &= \sum_{j=0}^{l-1} (H_{1|j} - H_{1|k+j}) \ . \end{aligned} \tag{3.64}$$

In case $l = k$ it holds that

$$\begin{aligned} I(k\,\text{past steps}; k\,\text{future steps}) &= \sum_{j=0}^{k-1} (H_{1|j} - H_{1|k+j}) \\ &= (H_1 - H_{1|k}) + (H_{1|1} - H_{1|k+1}) + \cdots + (H_{1|k-1} - H_{1|2k-1}) \ . \end{aligned} \tag{3.65}$$

For Markov processes of order k this reduces at least for entropic order $q = 1$ (cmp. eq. (3.78)) to the so-called 'finite-k-excess entropy'

$$I(k\ \text{past steps}; k\ \text{future steps}) = (H_1 - H_{1|k}) + (H_{1|1} - H_{1|k}) + \cdots + (H_{1|k-1} - H_{1|k}) \ . \tag{3.66}$$

3.3.4.2 Semi-infinite number of time steps

In case the whole past is known, the remaining uncertainty of k successive future time steps is $H_{k|\infty}$. The so-called 'k-redundancy' ([18], p.31)

$$\begin{aligned} I(\infty\,\text{past steps}; k\,\text{future steps}) &= \sum_{j=0}^{k-1} (H_{1|j} - H_{1|\infty}) \\ &= (H_{1|0} - H_{1|\infty}) + (H_{1|1} - H_{1|\infty}) + \ldots + (H_{1|k-1} - H_{1|\infty}) \\ &= H_k - kH_{1|\infty} \end{aligned} \tag{3.67}$$

is the available uncertainty reduction of k future time steps given the whole past.

3.3.4.3 Infinite number of time steps

In case both l and k tend to infinity, the remaining uncertainty of the future is $H_{\infty|\infty}$. The information contained in the whole past about the whole future is the so-called 'excess entropy'

$$\mathbf{E} = \sum_{j=0}^{\infty}(H_{1|j} - H_{1|\infty}) = \lim_{j\to\infty}(H_j - j \cdot H_{1|\infty}) \ , \tag{3.68}$$

which is the total uncertainty of the future minus the unavoidable uncertainty of the future, and because of this it is the total amount of possible reduction of uncertainty concerning joint prediction of the whole future. It holds that

$$\mathbf{E} = I(\text{full past; full future}) \ . \tag{3.69}$$

From this representation of the excess entropy as a mutual information the chosen name of 'excess' being in the same spirit as 'redundant', here infinitely many time steps in the future treated, can be understood.
One sees immediately that the excess entropy is the limit $k \to \infty$ of the k-redundancy in eq. (3.67). The excess entropy is also known under the name effective measure complexity (EMC) [32] or predictive information. It does not play a role concerning uncertainty reduction of the outcome of a random variable at one point of time in the future. Finiteness of the excess entropy is used to define the underlying process to be finitary.

3.3.5 Redundancy

We fix the notion 'redundancy' in the time series context with the definition

$$R_m(\epsilon) := H_1(\epsilon) - H_{1|m}(\epsilon) \ . \tag{3.70}$$

$R_m(\epsilon)$ is interpreted as the mean amount of information, by which the mean uncertainty of the immediate future event is reducable, when taking into account all information concerning the future from the last m observations. From the elementary replacement

$$R_m(\epsilon) = H_1(\epsilon) + H_m(\epsilon) - H_{m+1}(\epsilon) \tag{3.71}$$

it is seen that the redundancy is a mutual information of the immediate past joint-m-distribution and the immediate single future distribution. The mutual information in eq. (4.10) will include redundancy as a special case.

Some care concerning the notion 'redundancy' in the literature is in order: In ([18], eq.15), where because of underlying discrete distributions no resolution dependence is discussed, the redundancy is defined as our R_∞ and the so-called 'per-symbol m redundancy' is (translated to the above used notation) defined as $H_{1|m} - H_{1|\infty}$ ([18], eq.36), what will later in 5.3.1 be called the ignored memory.

3.4 Entropies of Markov processes

Discrete state space, discrete time and, as always, stationarity is assumed in this section. A Markov process of order $m = 1$ is defined by the property of eq. (2.12). An immediate consequence is that for Renyi order $q = 1$ it holds

$$H^{(q=1)}_{1|k}(\epsilon) = H^{(q=1)}_{1|1}(\epsilon) \quad \forall\, k \geq 1 \; . \tag{3.72}$$

The chain rule of entropies (cmp. (3.15))

$$H^{(q)}_{n}(\epsilon) = H^{(q)}_{1}(\epsilon) + H^{(q)}_{1|1}(\epsilon) + \ldots + H^{(q)}_{1|n-1}(\epsilon) \tag{3.73}$$

is then simplified to

$$H^{(q=1)}_{n}(\epsilon) = H^{(q=1)}_{1}(\epsilon) + (n-1) \cdot H^{(q=1)}_{1|1}(\epsilon) \; . \tag{3.74}$$

Eq. (3.74) can also be proven by the simplified factorization of joint probabilities from eq. (2.13) for discrete state space and time:

$$\begin{aligned} H^{(q=1)}_{n}(\epsilon) &= -\sum_{i_{k-n+1},\ldots,i_k} p_{i_{k-n+1},\ldots,i_k} \ln p_{i_{k-n+1},\ldots,i_k} \\ &= -\sum_{i_{k-n+1},\ldots,i_k} p_{i_{k-n+1},\ldots,i_k} \ln(p_{i_{k-n+1}} \cdot p_{i_{k-n+2}|i_{k-n+1}} \cdot \ldots \cdot p_{i_k|i_{k-1}}) \\ &= -\sum_{i_{k-n+1},\ldots,i_k} p_{i_{k-n+1},\ldots,i_k} [\ln p_{i_k} + (n-1) \cdot \ln p_{i_k|i_{k-1}}] \\ &= -\sum_{i_k} p_{i_k} \ln p_{i_k} - (n-1) \cdot \sum_{i_{k-1},i_k} p_{i_{k-1},i_k} \ln p_{i_k|i_{k-1}} \; . \end{aligned} \tag{3.75}$$

For Renyi order $q \neq 1$ eq. (3.72) doesn't hold for a Markov process of order $m = 1$ because the factorisation of the joint probabilities in conditional probabilities is not translatable into a representation of conditional entropies as functions of conditional probabilities as in the Shannon-case. Hence also a simplification of the factorisation of joint probabilities by the Markov property doesn't lead to a simplification in the conditioning of generalized entropies and the chain rule remains unsimplified for $q \neq 1$.

A Markov process of arbitrary order $m > 1$ is defined by eq. (2.54). An immediate consequence is that for Renyi order $q = 1$ it holds

$$H^{(q=1)}_{1|k}(\epsilon) = H^{(q=1)}_{1|m}(\epsilon) \qquad \forall\, k \geq m \; . \tag{3.76}$$

The chain rule of entropies then simplifies for $n \geq m$ to

$$\begin{aligned} H^{(1)}_{n}(\epsilon) &= H^{(1)}_{1}(\epsilon) + H^{(1)}_{1|1}(\epsilon) + \ldots + H^{(1)}_{1|m-1}(\epsilon) + \; (n-m) \cdot H^{(1)}_{1|m}(\epsilon) \\ &= H^{(1)}_{m}(\epsilon) + \; (n-m) \cdot H^{(1)}_{1|m}(\epsilon) \; , \end{aligned} \tag{3.77}$$

which can also be proven via the explicit factorization of joint probabilities from eq. (2.55) for discrete state space.

For Renyi order $q \neq 1$ again (3.76) does not hold and the Markov property does not help to simplify the chain rule of entropies. Vice versa, having a simplified chain rule for generalized entropies at hand does not allow for the conclusion of a Markov property as one perhaps might think, i.e.,

$$H^{(q\neq 1)}_{1|k}(\epsilon) = H^{(q\neq 1)}_{1|m}(\epsilon) \qquad \forall\, k \geq m \tag{3.78}$$

seems to be the Markov property of order m, but it is not. At least it is not equivalent to the space discrete version of eq. (2.12).

One can conclude that the Markov property simplifies the conditional entropies and the chain rule of entropies in case of Renyi order $q = 1$ (Shannon). For different Renyi orders also conditional entropies and a chain rule exist, and the Markov property simplifies the probabilities in the definition of the generalized entropies, but this simplification of probabilities cannot be transported into a simplification of the generalized conditional entropies and hence also not into a simplification of the chain rule of generalized entropies.

In terms of conditional mutual information (cmp. fig. 3.1) for a Markov process of order one it holds that [79]

$$I^{(q=1)}(X_{n+2}; X_n|X_{n+1}) = 0 \ . \tag{3.79}$$

3.5 Dynamical entropies and dimensions

3.5.1 Topological and metric entropy

In the 'flow'-picture (cmp. sec. 3.3.3.1) the *topological entropy rate* can be defined via

$$h_{top}(\mathbf{M}) := \lim_{\tau\to 0} \sup_{\mathcal{P}} \lim_{m\to\infty} \frac{\ln[N(\mathcal{P}^{(m)})]}{m\tau} \ , \tag{3.80}$$

if partitions are used. Here $\mathbf{M}$ is the dynamical map and $N(\mathcal{P}^{(m)})$ is the number of non-empty components (box counting) of the partition $\mathcal{P}^{(m)}$ derived from $\mathcal{P}$. In the 'map'-picture τ can be replaced by '1' and the corresponding limit is discarded from the formula. Then it can also be called *topological entropy* ([55], p.145 ff, changed notation).

It is a measure of how the number of distinct periodic orbits increases with the length of the period [80]. The topological entropy is invariant under continuous, invertible (but not necessarily differentiable) change of phase space variables. A comparison with above presented general information theory (compare sec. 3.3.3.2) yields

$$h_{top} = h_{\infty}^{(q=0)} \ . \tag{3.81}$$

In the 'flow'-picture the *metric entropy rate* can be defined via

$$h_{\mu} := \lim_{\tau\to 0} \sup_{\mathcal{P}} \lim_{m\to\infty} \frac{H(\mathcal{P}^{(m)}, \tau)}{m\tau} \ , \tag{3.82}$$

if partitions are used. H is the Shannon entropy and μ is the invariant probability measure for the underlying dynamical system. In the 'map'-picture τ can be replaced by '1' and the corresponding limit is discarded from the formula. Then it can also be called *metric entropy.*

The metric entropy can be interpreted as the gain of information concerning the location of the initial condition per iterate, assuming that arbitrarily fine resolution can be allowed. It is invariant under isomorphisms, i.e., one to one changes of variables (not necessarily continuous). The metric entropy is sometimes also called Kolmogorov-Sinai(KS) entropy or measure theoretic entropy, because the underlying measure is considered in the definition. It holds (compare sec. 3.3.3.2) that

$$h_{KS} \equiv h_{\mu} = h_{\infty}^{(q=1)} \ . \tag{3.83}$$

A comparison of topological and metric entropy rates yields generally

$$h_{top} \geq h_{\mu} \; . \tag{3.84}$$

Metric and topological entropies are positive for chaotic systems and zero for regular dynamics. For stochastic systems both quantities tend to infinity. As an exception stochastic dynamics with discrete and bounded support has finite h_{KS}, because of the existence of a finite finest resolution, for which improvement of resolution does not make sense.

It is finally pointed out that in contrast to the dynamical invariants discussed here the so-called 'ϵ-entropies' introduced before also allow for a quantification of stochastic dynamics.

3.5.2 A few comments on dimensions

Generalized dimensions are defined as [68]

$$D^{(q)} = \frac{1}{q-1} \left[\lim_{\epsilon \to 0} \frac{\ln \sum_{i=0}^{N(\epsilon)} p_i(\epsilon)^q}{\ln \epsilon} \right] . \tag{3.85}$$

Here $p_i(\epsilon)$ is the probability of the trajectory to visit the i'th box and $N(\epsilon)$ is the number of non-empty boxes. Dimensions for different parameter q correspond to different usage of the underlying probability distribution. It holds that [68]

$$D^{(q')} \leq D^{(q)} \quad \text{for} \quad q' > q \; . \tag{3.86}$$

3.5.3 Systematics of entropies and dimensions for different Renyi order

Renyi order q	Inf-th. entropy	Dynam. entropy	Dimension
0	Hartley ent. (Renyi(0) ent.)	Topological ent.	Box counting dim. Capacity dim.
1	Shannon ent. (Renyi(1) ent.)	Metric ent. Kolmogorov or KS ent. Measure theoret. ent. Information ent.	Information dim.
2	Renyi(2) ent.	Correlation ent.	Correlation dim.

3.6 Estimation from data

In the following two non-partition covering-based estimation methods and one partition-based estimation method are presented[6]. All of them can be applied to various (not all) Renyi orders q, but there exist preferences of the different methods for certain q-values. Special emphasis is put on fixed size methods based on the correlation sum. This section is closed by the method of estimation of the error in the estimation of the entropy needed for the optimization procedures in chap. 6.

---

[6] A distinction in covering-based methods and partition-based methods is problematic insofar as partitions are special coverings and hence talking about covering doesn't say anything about the question if it is a partition or not.

3.6.1 Fixed size methods

Essentially this non-partition covering-based estimation method is practically used in the present work. The following section introduces the central mathematical tool of this method.

3.6.1.1 Correlation integral and correlation sum

The generalized correlation integral is defined as

$$C_m^{(q)}(\epsilon) := \int_{\mathbf{x}} (P(\epsilon, \mathbf{x}))^{q-1} d\mu(\mathbf{x}) \equiv \langle (P(\epsilon))^{q-1} \rangle_\mu \, . \tag{3.87}$$

m is the dimension of the phase space or in later contexts the dimension of the embedding space, and

$$P(\epsilon, \mathbf{x}) := \int_{\mathcal{U}(\epsilon,\mathbf{x})} d\mu(\mathbf{y}) \, . \tag{3.88}$$

Here $\mathcal{U}(\epsilon, \mathbf{x})$ is the ϵ-neighborhood of $\mathbf{x}$. In case of existing densities it holds that

$$P(\epsilon, \mathbf{x}) = \int_{\mathcal{U}(\epsilon,\mathbf{x})} d\mathbf{y}\, p(\mathbf{y}) = \int_{\Gamma} d\mathbf{y}\, p(\mathbf{y}) \Theta(\epsilon - \|\mathbf{x} - \mathbf{y}\|) \, , \tag{3.89}$$

where Θ is the Heaviside-function. The generalized correlation sum estimates the generalized correlation integral 'consistently':

$$\hat{C}_m^{(q)}(\epsilon, N) = \frac{1}{N(N-1)^{(q-1)}} \sum_{i=1}^{N} \left[\sum_{j=1;j\neq i}^{N} \Theta(\epsilon - \|\mathbf{x}_i - \mathbf{x}_j\|) \right]^{q-1} \overset{N\to\infty}{\longrightarrow} C_m^{(q)}(\epsilon) \, . \tag{3.90}$$

Estimation via correlation sum is a special kernel estimation method with rectangular kernel, and hence it is a member of the class of nonparametric density estimation methods. For $q = 1$ the correlation sum becomes trivial: $\hat{C}_m^{(1)} = 1$. Important is the special case of $q = 2$ being free from systematic finite sample effects (bias) due to linearity. The correlation integral then takes the form of

$$\begin{aligned} C_m^{(2)}(\epsilon) &= \int_{\mathbf{x}} d\mu(\mathbf{x}) \int_{\mathbf{y}\in\mathcal{U}(\epsilon,\mathbf{x})} d\mu(\mathbf{y}) \\ &= \int_{\mathbf{x}} d\mu(\mathbf{x}) \int_{\mathbf{y}} d\mu(\mathbf{y}) \Theta(\epsilon - \|\mathbf{x} - \mathbf{y}\|) \, , \end{aligned} \tag{3.91}$$

and in case of existing densities it can be written as

$$C_m^{(2)}(\epsilon) = \int_{\mathbf{x}} d\mathbf{x}\, \rho(\mathbf{x}) \left[\int_{\mathbf{y}} d\mathbf{y}\, \rho(\mathbf{y}) \Theta(\epsilon - \|\mathbf{x} - \mathbf{y}\|) \right] \, . \tag{3.92}$$

The correlation sum for $q = 2$, also called Grassberger-Procaccia correlation sum estimates the correlation integral consistently:

$$\begin{aligned} \hat{C}_m^{(2)}(\epsilon, N) &= \frac{2}{N(N-1)} \sum_{i,j=1;i<j}^{N} \Theta(\epsilon - \|\mathbf{x}_i - \mathbf{x}_j\|) \\ &= \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{N-1} \sum_{j=1;j\neq i}^{N} \Theta(\epsilon - \|\mathbf{x}_i - \mathbf{x}_j\|) \right] \overset{N\to\infty}{\longrightarrow} C_m^{(2)}(\epsilon) \quad \in \; [0;1] \, . \end{aligned} \tag{3.93}$$

Estimation with the correlation sum is adapted to the natural measure, what is also circumscribed with the notion 'importance sampling'. The linear case of Renyi order $q = 2$ is the only case of estimation with the correlation sum, in which the bias vanishes [33]. The quantity

$$\hat{B}_m(\epsilon, N, \mathbf{x}_i) := \frac{1}{N-1} \sum_{j=1;j\neq i}^{N} \Theta(\epsilon - \|\mathbf{x}_i - \mathbf{x}_j\|) \tag{3.94}$$

($B_{\mathbf{x}_i}(\epsilon)$ in [72], p.1063) is the relative frequency for having a phase space vector in the ball of radius ϵ with central point $\mathbf{x}_i$. It is also called local density. From this the correlation sum is interpreted as an estimator for the mean probability that an arbitrary phase space vector with fixed index is lying inside of the ball of radius ϵ of any other of the phase space vectors with fixed index. In this sense the correlation sum estimates the mean phase space density of points, not to mix up with the probability density on the phase space, which is not estimated by the correlation sum, but by a quantity like $\hat{B}_m(\epsilon, N, \mathbf{x}_i)/\epsilon$.

3.6.1.2 Estimation of entropies and derived quantities via correlation sum

The estimator of joint Renyi entropies via correlation sum is defined as

$$\hat{H}_m^{(q)}(\epsilon, N) := \frac{1}{1-q} \ln \hat{C}_m^{(q)}(\epsilon, N) \ . \tag{3.95}$$

In order to justify this definition a comparison with eq. (3.35) shows that for our purposes at least a plausibility argument for $\hat{C}_m^{(q)}(\epsilon, N) \overset{N\to\infty}{\longrightarrow} \sum_{j_1,\ldots,j_m} p_{j_1,\ldots,j_m}{}^q$ ('consistency' in sense of estimation theory) is necessary:

$$\begin{aligned}
\hat{C}_m^{(q)}(\epsilon, N) &= \frac{1}{N(N-1)^{q-1}} \sum_{i=1}^{N} \left[\sum_{j=1;j\neq i}^{N} \Theta(\epsilon - \|\mathbf{x}_i - \mathbf{x}_j\|) \right]^{q-1} \\
&= \frac{1}{N(N-1)^{q-1}} \sum_{k=1}^{\#\,\text{boxes}\,(\epsilon)} \sum_{i\,:\,\mathbf{x}_i \in \text{box}\ k} \left[\sum_{j=1;j\neq i}^{N} \Theta(\epsilon - \|\mathbf{x}_i - \mathbf{x}_j\|) \right]^{q-1} \\
&\approx \frac{1}{N^q} \sum_{k=1}^{\#\,\text{boxes}\,(\epsilon)} \sum_{i\,:\,\mathbf{x}_i \in \text{box}\ k} \left[\sum_{j=1}^{N} \Theta(\epsilon - \|\tilde{\mathbf{x}}_k - \mathbf{x}_j\|) \right]^{q-1} \\
&= \sum_{k=1}^{\#\,\text{boxes}\,(\epsilon)} \frac{n_k(\epsilon)}{N} \cdot \left[\frac{1}{N} \sum_{j=1}^{N} \Theta(\epsilon - \|\tilde{\mathbf{x}}_k - \mathbf{x}_j\|) \right]^{q-1} \\
&\overset{N\to\infty}{\approx} \sum_{k=1}^{\#\,\text{boxes}\,(\epsilon)} (p_k(\epsilon))^q \ .
\end{aligned} \tag{3.96}$$

In the second step the underlying space was partitioned into disjoint boxes of about the same volume as the norm- and ϵ-dependent region (spheres for euclidean norm) selected by the Heaviside-function Θ. $\tilde{\mathbf{x}}_k$ is either chosen as the center of the box (which makes the first approximation rather crude) or it is chosen to be the mean value of the $\mathbf{x}_i \in$ box k

(which makes the final approximation rather crude). In both cases the crudeness is of course suppressed by reduction of ϵ. $n_k(\epsilon)$ is defined as the number of phase space vectors $\mathbf{x}_i$ in the k'th m-dimensional box of size given by ϵ. The final step in eq. (3.96) is the crucial one, because there the probabilities from the chosen boxes and from the Heaviside-function-selected regions are mixed. Since k counts all boxes through all dimensions the correspondence of the final line with $\sum_{j_1,\ldots,j_m} p_{j_1,\ldots,j_m}{}^q$ becomes clear.

The interesting case of $q = 2$ is given explicitly by

$$\hat{H}_m^{(2)}(\epsilon, N) = -\ln \hat{C}_m^{(2)}(\epsilon, N) \; . \tag{3.97}$$

A small correlation sum corresponds to a large entropy and accordingly to a large uncertainty and vice versa. The conditional entropy is estimated as

$$\begin{aligned} \hat{H}_{1|m}^{(q=2)}(\epsilon, N) &= \hat{H}_{m+1}^{(q=2)}(\epsilon, N) - \hat{H}_m^{(q=2)}(\epsilon, N) \\ &= -\ln \hat{C}_{m+1}^{(2)}(\epsilon, N) + \ln \hat{C}_m^{(2)}(\epsilon, N) \\ &= \ln \frac{\hat{C}_m^{(2)}(\epsilon, N)}{\hat{C}_{m+1}^{(2)}(\epsilon, N)} \; . \end{aligned} \tag{3.98}$$

3.6.1.3 Efficiency of estimation of entropies

For increasing the efficiency of the estimation of the entropies as an approximation not all pairs of the correlation sum are used at coarser resolutions (large ϵ), because neighbors[7] are found very often allowing for significant estimations. In order to allow for such abbreviated estimations in a correct way, a representatitive sample has to be selected, which can be carried through by a scramble-routine of the sampling points.

3.6.1.4 Estimation of dimension

Generalized dimensions are estimated in the scaling regime for rather small ϵ as

$$\hat{D}_m^{(q)} = \frac{1}{q-1} \frac{d \ln \hat{C}_m^{(q)}(\epsilon)}{d \ln \epsilon} \; . \tag{3.99}$$

Because of the scaling-behaviour, the left side of the equation is ϵ-independent. As a crude q-independent rule of thumb for the dimension holds $D_m = m$ (embedding dimension) in case of stochastic dynamics; $D_m \approx d$ (attractor dimension) for $m > d$ and $D_m = m$ for $m < d$ in case of deterministic dynamics.

3.6.2 Further classification of estimation methods for entropies and dimensions

Besides fixed size methods there exist other estimation methods which should be shortly mentioned.

---

[7] Two vectors are called neighbors, if their distance according to the given metric is smaller than a certain distance, which here is given by the resolution ϵ.

3.6.2.1 Fixed mass methods

The second non-partition covering-based method carries its name from demanding a fixed certain number of nearest neighbors (fixed mass) around a center point and hence adjusting adequately the radius ϵ of the sphere for neighbor search [3] in contrast to above introduced fixed size methods, which can also be seen as average pointwise mass methods. The scaling of the average radii of the neighborhoods as a function of the mass gives e.g. access to dimensional results. Fixed mass methods also belong to the class of kernel estimation methods. Usually this method is preferred for the case of $q = 1$.

3.6.2.2 Partition-based estimation methods (Generalized box counting methods)

This class of methods is based on the fact that the underlying space (state space or embedding space) is partitioned in usually equi-sized boxes. From estimations of probabilities in these boxes the results concerning entropies and dimensions are derived. Partition-based estimation methods belong to the more general class of maximum likelihood estimation methods.

Because of not being adjusted to the natural measure, these methods have large computational memory demands and hence are not very efficient. Furthermore there exists a bias in this method especially for $q = 1$. More precisely, for every finite sample size N there exists a resolution ϵ such that for higher resolution the entropies are estimated too small. Finally these methods have larger variance compared to fixed size methods.

An improvement of naive partition-based estimation methods was tried in [27], where the equi-sized boxes were replaced by boxes with size adapted to the natural measure, what is in the spirit of the fixed mass approach. Practically this method did not prevail because of the huge necessary efforts of finding a suitable partition especially if it is worked in higher dimensions.

Partition-based estimation methods should not be called solely 'box-counting methods' since in combination with the notion 'box-counting dimension' this notion is exhausted already for counting *of* boxes and in partition-based estimation methods it is dealt with counting of outcomes *in* boxes.

3.6.3 Statistical error of entropy estimates via correlation sum

It is introduced the random variable $N_m(\epsilon, \mathbf{x}_i)$ for the number of ϵ-neighbors of the vector $\mathbf{x}_i$, which is distributed according to a Binomial distribution. From data an outcome of this random variable is estimated by

$$\hat{n}_m(\epsilon, N, \mathbf{x}_i) := \sum_{k=1;k\neq i}^{N} \Theta(\epsilon - \|\mathbf{x}_i - \mathbf{x}_k\|) \ . \tag{3.100}$$

With this estimator the correlation sum can be written as

$$\hat{C}_m^{(2)}(\epsilon, N) = \frac{1}{N(N-1)} \sum_i \hat{n}_m(\epsilon, N, \mathbf{x}_i) \ . \tag{3.101}$$

For small ϵ the distribution of the random variable $N_m(\epsilon, \mathbf{x}_i)$ can be approximated by a Poisson distribution, i.e.,

$$\mathrm{Var}(N_m(\epsilon, \mathbf{x}_i)) = E(N_m(\epsilon, \mathbf{x}_i)) \; , \tag{3.102}$$

and therefore

$$\Delta n_m(\epsilon, N, \mathbf{x}_i) \approx \sqrt{\hat{n}_m(\epsilon, N, \mathbf{x}_i)} \; . \tag{3.103}$$

From the standard rules for error propagation (additivity of the variances) and the approximate relation (3.103), the statistical error of the correlation sum is estimated by

$$\Delta C_m^{(2)}(\epsilon, N) = \frac{1}{N(N-1)} \sqrt{\sum_i (\Delta n_m(\epsilon, N, \mathbf{x}_i))^2} \approx \frac{1}{N(N-1)} \sqrt{\sum_i \hat{n}_m(\epsilon, N, \mathbf{x}_i)} \; . \tag{3.104}$$

Thus it can be computed by using the non-normalized correlation sums, which are needed anyway to estimate entropies. From eq. (3.97) the statistical error of the Renyi entropy can be calculated as

$$\begin{aligned} \Delta H_m^{(2)}(\epsilon, N) &= \frac{\Delta C_m^{(2)}(\epsilon, N)}{\hat{C}_m^{(2)}(\epsilon, N)} \\ &\approx \frac{1}{\sqrt{\sum_i \hat{n}_m(\epsilon, N, \mathbf{x}_i)}} \; , \end{aligned} \tag{3.105}$$

and the statistical error of the usual conditional entropy is obtained from

$$\Delta H_{1|m}^{(2)}(\epsilon, N) = \sqrt{\Delta H_{m+1}^{(2)}(\epsilon, N)^2 + \Delta H_m^{(2)}(\epsilon, N)^2} \tag{3.106}$$

$$\approx \sqrt{\frac{1}{\sum_i \hat{n}_{m+1}(\epsilon, N, \mathbf{y}_i)} + \frac{1}{\sum_i \hat{n}_m(\epsilon, N, \mathbf{x}_i)}} \; . \tag{3.107}$$

Further error propagation, e.g., for the estimated redundancy

$$\Delta R_m(\epsilon, N) = \sqrt{\Delta H_1(\epsilon, N)^2 + \Delta H_{m+1}(\epsilon, N)^2 + \Delta H_m(\epsilon, N)^2} \tag{3.108}$$

is possible in the same way.

The assumptions entering the arguments for usual error propagation are:

1. Independence of the random variables
2. Gaussian error statistics
3. Errors are small so that nonlinear expressions can be approximated by first order Taylor expansions around the mean.

Item 1 is violated, since if $\|\mathbf{x}_i - \mathbf{x}_{i'}\| < \epsilon$, then the phase space points have overlapping neighborhoods and $N_m(\epsilon, \mathbf{x}_i)$ and $N_m(\epsilon, \mathbf{x}_{i'})$ are not independent of each other. The violation of this assumption, however, becomes the less relevant the smaller ϵ, since then the overlap of neighborhoods decreases. Item 2 is violated, since the error statistics of our basic random variables $N_m(\epsilon, \mathbf{x}_i)$ is explicitly non-Gaussian. This violation becomes the stronger the smaller the values of $\hat{n}_m(\epsilon, N, \mathbf{x}_i)$ become, i.e., for small ϵ. In spite of those arguments, usual error propagation is used as an approximation of the true errors of the estimation.

3.7 Interrelation of the correlation integral, entropies and dimensions

First, it is defined

$$A_m^{(q)}(\epsilon) := \sum_{k=1}^{\#\,\text{boxes}\,(\epsilon)} (p_k(\epsilon))^q \ , \tag{3.109}$$

where m is the dimension of the boxes. From eq. (3.35) with a multidimensional counting procedure by one index k it is derived

$$H_m^{(q)}(\epsilon) = \frac{1}{1-q} \ln A_m^{(q)}(\epsilon) \ , \tag{3.110}$$

or

$$A_m^{(q)}(\epsilon) = e^{(1-q)H_m^{(q)}(\epsilon)} \ , \tag{3.111}$$

and eq. (3.85) can be written as

$$D_m^{(q)} = \frac{1}{q-1}\left[\lim_{\epsilon\to 0} \frac{\ln A_m^{(q)}(\epsilon)}{\ln \epsilon}\right] \ . \tag{3.112}$$

As an ansatz for small ϵ it is chosen

$$A_m^{(q)}(\epsilon) = \epsilon^{(q-1)\cdot D_m^{(q)}} \cdot f(m,q,\epsilon) \quad \text{with} \quad f(m,q,\epsilon) \propto \epsilon^p \ . \tag{3.113}$$

Inserting this into eq. (3.112) the following line of implications can be carried out:

$$\begin{aligned}
D_m^{(q)} =& \frac{1}{q-1}\left[\lim_{\epsilon\to 0} \frac{(q-1)\cdot D_m^{(q)} \cdot \ln\epsilon + \ln f(m,q,\epsilon)}{\ln\epsilon}\right] && (3.114)\\
\Rightarrow \quad & \lim_{\epsilon\to 0} \frac{\ln f(m,q,\epsilon)}{\ln \epsilon} = 0 && (3.115)\\
\Rightarrow \quad & p = 0 && (3.116)\\
\Rightarrow \quad & A_m^{(q)}(\epsilon) = \epsilon^{(q-1)\cdot D_m^{(q)}} \cdot f(m,q) \qquad \text{for small } \epsilon \ . && (3.117)
\end{aligned}$$

The function f needs to be independent of ϵ. Simple A-elimination in eq. (3.110) yields with

$$H_m^{(q)}(\epsilon) = -D_m^{(q)} \ln\epsilon + \frac{\ln f(m,q)}{1-q} \quad , \quad q \neq 1 \tag{3.118}$$

a relation for the entropy as a function of the dimension, which is still quite flexible because of the non-fixed function f. The formula [36]

$$H_m(\epsilon) \approx -D\ln\epsilon + hm + \text{const} \ , \tag{3.119}$$

being in accordance with the relation before, is more detailed concerning an implicit suggestion of an appearance of the entropy rate, but has the status of an ansatz. For the conditional entropy it holds that

$$H_{1|m}^{(q)}(\epsilon) = H_{m+1}^{(q)}(\epsilon) - H_m^{(q)}(\epsilon) = -D_{m+1}^{(q)} \ln\epsilon + \frac{\ln f(m+1,q)}{1-q} + D_m^{(q)}\ln\epsilon - \frac{\ln f(m,q)}{1-q} \ . \tag{3.120}$$

For dissipative dynamical systems and sufficiently high m the dimension terms cancel. Thus a relation of a conditional entropy rate $h = \frac{\Delta H}{\tau}$ and the dimension D is impossible. Substituting eq. (3.111) into eq. (3.112) one obtains

$$D_m^{(q)} = -\lim_{\epsilon \to 0} \frac{H_m^{(q)}(\epsilon)}{\ln \epsilon} \; . \tag{3.121}$$

For $q \neq 1$ this is also obtained from eq. (3.118) by resolving for the dimension

$$D_m^{(q)} = \left(-\frac{H_m^{(q)}(\epsilon)}{\ln \epsilon} + \frac{\ln f(m,q)}{\ln \epsilon^{1-q}} \right) \tag{3.122}$$

and performing the limit $\epsilon \to 0$. Eq. (3.121) is a simple and universal representation of dimensions from joint entropies valid for *all stochastic* and also dissipative deterministic cases! The dimension is obtained from operations on the total uncertainty. In the framework of deterministic dynamics this relation is given without m in ([68], p.107).

On the basis of relation (3.121) for the dimension as a function of entropies, an access to an ϵ-dependent dimension is suggested by

$$D_m^{(q)}(\epsilon) = -\frac{H_m^{(q)}(\epsilon)}{\ln(\epsilon)} \; . \tag{3.123}$$

Then a (of course only) qualitative analogy of this dimension increasing in ϵ for sufficiently small $\epsilon < 1$ with the increase of the number of relevant coupling constants in the coarse graining of the renormalization group flow of Wilson could be suggested as a contribution to analogies of dynamical systems with statistical physics concerning effective behaviour.

From the fact that the generalized correlation sum estimates consistently the generalized correlation integral in eq. (3.90) as well as (approximately) $\sum_{k=1}^{\#\,\text{boxes}\,(\epsilon)} (p_k(\epsilon))^q$ in eq. (3.96) it is deduced that

$$C_m^{(q)}(\epsilon) \approx A_m^{(q)}(\epsilon) = \sum_{k=1}^{\#\,\text{boxes}\,(\epsilon)} (p_k(\epsilon))^q \; , \tag{3.124}$$

where $C_m^{(q)}(\epsilon)$ is the generalized correlation integral. For all ϵ

$$C_m^{(q)}(\epsilon) \approx e^{(1-q)\cdot H_m^{(q)}(\epsilon)} \tag{3.125}$$

and for small ϵ

$$C_m^{(q)}(\epsilon) \approx \epsilon^{(q-1)\cdot D_m^{(q)}} \cdot f(m,q) \tag{3.126}$$

hold. Knowing those stronger formulations for the correlation integral, the proportionality

$$C_m^{(q)}(\epsilon) \propto \epsilon^{(q-1)\cdot D_m^{(q)}} e^{(1-q)\cdot H_m^{(q)}(\epsilon)} \tag{3.127}$$

in ([46], p.227) is in the author's opinion not correct, seen from combining eq.'s (3.125) and (3.126), which are two different representations of the correlation integral via either the entropy or the dimension.

3.8 Entropies in case of continuous time

Observing from eq. (3.57) that the entropic quantities are related via quotients of differences (seemingly discrete versions of derivatives), a time-continuous formulation of the relationships of entropic quantities should naturally also be at hand. Furthermore, the discussion of changes in the sampling rate in sec. 4.4.5 leads in the limit case of upsampling to the case of continuous time.

First, entropic formulas with explicit discretization time dependence close to those of sec. 3.3, but in formulation of [31], the central paper behind this section, are shown in sec. 3.8.1. The limit of infinitesimal time step length is performed after having performed the limit of infinite time. In sec. 3.8.2 the relations of entropic quantities are presented for non-infinitesimal time step length and finite time. Sec. 3.8.3 is the central aim of this part. It is the generalization of sec. 3.8.2 for the limit of infinitesimal time step length as well as the generalization of sec. 3.8.1 with omitted limit of infinite time. Attempts for such formulations are not available in the literature until now. Before concluding in sec. 3.8.5, the developed ideas are applied to example dynamics of the Roessler and Lorenz system as well as the Ornstein-Uhlenbeck process in sec. 3.8.4.

3.8.1 Usual procedure

The starting point is the definition of the joint entropy

$$H(\mathcal{P},\tau,t) = H(\mathcal{P},\tau,m := t/\tau) = -\sum_{\omega_0,\dots,\omega_{m-1}} P(\omega_0,\dots,\omega_{m-1}) \ln P(\omega_0,\dots,\omega_{m-1}) \; . \tag{3.128}$$

The symbol τ, being the time step length, appears on the right side only implicitly as the time between successive realizations and the partition $\mathcal{P}$ appears on the right side implicitly in the range of values ω_i can take. In order for the following expressions to make sense it has to be demanded that the partition for t_1 is consistent with the partition for $t_2 > t_1$. It is possible to define the rate

$$\tilde{h}(\mathcal{P},\tau,t) := \frac{H(\mathcal{P},\tau,t)}{t} \; . \tag{3.129}$$

The entropy per unit time with respect to partition $\mathcal{P}$ is then defined [31] as the limit:

$$\tilde{h}(\mathcal{P},\tau) := \lim_{t\to\infty} \tilde{h}(\mathcal{P},\tau,t) \; . \tag{3.130}$$

Starting from eq. (3.128) it is also possible to define another rate

$$h(\mathcal{P},\tau,t) := \frac{H(\mathcal{P},\tau,t+\tau) - H(\mathcal{P},\tau,t)}{\tau} \tag{3.131}$$

and the corresponding limit:

$$h(\mathcal{P},\tau) := \lim_{t\to\infty} h(\mathcal{P},\tau,t) \; . \tag{3.132}$$

In general it holds that

$$\tilde{h}(\mathcal{P},\tau) = h(\mathcal{P},\tau) \quad \text{, but} \qquad \tilde{h}(\mathcal{P},\tau,t) \neq h(\mathcal{P},\tau,t) \; . \tag{3.133}$$

The $\epsilon\tau$-entropy rate is derived by

$$h(\epsilon,\tau) = \inf_{\mathcal{P}:diam(P_i)\leq\epsilon} h(\mathcal{P},\tau)\ . \tag{3.134}$$

On the other hand, having obtained $H_m(\epsilon,\tau)$ from $H(\mathcal{P},\tau,m)$, e.g., again via infimum

$$H_m(\epsilon,\tau) = \inf_{\mathcal{P}:diam(P_i)\leq\epsilon} H(\mathcal{P},\tau,m)\ , \tag{3.135}$$

by the limit of infinite time from the conditional entropy rate also $\epsilon\tau$-entropy rates can be obtained [15] [8]:

$$h_\infty(\epsilon,\tau) = \lim_{m\to\infty} h_m(\epsilon,\tau) = \lim_{m\to\infty}\frac{1}{\tau}[H_{m+1}(\epsilon,\tau) - H_m(\epsilon,\tau)]\ . \tag{3.136}$$

From a physicists point of view the $\epsilon\tau$-entropy rates of eq. (3.134) and of eq. (3.136) should be treated as identical. Depending on the state space resolution ϵ and time resolution τ, $h(\epsilon,\tau)$ and $h_\infty(\epsilon,\tau)$ both are the average rate of information produced by the system or the information per time step needed for prediction of the immediate future time step ahead if infinite time conditioning is imposed. According to eq. (3.61) the Kolmogorov-Sinai entropy rate for processes in continuous time is obtained via

$$h_{KS} = \lim_{\epsilon\to 0,\tau\to 0} h_\infty(\epsilon,\tau) = \lim_{\epsilon\to 0,\tau\to 0} h(\epsilon,\tau)\ . \tag{3.137}$$

3.8.2 Discrepancies from usual differentiation and integration structure of entropies in discrete time

Explicit space-resolution dependence exists for all following quantities, but is avoided here in notation. Already being introduced in eq. (3.57) the following formula is given again here for better visibility of the structure:

$$h_m(\tau) = \frac{\Delta H_m(\tau)}{\tau} = \frac{H_{m+1}(\tau) - H_m(\tau)}{\tau} = \frac{H_{1|m}(\tau)}{\tau}\ . \tag{3.138}$$

Two interpretations are possible: As finite-m-entropy rate it is the change of the uncertainty under elongation of the time interval $m \cdot \tau$ divided by the change of time and so it is the slope of $H_m(\tau)$ in m. On the other hand, it is interpreted as a special conditional entropy per time step length. In the same sense as before it is defined

$$\begin{aligned} g_m(\tau) &:= \frac{\Delta h_m(\tau)}{\tau} = \frac{h_{m+1}(\tau) - h_m(\tau)}{\tau} = \frac{h_{1|m}(\tau)}{\tau} \\ &= \frac{H_{m+2}(\tau) - 2H_{m+1}(\tau) + H_m(\tau)}{\tau^2} = \frac{\Delta H_{1|m}(\tau)}{\tau^2}\ . \end{aligned} \tag{3.139}$$

$g_m(\tau)$ is either interpreted as the τ-dependent curvature of H_m as function of m (unit: $[g_m] = \frac{1}{s^2}$) or it is interpreted as the change of the special conditional entropy per time step

---

[8] Some formulas and quantities already met in sec. 3.3.3 are repeated here, in order to avoid to tear in parts the line of thought intended here.

length under increase of conditioning divided by the time step length.

Above definitions allow for the representation as sums (or integrals) of

$$h_m(\tau) = (g_{m-1}(\tau) + ... + g_0(\tau)) \cdot \tau + h_0(\tau) \tag{3.140}$$

and

$$\begin{aligned} H_m(\tau) &= (h_{m-1}(\tau) + ... + h_0(\tau)) \cdot \tau \\ &= (g_{m-2}(\tau) + 2g_{m-3}(\tau) + ... + (m-1)g_0(\tau)) \cdot \tau^2 + mh_0(\tau) \cdot \tau \ . \end{aligned} \tag{3.141}$$

The crucial point in this section as preparation for the following is that *all* quantities are *explicitly τ-dependent*, what accounts for the headline of this section.

The idea of connection of the entropic quantities by derivatives with finite τ is also supported in [18].

3.8.3 Time-continuous information theory

In sec. 3.8.1 the succession of limit procedures for accessing the KS entropy introduced in [31] was redisplayed. Starting with the same formula (3.128), it is a naturally arising question what happens if the limits of $t \to \infty$ and $m \to \infty$ are not performed, but instead the limit $\tau \to 0$ is investigated. This does not lead to dynamical invariants, but instead to prediction-relevant (finite t, m) relations of information theory in the time-continuous case. The relevance for prediction lies in case of existence of the corresponding entropic quantities in the determination of threshold-values for maximal uncertainties with respect to increased sampling rates.

Partition dependence of (3.128) can be translated into a resolution dependence by

$$H(\epsilon, \tau, t = m\tau) = \inf_{\mathcal{P}:diam(P_i) \leq \epsilon} H(\mathcal{P}, \tau, t) \ , \tag{3.142}$$

but since only time aspects are discussed here, space resolution dependence is suppressed in notation (though existing) during the following qualitative discussion and it remains the entropy $H(\tau, t)$.

Starting with sec. 3.8.2 a naturally arising question is, what happens in the case of performing the limit for $\tau \to 0$. A small change of notation from $H_m(\tau)$ to $H(\tau, t = m \cdot \tau)$ leads to the same starting point of $H(\tau, t)$.

Since $H_{m=1}(\tau)$ is the uncertainty of one random variable, it should be independent of τ. This results in

$$H(\tau, t = \tau) = \text{const} \quad \text{for} \quad \tau > 0 \ , \tag{3.143}$$

where a double τ - dependence on the left side of the equation leads surprisingly to τ-independence of the whole expression.

For fixed t now the limit $\tau \to 0$ has to be performed. If the limit exists, then[9]

$$H(\tau = 0, t) = \lim_{\tau \to 0} H(\tau, t) \tag{3.144}$$

---

[9] The notation $H(t) := H(\tau = 0, t)$ is problematic, since the same procedure is also natural for $H(\tau) := \lim_{t \to \infty} H(\tau, t)$, what was already used in (3.132) for the entropy rate. Then introducing numbers for t and τ the confusion would be perfect.

is the time-continuous joint entropy. Otherwise a time-continuous treatment of the joint uncertainty is not possible for the process at hand. Taking on the other hand the other involved variable equal to zero it is defined

$$H(\tau, t = 0) \equiv 0 \; . \tag{3.145}$$

It is inferred

$$H(\tau = 0, t = 0) = 0 \; . \tag{3.146}$$

From $H_m(\tau) = H(\tau, t = m\tau)$ being (not necessarily strong) monotonously increasing in m it is inferred that also in the limit of $\tau \to 0$ the entropy $H(\tau = 0, t)$ is (not necessarily strong) monotonously increasing in t. An interesting question is the finiteness of $H(\tau = 0, t)$ for finite t for various process classes, which will be answered numerically by treating examples in sec. 3.8.4 shown in fig.'s 3.6, 3.7, 3.14 and 3.15. For fixed t the performed limit causes

$$\lim_{\tau \to 0} m = \lim_{\tau \to 0} \frac{t}{\tau} = \infty \; . \tag{3.147}$$

From eq.'s (3.128) and (3.147) it has to be inferred that $H(\tau = 0, t)$ has to be understood in terms of path integrals.

Eq. (3.138) in notation suitable for the purpose of a time-continuous formulation reads

$$h(\tau, t) := \frac{H(\tau, t + \tau) - H(\tau, t)}{\tau} \; . \tag{3.148}$$

If the limit exists, then

$$h(\tau = 0, t) = \lim_{\tau \to 0} h(\tau, t) \tag{3.149}$$

is the finite time entropy rate in the time-continuous case. The discrepancy from the usual differentiation is that the function to be differentiated *depends explicitly* on the parameter of the differentiation. This is unusual, but not untreatable. In case of existence of $H(\tau = 0, t)$ in eq. (3.144) for the corresponding derivative it holds that

$$h(\tau = 0, t) = \lim_{\Delta t \to 0} \frac{H(0, t + \Delta t) - H(0, t)}{\Delta t} \; , \tag{3.150}$$

i.e., $h(0, t)$ from (3.149) is obtained as a usual derivative from $H(0, t)$. $h(\tau = \Delta t, t)$ is found to be a perturbation of the usual quotient of differences $\frac{H(0,t+\Delta t) - H(0,t)}{\Delta t}$, but the limit of $\Delta t \to 0$ is the same in both cases. From (3.148) and (3.145) it is derived

$$\tau \cdot h(\tau, t = 0) = H(\tau, t = \tau) = \text{const} \; . \tag{3.151}$$

One concludes that for arbitrary dynamics

$$\lim_{\tau \to 0} h(\tau, t = 0) = h(\tau = 0, t = 0) = \infty \quad \text{if} \quad H(\tau, t = \tau) > 0 \; . \tag{3.152}$$

This singularity in general should make necessary to introduce generalized functions for h. Since also in the time-continuous case there is no reason to assume that increase of conditioning could in any case increase the uncertainty it should be given that $h(\tau = 0, t)$ is monotonously decreasing in $t \in [0, \infty]$. The behaviour of h is shown in the examples of sec. 3.8.4 in fig.'s 3.8, 3.9 and 3.17.

Eq. (3.139) in notation suitable for the purpose of a time-continuous formulation reads

$$g(\tau, t) := \frac{h(\tau, t+\tau) - h(\tau, t)}{\tau} = \frac{H(\tau, t+2\tau) - 2H(\tau, t+\tau) + H(\tau, t)}{\tau^2} \ . \tag{3.153}$$

In case of existence of the corresponding quantities,

$$g(\tau = 0, t) = \lim_{\tau \to 0} g(\tau, t) \tag{3.154}$$

is the curvature of the joint entropy in the time-continuous case.

Eq. (3.140), written in a form closer to the time-continuous notation, reads

$$h(\tau, t = m\tau) = (g(\tau, t = (m-1) \cdot \tau) + ... + g(\tau, 0)) \cdot \tau + h(\tau, 0) \ . \tag{3.155}$$

With the aim of representing $h(\tau = 0, t)$ from the function g it is now necessary to enlarge the domain of definition of the function g for finite τ from discrete t to continuous t and the simplest way of doing so is to linearly interpolate between the values of the function g at t for multiples of τ. Other more smooth interpolation procedures could also be allowed. Now from decoupling the two τ-dependences it is possible to obtain the following integral representation:

$$\begin{aligned} h(\tau = 0, t = m\tau) \quad &= \quad \lim_{\tau \to 0} [\sum_{k=0}^{m-1} \tau \cdot g(\tau, t = k\tau) + h(\tau, 0)] \\ &\overset{m\tau = m'\tau'}{=} \lim_{\tau \to 0} \lim_{\tau' \to 0} [\sum_{k=0}^{m'-1} \tau' \cdot g(\tau, t = k\tau') + h(\tau, 0)] \\ &= \quad \lim_{\tau \to 0} [\int_0^t ds \, g(\tau, s) + h(\tau, 0)] \ . \end{aligned} \tag{3.156}$$

In order to avoid the possibility of '$-\infty + \infty$' it is preferred to first sum the finite expressions before performing the limit.

Also eq. (3.141), written in a form closer to the time-continuous notation by

$$\begin{aligned} H(\tau, t = m\tau) &= (h(\tau, t = (m-1)\tau) + ... + h(\tau, 0)) \cdot \tau \\ &= (g(\tau, t = (m-2)\tau) + 2g(\tau, t = (m-3)\tau) + ... \\ &\qquad + (m-1)g(\tau, 0)) \cdot \tau^2 + mh(\tau, 0) \cdot \tau \ , \end{aligned} \tag{3.157}$$

leads in the time-continuous limit to a representation of the joint entropy via integral:

$$\begin{aligned}
H(\tau=0,[0,t]) &\equiv H(\tau=0,t) \\
&= \lim_{\tau\to 0}[\sum_{k=0}^{m-1}\tau\cdot h(\tau,t=k\tau)] \\
&\overset{m\tau=m'\tau'}{=} \lim_{\tau\to 0}\lim_{\tau'\to 0}[\sum_{k=0}^{m'-1}\tau'\cdot h(\tau,t=k\tau')] \\
&= \lim_{\tau\to 0}\int_0^t du\, h(\tau,u) \\
&= \int_0^t du\, h(\tau=0,u) \\
&= \int_0^t du\, \left[\lim_{\tau\to 0}[\int_0^u ds\, g(\tau,s)+h(\tau,0)]\right] . \qquad (3.158)
\end{aligned}$$

Interchangeability of the limit procedures differentiation and integration is needed.

For getting closer to prediction it is now necessary to open a framework with finite conditioning time in the past and a finite time ahead concerning which to predict. The joint conditional entropy reads

$$H_{m|n}(\tau)=H_{m+n}(\tau)-H_n(\tau) . \qquad (3.159)$$

Converted to a notation suitable for the time-continuous case with

$$t_1=m\tau \quad \text{and} \quad t_2=n\tau \qquad (3.160)$$

it looks like

$$H(\tau,t_1|t_2)=H(\tau,t_1+t_2)-H(\tau,t_2) . \qquad (3.161)$$

From the chain rule

$$H(\tau,t_1=m\tau)=\tau\cdot(h(\tau,0)+\ldots+h(\tau,(m-1)\tau)) \qquad (3.162)$$

it is derived

$$H(\tau,t_1=m\tau|t_2=n\tau)=\tau\cdot(h(\tau,n\tau)+\ldots+h(\tau,(m+n-1)\tau)) , \qquad (3.163)$$

and using eq. (3.158) finally a representation of $H(\tau=0,t_1|t_2)$ via rate-quantities in the limit of $\tau\to 0$ is possible as

$$H(\tau=0,t_1|t_2)=\int_{t_2}^{t_1+t_2} du\ h(\tau=0,u) . \qquad (3.164)$$

Such quantities are interesting as a maximum uncertainty under conditioning for arbitrarily high sampling rate.

3.8.4 Example calculations for time-continuous entropies

3.8.4.1 Roessler system (deterministic chaotic dynamics)

The Roessler system is given by

$$\begin{aligned}\dot{x} &= -y - z \\ \dot{y} &= x + ay \\ \dot{z} &= b + z(x - c)\ . \end{aligned} \tag{3.165}$$

The parameters are chosen as $a = b = 0.2$ and $c = 5.7$. The quadratic term 'zx' is the only nonlinearity. The largest Lyapunov exponent of the Roessler attractor is $\lambda \approx 0.07$ and the fractal dimension is $D^{(2)} = 1.99 \pm 0.07$.

In fig. 3.4 the joint entropy of the x-coordinate of the Roessler system is given as a function of resolution ϵ, time step length τ and time t. Convergence of the joint entropy for decreasing time step length τ can be seen. The joint entropy H depends logarithmically on the resolution ϵ, from which the dimension is obtained as a slope of H with $\ln \epsilon$ in the limit of infinitesimal ϵ according to eq. (3.121) with a value as predicted. It has to be pointed out that in comparison with the usually used algorithms for the determination of the dimension here an alternative access to the determination of the dimension exists, from which it can be read off as a slope.

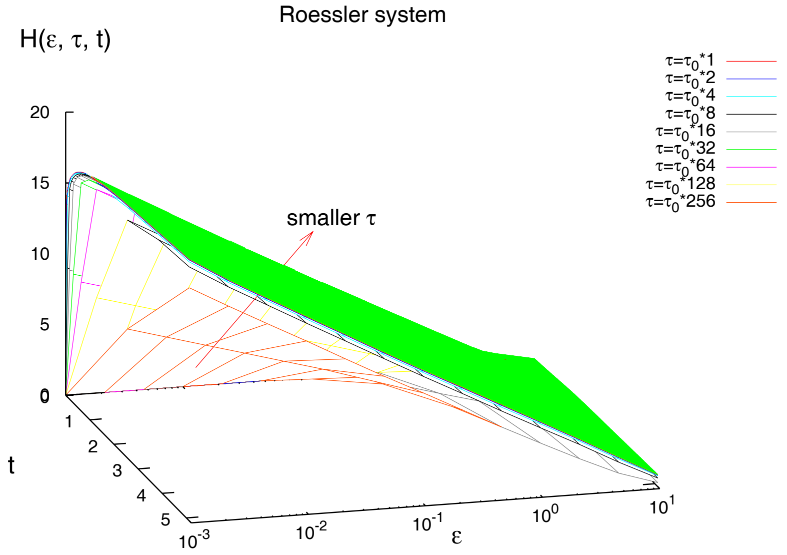


Figure 3.4: Joint entropies of the Roessler system as a function of the time t and the resolution ϵ for various $\tau = \tau_0 \cdot 2^n$; $n = 0, ..., 8$ $(\tau_0 = 0.01)$ for the x-coordinate of the Roessler dynamics; $3 \cdot 10^6$ data points; $t = 5$ corresponds to approximately 3 circulations in the attractor.

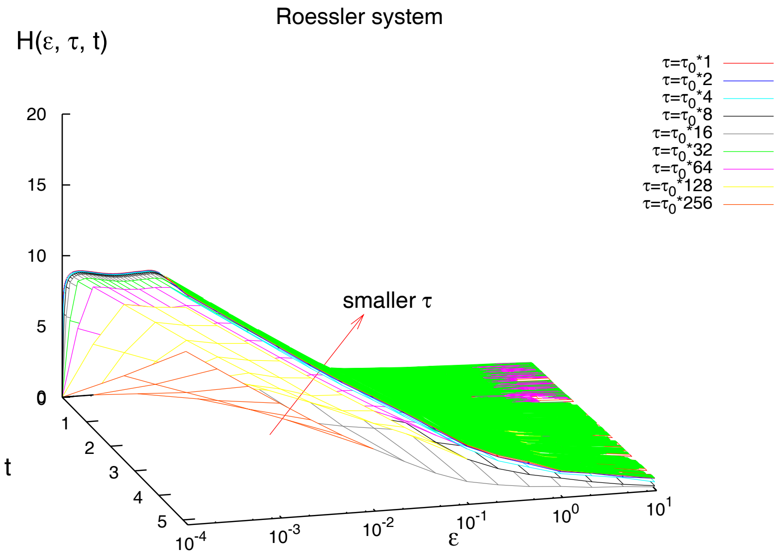


Figure 3.5: Joint entropies of the Roessler system as a function of the time t and the resolution ϵ for various $\tau = \tau_0 \cdot 2^n$; $n = 0, ..., 8$ $(\tau_0 = 0.01)$ for the z-coordinate of the Roessler dynamics; $3 \cdot 10^6$ data points.

It is possible to see in fig. 3.5 (in comparison with fig. 3.4 one smaller decade of resolutions is shown) that the z-component of the Roessler attractor carries less uncertainty for the same values of (ϵ, τ, t) than the x-component. Nevertheless the dimension of the attractor is captured also by the joint entropy of the z-component.

In fig. 3.6 the slice of fig. 3.4 for resolution $\epsilon = 10^{-2}$ with extended time t is shown. It is found that asymptotically in t the joint uncertainty of the Roessler system increases linearly, but very slowly. The measured slope for smallest τ in between $t = 20$ and $t = 40$ is 0.08. It estimates the KS entropy rate

$$h_{KS} = \lim_{\epsilon \to 0} \lim_{\tau \to 0} \lim_{t \to \infty} h(\epsilon, \tau, t) \ . \tag{3.166}$$

The deviation from the expected value 0.07 can be found in the fact that still too small t or too large ϵ are used for the estimation. It can be concluded from the small slope that the uncertainty of the first few points is larger than the uncertainty of a rather long motion in the attractor given the first few points.

In fig. 3.7 the slice of fig. 3.5 for resolution $\epsilon = 10^{-2}$ with extended time t is shown. Compared to fig. 3.6 a much smaller initial uncertainty H_1 is observed for the z-component. This is understood from the high probability of the z-component to stay at zero. Furthermore in contrast to the x-component a wave-like structure of the joint entropy for smaller time t is observed for the z-component. A non-monotonous entropy rate h as a function of t has to be inferred. This unintuitive signature was robustly reproduced under various parameter values and confidence in this result is caused by the fact that asymptotically for smallest available τ the slope of 0.08 is found, which serves as a correct estimation of the KS entropy also from the

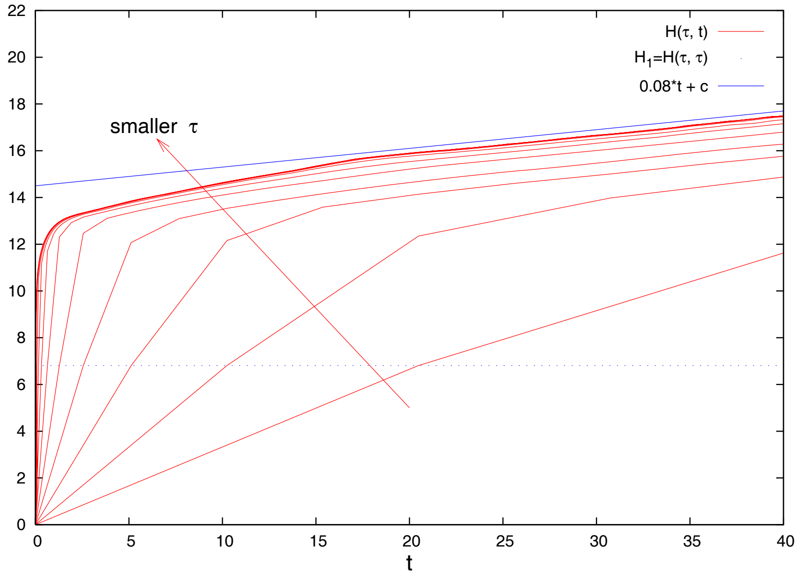


Figure 3.6: Joint entropy of the x-coordinate of the Roessler system for fixed resolution $\epsilon = 10^{-2}$. With respect to time t it is an extended slice of fig. 3.4 for fixed resolution ϵ.

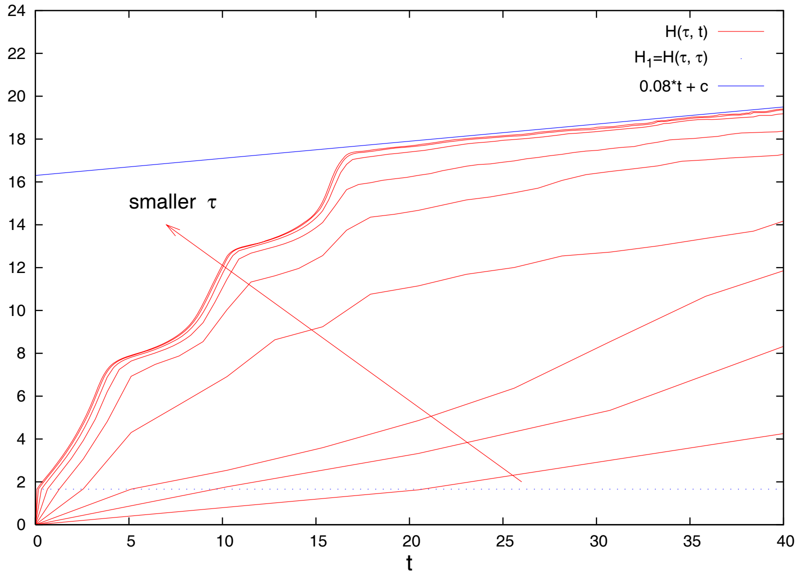


Figure 3.7: Joint entropy of the z-coordinate of the Roessler system for fixed resolution $\epsilon = 10^{-2}$. With respect to time t it is an extended slice of fig. 3.5 for fixed resolution ϵ.

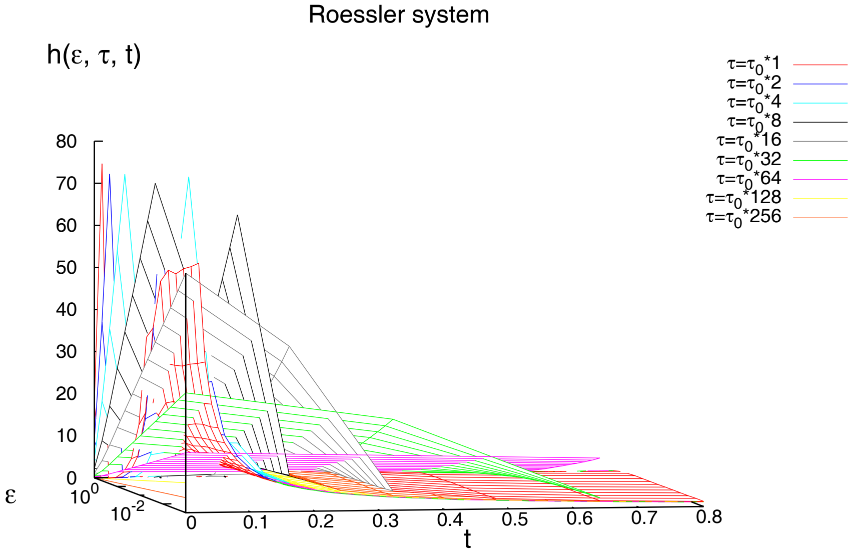
z-component of the Roessler system. Since also (from fig. 3.5) the dimension of the Roessler dynamics is correctly extractable the result should not be a numerical artefact. Nevertheless we do not expect that it is true physics in the sense of increasing uncertainty by enlarged conditioning. According to Shannon entropies, for which additional conditioning cannot increase uncertainty, this behaviour is forbidden. A possible explanation of this behaviour can be assigned to effects of Renyi order $q = 2$, because in connection with eq.'s (3.17) and (3.18) a reduction of generalized entropies with increased conditioning remained as an open problem. With this reasoning the example carries the potential of tearing apart the interpretation of uncertainty from $q \neq 1$ - entropies. Another possible explanation that the wave-like structure corresponds to a finite sample effect being responsible for non-ergodicity has to be treated as rather improbable, because also for smallest resolution (largest ϵ) it was not possible to detect a qualitative change in the sense of smoothing of the wave-like structure of the joint entropy under drastical enlargement of the length of the dataset.

As indicated in sec. 3.8.3 the main intention of this chapter does not concern the invariants of dynamical systems obtained in the limit cases of ϵ and t, but especially the limit of $\tau \rightarrow 0$ for finite ϵ and t, which contains the informational characteristic of the dynamics. The corresponding area $H(\epsilon, \tau = 0, t)$ is of importance with respect to prediction, because optimal prediction is performed at a certain finite optimal resolution, the area for $\tau = 0$ is approximated by the results of a time series with finite τ and especially because prediction deals with finite time. The difference of H as a function of t for $\tau = 0$ of the x- and z-component of the Roessler system is not of importance for the dynamical invariants, but of importance for their usage in tasks of prediction. In general, demands with respect to the finite part of the area can be used as selection criteria for observables of prediction.

With consideration of the fact that the values of h_{KS} and D were estimated for finite large t and finite small ϵ instead of in the true limit it can be concluded that the same values are obtained from the x- and the z-component of the chaotic Roessler system in the limit cases, in accordance with the theorem of Takens. On the other hand it can be clearly seen that for finite time t and finite resolution ϵ the joint uncertainty is allowed to behave qualitatively different for different observables as the x- and the z-component of the chaotic Roessler system without contradicting the theorem of Takens.

In fig. 3.8 the entropy rate of the Roessler system is shown. For every τ the entropy rate as a function of t and ϵ is represented by a surface. Those surfaces are interleaved for different τ. Depending on the ranges shown along the axes different results become apperent: With the example of the Roessler system even for deterministic dynamics the posed divergence in eq. (3.152) of $h(\tau = 0, t = 0) = \infty$ can be verified for sufficiently small ϵ observed in particular in the upper panel of fig. 3.8. It is possible to see that $h(\epsilon, \tau, t = 0)$ for finite τ increases logarithmically in ϵ with a slope depending on τ. On the other hand in the lower panel in accordance with the result from the joint entropy it is possible to see that the entropy rate reaches a small finite value for larger times t, which is the KS entropy rate.

A power law decay of the entropy rate can be seen for small t in fig. 3.9. Furthermore the transition to constant entropy rate larger than zero for large time t is an indication for a finite KS entropy rate in accordance with the literature.



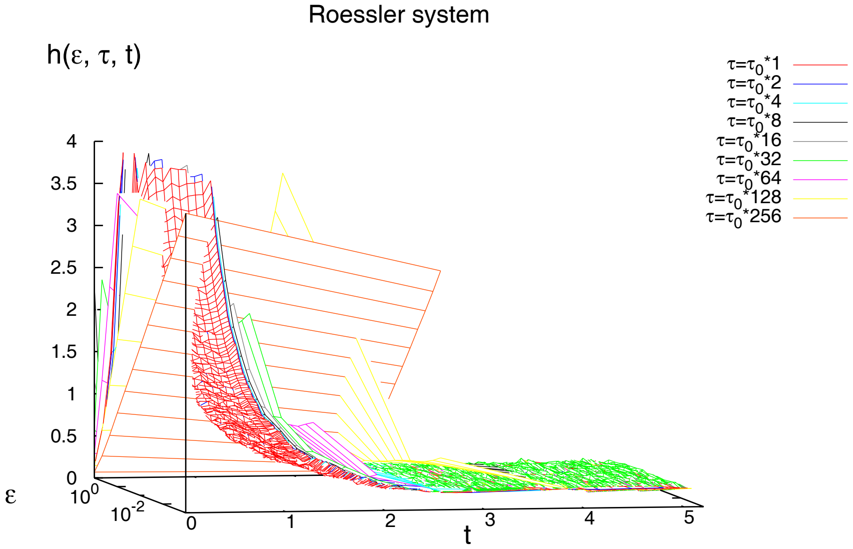


Figure 3.8: Entropy rate as a function of resolution ϵ, time t and varying time interval τ for the x-coordinate of the Roessler system with parameters for chaotic dynamics. Upper panel: Focus on resolution dependence of divergent behaviour for small time t. Lower panel: Focus on larger times for the Kolmogorov-Sinai entropy rate.

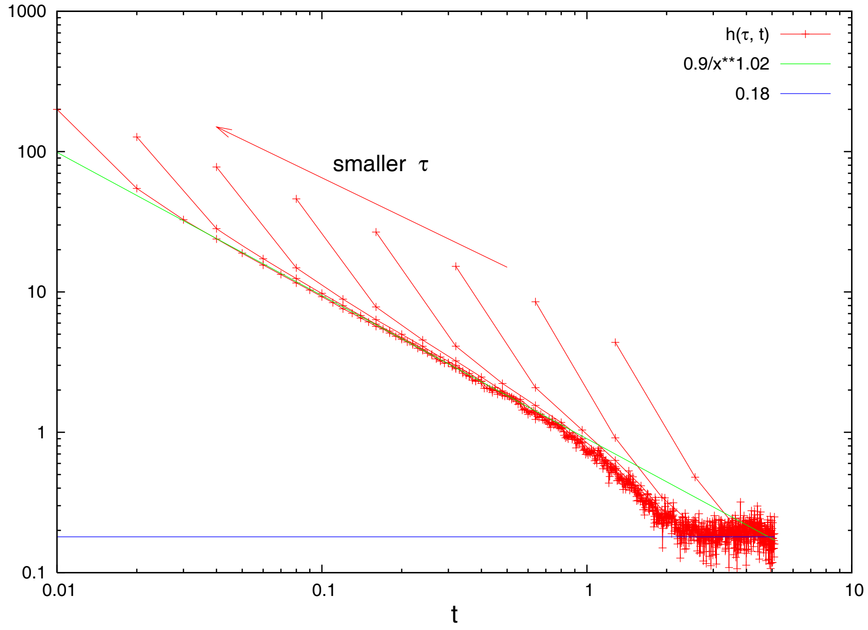


Figure 3.9: Entropy rate as a function of time and decreasing time interval τ for fixed resolution ($\epsilon = 10^{-2}$ in this case) for the x-coordinate of the Roessler system in logarithmic representation. The values of h at $t = 0$ are omitted.

3.8.4.2 Roessler system with limit cycle (deterministic regular dynamics)

The parameters chosen for the Roessler dynamics in this section are $a = b = 0.2$ and $c = 8.0$. The dynamics becomes regular with an expected dimension of the attractor as $D = 1$ and an expected KS entropy rate of $h_{KS} = 0$. The trajectory is shown in fig. 3.10.

In fig. 3.11 the joint entropy of the x-coordinate of the regular Roessler system for the given parameter values is shown as a function of resolution, time and time step length. From the slope of H with $\ln \epsilon$ it is verified $D \approx 1$.

A slice of fig. 3.11 extended in time is shown in fig. 3.12. For large time the entropy doesn't increase anymore, what indicates regularity. The kink in the upper curve (for smallest τ) in the entropy plot marks the point of time at which the loop in the Roessler attractor of the regular dynamics is closed. The fact that this time of increasing entropy ($t \approx 22$) is larger than the time for one circulation ($t \approx 2.2$) in the Roessler attractor shows a period of approximately 10 for the chosen parameter values. In the right panel of fig. 3.10 more than 10 lines are visible. This should be traced back to the circumstance that secants are plotted.

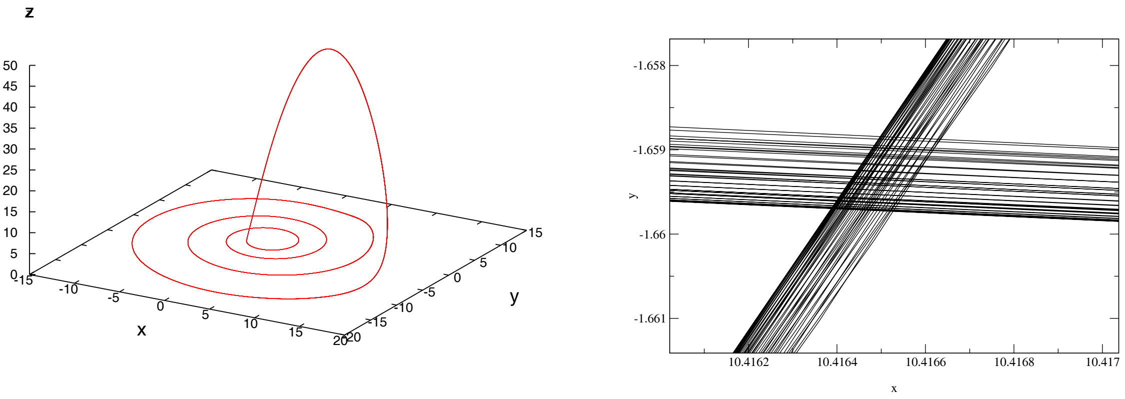


Figure 3.10: Left panel: Regular Roessler dynamics. Right panel: Zoom of a projection into the xy-plane, which shows that the period within the attractor is larger than one. An explanation of the right panel with a long transient can be excluded.

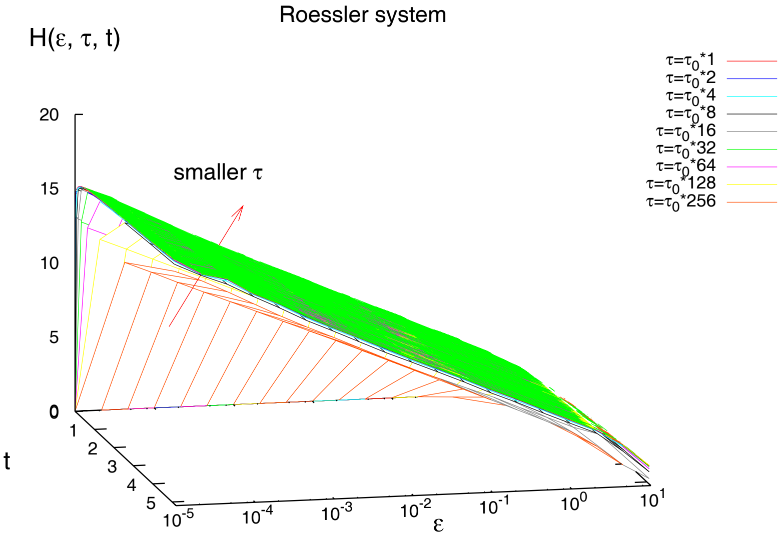


Figure 3.11: Joint entropies of the regular Roessler system as a function of the time t and the resolution ϵ for various $\tau = \tau_0 \cdot 2^n$; $n = 0, ..., 8$ $(\tau_0 = 0.01)$ for the x-coordinate of the Roessler dynamics; $3 \cdot 10^6$ data points.

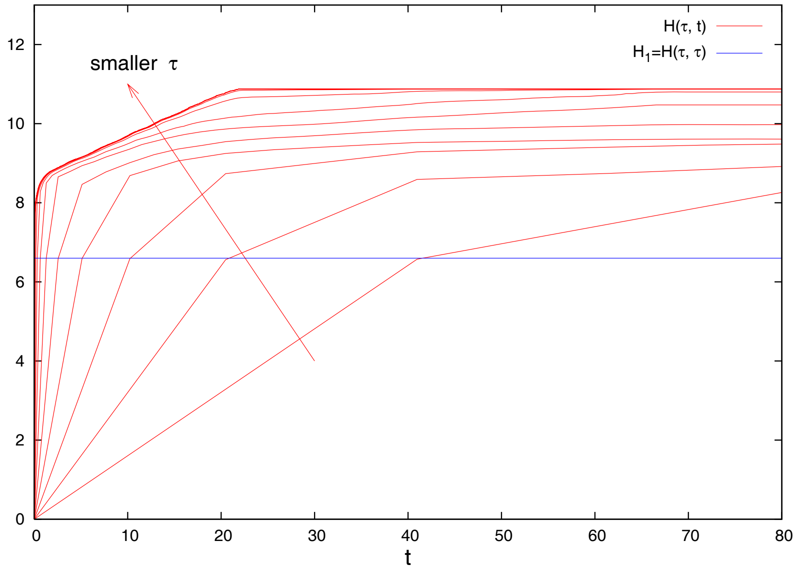


Figure 3.12: Joint entropy as a function of time t for the regular Roessler dynamics at fixed resolution $\epsilon = 10^{-2}$.

3.8.4.3 Lorenz system (deterministic chaotic dynamics)

The implemented equations for the Lorenz dynamics are

$$\begin{aligned} x_{n+1} &= x_n + \sigma(-x_n + y_n)\Delta t \\ y_{n+1} &= y_n + (-x_n z_n + r x_n - y_n)\Delta t \\ z_{n+1} &= z_n + (x_n y_n - b z_n)\Delta t \end{aligned} \tag{3.167}$$

with the usual parameters

$$r = 28.0 \quad , \quad \sigma = 10.0 \quad , \quad b = \frac{8}{3} \; . \tag{3.168}$$

The fractal dimension of the Lorenz attractor is about $D^{(2)} \approx 2.06$ and the largest Lyapunov exponent is $\lambda \approx 0.91$.

In fig. 3.13 [10] the joint entropy of the Lorenz system is shown as a function of the resolution, time step length and time. As for the Roessler system it is possible to see the convergence of the joint entropy for decreasing time step length τ and a logarithmic dependence of the joint entropy on the resolution ϵ with a slope giving the expected dimension in the limit of infinitesimal small ϵ.

---

[10] If the conditional entropies in the panels of fig. 3.2 would be divided by the time step length of the underlying time series, then the resulting quantities would in principle correspond to the slopes of H in time-direction of the segments for non-zero τ as a function of the resolution ϵ in fig. 3.13 with the parameter t (given as m).

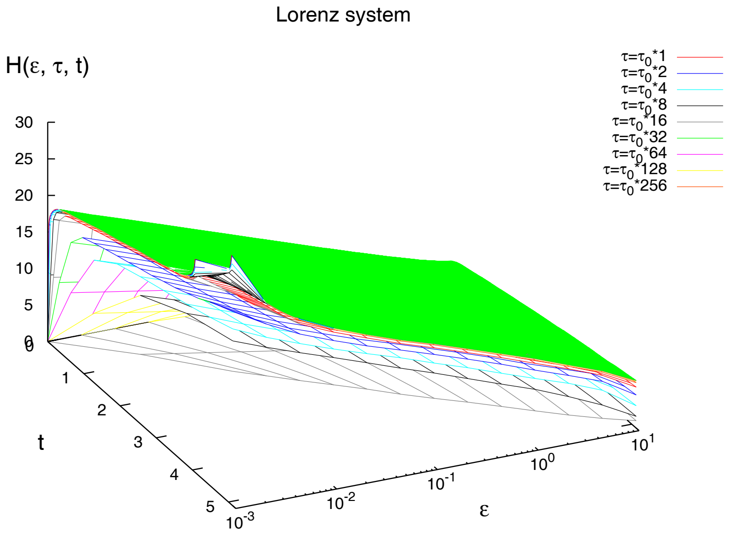


Figure 3.13: Joint entropy of the Lorenz system as a function of the time t and the resolution ϵ for various $\tau = \tau_0 \cdot 2^n$; $n = 0, ..., 8$ $(\tau_0 = 0.01)$ for the x-coordinate of the Lorenz dynamics; $3 \cdot 10^6$ data points.

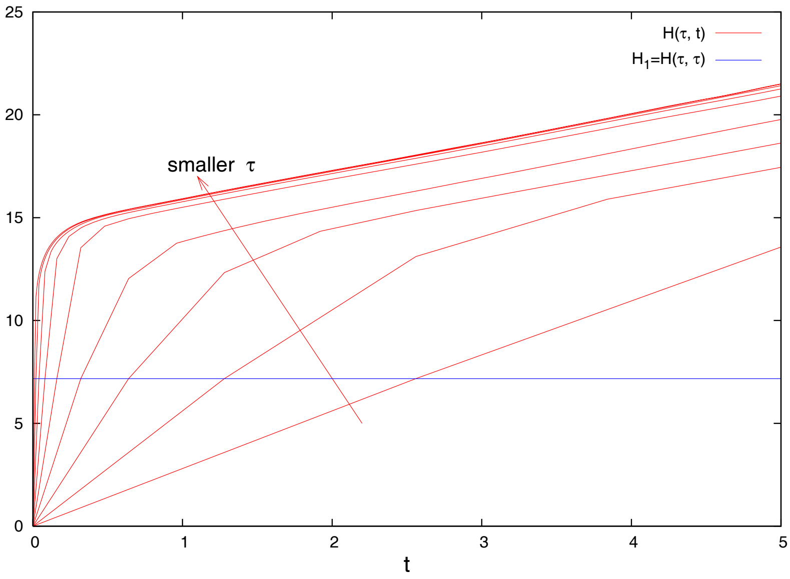


Figure 3.14: Joint entropy of the Lorenz dynamics for resolution $\epsilon = 10^{-2}$ (Slice of fig. 3.13); $3 \cdot 10^6$ data points; $\tau_0 = 0.01$.

In fig. 3.14 a slice of fig. 3.13 for fixed resolution is shown. An asymptotically linear behaviour in time can be found. The fact, that the slope is steeper than that of the chaotic Roessler attractor of fig. 3.6 is in accordance with the larger number of nonlinear terms and the known largest Lyapunov exponents. The slightly larger slope compared to the known KS entropy of the Lorenz system can be explained with estimation at rather large ϵ.

The appearance of finite sample fluctuations in entropy estimation in fig. 3.13 for small ϵ, small τ and large t under the same estimation conditions sooner in the Lorenz-case than in the Roessler-case is in accordance with the enhanced uncertainty H of the Lorenz dynamics from the higher largest Lyapunov exponent.

Furthermore fig. 3.14 indicates that the joint entropies for different finite τ-values do not coincide for asymptotically large times t even though this cannot finally be proven in a plot for finite t. The slopes seem to reach the same value in all cases and hence the estimation of the KS entropy rate should be rather independent of the choice of τ, but the behaviour of $H(\tau = 0, t)$ for finite t is better resolved for smaller τ, and this in general is rather important with respect to prediction.

3.8.4.4 1-dim. Ornstein-Uhlenbeck process (linear stochastic dynamics)

In sec. 2.2.2.3 the Ornstein-Uhlenbeck process was introduced and with eq. (2.67) it can be numerically implemented.

In fig. 3.15 the numerical analysis indicates the non-existence of the limit $\lim_{\tau\to 0} H(\epsilon, \tau, t)$ for sufficiently small ϵ for finite t except at $t = 0$ for time-continuous stochastic dynamics. An explanation of this behaviour can be found in the fact that the nowhere differentiable, hence fractal curve of a time-continuous stochastic process is infinitely long on (whatever small) finite time intervals. It is natural to expect this to have infinite uncertainty.

Whereas especially in the case of Roessler dynamics in the deterministic case the value of the KS entropy rate was approximately seen as a slope at sufficiently small finite ϵ also for finite τ, in the stochastic case the slope of H with respect to t asymptotically in t for sufficiently small finite ϵ is explicitly a function of τ. Furthermore fig. 3.15 supports $h_{KS} = \infty$ for the stochastic case.

From fig. 3.16 it is obtained that in the transition regime of large ϵ the behaviour of non-existing $\lim_{\tau\to 0} H(\tau, t)$ is suppressed. The plots do not allow for the decision of the question if the transition from finite to infinite occurs at the resolution of the size of the system or at a smaller value for ϵ. Decreasing the variance of the input noise of the stochastic dynamics to all smaller finite values results in qualitatively the same behaviour at smaller ϵ (not shown).

According to eq. (3.121) fig. 3.16 yields finite dimensions for finite τ increasing in t for stochastic dynamics, even though the limit $\epsilon \to 0$ can of course not be seen. This is in accordance with the stochastic statement given in sec. 3.6.1.4. In the limit of $\tau \to 0$, what is effectively like the limit of $m \to \infty$, fig. 3.16 suggests an infinite dimension for stochastic dynamics.

In fig. 3.17 the entropy rate as a function of time for varying time step length is shown and as well as for the deterministic case in fig. 3.8 also in the stochastic case the even more expected validity of $h(0, 0) = \infty$ can finally be verified.

The results of this section are qualitatively robust against changes of the parameter α of the Ornstein-Uhlenbeck process and hence qualitatively robust against changes of the autocorrelation of the process.

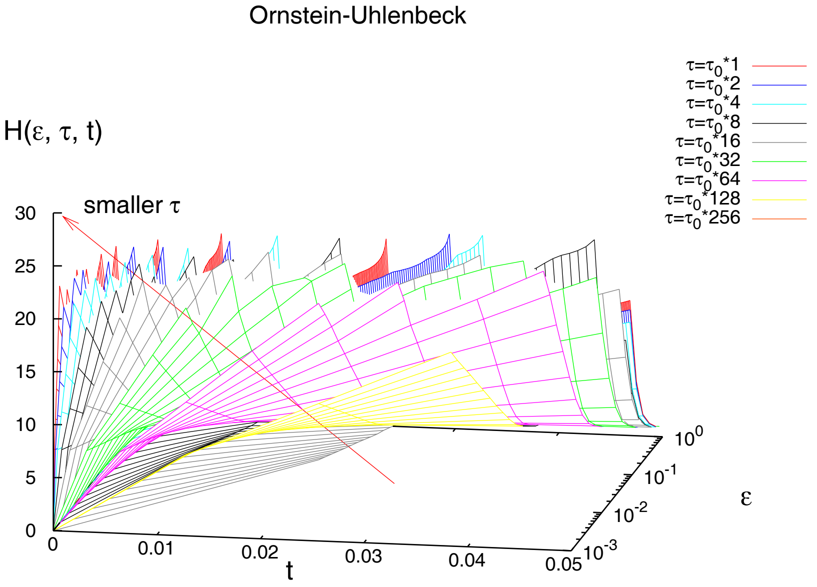


Figure 3.15: Joint entropy of the Ornstein-Uhlenbeck process. Parameters: $\alpha = 0.025$; $\tau_0 = 10^{-4}$; 150000 data points.

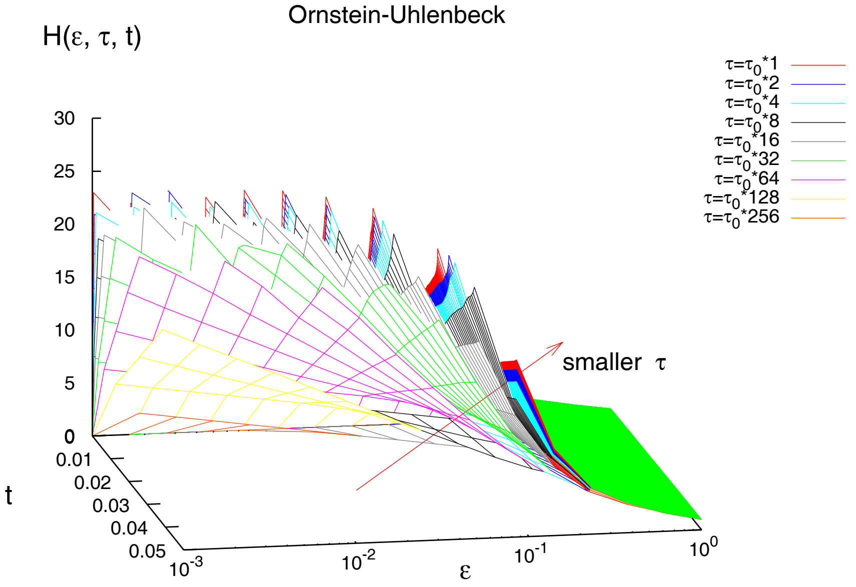


Figure 3.16: The same plot as in fig. 3.15 from another point of view.

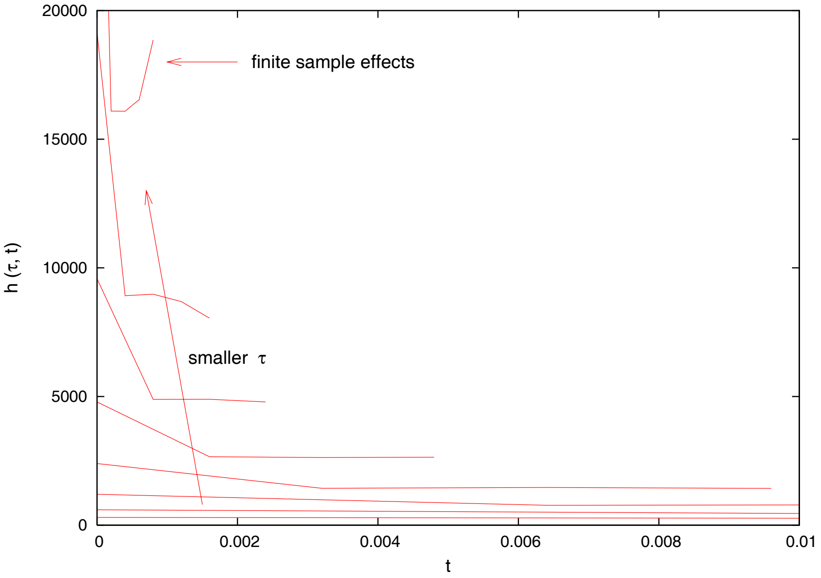


Figure 3.17: Increasing entropy rate as a function of t for decreasing τ for the Ornstein-Uhlenbeck process.

3.8.4.5 Results of numerical calculations

In the deterministic case $H(\epsilon, \tau = 0, t)$ is finite for all finite ϵ and t. $h(\epsilon, \tau = 0, t)$ becomes constant for large t and sufficiently small ϵ. In the stochastic case $H(\epsilon, \tau = 0, t)$ is infinite for all $t > 0$ if ϵ is below some threshold, but finite for ϵ above the threshold, i.e., effectively like in the deterministic case. This result is qualitatively independent of the value of the width of the input noise as long as the width is not exactly zero, corresponding to determinism and introducing a qualitative change, which is discussed more general in appendix A. This behaviour of entropies in the limit $\tau \to 0$ seems to offer a new possibility for distinction of chaos and noise.

In the limit case $\epsilon \to 0$ the partial derivative of the joint entropy H with respect to $\ln \epsilon$ yields (minus) the dimension D and if furthermore the limit $t \to \infty$ is carried out, the partial derivative of H with respect to t yields the KS entropy rate. In the deterministic case known values were verified in examples. In the stochastic case it is seen that both partial derivatives become infinite in the limit cases with respect to ϵ and t given $\tau \to 0$. The total differential of the joint entropy H can be written as

$$dH = \frac{\partial H}{\partial \ln \epsilon} d \ln \epsilon + \frac{\partial H}{\partial t} dt \, , \tag{3.169}$$

which for sufficiently small ϵ and sufficiently large t becomes[11]

$$dH = -D \, d \ln \epsilon + h_{KS} \, dt \, . \tag{3.170}$$

---

[11] Here it is possible to see that directional derivatives of H in general mix the properties of dimension and entropy rate.

In the deterministic case the gain in higher sampling rates saturates, such that a threshold sampling rate can be postulated, above which almost nothing more can be learned, i.e., the continuous limit can be well approximated by discrete sampling with sufficiently high sampling rate in this case. A criterion for an optimal sampling rate could be formulated. In the stochastic case the explicit τ-dependence of as well the entropy rates as the dimensions can be seen.

3.8.5 Conclusion and outlook

The unusual and seemingly pedantism concerning the notation of entropies in this work led for the time-continuous case of information theory to quantities which in the limit case of $\tau \to 0$ can be understood as successive time-derivatives. This is a situation reminding a bit of velocities and acceleration in elementary mechanics. However, with the explicit τ-dependences there are some further subtleties weakening this analogy. In analysis usually integrals are defined as limits of sums of areas with decreasing width. The central difference in comparison with usual analysis and the integrals used here is that the shape and size of the area over which to integrate changes with the step width what in analysis usually is not the case. In the deterministic case this change converges, such that usual analysis is the limit.

This section intends to give theoretical insights into the structures behind the entropic quantities. The natural starting point should be the time-continuous entropies, because the limit $\tau \to 0$ contains the proper informational characteristics of the dynamics. Then it is realized that for a numerical implementation a time discretization is unavoidable and then it is even more realized that for sufficiently small time step length under suitable conditions the approximation is quite well. But nevertheless the basis is the time-continuous case as e.g. in the theory of differential equations, where the mathematical general statements also do not talk about discretization. Until now the author missed this point of view concerning entropies in the presented works.

It was concentrated essentially on the function $H^{(q)}(\epsilon, \tau, t)$ with numerical access to the case $q = 2$. Knowing how to read such a plot *all* information concerning entropy rates (also the KS entropy rate) and dimensions of the dynamics can be extracted. In case of the Roessler dynamics $h(\epsilon, \tau, t)$ as derivative (and discrete pendants) of H by t was shown explicitly. It would also be justified to do the same thing with respect to the dimension as $D(\epsilon, \tau, t)$. A precondition would be the acceptance of eq. (3.123) or the construction of a suitable resolution-dependent alternative from the derivative by $\ln \epsilon$ and the explicit introduction of the existing τ-dependence in such formulas.

It has to be stressed again that the essential focus of this analysis is not the limit case of $t \to \infty$ with the determination of dynamical invariants, which was discussed carefully in [31]. Concerning those invariants the time-continuous case only gives a new representation of known results and can receive crosschecks from the limit cases with respect to ϵ and t. Instead the case of finite time t is of mayor interest here, because this is the part one deals with, when questions concerning prediction are addressed. Observables with steeper slope of H for rather small t are preferable with respect to prediction.

Furthermore this section tries to contribute to a symmetrization of the treatment of entropies concerning phase space, where continuity is already realized in formulas, and time, where continuity is still undiscussed.

The thoughts presented here are only the first step in a line of generalizations. The crucial step should finally be the connection of a path integral formulation with the formulas given in sec. 3.8.3. Furthermore the whole procedure would have to be carried to the perforated case (introduced in chap. 4) for accessing interval structures.

Chapter 4

Entropies on perforated structures

4.1 Motivation and aim

In the time series analytical framework of information theory of the former chapter only entropies were discussed, in which all successive time steps of a section of a time series were involved and none of the time steps was omitted. Concerning conditioning this is in accordance with the usual treatment of Markov processes of higher order.

From the idea, that it could be necessary to use as input for the prediction of wind speed not only the last measurements but different scales of time, supported also by the study of multiscale cascade processes of turbulence and by the usual usage of different grid sizes at different positions in simulations of computational fluid dynamics, it seems useful to introduce Markov processes of higher order, in which *not* the whole range of the Markov model given by the order m influences the conditioning, but only parts of it. This motivates the following definition:
A Markov process of order m shall be called *perforated* [1], if

$$\exists \quad \mathbf{J} := \{j_1, ..., j_{|\mathbf{J}|} : j_1 < ... < j_{|\mathbf{J}|} = m\} \subsetneq \{1, ..., m\} \tag{4.1}$$

such that the associated conditional probabilities fulfil

$$p_{i_k|i_{k-m},...,i_{k-1}} = p_{i_k|i_{k-j_{|\mathbf{J}|}},...,i_{k-j_1}} \quad \forall i_k \ . \tag{4.2}$$

Here i_k denotes the possible outcomes of the random variable at time k.

Since a Markov process of higher order can only be perforated if it is of order larger than one, the supplement 'of higher order' is suppressed in the context of 'perforated Markov processes'. The order of a perforated Markov process is given by max($\mathbf{J}$), but for a unique characterization of a perforated Markov process the specification of the whole structure of conditioning is necessary.

The conditional entropies corresponding to perforated Markov processes, which in the Shannon-case are directly calculable from the conditional probabilities, are objects of interest in this chapter.

---

[1] It has to be remarked that the notion 'perforated' appearing very often in this work is an invented notion in the context of Markov processes of higher order, which is not taken from the literature. If this does not coincide with the taste of the reader, it could consequently be replaced under acceptance of rearrangement within sentences by 'with omissions'.

Furthermore not only the non-equidistant omittance of time steps in conditions typically in the past, but instead general *perforated structures* [2] including omissions also in the part of assessed time steps typically in the future, are considered with respect to entropies. Hence the idea of perforated entropies also applies to a description of variable future prediction time, arbitrary joint prediction and different sampling rates.

Because of this, the aim of this chapter is the construction of a new unified notation of Renyi information theory on the basis of perforated structures, being appropriate for all above stated applications and the respective discussion of those cases in terms of the new entropic notation.

4.2 General definitions of entropies and derived quantities on perforated structures

Starting from the usual time series analytical joint entropy

$$H_m^{(q)}(\epsilon) = \frac{1}{1-q} \ln \sum_{i_1,\ldots,i_m} p_{i_1,\ldots,i_m}{}^q \quad , \tag{4.3}$$

it is possible to generalize this formula for the joint entropy by the following definition to

$$H_{\mathbf{K};\alpha}^{(q)}(\epsilon) := \frac{1}{1-q} \ln \sum_{i_k : k \in \mathbf{K}} (p_{i_{k_1},\ldots,i_{k_{|\mathbf{K}|}}}(\alpha))^q \quad , \qquad \alpha \in \mathbb{Q} \; . \tag{4.4}$$

Instead of indices corresponding to successive time steps, this generalization allows for entropies corresponding to a possibly non-successive set of time indices $\mathbf{K}$ given by the underlying perforated structure. Furthermore the index α gives a factor which is multiplied on the elementary time interval given by the time series to determine a new time interval, to which the calculation of the entropy refers. This parameter is especially introduced with regard to sampling changes. The Renyi order q of entropies and the resolution dependence ϵ, if not explicitly needed, are omitted in the following formulas of this chapter, in order to allow for a larger attention to variability concerning time dependence.

Supposed besides $\mathbf{K}$ another set $\mathbf{J}$ of indices with $\mathbf{K} \cap \mathbf{J} = \emptyset$, then the conditional entropy is defined as[3]

$$H_{[\mathbf{K}|\mathbf{J}];\alpha} := H_{\mathbf{K}\cup\mathbf{J};\alpha} - H_{\mathbf{J};\alpha} \; . \tag{4.5}$$

If special elements are given for $\mathbf{K}$ and $\mathbf{J}$ this quantity looks like

$$H_{[\{k_1,k_2,\ldots\}|\{j_1,j_2,\ldots.\}];\alpha} \; , \tag{4.6}$$

where the k_i and j_i are numbers denoting the time steps. Of course

$$H_{[\mathbf{K}|\emptyset];\alpha} = H_{\mathbf{K};\alpha} \quad \forall \mathbf{K}, \alpha \tag{4.7}$$

---

[2] A visualisation of perforated structures will be given in fig. 4.1 in the framework of estimation of perforated entropies by generalized delay vectors.

[3] The introduced squared bracket stresses the fact that the changed sampling rate from α does not only concern the conditioning set $\mathbf{J}$, but also the set $\mathbf{K}$, for which the uncertainty is assessed.

and the identity

$$H_{[\emptyset|\mathbf{J}];\alpha} \equiv 0 \quad \forall \mathbf{J}, \alpha \tag{4.8}$$

hold. If $\alpha = 1$, i.e., a change of the elementary time interval is not discussed, the index α is omitted:

$$H_{\mathbf{K}|\mathbf{J}} := H_{[\mathbf{K}|\mathbf{J}];\alpha=1} \ . \tag{4.9}$$

If $\min(\mathbf{K}) > \max(\mathbf{J})$ then this conditional entropy is suitable for prediction. If $\min(\mathbf{J}) > \max(\mathbf{K})$ then a postdiction situation is at hand. Otherwise the conditional entropy yields a mixed uncertainty assessment. The mutual information (of joint random variables) is defined as

$$I_{\mathbf{K};\mathbf{J}} := H_{\mathbf{K}} + H_{\mathbf{J}} - H_{\mathbf{K}\cup\mathbf{J}} = H_{\mathbf{K}} - H_{\mathbf{K}|\mathbf{J}} \ . \tag{4.10}$$

Supposed the set $\mathbf{K}$ contains exactly one element $k > 0$ and $\max(\mathbf{J}) \leq 0$, then the conditional entropy reads $H_{\{k\}|\mathbf{J}}$ and the generalized redundancy or mutual information reads $R_{\{k\};\mathbf{J}} \equiv I_{\{k\};\mathbf{J}} = H_{\{k\}} - H_{\{k\}|\mathbf{J}}$, which will be important quantities concerning the search for the optimal perforated Markov model in chap. 6. Given pairwise disjoint $\mathbf{K}_i$, corresponding to fig. 3.1 conditional mutual information is defined by

$$I_{\mathbf{K}_1;\mathbf{K}_2|\mathbf{K}_3} := H_{\mathbf{K}_1|\mathbf{K}_3} - H_{\mathbf{K}_1|\mathbf{K}_2\cup\mathbf{K}_3} \ , \tag{4.11}$$

and triple mutual information as

$$I_{\mathbf{K}_1;\mathbf{K}_2;\mathbf{K}_3} := H_{\mathbf{K}_1} + H_{\mathbf{K}_2} + H_{\mathbf{K}_3} - H_{\mathbf{K}_1\cup\mathbf{K}_2} - H_{\mathbf{K}_2\cup\mathbf{K}_3} - H_{\mathbf{K}_3\cup\mathbf{K}_1} + H_{\mathbf{K}_1\cup\mathbf{K}_2\cup\mathbf{K}_3} \ . \tag{4.12}$$

The latter quantity and generalizations to multiple mutual information are needed in sec. 4.4.4 for calculating redistribution of information under omission of time components.

For all entropic quantities indexed by sets, translation invariance is usable to eliminate one parameter. On the other hand, this elimination of one component by using this translation invariance would cause problems in the ability of distinction of for example $H_{\{0\}|\{-1,1\}}$ and $H_{\{1\}|\{-1,0\}}$, which are different conditional entropies. The need for a distinction of such quantities would enforce the introduction of at least one new parameter, what compensates the gain in notational reduction by having used the translation invariance. This is the reason, why this new notational suggestion for entropies in time series analysis remains with unused translation invariance.

In the case of nonperforation (referring to a fixed sampling) it is not necessary to give explicitly a set of indices, but instead a number is enough for a unique description of the entropies. If the cardinal number of $\mathbf{K}$ is $|\mathbf{K}| = m$ and $\mathbf{K}$ denotes nonperforated, i.e., successive time-indices, then the joint entropy is redefined as

$$H_{m;\alpha} := H_{\mathbf{K};\alpha} \ . \tag{4.13}$$

From eq. (4.8) it follows

$$H_{0;\alpha} = 0 \ . \tag{4.14}$$

If furthermore $\mathbf{J}$ is nonperforated with $|\mathbf{J}| = n$ and $\min(\mathbf{K}) = \max(\mathbf{J}) + 1$, then

$$H_{[m|n];\alpha} := H_{m+n;\alpha} - H_{n;\alpha} = H_{\mathbf{K}\cup\mathbf{J};\alpha} - H_{\mathbf{J};\alpha} = H_{[\mathbf{K}|\mathbf{J}];\alpha} \ . \tag{4.15}$$

Conditional entropy rates for variable sampling are defined as

$$h_{n;\alpha} := \frac{H_{[1|n];\alpha}}{\alpha \cdot \tau} \ . \tag{4.16}$$

Here τ is the elementary time step of the time series. The mutual information is defined as

$$I_{[m;n];\alpha} := H_{m;\alpha} - H_{[m|n];\alpha} \ . \tag{4.17}$$

The definition of redundancy reads

$$R_{n;\alpha} := I_{[1;n];\alpha} = H_{1;\alpha} - H_{[1|n];\alpha} \ , \tag{4.18}$$

and the redundancy rate

$$r_{n;\alpha} := \frac{R_{n;\alpha}}{\alpha \cdot \tau} \ . \tag{4.19}$$

It is redefined

$$H_m := H_{m;\alpha=1} \ , \tag{4.20}$$

and it holds

$$H_{1;\alpha} = H_1 = H_{\{k\}} \quad \forall \alpha, k \ . \tag{4.21}$$

The conditional entropy for original sampling as already used in sec. 3.3.4 is redefined by

$$H_{m|n} := H_{m+n} - H_n = H_{[m|n];\alpha=1} \ , \tag{4.22}$$

and the conditional entropy rate is defined as

$$h_n := \frac{H_{1|n}}{\tau} = h_{n;1} \ . \tag{4.23}$$

In the nonperforated case in original sampling the mutual information is defined as

$$I_{m;n} := H_m - H_{m|n} = I_{[m;n];\alpha=1} \ , \tag{4.24}$$

from which the redundancy can be derived by

$$R_n := I_{1;n} = H_1 - H_{1|n} \ . \tag{4.25}$$

Finally a definition for the redundancy rate reads

$$r_n := \frac{R_n}{\tau} \ . \tag{4.26}$$

4.3 Estimation of entropies in perforated case

In sec. 3.6 estimation of entropies was performed on the basis of measurement vectors. Delay vectors are a generalization of this with perforation such that equidistant components are kept. A further generalization with arbitrary non-equidistant perforation in the sense of general perforated structures leads to what will be called generalized delay vector. The concept of a delay vector and its generalization are illustrated in fig. 4.1.

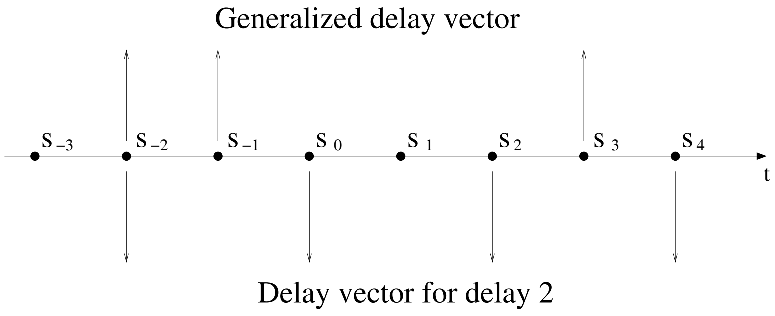


Figure 4.1: Delay vector and its generalization

Hence in the perforated case the correlation sum is calculated with generalized delay vectors $P_{\mathbf{J}}\mathbf{x}_i$ according to the perforation structure by

$$\hat{C}^{(2)}_{\mathbf{J}}(\epsilon, N) = \frac{2}{N(N-1)} \sum_{i=1,j=1:j<i}^{N} \Theta(\epsilon - |P_{\mathbf{J}}\mathbf{x}_i - P_{\mathbf{J}}\mathbf{x}_j|) \ . \tag{4.27}$$

$P_{\mathbf{J}}$ is the projection operator onto the perforation structure given by the set $\mathbf{J}$. The perforated joint Renyi entropies are estimated via correlation sum according to

$$\hat{H}^{(2)}_{\mathbf{J}}(\epsilon, N) = -\ln \hat{C}^{(2)}_{\mathbf{J}}(\epsilon, N) \ . \tag{4.28}$$

Further derived perforated quantities are based on those joint entropies. Since practically the estimations are performed for Renyi order $q = 2$ maximal generality in the given formulas is renounced. The error in estimation of perforated entropies is obtained as a straightforward generalization of eq. (3.105) by

$$\Delta H^{(2)}_{\mathbf{J}}(\epsilon, N) = \frac{\Delta C^{(2)}_{\mathbf{J}}(\epsilon, N)}{\hat{C}^{(2)}_{\mathbf{J}}(\epsilon, N)} \ . \tag{4.29}$$

From this equation the error of the redundancy $\Delta R_m(\epsilon, N)$ in eq. (3.108) can be generalized to the corresponding error in the perforated case

$$\Delta R_{\{f\};\mathbf{J}}(\epsilon, N) = \sqrt{\Delta H_1(\epsilon, N)^2 + \Delta H_{\{f\}\cup\mathbf{J}}(\epsilon, N)^2 + \Delta H_{\mathbf{J}}(\epsilon, N)^2} \ , \tag{4.30}$$

which will appear in the central criterion in eq. (6.3).

4.4 Special cases of generalized prediction and modelling situations

4.4.1 Variable future prediction times

4.4.1.1 Conditional entropies and redundancy

In the following it is adopted the convention that the set-element 0 corresponds to the presence. The past obtains negative and the future obtains positive indices, where f is the number of time steps to the future prediction time. The mean uncertainty in one single time step is[4]

$$H_{\{f\}} \equiv H_{\{0\}} = H_1 = h_0 \cdot \tau \; . \tag{4.31}$$

For the following conditional entropy the mean joint uncertainty $H_{\{0,f\}}$ of two measurements, which are executed with the time distance of f steps from one another, is needed:

$$H_{\{f\}|\{0\}} = H_{\{0,f\}} - H_{\{0\}} \quad , \qquad H_{\{f=1\}|\{0\}} = H_{1|1} = h_1 \cdot \tau \; . \tag{4.32}$$

$H_{\{f\}|\{0\}}$ is the mean conditional uncertainty of the underlying random variable at some time step, if the outcome of the random variable is known f time steps before. Further conditioning leads to

$$H_{\{f\}|\{-1,0\}} = H_{\{-1,0,f\}} - H_{\{-1,0\}} \quad , \qquad H_{\{f=1\}|\{-1,0\}} = H_{1|2} = h_2 \cdot \tau \; , \tag{4.33}$$

and

$$H_{\{f\}|\{-(m-1),\ldots,0\}} = H_{\{-(m-1),\ldots,0,f\}} - H_{\{-(m-1),\ldots,0\}} \; , \; H_{\{f=1\}|\{-(m-1),\ldots,0\}} = H_{1|m} = h_m \cdot \tau \; . \tag{4.34}$$

In fig. 4.2 the usual ϵ-dependence is shown for the new conditional entropies for the simple AR(1) process for different future prediction times. One can see that of course the unconditioned entropy is the same for all future prediction times and that the effect of reducing uncertainty by conditioning tends to disappear for increasing future prediction time. Additional conditioning ($m \geq 2$) does not further reduce the uncertainty as expected for the AR(1).

The generalized redundancy in the case of variable future prediction time reads

$$R_{\{f\};\{-(m-1),\ldots,0\}} \equiv I_{\{f\};\{-(m-1),\ldots,0\}} = H_1 - H_{\{f\}|\{-(m-1),\ldots,0\}} \; . \tag{4.35}$$

In fig. 4.3 the information, which is available for prediction, i.e., the redundancy, is shown as a function of the future prediction time steps for variable parameter values of the AR(1) process. The result is a parameter-dependent exponential decay of the available information for prediction as a function of the future prediction time, where the decay is stronger for lower correlation. This is essentially the same behaviour as given by the autocorrelation function of the AR(1) process in eq. (2.66).

4.4.1.2 Decrease of mutual information when increasing the future prediction time

If the presence is known and prediction of only the first step into the future is intended, then the corresponding uncertainty is given by the following conditional entropy as

$$H_{\{1\}|\{0\}} = H_2 - H_1 = H_1 - I_{\{1\};\{0\}} = H_1 - I_{1;1} = H_1 - R_1 \; . \tag{4.36}$$

If otherwise the presence is known, but prediction of only the second step in the future is intended, then the uncertainty is

$$H_{\{2\}|\{0\}} = H_{\{0,2\}} - H_1 = H_1 - I_{\{2\};\{0\}} \; . \tag{4.37}$$

---

[4] In order to avoid confusion it is stressed here again explicitly that $\{f\}$ is not meant to be the set of all possible values a variable f can take, but instead the set consisting of one element f

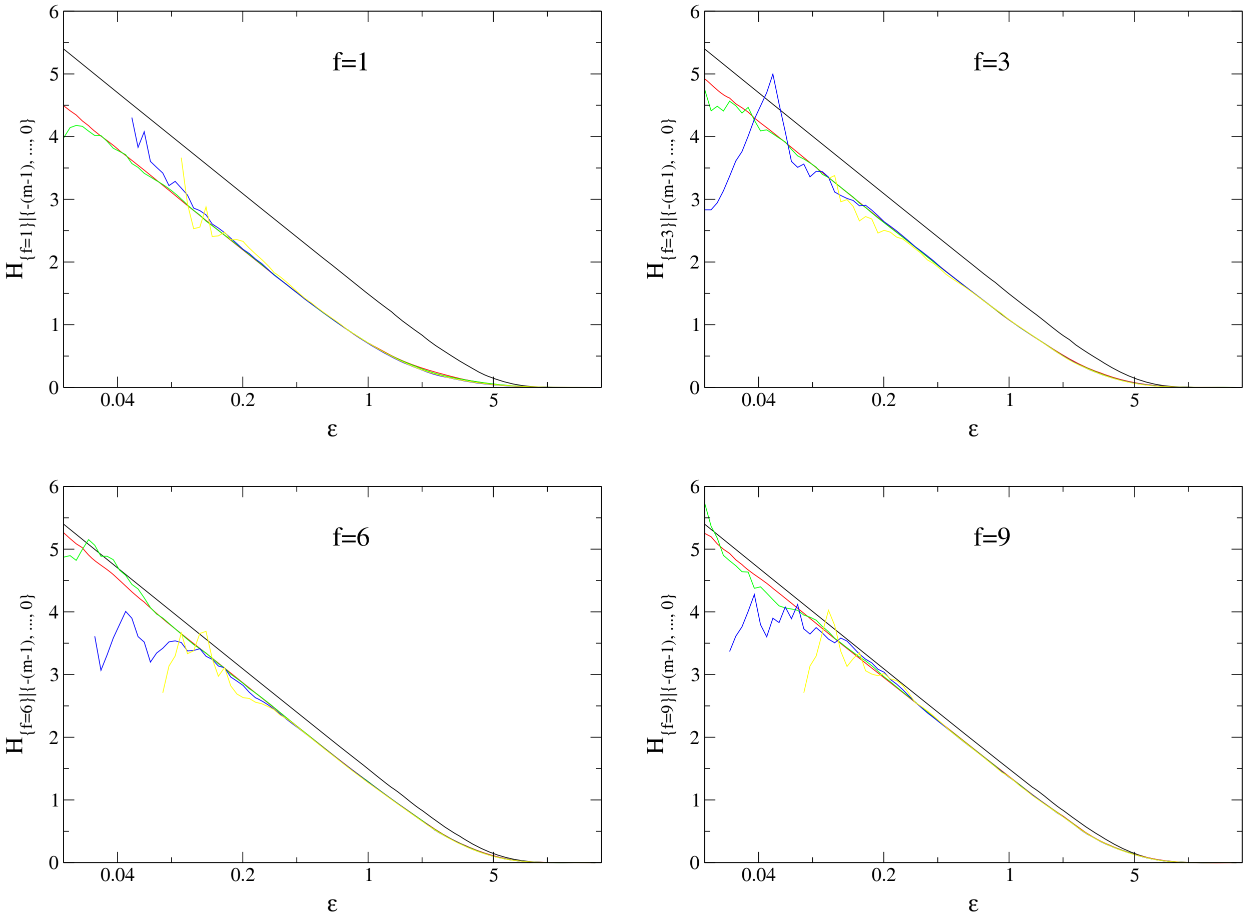


Figure 4.2: Conditional entropies of the AR(1) process with parameter value $a = 0.92$ and standard Gaussian noise for four different future time steps f. In each case the conditional entropies $H_{\{f\}|\{-(m-1),\ldots,0\}}$ are identical for all $m \geq 1$ apart from finite sample effects, as it is expected for an AR(1) process.

The information lost for prediction, if passing over from prediction of the first time step ahead to prediction of the second time step ahead given only the presence, is the difference of above quantities, representable as the difference of joint entropies or the difference of mutual informations:

$$\begin{aligned} H_{\{2\}|\{0\}} - H_{\{1\}|\{0\}} &= H_{\{0,2\}} - H_2 \\ &= I_{\{1\};\{0\}} - I_{\{2\};\{0\}} \ . \end{aligned} \tag{4.38}$$

The result here is that analytical calculations of valuable simple quantities become possible because of the renewed entropic notation. The discussion of variable future prediction times will be a special case of sec. 4.4.3.

4.4.2 Arbitrary joint prediction

Here shortly the generalization of sec. 3.3.4 to the case of non-block future is mentioned explicitly. Relevant quantities are for various $f_k > 0$ of the type $H_{\{f_1,\ldots,f_i\}|\{-(m-1),\ldots,0\}}$ for

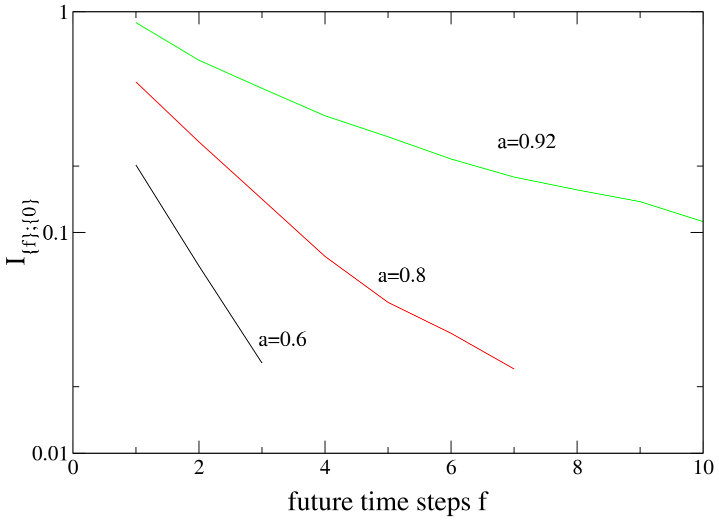


Figure 4.3: Information available for prediction as a function of variable future time steps for different parameter values of the AR(1) process for sufficiently small ϵ, where the informational structure is fully developed without any disturbance by finite size effects.

nonperforated **J** and perforated **K** in eq. (4.9). No matter, if $f_1 < f_2$ or $f_1 > f_2$, it holds

$$H_{\{f_1,f_2\}|\{-(m-1),\ldots,0\}} = H_{\{f_1\}|\{-(m-1),\ldots,0\}} + H_{\{f_2\}|\{-(m-1),\ldots,0,f_1\}} \ . \tag{4.39}$$

As in the nonperforated case the total conditional uncertainty is again decomposable into a sum of partial conditional uncertainties. Such decompositions are of importance to understand finally a perforated time-continuous case. The total uncertainty is again decomposable into an uncertainty-reducing mutual information term and a remaining uncertainty after conditioning:

$$H_{\{f_1,f_2\}} = I_{\{f_1,f_2\};\{-(m-1),\ldots,0\}} + H_{\{f_1,f_2\}|\{-(m-1),\ldots,0\}} \ . \tag{4.40}$$

4.4.3 Perforation in conditioning of entropies

4.4.3.1 Motivation

The uncertainty of the outcome of a random variable X (possibly corresponding to a joint distribution) can in principle be reduced by conditioning it on the outcome of other random variables Y and Z. This reduction of uncertainty of X by the outcome of Y is trivially not availabe if the random variable Y or its outcome is skipped by perforation. Now the question arises if this uncertainty reduction by Y is lost under perforation or if this perforation changes the uncertainty reduction of X by the outcome of Z? Another question is, how to quantitatively describe a possible change of information? Especially in prediction situations such a kind of question arises and will be of crucial importance with respect to chap. 6.

4.4.3.2 Simplest example

In the framework of time series it is now considered the single prediction of the immediate future time step and the change of the distribution of information for prediction under the operation of perforation in the first time steps of the past. Here again the convention is imposed that the time step for the presence is denoted by zero, future time steps by positive and past time steps by negative numbers.

In the nonperforated case the presence contains the information $I_{1;1} = H_1 - H_{1|1} = 2H_1 - H_2$ about the immediate future time step. The first time step in the past contains the information $I_{\{1\};\{-1\}|\{0\}} = H_{\{1\}|\{0\}} - H_{\{1\}|\{-1,0\}} = H_{1|1} - H_{1|2} = H_2 - H_1 - H_3 + H_2$ about the immediate future time step.

Now it is considered the case that the outcome of the random variable at the presence is not available, a special perforated situation. Instead of yielding the information $H_1 - H_{1|1}$, the presence yields no information for prediction, and the first time step in the past now yields the information $I_{\{1\};\{-1\}} = H_{\{1\}} - H_{\{1\}|\{-1\}} = 2H_1 - H_{\{-1,1\}}$ for prediction of the immediate future time step.

Under the operation of perforation the information for prediction in the time steps of the further past change. In the above example the information for prediction in the first time step of the past about the first future time step changes by the triple mutual information, if the presence is omitted:

$$\begin{aligned}(2H_1 - H_{\{-1,1\}}) &- (2H_2 - H_1 - H_3)\\ &= 3H_1 - 2H_2 - H_{\{-1,1\}} + H_3\\ &\overset{(4.12)}{=} I_{\{-1\};\{0\};\{1\}} \ .\end{aligned} \tag{4.41}$$

It has to be stressed again that the triple mutual information is not necessarily positive, i.e., the information for prediction of the first time step in the future from the first time step in the past can at first sight unintuitively *decrease* by omission of the presence. From the generalized Henon map

$$y_{n+1} = a - y_{n-K+2}^2 - c y_{n-K+1} \tag{4.42}$$

with $K = 3$, which for arbitrary K will be discussed more carefully in sec. 6.4.2.1, the possibility of negativity of the triple mutual information in dynamical and prediction situations was verified numerically.

The non-predictable information in general increases under the operation of perforation seen from

$$H_{1|\infty} = H_{\{1\}|\{-\infty,...,0\}} \le H_{\{1\}|\{-\infty,...,-1\}} \ , \tag{4.43}$$

because increasing the conditioning reduces uncertainty, at least for Renyi order $q = 1$.

Now the questions of this section are approached in a much more general way. The results obtained just before can be found as special cases of the following formulations.

4.4.3.3 Joint perforation of arbitrary non-causal joint conditioning in joint entropies

For convenience again the set of all finite discrete points of time for given interval τ $\{t_i = i \cdot \tau : i \in \mathbb{Z}\}$ is represented by the set of values their index can take and hence by the set $\mathbb{Z}$. Furthermore it is assumed

$$\mathbf{K} := \{k_1, ..., k_{|\mathbf{K}|}\} \subset \mathbb{Z} \ , \tag{4.44}$$

corresponding to a finite arbitrary set of discrete times in a time series for which the joint uncertainty $H_{\mathbf{K}}$ should be assessed. It is defined another finite set corresponding to discrete times

$$\mathbf{L} := \{l_1, ..., l_{|\mathbf{L}|}\} \subset \mathbb{Z} \quad \text{with} \quad \mathbf{L} \cap \mathbf{K} = \emptyset \,, \tag{4.45}$$

at which the outcome of the random variables is firstly assumed to be known, so that conditioning on $\mathbf{L}$ can be performed. The uncertainty $H_{\mathbf{K}}$ belonging to $\mathbf{K}$ is reduced to $H_{\mathbf{K}|\mathbf{L}}$. In other words: Observation of outcomes of the random variables at the discrete times of $\mathbf{L}$ reduces the uncertainty at the times of $\mathbf{K}$ by $H_{\mathbf{K}} - H_{\mathbf{K}|\mathbf{L}}$. Further observations of random variables at the times of

$$\mathbf{J} = \{j_1, j_2, ...\} := \mathbb{Z} \backslash (\mathbf{K} \cup \mathbf{L}) \tag{4.46}$$

do in general further reduce the uncertainty at the time steps of $\mathbf{K}$ by further conditioning. The question now arising is the following: Skipping by perforation the knowledge of outcomes at $\mathbf{L}$, what is the change of uncertainty reduction concerning the joint uncertainty at $\mathbf{K}$ successively by all j_i respectively, if compared to the case when $\mathbf{L}$ was known? Shortly formulated: What happens with the uncertainty reduction $I_{\mathbf{K};\mathbf{L}} = H_{\mathbf{K}} - H_{\mathbf{K}|\mathbf{L}}$ from knowledge of $\mathbf{L}$, when outcomes at $\mathbf{L}$ are not available? Is it completely lost or does it influence the uncertainty reduction by knowledge of $\mathbf{J}$? In general the latter suggestion is correct, seen by the formula

$$\begin{aligned}
&H_{\mathbf{K}} - H_{\mathbf{K}|\mathbf{L}} \qquad (\widehat{=} \text{ Uncertainty reduction of } \mathbf{K} \text{ from } \mathbf{L} \text{ ;unavailable from } \mathbf{L}\text{, if } \mathbf{L} \text{ skipped})\\
&= (H_{\mathbf{K}} - H_{\mathbf{K}|\{j_1\}}) - (H_{\mathbf{K}|\mathbf{L}} - H_{\mathbf{K}|(\mathbf{L}\cup\{j_1\})}) \qquad (\widehat{=} \ * \text{ at } j_1)\\
&\quad +(H_{\mathbf{K}|\{j_1\}} - H_{\mathbf{K}|\{j_1,j_2\}}) - (H_{\mathbf{K}|(\mathbf{L}\cup\{j_1\})} - H_{\mathbf{K}|(\mathbf{L}\cup\{j_1,j_2\})}) \qquad (\widehat{=} \ * \text{ at } j_2)\\
&\quad +(H_{\mathbf{K}|\{j_1,j_2\}} - H_{\mathbf{K}|\{j_1,j_2,j_3\}}) - (H_{\mathbf{K}|(\mathbf{L}\cup\{j_1,j_2\})} - H_{\mathbf{K}|(\mathbf{L}\cup\{j_1,j_2,j_3\})}) \qquad (\widehat{=} \ * \text{ at } j_3)\\
&\quad +...\\
&\quad +H_{\mathbf{K}|[\mathbb{Z}\,\backslash(\mathbf{K}\cup\mathbf{L})]} - H_{\mathbf{K}|[\mathbb{Z}\,\backslash\mathbf{K}]} \quad (\widehat{=} \text{ Increase of non-accessible inf. about } \mathbf{K} \text{ due to perf. of } \mathbf{L})\,,
\end{aligned} \tag{4.47}$$

in which [$*$:= "Change of uncertainty reduction of $\mathbf{K}$ due to perforation of $\mathbf{L}$"] is used.

The validity of the whole formula is proven by finding the so-called 'telescope sums' and by using definition (4.46). The statements on the right side of all lines commenting the part of the formula on the left side, become clear from the fact that inside of the parentheses uncertainty reductions of $\mathbf{K}$ by further conditioning are calculated and the difference of the parentheses gives a comparison of uncertainty reductions under different conditioning situations concerning the set $\mathbf{L}$.

A non-causal representation could for example be useful for a random walk with memory in the dynamical law, where the trajectory is fixed in the past and in the future and one asks for uncertainties in between.

4.4.3.4 Specifications

1. For prediction causality is imposed. Because measurements of the future are intrinsically not available, conditioning on the future does not reduce the uncertainty. This leads to demanding $\min(\mathbf{K}) > \max(\mathbf{L})$ and $\min(\mathbf{K}) > \max(\mathbf{J})$. Imposing the convention to denote the presence with zero, future time steps with positive and past time steps with negative numbers, this leads to $\min(\mathbf{K}) > 0$, $\max(\mathbf{L}) \leq 0$ and $\max(\mathbf{J}) \leq 0$.
2. Succession of conditioning from presence to further past. This is of course the natural way to proceed, and this succession is intrinsically fixed by working only with conditional entropies from $H_{m+1} - H_m$, but in the now more generalized notational framework this is not a priori fixed anymore and has to be declared explicitly.
3. Only the uncertainty in one single time step is assessed: $\mathbf{K} = \{k\}$.
4. Perforation is performed in only one single time step: $\mathbf{L} = \{l\}$.
5. Uncertainty assessment of the immediate future time step ahead: $k = 1$.
6. Perforation at the presence: $l = 0$.

Under those conditions from eq. (4.47) one gets the special case

$$\begin{aligned} H_1 - H_{1|1} = \\ & (H_{\{1\}} - H_{\{1\}|\{-1\}}) - (H_{1|1} - H_{1|2}) \\ & + (H_{\{1\}|\{-1\}} - H_{\{1\}|\{-2,-1\}}) - (H_{1|2} - H_{1|3}) \\ & + (H_{\{1\}|\{-2,-1\}} - H_{\{1\}|\{-3,-2,-1\}}) - (H_{1|3} - H_{1|4}) \\ & + ... \\ & + H_{\{1\}|\{-\infty,...,-1\}} - H_{1|\infty} \ . \end{aligned} \tag{4.48}$$

The information for prediction, which is not available from the presence, because of perforation of the presence, is splitted into a part, which changes the information for prediction in the time steps of the further past and another part, which increases the non-predictable information.

The left side of eq. (4.48) is the amount of information, which has to be redistributed, because of non-observation of the presence. In the first line of the right side it is written the change of information for prediction of the immediate future time step in the first step of the past, and in the n'th line it is written the change of information for prediction of the immediate future time step in the n'th step of the past. The last line gives the change of really lost information for prediction of the immediate future time step under perforation of the presence.

The central message here is that with the new introduced notation this splitting and change of information for prediction is explicitly calculable for various prediction situations in the time series framework. This important result was only accessible after having introduced the generalized notation of information theory.

Examination of formula (4.48) makes clear that perforation somewhere in the past does not change the informational content of components closer to the presence and does generally only influence the informational content of components in the further past. This statement is closely connected to the second item of the listing of specifications above.

In the example of the AR(1) process it holds

$$H_{1|m} = H_{1|1} \qquad \forall m \geq 1 \tag{4.49}$$

and

$$H_{\{1\}|\{-m,\dots,-1\}} = H_{\{1\}|\{-1\}} \qquad \forall m \geq 1 \; . \tag{4.50}$$

Under perforation of the presence in the case of the AR(1) process the redistribution formula collapses into

$$\begin{aligned} H_1 - H_{1|1} = \quad & H_1 - H_{\{1\}|\{-1\}} && (\widehat{=} \text{ Increase of information in the first past step}) \\ & + H_{\{1\}|\{-1\}} - H_{1|1} && (\widehat{=} \text{ Lost information}) \; . \end{aligned} \tag{4.51}$$

4.4.3.5 Joint perforation vs. successive single perforation; Commutativity of successive single perforation

In this section it is intended to give some hints about the behaviour of information for prediction under different ordering of the operation of perforation. There is no rigorous proof given here for the equivalence of different succession of arbitrary perforation. Instead a simple example of perforation near the presence is chosen to illustrate the behaviour of information for prediction under perforation. The following table shows the information for prediction of the immediate future time step (denoted by 1) broken down to the single time steps of the past under different perforation conditions.

State of perforation	Information for predicion at time step 1	from time step
nonperforated	$H_{\{1\}\|\emptyset} - H_{\{1\}\|\{0\}} = I_{1;1}$	0
	$H_{\{1\}\|\{0\}} - H_{\{1\}\|\{-1,0\}} = I_{\{1\};\{-1\}\|\{0\}}$	-1
	$H_{\{1\}\|\{-1,0\}} - H_{\{1\}\|\{-2,-1,0\}} = I_{\{1\};\{-2\}\|\{-1,0\}}$	-2
0,-1 perforated together	0	0
	0	-1
	$H_{\{1\}\|\emptyset} - H_{\{1\}\|\{-2\}} = I_{\{1\};\{-2\}}$	-2
0 perforated	0	0
	$H_{\{1\}\|\emptyset} - H_{\{1\}\|\{-1\}} = I_{\{1\};\{-1\}}$	-1
	$H_{\{1\}\|\{-1\}} - H_{\{1\}\|\{-2,-1\}} = I_{\{1\};\{-2\}\|\{-1\}}$	-2
-1 perforated after 0 perforated	0	0
	0	-1
	$H_{\{1\}\|\emptyset} - H_{\{1\}\|\{-2\}} = I_{\{1\};\{-2\}}$	-2
-1 perforated	$H_{\{1\}\|\emptyset} - H_{\{1\}\|\{0\}} = I_{1;1}$	0
	0	-1
	$H_{\{1\}\|\{0\}} - H_{\{1\}\|\{-2,0\}} = I_{\{1\};\{-2\}\|\{0\}}$	-2
0 perforated after -1 perforated	0	0
	0	-1
	$H_{\{1\}\|\emptyset} - H_{\{1\}\|\{-2\}} = I_{\{1\};\{-2\}}$	-2

One can see the change of information for prediction under perforation and the different ways information takes for finally reaching the same state. The table says that as well joint perforation and successive perforation of corresponding time steps is equivalent as also commutativity in perforation is valid. The central argument behind the table is rather trivial.

4.4.3.6 Delay dynamics

In general delay dynamics is characterized by the fact that only one time step in the far past and frequently also the presence influence the future. An example is the discretized Mackey-Glass dynamics given by

$$x_{n+1} = (1 - b\tau)x_n + \frac{ax_{n-k}}{1 + x_{n-k}^{10}}\tau \ , \tag{4.52}$$

which is discussed in sec. 6.4.3.

An entropic description of this kind of dynamics needs strong joint perforation with little resulting conditioning. The new notation introduced in this chapter made accessible a discussion of delay dynamics in terms of information theory.

4.4.3.7 Is there a redistribution of information under perforation according to the process characteristics?

In the nonperforated case the outcome of the random variable at time step $-m$ of the past reduces the uncertainty of the outcome of the random variable at the first step of the future by $H_{1|m} - H_{1|m+1}$. In case of omission of one uncertainty reducing random variable in the past one could now be tempted to redistribute this uncertainty reduction of the future in the further past according to exactly the conditional entropy differences of the nonperforated case. The tantalizing thing with this idea is that under omission of outcomes of random variables in the past all information redistributions would in principle be describable with *only* the set of all $H_{1|m}$'s. Finally it was possible to *falsify* this idea.

One of the falsifying arguments is obtained again from the AR(1) process. Omitting the presence causes that $H_1 - H_{1|1}$ has to be redistributed into the further past. According to above idea it should be $\frac{(H_1 - H_{1|1})\cdot H_{1|1}}{H_1}$ lost for prediction and $\frac{(H_1 - H_{1|1})^2}{H_1}$ the information for prediction available in the next further past time step. This is not allowed, because qualitatively $\frac{(H_1 - H_{1|1})^2}{H_1}$ carries an explicit ϵ, i.e., a resolution dependence in the denominator. On the other hand from eq. (4.51) it has to be equal to $H_1 - H_{\{1\}|\{-1\}}$. This is not resolution-dependent (for sufficiently small ϵ), which is provable from continuous conditional entropies in generalized notation, and hence the idea of redistribution of information under perforation according to processcharacteristics is falsified. This example is mentioned here to show that seemingly intuitively correct ideas can be quite misleading in information theory.

4.4.4 Connection of multiple mutual information and perforation

The section is intended to give an idea, that the central quantities for redistribution of information for prediction under perforation of the past or presence are *multiple mutual informations* introduced in sec. 3.1.6. This is illustrated in the following for the simple case of perforation of only the presence (time step 0) and prediction of only the outcome of the random variable at the first step in the future (time step 1).

Under perforation of the presence the change of information for prediction in the first step of the past (time step -1) is the triple mutual information of single element sets at the time steps -1, 0 and 1

$$I_{\{-1\};\{0\};\{1\}} = H_1 - H_{\{1\}|\{-1\}} - (H_{1|1} - H_{1|2}) \; . \tag{4.53}$$

The proof for eq. (4.53) was already given in sec. 4.4.3.2. The right side of eq. (4.53) is exactly the first line on the right side of eq. (4.48).

Under perforation of the presence the change of information for prediction in the second step of the past (time step -2) is the following difference of multiple mutual informations of single element sets. Using eq. (3.30) one gets the proof:

$$\begin{aligned}
&I_{\{-2\};\{0\};\{1\}} - I_{\{-2\};\{-1\};\{0\};\{1\}} \\
&\quad = 3H_1 - H_2 - H_{\{-2,0\}} - H_{\{-2,1\}} + H_{\{-2,0,1\}} \\
&\qquad - 4H_1 + 3H_2 + 2H_{\{-2,0\}} - H_{\{-2,1\}} - 2H_3 - H_{\{-2,0,1\}} - H_{\{-2,-1,1\}} + H_4 \\
&\quad = -H_1 + 2H_2 + H_{\{-2,0\}} - 2H_3 - H_{\{-2,-1,1\}} + H_4 \\
&\quad = H_{\{-1,1\}} - H_1 - (H_{\{-2,-1,1\}} - H_2) - [(H_3 - H_2) - (H_4 - H_3)] \\
&\quad = H_{\{1\}|\{-1\}} - H_{\{1\}|\{-2,-1\}} - (H_{1|2} - H_{1|3}) \; .
\end{aligned} \tag{4.54}$$

The last line is exactly the second line on the right side of eq. (4.48).

Under perforation of the presence the change of information for prediction in the third step of the past (time step -3) is the following combination of multiple mutual informations of single element sets. It is somewhat lenghty, but provable in exactly the same spirit as above that

$$\begin{aligned}
&I_{\{-3\};\{0\};\{1\}} - I_{\{-3\};\{-2\};\{0\};\{1\}} - I_{\{-3\};\{-1\};\{0\};\{1\}} + I_{\{-3\};\{-2\};\{-1\};\{0\};\{1\}} \\
&\qquad\qquad = H_{\{1\}|\{-2,-1\}} - H_{\{1\}|\{-3,-2,-1\}} - (H_{1|3} - H_{1|4}) \; .
\end{aligned} \tag{4.55}$$

The last line is exactly the third line on the right side of eq. (4.48).

The results call for a proposal that the change of information in the n'th step of the past (time step $-n$) for prediction of the first step in the future under perforation of the presence is given by

$$\begin{aligned}
&I_{\{-n\};\{0\};\{1\}} - \sum_{i=-n+1}^{-1} I_{\{-n\};\{i\};\{0\};\{1\}} + \sum_{i_1,i_2=-n+1\,:\,i_1<i_2}^{-1} I_{\{-n\};\{i_1\};\{i_2\};\{0\};\{1\}} \\
&\qquad \pm \ldots + (-1)^{n-1} I_{\{-n\};\{-n+1\};\ldots;\{0\};\{1\}} \; .
\end{aligned} \tag{4.56}$$

The result of above calculations is, that multiple mutual informations are the basic construction elements for redistribution in the further past of information for prediction into the future, if the operation of perforation is carried out.

4.4.5 Changed sampling: Downsampling and upsampling

4.4.5.1 Downsampling

Downsampling is the reduction of the sampling rate. Instead of taking every data point of a time series, only every n'th data point is taken into account for the analysis. The aim here is now to represent all entropic quantities of the downsampled time series as functions of only the entropies of the original time series. Finally it should be possible to give a functional relationship of entropies for different time resolutions.

In the following the special case of downsampling by a factor of $\alpha = 2$ is analyzed theoretically, but all other values $\alpha > 0$ would also be possible. First, the downsampled joint 'block-type' entropies are written as functions of non-downsampled entropies: From eq. (4.14)

$$H_{0;2} = 0 \tag{4.57}$$

and from eq. (4.21)

$$H_{1;2} = H_{1;1} \tag{4.58}$$

immediately follow. The latter is essentially the same as eq. (3.143). From eq. (4.13), adaequate numbering change under sampling change, eq. (4.5) and again eq. (4.13) one gets

$$\begin{aligned} H_{2;2} &= H_{\{0,1\};2} \\ &= H_{\{-1,1\};1} \\ &= H_{\{-1,0,1\};1} - H_{[\{0\}|\{-1,1\}];1} \\ &= H_{3;1} - H_{[\{0\}|\{-1,1\}];1} \ . \end{aligned} \tag{4.59}$$

One can see that expressing downsampled nonperforated joint entropies by entropies refering to the original sampling leaves the space of nonperforated joint entropies and also causal nonperforated one-step-future conditional entropies are not enough for the description. One can see that with $H_{[\{0\}|\{-1,1\}];1}$ a nonperforated description needs *non-causal conditional* entropies, where conditioning appears into the future as well as into the past. For Shannon entropies the derivation of eq. (4.59) by probabilities reads

$$\begin{aligned} H_{2;2} &= -\sum_{i_{-1},i_1} p_{i_{-1},i_1} \ln p_{i_{-1},i_1} \\ &= -\sum_{i_{-1},i_0,i_1} p_{i_{-1},i_0,i_1} \ln p_{i_{-1},i_1} \\ &= -\sum_{i_{-1},i_0,i_1} p_{i_{-1},i_0,i_1} \ln \frac{p_{i_{-1},k_0,i_1}}{p_{k_0|i_{-1},i_1}} \quad , \quad k_0 \text{ fixed} \ . \end{aligned} \tag{4.60}$$

k_0 has to be chosen, such that divergences from vanishing denominator or vanishing argument of the logarithm are avoided. As the crucial step now the fixed k_0 is replaced by a variable index running exactly according to the summation index, i.e.,

$$\begin{aligned} H_{2;2} &= -\sum_{i_{-1},i_0,i_1} p_{i_{-1},i_0,i_1} \ln \frac{p_{i_{-1},i_0,i_1}}{p_{i_0|i_{-1},i_1}} \\ &= -\sum_{i_{-1},i_0,i_1} p_{i_{-1},i_0,i_1} \ln p_{i_{-1},i_0,i_1} + \sum_{i_{-1},i_0,i_1} p_{i_{-1},i_0,i_1} \ln p_{i_0|i_{-1},i_1} \\ &= H_{3;1} - H_{[\{0\}|\{-1,1\}];1} \ . \qquad \blacksquare \end{aligned} \tag{4.61}$$

Further finally

$$H_{3;2} = H_{\{-1,0,1\};2} = H_{\{-2,0,2\};1} = H_{5;1} - H_{[\{-1,1\}|\{-2,0,2\}];1} \tag{4.62}$$

holds. For $m \geq 3$ in $H_{m;2}$ a nonperforated representation in the original sampling of the nonperforated downsampled joint entropy enforces the appearance of *non-causal conditional joint* entropies.

For discussion of the downsampled conditional 'block-type' one-future-time-step entropies, it is started with the trivial case of empty set conditioning:

$$h_{0;2} \cdot 2\tau = H_{1;2} = H_{[1|0];2} = H_{[\{i\}|\emptyset];2} = H_{[\{i\}|\emptyset];1} = H_{[1|0];1} = H_1 = h_{0;1} \cdot \tau \; . \tag{4.63}$$

One derives for the information rates

$$h_{0;2} = \frac{h_{0;1}}{2} \; . \tag{4.64}$$

For larger conditioning one gets

$$\begin{aligned} h_{1;2} \cdot 2\tau = H_{[1|1];2} &= H_{\{0,1\};2} - H_{\{1\};2} \\ &= H_{\{-1,1\};1} - H_{\{1\};1} \\ &= H_{3;1} - H_{[\{0\}|\{-1,1\}];1} - H_{\{1\};1} \end{aligned} \tag{4.65}$$

and

$$\begin{aligned} h_{2;2} \cdot 2\tau = H_{[1|2];2} &= H_{\{-3,-1,1\};1} - H_{\{-3,-1\};1} \\ &= \left[H_{5;1} - H_{[\{-1,1\}|\{-2,0,2\}];1}\right] - \left[H_{3;1} - H_{[\{0\}|\{-1,1\}];1}\right] \; . \end{aligned} \tag{4.66}$$

For the limit of infinite conditioning it holds[5]

$$\begin{aligned} h_{\infty;2} \cdot 2\tau &= H_{[1|\infty];2} \\ &= \lim_{m\to\infty} \left[H_{\{-m,\ldots,-3,-1,1\};1} - H_{\{-m,\ldots,-3,-1\};1}\right] \\ &= \lim_{m\to\infty} \left[H_{m;1} - H_{m-2;1}\right] + \lim_{m\to\infty} \left[H_{[\{-m+1,\ldots,-2\}|\{-m,\ldots,-1\}];1} - H_{[\{-m+1,\ldots,0\}|\{-m,\ldots,1\}];1}\right] . \end{aligned} \tag{4.67}$$

Since the new non-causal conditional entropies are in general not expressible as functions of causal conditional entropies alone, one can see again that the representation of downsampled entropies as functions of entropies of the original sampling needs an enlargement of the space of entropies. Downsampled entropies $H_{m;\alpha>1}$'s and $H_{[1|n];\alpha>1}$'s are not representable as functions of $H_{m;1}$'s and $H_{[1|n];1}$'s, what a posteriori gives the most crucial justification for having expanded the definition of entropies in time series analysis, and what eliminates the treatment of H_m's and $H_{1|n}$'s as the fundamental entropic quantities, because they do not span the space of needed entropic quantities. Of course the framework of Renyi entropies was left at no point with the given generalizations.

---

[5] The final complicated line is again necessary for a representation of the downsampled quantity in terms of nonperforated entropic quantities in the original sampling.

In case of *deterministic* dynamics it holds

$$h_{\infty;2} \cdot 2\tau = H_{[1|\infty];2} = 2H_{[1|\infty];1} = 2h_{\infty;1} \cdot \tau \ , \tag{4.68}$$

what can of course not be deduced from eq. (4.67), which is valid in general. One derives that

$$h_{\infty;2} = h_{\infty;1} \ . \tag{4.69}$$

The entropy rate, which is independent of the time interval in the deterministic case, makes the entropy rate being the same for the infinitesimal as for the finite case. This causes the limit $\lim_{\tau \to 0}$ in the case of $\epsilon\tau$- entropies to be unnecessary in the deterministic case.

Eq. (4.69) and eq. (4.64) are not in contradiction because behaviour of entropy rates is different at different time. This and the validity of eq. (4.69) were already seen in the derivatives of the curves in fig. 3.14.

Here very much care in comparison with usual (but in the author's opinion inconsistent) notation is necessary! If the quantities h_∞ are treated as entropies instead of as entropy rates, what is *not* supported by this work, one ends up with $h_{\infty;2} = 2h_{\infty;1}$ in contradiction to the result above and potentially leading to confusion, if not treated with care.

4.4.5.2 Upsampling

From statistical physics is known that the renormalization group is truely only a semigroup, which supports only the direction of coarse graining and not the direction of refinement. The situation here is similar in the sense that given a time series belonging to a certain time scale, passing over to higher time resolutions is not possible without using interpolation procedures. The corresponding entropies are based only on approximated values of the interpolated time series. It is obvious that

$$H_{1;\frac{1}{n}} = H_{1;1} \equiv H_1 \tag{4.70}$$

and

$$H_{2;\frac{1}{n}} \equiv H_{\{0,1\};\frac{1}{n}} = H_{\{0,\frac{1}{n}\};1} \ . \tag{4.71}$$

A possible statement concerning upsampled entropies from interpolation is that practical estimation of joint entropies as $H_{2;\frac{1}{n}}$ of two in general correlated random variables by delay vectors of length two from *linear* interpolation

$$\mathbf{s}_{i;\frac{1}{n}}(k) = (\frac{k}{n}s_i + \frac{n-k}{n}s_{i+1}, \frac{k-1}{n}s_i + \frac{n-k+1}{n}s_{i+1}) \tag{4.72}$$

in general yields a slower convergence in the approximation of H_1 with increasing n than estimation by delay vectors of length two from a *higher order* interpolation polynomial on a longer section of the time series. Whereas in this sense small upsampling with advanced interpolation can possibly lead to useful approximations, increasing crudeness of the approximation makes the limit of large upsampling from interpolation of course practically irrelevant.

There is another possibility in which upsampling could become practically relevant: Given the possibility of extremely long data sets from an extremely high sampling rate in the course of contemporary enlargement of computational resources. The data analyst then initiates his analysis in the first instance on a drastically downsampled dataset. For ordinary applications this is completely sufficient. However, according to special requirements the data analyst has the option to zoom into the data and perform true 'upsampling'.

The main intention of this section is to make explicit the link of the limit of the opposite of downsampling, i.e. the limit of upsampling, with the content of sec. 3.8 concerned with entropies in the time-continuous case.

4.4.6 Example calculation of AR(2) for downsampling

In the following two different approaches to increase the time step length are introduced and it is tried to compare them. First, *downsampling* is carried out. As a first step, a version of the equation of motion of the AR(2) with shifted indices is inserted into the original equation of motion to yield

$$\begin{aligned} X_{n+2} &= a_0 X_{n+1} + a_1 X_n + \xi_{n+1} \\ &= a_0(a_0 X_n + a_1 X_{n-1} + \xi_n) + a_1 X_n + \xi_{n+1} \\ &= a_0(a_0 X_n + a_1(\frac{X_n - a_1 X_{n-2} - \xi_{n-1}}{a_0}) + \xi_n) + a_1 X_n + \xi_{n+1} \\ &= (a_0^2 + 2a_1) X_n - a_1^2 X_{n-2} + \psi_n \end{aligned} \tag{4.73}$$

with

$$\psi_n := \xi_{n+1} + a_0 \xi_n - a_1 \xi_{n-1} \; , \tag{4.74}$$

which is *correlated noise*, because

$$\langle \psi_n \psi_{n+2} \rangle = -a_1 \; . \tag{4.75}$$

Hence eq. (4.73) does not describe an AR(2). Redefining ψ_n by a simpler effective process (MA(2)) having the same correlation property by

$$\psi_n := \zeta_n - a_1 \zeta_{n-2} \quad \text{with} \quad \zeta_i \sim \mathcal{N}(0, \sigma_\xi^2) \tag{4.76}$$

makes the process in eq. (4.73) to be an ARMA(2,2) process. Iteratively inserting the process into itself leads to

$$\psi_n = \zeta_n - a_1 \psi_{n-2} - a_1^2 \psi_{n-4} - a_1^3 \psi_{n-6} - \dots \; . \tag{4.77}$$

Using

$$\psi_n = X_{n+2} - (a_0^2 + 2a_1) X_n + a_1^2 X_{n-2} \tag{4.78}$$

leads to

$$X_{n+2} = \zeta_n + (a_0^2 + a_1) X_n + a_1 a_0^2 X_{n-2} + a_1^2 a_0^2 X_{n-4} + a_1^3 a_0^2 X_{n-6} + a_1^4 a_0^2 X_{n-8} + \dots \; . \tag{4.79}$$

With (2.79) it is extracted for example

$$a_0(2\Delta t) = a_0^2(\Delta t) + a_1(\Delta t) = \frac{12(\gamma^2 - 16\omega_0^2)\Delta t^2 + 4\omega_0^4 \Delta t^4}{(2 + \gamma \Delta t)^2} \; . \tag{4.80}$$

The main result of the calculation is a stochastic process with *infinite memory*.

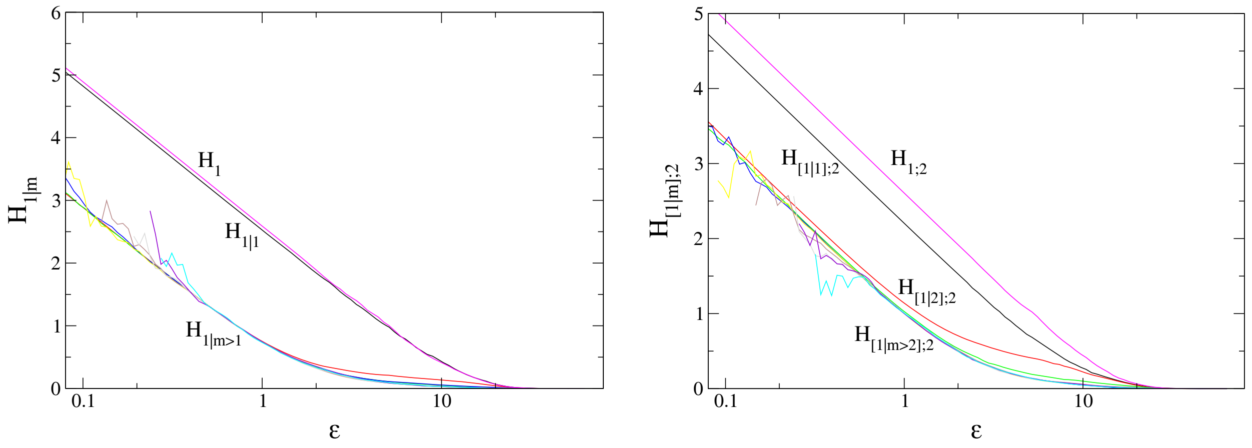


Figure 4.4: Conditional entropies of the AR(2) process with parameters $a_0 = 0.7; a_1 = -0.99$; $\sigma_\xi = 1$ for a dataset of 150000 data points. Left panel: Usual conditional entropies. Right panel: Downsampled conditional entropies.

The second possibility to increase the time step length consists in working a priori with doubled sampling time. Again using (2.79) this leads to an AR(2) process with coefficients

$$a_0(2\Delta t) = \frac{2 - \omega_0^2(2\Delta t)^2}{1 + \gamma\Delta t} \quad \text{and} \quad a_1(2\Delta t) = \frac{-1 + \gamma\Delta t}{1 + \gamma\Delta t} \,, \tag{4.81}$$

i.e., a Markov process of *second order*.

Hence two different approaches to a reduced rate of sampling in the approximation of the stochastically perturbed damped harmonic oscillator lead to very different memory behaviour. The task of a small numerical analysis is now to find out which of the proposed schemes is more reliable concerning memory behavior under increase of the time step length. For the numerical analysis $a_0(\Delta t)$ and $a_1(\Delta t)$ are chosen suitably and a sufficiently long dataset is created. For convenience b in eq. (2.75) is chosen such that $\xi_n \sim \mathcal{N}(0, 1)$. Then numerically downsampling by a factor of two is performed, i.e., every second data point is omitted from the dataset. As a final step the conditional entropies are estimated numerically.

A completely systematical analysis of the area of stability of the AR(2) concerning extension of the number of time steps containing memory under downsampling was not performed, but empirically best results were obtained for parameter values $0.5 < a_0 < 0.7$ and $-1 < a_1 < -0.95$. The left panel of fig. 4.4 shows the conditional entropies in the usual non-downsampled case. Whereas the first step of conditioning reduces almost no uncertainty, the second step of conditioning reduces uncertainty strongly and further steps do not decrease the uncertainty anymore. The downsampled case is shown in the corresponding right panel. The result is rather weak, but besides uncertainty reduction in the first two steps of conditioning a further small uncertainty reduction in higher conditioning is clearly visible. If started with suitable parameter values for the stochastically perturbed harmonic damped oscillator, e.g., if chosen $\omega_0 = 1.0$, $\gamma = 0.1$ and $\Delta t = 0.1$, then $a_0 = 1.9801$ and $a_1 = -0.99005$ is derived and in this area of the parameter space longer memory from downsampling is not seen. The quality of the result is parameter-dependent. Nevertheless the first suggested approach has to be preferred, because at least for some parameter values memory beyond second order is seen. The second procedure has to be treated as an approximation.

This example illustrates that originally Markovian dynamics after performing downsampling, i.e., coarse graining in time, can possibly lead to non-Markovian dynamics, i.e., dynamics with in principle memory extension to infinite time. Simple operations cause transition to infinite memory. Practically a truncation of irrelevant memory is necessary and this shows the importance for the need of a tool for suitable memory truncation as will be presented in chap. 6.

4.5 Comment on dimensions in perforated case

In the perforated case the embedding dimension is given by the length of the generalized delay vector, i.e., the cardinal number of $\mathbf{J}$. This is the same as in the nonperforated case, where because of missing structure instead of $\mathbf{J}$ only its cardinal number, i.e., the embedding dimension, is given.

A special perforation of points of time as carried out in chap. 4 corresponds to a special projection in the embedding space (cmp. eq. (4.27)). Eq. (3.121) can be generalized to a formula for the dimension corresponding to projections in the embedding space:

$$D_{\mathbf{J}}^{(q)} = -\lim_{\epsilon \to 0} \frac{H_{\mathbf{J}}^{(q)}(\epsilon)}{\ln \epsilon} \; . \tag{4.82}$$

In case of infinitesimal ϵ eq. (4.82) should yield in the spirit of [64] for $|\mathbf{J}| > D$ (D: dimension of underlying dynamics)

$$D_{\mathbf{J}}^{(q)} = D \; , \tag{4.83}$$

no matter which components chosen in $\mathbf{J}$. For $|\mathbf{J}| < D$

$$D_{\mathbf{J}}^{(q)} = |\mathbf{J}| \tag{4.84}$$

is expected. The reason can be found in descriptions of fractals from nonlinear dynamics by SRB measures.

On the other hand for finite ϵ different projections in the embedding space given by different $\mathbf{J}$ in general yield *different* effective dimensions $D_{\mathbf{J}}^{(q)}$ also if the different $\mathbf{J}$ are of the same cardinality, because the attractor is then differently reconstructed. Moreover repellers, which show fractality in more than one direction have the property of different effective dimensions $D_{\mathbf{J}}^{(q)}$ for different projections in the embedding space also for infinitesimal ϵ.

Chapter 5

Origin and description of memory

5.1 The problem

One of the main questions of this work is where to truncate memory in order to make optimal predictions. At this point, it seems to be in order to discuss shortly the nature of memory.

Memory on the one hand is said to have the property that knowing a certain state in the far past can reduce somehow uncertainty of the future. So there is some kind of farreaching interaction. On the other hand physicists should always be aware of the most fundamental statements in their theories, and those are the physical principles. Besides for example principles of causality and relativity one of the most fundamental principles in physics is the *principle of locality* of all fundamental kinds of interactions. So it is necessary to address shortly the question if there is a contradiction of the existence of long range memory and the principle of locality and if not, how to combine them consistently. How does memory arise, if all actions happen local in time and space? The reasoning will follow mostly along the usual way of emergence of nonlocal interactions.

The implicitness in the treatment of the notion 'memory' with no mention of anything unnatural about it and the absence of any comment on the principle of locality in contexts of e.g. delay dynamics caused the necessity of some statements on the subject in this chapter.

5.2 Origin and nature of memory

5.2.1 Resolution, effective degrees of freedom and Markovianity

The following arguments are purely classical and do not involve any quantum effects.

Starting from highest resolution (at a presumed fundamental scale) in principle a description of nature has to be allowed with all degrees of freedom involved and hence all interactions are local and absolutely no memory appears. This corresponds in principle to very high-dimensional measurement vectors at one time. Every system is securely Markovian (of first order!) if all participating degrees of freedom are taken into account.

For highest resolution there can still be memory effects if the resolvability of degrees of freedom is unused, i.e., if yet simultaneous measurements of inadaequately few or even only one quantity are performed.

For coarser resolution the resolvability of degrees of freedom is restricted and this is the reason, why for coarser resolution not necessarily memory can always be avoided! If it is possible to take into account sufficiently many relevant degrees of freedom the situation could remain Markovian.

For coarsest resolution the resolvability of degrees of freedom is very bad and taking into account several effective degrees of freedom in general doesn't help anymore for avoiding memory. The effect of coarse graining on the increase of memory depth can be seen in sec. 6.2.2.1.

Summarizing, memory arises if finally the description of the system is tried by inadaequately few quantities. Practically measurable quantities are never truely microscopic, but always coarse grained in some sense. Hence in time series analysis the appearance of memory is a common phenomenon.

This is the right place to comment on a statement in ([29], p.45): In the framework of time-continuous stochastic processes it is claimed that "there is really no such thing as a Markov process". From this one probably has to conclude that non-Markovianity should be an intrinsic feature of nature also in classical descriptions and this is a point of view where I would like to disagree!

Fixing the length of the describing state vector, i.e., the dimension of the state space to a finite (rather low) number, in the limit of the treated time scale going to zero, I expect that non-Markovianity will in general appear, but this is a consequence of the effective description. The dimension of the state space not fixed, but allowing it to be suitably high while the treated time scale goes to zero for time continuity in stochastic processes, non-Markovianity will in principle (of course not practically) always be avoided! In this sense classically the Markov process is the natural always existing process and the non-Markovianity comes into play only as a (practically unavoidable) artefact of a deficitary, because low dimensional, effective description of nature.

In order to be practically able to describe physical systems, it is necessary to utilize some reduction method to combine degrees of freedom and generate effective dynamical variables or to start modeling from scratch in a rather crude way. Both concepts are capable to introduce memory and are shortly explained now.

5.2.2 Methods of complexity reduction

5.2.2.1 Renormalization group

First of all there are the renormalization group methods [25] introduced by Wilson. Degrees of freedom are integrated out and the development of coupling constants under the coarse graining procedure is observed to change the described dynamics.

5.2.2.2 Projection methods

A more crude method of complexity reduction of more practical type is projection. In nonequilibrium statistical physics projection-operator methods of Mori and Zwanzig are well established [81]. Furthermore in general a measurement procedure is a special case of projection. Practical application of time series analysis using successive measurements as well as the corresponding statements of Takens [71] and Sauer et al. [65] are hence built on the notion of projection.

5.2.3 A priori memory-including modelling

A general dynamical law in a deterministic framework being a priori memory-including could be formulated as

$$x_{n+1} = f(x_n, x_{n-k}) \qquad k > 0 \; . \tag{5.1}$$

A simple example of a logistic equation with memory can be found in [24]. Time discretization of delay differential equations leads to such types of equations. An example is the Mackey-Glass delay differential equation discussed in sec. 6.4.3. In a stochastic setting every Markov process of truely higher order (sec. 2.2.3.1) would be an example for dynamics being a-priori memory-including.

A priori memory-including modelling should finally be seen as the consequence of a usually not explicitly given projection procedure.

5.2.4 Conclusion

Nature on a fundamental scale does not know anything like memory! Fundamental interactions also do not use something like memory. There is no intrinsic mechanism[1] in nature, which introduces memory in a natural way!

Descriptions with too few effective degrees of freedom introduce memory and this is in the author's opinion the only reason for the appearance of non-Markovianity and the cause for need of questions concerning optimal Markov approximations. So the discussion on truncating memory is not a discussion of aspects which intrinsically arise in nature, but is a discussion on problems, which appear by a rather inordinately crude description of nature.

5.3 Entropic assessment of memory

In order to describe a system with few effective degrees of freedom the formerly developed framework of information theory is used to provide suitable quantities for memory. Sometimes the notion 'memory' is used for the longest time in the past to influence the future. Here exclusively the informational content is meant. Resolution dependence is omitted in the following.

5.3.1 Quantities in block-case

For the description of ignored memory or loss of information (for prediction) after a Markov approximation of order m

$$Q_m := H_{1|m} - H_{1|\infty} \tag{5.2}$$

---

[1] Seemingly intrinsic mechanisms for generation of memory possibly treat a part of the system as a black box stealthily as e.g. signal propagation in feedback-systems and henceforth hide the effectiveness of the description of the system.

is suggested. From eq. (3.70) it holds that

$$Q_m + R_m + H_{1|\infty} = H_1 \; . \tag{5.3}$$

Hence ignored memory and redundancy are complementary quantities. From eq. (5.2) the ignored memory Q_∞ in the case of Markov order ∞, which terminates Markovianity, vanishes as well as the redundancy R_0 in the case of Markov approximation of order zero (complete truncation) from eq. (3.70), i.e.,

$$Q_\infty = R_0 = 0 \; . \tag{5.4}$$

From eq. (5.3) and eq. (5.4) it is derived that for a Markov approximation of order zero the ignored memory Q_0 as well as the maximal redundancy R_∞ for Markov order ∞ equate the total memory $H_1 - H_{1|\infty}$:

$$Q_0 = H_1 - H_{1|\infty} \quad (\quad = R_\infty) \; . \tag{5.5}$$

It is not a contradiction that the case of maximal index, i.e. Q_∞, does not correspond to the total memory, because Q_m is defined to be the *ignored* memory.

5.3.2 Generalization to perforated case

The quantity Q_m generalizes under transition from a Markov model of order m to a perforated Markov model for variable future to

$$Q_{\{f\};\mathbf{J}} := H_{\{f\}|\mathbf{J}} - H_{\{f\}|\mathbb{Z}_0^-} \tag{5.6}$$

with $\mathbf{J} \subset \mathbb{Z}_0^- := \mathbb{Z}^- \cup \{0\}$. In case $f = 1$ this simplifies to

$$Q_\mathbf{J} \equiv Q_{\{1\};\mathbf{J}} = H_{\{1\}|\mathbf{J}} - H_{1|\infty} \; . \tag{5.7}$$

Since

$$R_{\{f\};\mathbf{J}} = H_{\{f\}} - H_{\{f\}|\mathbf{J}} \; , \tag{5.8}$$

one obtains

$$Q_{\{f\};\mathbf{J}} + R_{\{f\};\mathbf{J}} + H_{\{f\}|\mathbb{Z}_0^-} = H_1 \tag{5.9}$$

and for $f = 1$

$$Q_{\{1\};\mathbf{J}} + R_{\{1\};\mathbf{J}} + H_{1|\infty} = H_1 \tag{5.10}$$

as generalizations of eq. (5.3).

Chapter 6

Optimal Markov approximations

6.1 Introduction

In chap. 4 there was introduced a generalized notation of information theory in time series analysis, which made accessible a calculation of jointly conditioned joint entropies under perforated conditions. It is not only the task of this work to represent the structure of information for prediction in the past under generalized conditions, but it is especially the task to provide a justification for a truncation of information for prediction in conditioning, such that for practical prediction tasks reasonably few components of the past are taken into account in order to increase the power of nearest neighbor search methods. A solution is tried in this chapter.

With the aim of prediction it seems useful in the sense of information theory to include all components of the past, which contain information about the time step to predict. In general this could be infinitely many time steps. For real prediction purposes infinitely many time steps[1] cannot be taken into account and the need for a usual Markov approximation[2] or some kind of advanced Markov approximation is necessary. The question arising is how information theory can be used in the selection of relevant memory-containing components, with which to perform an optimal prediction.

The idea is now the following: Since the entropies are obtained from data sets by estimation (sec.'s 3.6 and 4.3), they underly some statistical error. This statistical error present in joint entropies can be transmitted to conditional entropies and to derived quantities like redundancies. The important quantity is the mutual information of the joint random variable of the chosen past components with the random variable of the future, because this is the mean available uncertainty reduction about the future random variable. The statistical error of this quantity is chosen to play the central role in different component selection criteria concerning the optimal (possibly perforated) Markov approximation.

In contrast to the assessment of arbitrary joint uncertainties developed in chap. 4, in the

---

[1] In the sense of time steps of conditioning in a Markov model on one side and component of a generalized delay vector in the estimation of entropies corresponding to a Markov model on the other side, in this section the notions 'component' and 'time steps' are used interchangeably.

[2] The usage of the notion 'Markov approximation' throughout this work strongly depends on the acceptance of a concept of 'Markovianity of higher order' introduced in sec. 2.2.3.1. In a framework, where the notion of 'Markovianity' is restricted to the case of memory of one time step, the terminology would surely be different.

following the numerical procedure is restricted to the case of one future time step according to perforated Markov processes.

6.2 Optimal usual Markov approximation

6.2.1 A novel criterion

Under the constraint of truncation of memory in a nonperforated sense, the maximal useful memory to take into account is restricted by the condition that the statistical error of the redundancy has to be smaller than the ignored memory. Otherwise the ignored memory is anyway not anymore resolvable. This results to the following ad hoc criterion:

$$m_{opt}(\epsilon) = \max\{m \in \mathbb{N}_0 : \Delta R_m(\epsilon) < Q_m(\epsilon)\} \ . \tag{6.1}$$

The estimation of $\Delta R_m(\epsilon)$ is given by eq. (3.108) and $Q_m(\epsilon)$ is defined in eq. (5.2). In case of variable future this becomes

$$m_{opt}(\epsilon) = \max\{m \in \mathbb{N}_0 : \Delta R_{\{f\};\{-(m-1),\dots,0\}}(\epsilon) < Q_{\{f\};\{-(m-1),\dots,0\}}(\epsilon)\} \ . \tag{6.2}$$

6.2.2 Examples

6.2.2.1 Stochastic example: Autoregessive process

In fig. 6.1 the result of applying the criterion of eq. (6.1) on a dataset of an AR(3) process is shown. For mean resolutions the memory depth $m = 3$ is exactly found. For higher resolutions, i.e. smaller ϵ, the statistical error increases and the estimation of $H_{1|\infty}(\epsilon)$ is performed by entropies with smaller m such that a shorter Markov approximation is automatically enforced by the algorithm. For large ϵ a farther ranging memory is detected. In order to understand this from the estimation of entropy a zoom is performed and presented in fig. 6.2. It is possible to observe a splitting of entropy levels even though from the underlying process this should only be the case for conditioning up to three time steps. Coarse graining effects for low resolution cause originally unexpected further splitting of entropies with higher conditioning, because the transition to coarser resolution acts effectively as some kind of partial perforation, which makes clear the tendency to shift information about the future to the further past.

6.2.2.2 Deterministic example: Hénon map

In fig. 6.3 the result of applying the criterion of eq. (6.1) on a dataset of Hénon dynamics is presented. The statistical errors of entropy estimation are extremely small if compared to the stochastic case in fig. 6.1. This causes the criterion to be very information-dominated. It is observed that the first and second step of conditioning contribute largely to memory, but also rather small nonvanishing contributions in further conditioning on the whole ϵ-range are visible in fig. 6.4. A tendency of this effect to decrease for smaller ϵ cannot be seen.

In consequence of its non-hyperbolicity, for the Hénon map there does not exist a Markov partition, i.e., it does not exist a partition that makes the corresponding process a topological Markov chain [6]. A topological Markov chain is defined by the property that

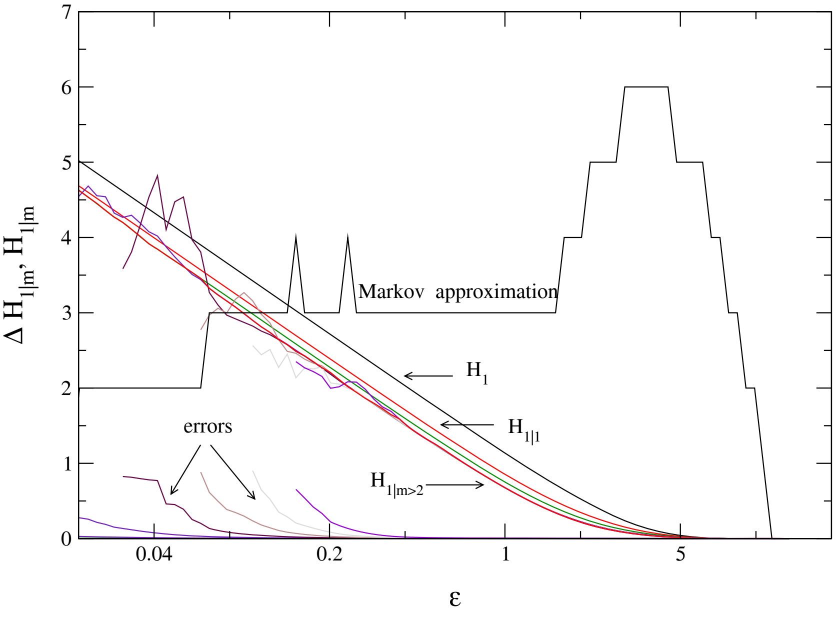


Figure 6.1: Conditional entropies, statistical errors and a suggestion for a Markov approximation for a dataset of length 50000 of an AR(3) process with coefficients $a_0 = 0.2; a_1 = 0.3; a_2 = 0.4$

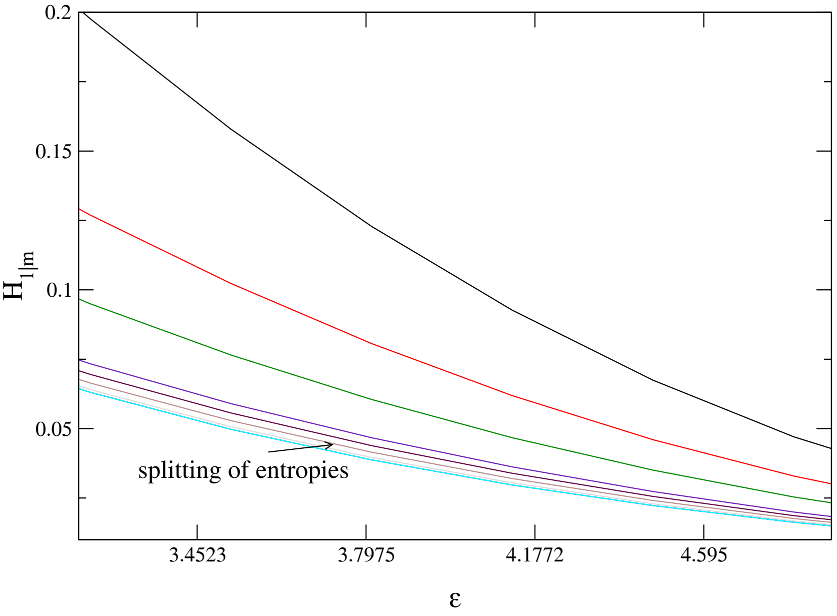


Figure 6.2: Zoom of fig. 6.1.

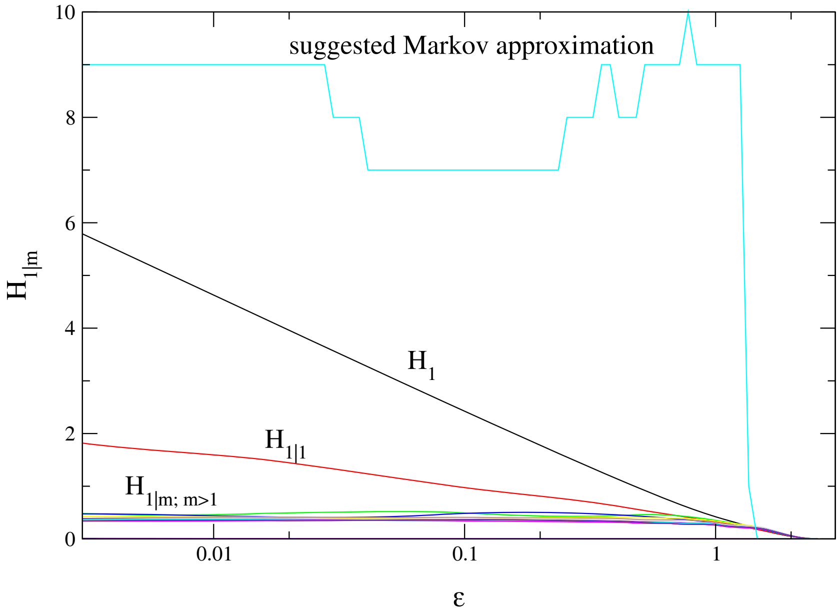


Figure 6.3: Conditional entropies and a suggestion for a Markov approximation for a dataset of length 100000 of the Hénon map with standard parameters $a = 1.4$ and $b = 0.3$. The statistical errors in the estimation of the entropies are so small that they are unvisible in this representation.

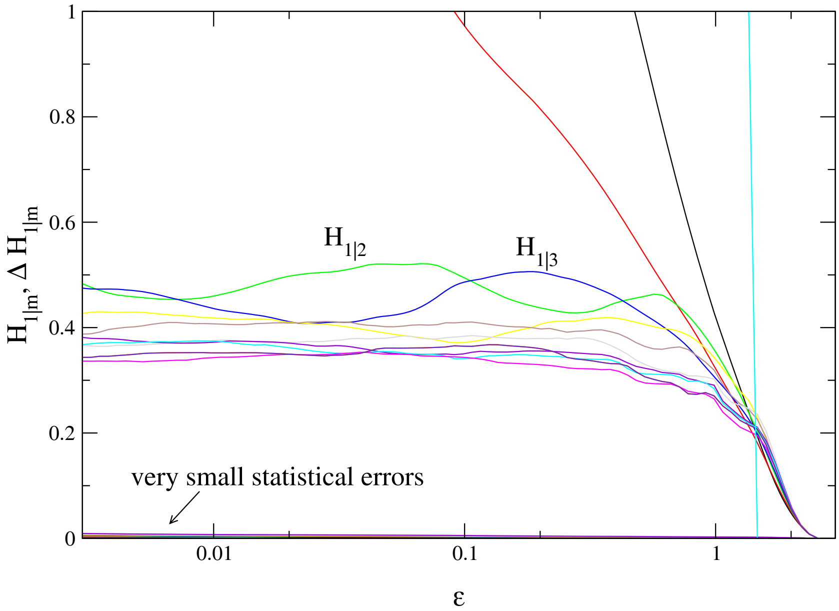


Figure 6.4: Zoom of fig. 6.3 with respect to the ordinate.

$p(i_N|i_0,...,i_{N-1}) = 0$ if and only if $p(i_N|i_{N-1}) = 0$ or $p(i_{N-1}|i_0,...,i_{N-2}) = 0$. Non-existence of the Markov partition is equivalent to non-existence of a finite set of forbidden basis sequences, which are parts of all forbidden sequences, i.e., for every $n_0 \in \mathbb{N}$ there exists a forbidden sequence of length $n \geq n_0$ that is not decomposable in a way that at least one subsequence of length smaller than n_0 is already forbidden. From the non-existence of a Markov partition it is derivable that for every resolution the Hénon map cannot be a Markov process and this explains the existence of long range memory and the result for the suggested Markov approximation.

Furthermore it is observed in fig. 6.4 a violation of monotonicity of the conditional entropies in the conditioning length m. This is not a statistical effect, but reproducible on a wide range of dataset lengthes. The reason is supposed to be found in the approximation of partition-based estimation of Shannon entropies, for which monotonicity is observed, by covering-based estimation of Renyi order $q = 2$ - entropies.

6.2.3 Concluding remarks

It is immediately clear that such a strategy cannot be optimal for delay dynamics with dependences on the further past because almost no memory is available for the immediate past and having reached the delay the statistics is already very bad.

Since especially wind speed predictions underlying turbulent effects are expected to show multiscale dependences, a generalization of above method to the perforated case is necessary. Finding an optimal usual Markov approximation is a special case of finding the optimal perforated Markov approximation introduced in the following.

6.3 Optimal perforated Markov approximation

6.3.1 A novel criterion

For finding the resolution-dependent optimal perforated Markov approximation, i.e., the optimal conditioning sets $\mathbf{J}^*(\epsilon)$, as a central criterion it is demanded

$$\boxed{Q_{\{f\};\mathbf{J}}(\epsilon) + b \cdot \Delta R_{\{f\};\mathbf{J}}(\epsilon) = \min \ , \qquad (6.3)}$$

where for given ϵ the minimum is taken over all possible conditionings $\mathbf{J} \subset \mathbb{Z}_0^-$. Repeating eq. (5.6) the ignored memory in the perforated case is defined as

$$Q_{\{f\};\mathbf{J}}(\epsilon) := H_{\{f\}|\mathbf{J}}(\epsilon) - H_{\{f\}|\mathbb{Z}_0^-}(\epsilon) \ , \qquad (6.4)$$

and $\Delta R_{\{f\};\mathbf{J}}(\epsilon)$ is the statistical error of the redundancy obtained from eq. (4.30). The parameter b to be determined accounts for the weight of the statistical error of the redundancy in the criterion. However, all results of chap. 6 are based on the choice $b = 1$.[3] If the solution for a certain ϵ is not unique, it is taken in a second step the set $\mathbf{J}(\epsilon)$ as $\mathbf{J}^*(\epsilon)$ with

$$\min(\mathbf{J}(\epsilon)) = \max \qquad (6.5)$$

---

[3] A short discussion on the balance from comparison with an empirical method in the framework of prediction can be found in sec. 7.2.3.

among the preselected ones. The chosen criterion is justified as follows:

- The ignored memory $Q_{\{f\};\mathbf{J}}(\epsilon)$, i.e., ignored potentially usable information has to be rather small. This alone as a criterion would in principle allow the optimal perforated Markov model to contain infinitely many time steps of conditioning, which is the task to avoid with regard to prediction purposes.

- The reduction of uncertainty should be confident, i.e., the statistical error of the redundancy $\Delta R_{\{f\};\mathbf{J}}(\epsilon)$ should be small. The statistical error of the redundancy is mainly influenced by the statistical errors of the joint entropies $H_{\mathbf{J}}(\epsilon)$ and $H_{\mathbf{J}\cup\{f\}}(\epsilon)$ and this increases with larger $|\mathbf{J}|$ (and smaller ϵ), because less neighbors are found for estimation of the correlation sum under more restrictive conditions. This alone as a criterion would prefer an empty conditioning.

Somehow the two quantities have to be composed together for a criterion to get the optimum. Multiplication makes no sense, because then components are preferred if one of the quantities is zero, for instance the ignored memory, which cannot be the wanted case.

A weighted sum of both quantities is suggested by (6.3). Increasing the value of the ignored memory by multiplying it with a number larger than one has the disadvantage that wrongly too low estimations in the small ϵ- and large $|\mathbf{J}|$-regime of the corresponding conditional entropies are not suppressed by the statistical error term and that such wrongly selected optimal perforated Markov models increase to dominate the result. Increasing the value of the statistical error by multiplying it with a number larger than one has the disadvantage that typically perforated structures with many elements are suppressed for being optimal, even if many elements contain a reasonable amount of information for prediction.

The chosen criterion will obtain the justification by its ability to recover known models behind sufficiently large data sets in a suitable intermediate interval of resolutions shown in sec. 6.4. That way its appropriateness for the determination of suitable generalized Markov approximations will be further substantiated.

A criterion of the type of that used for the usual Markov approximation is not applicable for finding the optimized perforated Markov model, because there is no single number which could be compared for all perforated Markov models.

A tool for finding the optimal perforated Markov model could also be useful outside of the framework of time series analysis, e.g. in spin-spin interaction.

6.3.2 Simplified approximative representation of the criterion

The criterion in eq. (6.3)

$$\begin{aligned} Q_{\{f\};\mathbf{J}}(\epsilon) + b \cdot \Delta R_{\{f\};\mathbf{J}}(\epsilon) &= H_{\{f\}|\mathbf{J}}(\epsilon) - H_{\{f\}|\mathbb{Z}_0^-}(\epsilon) + b \cdot \sqrt{\Delta H_1^2(\epsilon) + \Delta H_{\{f\}|\mathbf{J}}^2(\epsilon)} \\ &= \min \end{aligned} \tag{6.6}$$

can be approximated by

$$H_{\{f\}|\mathbf{J}}(\epsilon) + b \cdot \Delta H_{\{f\}|\mathbf{J}}(\epsilon) = \min \ , \tag{6.7}$$

because $H_{\{f\}|\mathbb{Z}_0^-}(\epsilon)$ and $\Delta H_1(\epsilon)$ are independent of $\mathbf{J}$ and hence constant for given resolution ϵ. The approximative character comes from the fact that in the error propagation the statistical error $\Delta H_1(\epsilon)$ does not appear as a constant summand in eq. (6.6), but squares of the errors are summed before the root is taken, i.e., eq. (6.7) is not exactly the same as eq. (6.3), but a very good approximation, since $\Delta H_1(\epsilon)$ is in general small compared to $\Delta H_{\{f\}|\mathbf{J}}(\epsilon)$.

The interpretation of this approximation of the criterion is that the value of the conditional entropy plus its statistical error has to be minimal, i.e., optimality is not reached in the case of a low estimation of a conditioned uncertainty, if this is bought by a large statistical error. The statistical uncertainty in the estimation of the uncertainty of random variables has to be considered. The simplicity and seriousness of the criterion becomes clearly obvious.

6.3.3 Implementational remarks

The basis of the introduced algorithms is guided by the TISEAN program d2 [40]. It estimates resolution-dependent conditional entropies for the nonperforated case. According to the needs of the criterion in eq. (6.3) the new algorithms estimate entropies for perforated conditionings, in which the case of variable single future is already included by perforation immediately behind the first component. Furthermore the resolution-dependent explicit estimation of the statistical error in the estimation of the entropies is implemented, which was already used in the nonperforated case in sec. 6.2.

Different loop-structures on sets of possible conditionings are offered:

1. Full loop on 2^m different conditionings restricted only by maximal number m of past time steps. This variant is computationally extremely expensive.
2. Loop on different conditionings restricted by the maximal number of past time steps, but also by the maximal cardinality of the conditioning set $\mathbf{J}$. The estimation of $H_{1|\infty}(\epsilon)$ resp. $H_{\{f\}|\mathbb{Z}_0^-}(\epsilon)$ is not anymore available in this case.
3. Loop on different conditionings, which reaches further past by a priori omission of past time steps with accumulated omissions in the further past and nearly all time steps taken along in the immediate past. This strategy is based on the assumption of a rather low conditional mutual information of a time step in the past and the future time step conditioned on a time step in the past close to the first. This should be practically reasonable for most cases, but it is expected that examples are constructable where this assumption is completely wrong.

The result is always a ranking of the best resolution-dependent perforated Markov models, where finally only the optimum is shown in this work.

According to the discrepancy of the criteria in sec. 6.3.1 and sec. 6.3.2 different choices concerning the treatment of the resolution-dependent generalization of the Kolmogorov-Sinai entropy for variable future $H_{\{f\}|\mathbb{Z}_0^-}(\epsilon)$ or the specialization $H_{1|\infty}(\epsilon)$ for the case of the immediate future time step are possible for the resolution-dependent full loop over all conditionings restricted only by the maximal number m of past time steps:

1. $H_{\{f\}|\mathbb{Z}_0^-}(\epsilon)$ is estimated from the conditional entropy with highest finite conditioning for which the statistical error of the estimation of the entropy is below some fixed threshold. The consequence is that with increasing resolution the quantity $H_{\{f\}|\mathbb{Z}_0^-}(\epsilon)$ is estimated from smaller and smaller conditionings. According to fig. 3.2 this variant

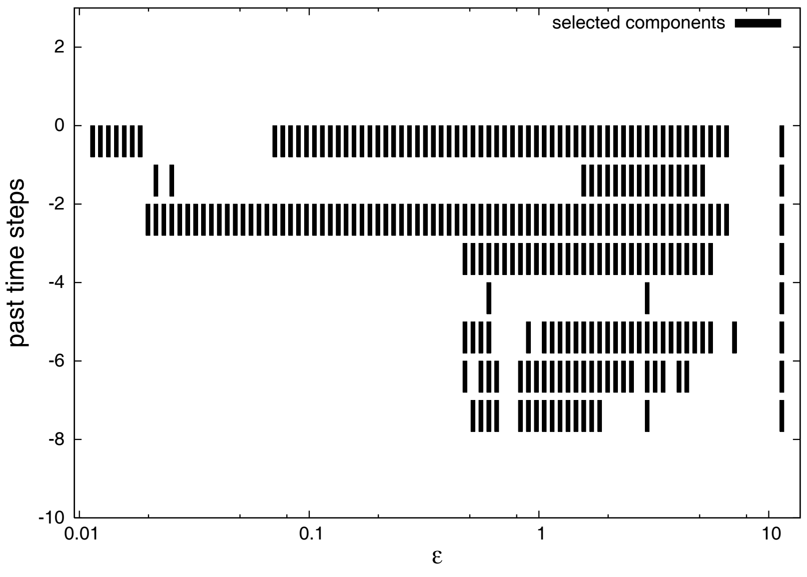


Figure 6.5: Resolution-dependent optimal perforated Markov model for a dataset of 40000 data points of an AR(3) process with $a_0 = a_2 = 0.4$. The optimal conditioning structures can be found in vertical direction. $H_{1|\infty}(\epsilon)$ is estimated according to the first item of the second enumeration in sec. 6.3.3.

performs rather well in the stochastic case. However, for deterministic dynamics this is not the case, because for certain resolutions huge jumps in the estimated quantity are the result. This is the reason, why this variant cannot be part of a universal model selection tool.

2. $H_{\{f\}|\mathbb{Z}_0^-}(\epsilon)$ is fixed to zero. The disadvantage of this case is that in contrast to the previous case a threshold against wrongly too low estimations of conditional entropies is missing and this leads to the possibility of erroneous assessment of optimality to conditioning structures especially for higher resolution. This possibility is excluded for the usual Markov approximation because the criterion in eq. (6.2) explicitly needs an estimation of $H_{\{f\}|\mathbb{Z}_0^-}(\epsilon)$ for the determination of $Q_{\{f\};\{-(m-1),\ldots,0\}}(\epsilon)$.

6.4 Application to different given dynamics in the perforated case

6.4.1 AR processes

The dynamical law of AR processes is given in eq. (2.56). For the first analysis a simple autoregressive process with parameters $a_0 = a_2 = 0.4$ is chosen. Parameters not mentioned are understood to be zero. A full search for the ϵ-dependent optimal conditioning structure according to the criterion stated in eq. (6.3) is carried out. The result is shown in fig. 6.5.

A first result is that the found optimal conditioning structure is resolution-dependent. Interpreting fig. 6.5, it is possible to extract four regimes: For high resolution, i.e. small ϵ, the

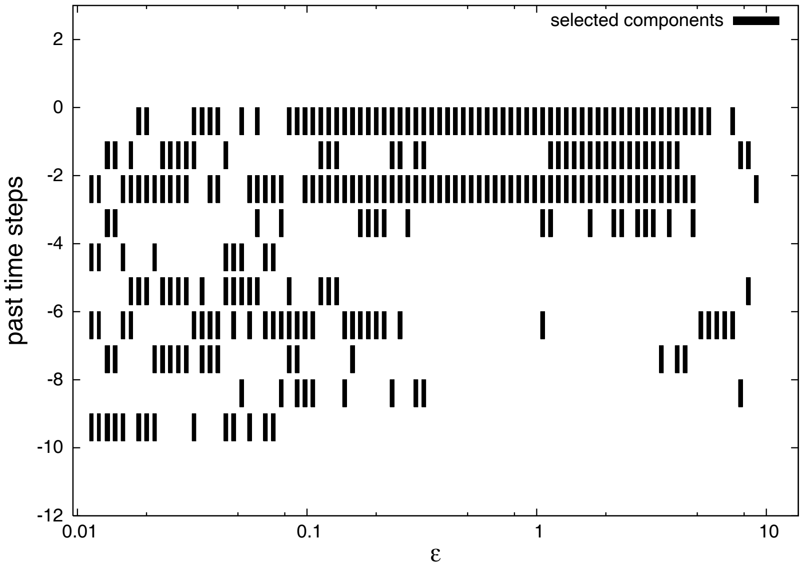


Figure 6.6: Resolution-dependent optimal perforated Markov model for a dataset of length 40000 of an AR(3) process with $a_0 = a_2 = 0.4$ under the restriction of maximal 4 conditioning components and non-estimated $H_{1|\infty}(\epsilon)$ according to the second item of the second enumeration in sec. 6.3.3.

statistical errors of the entropy estimations are rather large, and the criterion is dominated by the statistical error of the redundancy, which causes perforated structures with fewer elements to be detected as optimal. For intermediate resolutions, the most interesting part of the plot, the model behind the dataset is found, because the statistical error is sufficiently small and the resolution is sufficiently large that the information term dominates the criterion without being disturbed by either statistical or resolution effects. Nevertheless, in this domain of resolutions the statistical error of the redundancy has the task to exclude all conditionings longer than necessary among those which are equal and optimal from the informational point of view. For even coarser resolution, there is the domain of coarse graining splittings. The statistical error typically does not play a role anymore and the information term becomes influenced by resolution effects. Even though the components do not carry information from the dynamical law, analyzing the dataset with coarse resolution they are frequently chosen to appear in the optimal perforated Markov model. This is the same effect as shown already for the usual Markov approximation of the autoregressive process in sec. 6.2.2.1. For lowest resolution the correlation sum is one, entropies and statistical errors are zero or almost zero. Hence an interpretation of the result for coarsest resolution is not carried out.

Under restriction of the maximal cardinality $|\mathbf{J}|$ of the conditioning set not to be larger than 4 and with this renouncing on an estimation of the model-independent constant $H_{1|\infty}(\epsilon)$ the result is shown in fig. 6.6. In this case the situation is slightly different: The development of the coarse graining part is suppressed by the restriction of the number of allowed time steps in the conditioning of the perforated Markov model. The entropy $H_{1|\infty}(\epsilon)$ unestimated and set to zero, a lower threshold of the estimation of all conditional entropies is missing especially for the smaller ϵ where the fluctuations of the entropy estimations to smaller values lead to fluctuations in the choice of the optimal perforated Markov model. It is remarkable that in spite of those qualitative differences in comparison with fig. 6.5 for intermediate resolutions the result is identical: The model behind the dataset can be found.

The result for another more subtle autoregressive process of order seven with model input (0, 4, 6), generated for the investigation of variable dataset length, will be shown in fig. 6.13. The regimes found in fig. 6.5 are again visible especially for sufficiently long data sets and the model behind the dataset is qualitatively detectable at intermediate resolutions.

Instead of indicating only memory-containing components e.g. the program 'ar-model' of TISEAN [40] also estimates the coefficients and makes statements concerning the weight of the memory in the components of the past. This property is reached by being applicable only for linear stochastic processes. The advantage of the introduced tool is that it is not only usable on data sets with known nice properties as e.g. linearity for the AR process, but it is applicable on every dynamics with potential memory effects, especially also for nonlinear dynamics, since it is based only on information theory. This is analyzed with the following example.

6.4.2 Generalized Hénon map

6.4.2.1 Dynamical law, conditional entropies and optimal perforated Markov models

In [4] it was introduced a generalized Hénon map given by

$$y_{n+1} = a - y_{n-K+2}^2 - c y_{n-K+1} \ . \tag{6.8}$$

This map contains longer memory than the usual Hénon map of eq. (3.56) , which is obtained from the generalized Hénon map in the case of $K = 2$ from the transformation $y = ax$, $a = \alpha$ and $c = -\beta$. The nonlinearity still arises from one single quadratic term. The coefficients are chosen to be $a = 1.76$ and $c = 0.1$. From comparison of the coefficients 1 vs. c of the non-constant terms of eq. (6.8) it is possible to see that the linear term is suppressed in importance.

In fig. 6.7 it is verified that the squared term contributes most to memory. The memory in the first time step of the past in fig. 6.7 could be explained by modifying a shift of eq. (6.8) to

$$y_{n-K+1}^2 = a - y_n - c y_{n-K} \ . \tag{6.9}$$

Inserting this back into eq. (6.8) yields

$$y_{n+1} = a - y_{n-K+2}^2 - c(\pm\sqrt{a - y_n - c y_{n-K}}) \ , \tag{6.10}$$

which shows the explicit dependence on the presence no matter which sign is chosen. Since the usual nonperforated conditional entropies are constructed in a way that multiple mutual information of the future time step with the past is shifted as much as possible to the presence, the explicit dependence on the presence can be seen in the conditional entropies of fig. 6.7 even though the map alone seems to indicate something different.

In fig. 6.8 the results for the optimal perforated Markov model of the generalized Hénon map with delay $K = 5$ and different maximal cardinality of the conditioning set are given. Working with increasing maximal cardinality of the conditioning set makes the detection of

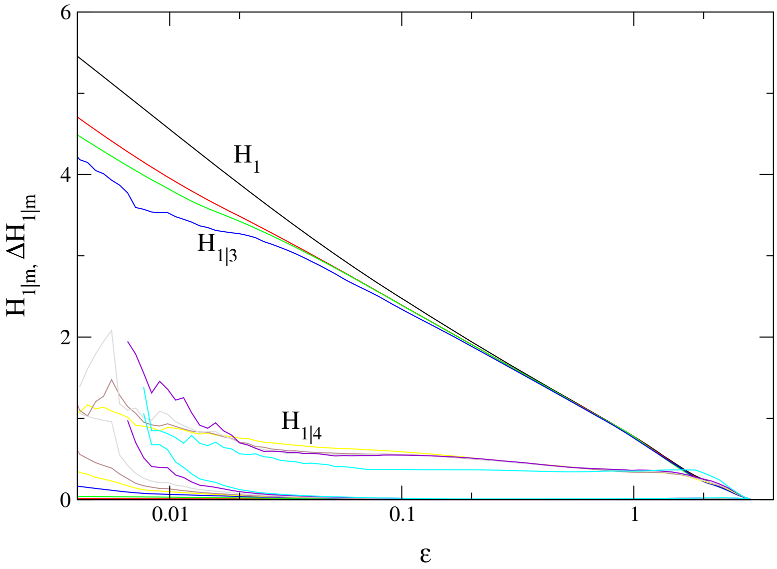


Figure 6.7: Conditional entropies and the statistical errors in their estimation of the generalized Hénon map for delay $K = 5$.

the succession of importance of the conditioning time steps possible. However, the optimal perforated Markov model according to the given criterion is only obtained after saturation of the chosen conditioning with further increased maximal cardinality of the conditioning set.

In case of maximal two conditioning time steps the result is that at least for small ϵ the immediate memory structure of the map is found and the important time step of conditioning at 'delay minus one' from the future time step is securely selected. The reason for the change of importance of the second conditioning time step can again be found in coarse graining effects. For maximal cardinality $|\mathbf{J}| = 4$ the immediate memory of the map is found, but also other time steps are selected. The reason for further memory than expected from the structure of the map can possibly be traced back to the non-existence of Markov partitions as already argued for the usual Hénon map in sec. 6.2.2.2. Finally it is possible to see in the lower panel of fig. 6.8 that for six allowed conditioning time steps, not the maximal number of time steps is fully exhausted for every resolution. This shows that the criterion selects instead of filling the possible sets of conditioning in an arbitrary sense. A coarse graining regime for coarser resolutions becomes again visible.

To conclude, for chaotic deterministic dynamics the statistical errors in the estimation of entropies for comparable dataset length are much smaller than in the stochastic case. This results from the in general rather restricted extension of the attractor, which is a rather low-dimensional object in phase space and hence from the increased probability of finding neighbors in the algorithm for the correlation sum. The criterion for the selection of the optimal perforated Markov model is dominated by the informational term for a large interval of resolutions extended towards smaller ϵ. Fluctuation effects for small ϵ in the search for the optimal perforated Markov model are thus strongly suppressed in the deterministic case. As a consequence, there is a tendency to select much more conditioning time steps in the optimal perforated Markov model with smallest informational contribution and this is what is found in the case of the generalized Hénon map.

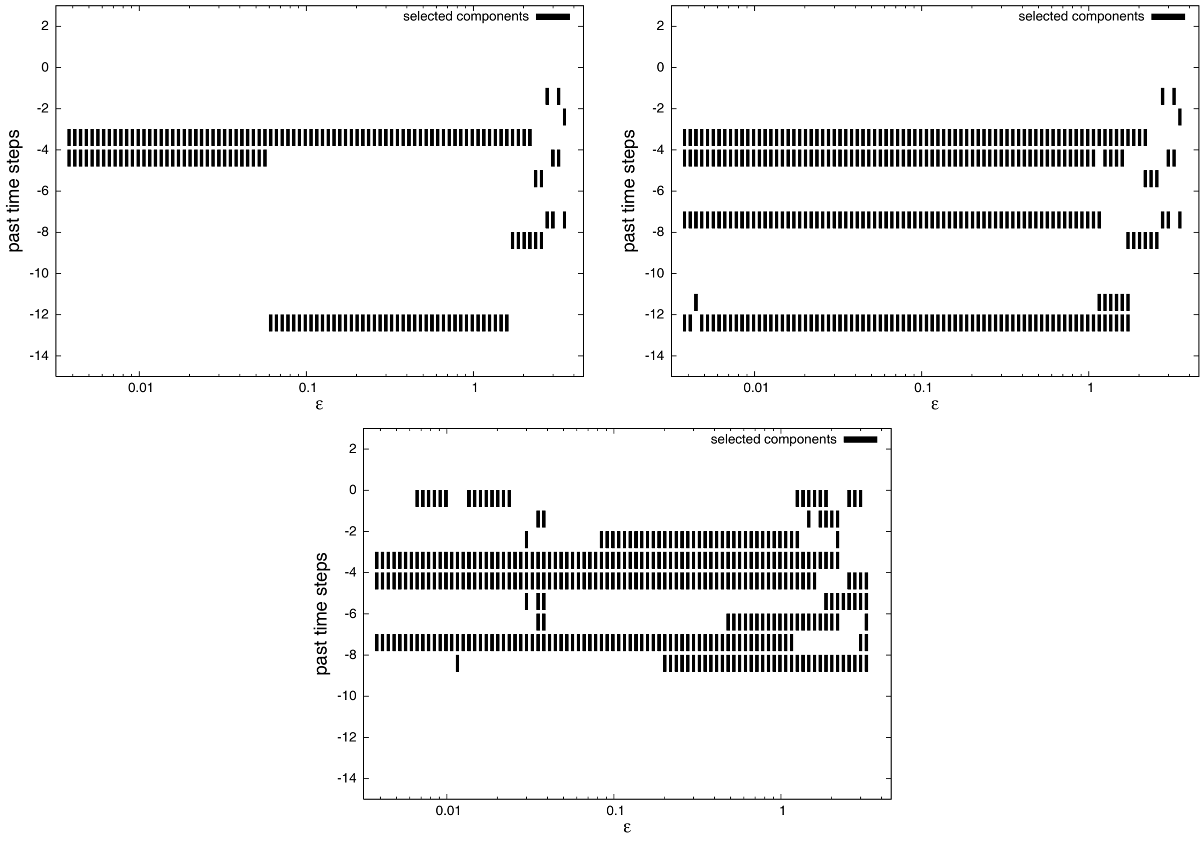


Figure 6.8: Resolution-dependent optimal perforated Markov models for the generalized Hénon map with delay $K = 5$. Upper left panel: Maximal cardinality of the conditioning: 2, dataset length: 50000. Upper right panel: Maximal cardinality of the conditioning: 4, dataset length: 50000. Lower panel: Maximal cardinality of the conditioning: 6, dataset length: 25000.

6.4.2.2 A priori omission for far memory

The only past time steps taken into account for the following calculation are: 0, -1, -2, -4, -6, -8, -11, -14, -17, -21, -25, -29, -34. This means, a strategy is followed, where in the past close to the presence a higher density of time steps is taken into account, which are tested for potential memory. Knowing the exponential increase of needed computational time with the number of time steps in the past this strategy supports reaching the further past concerning detection of memory with the accepted disadvantages of smearing effects and decrease of accuracy.

Fig. 6.9 shows that for various rather large delays of the generalized Hénon map there is an island for intermediate resolutions where the order of the delay is found. If the delay of the dynamics is not element of the set of time steps in the past taken into account as in the cases of delay $K = 20$ and $K = 30$, the detection of the delay is nevertheless possible, because the multiple mutual information of neighboring past time steps with the future has to be assumed as rather high. In all cases the output gives a good hint for a rough estimation of the correct delay depth.

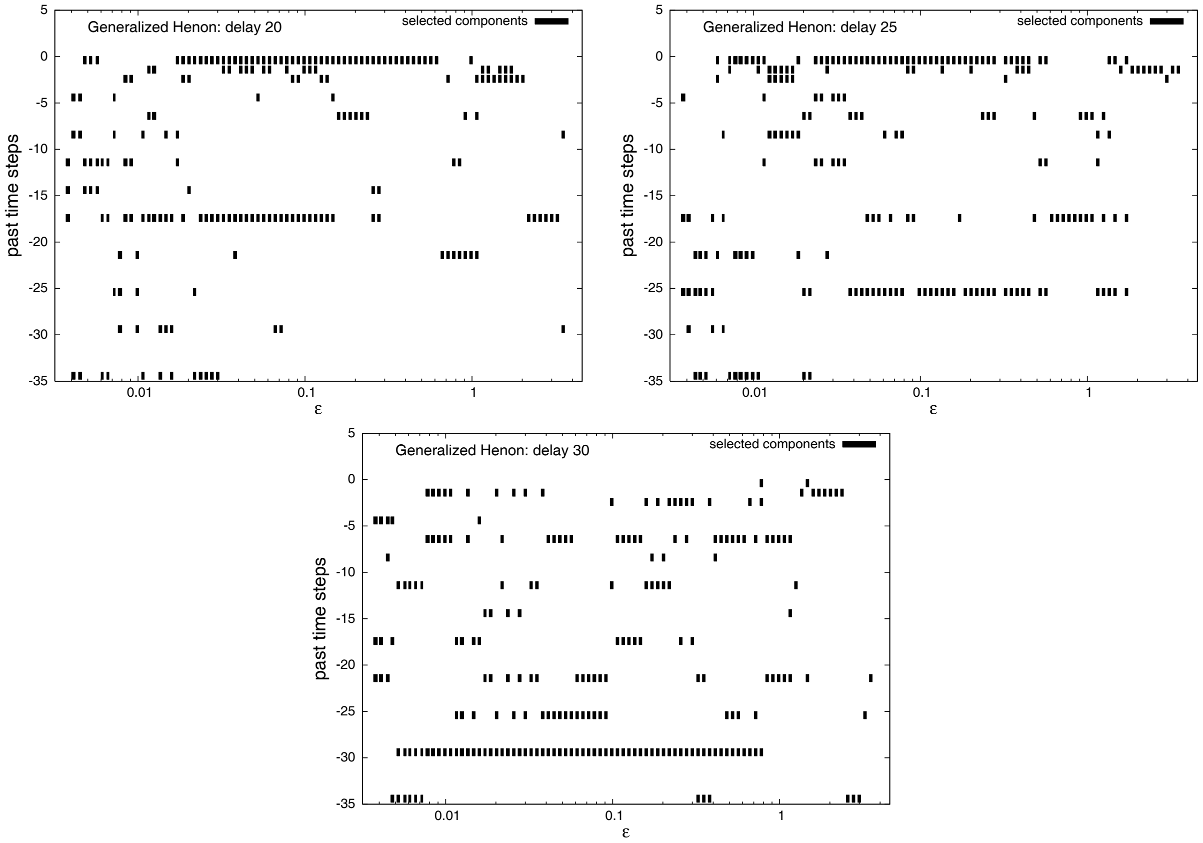


Figure 6.9: Resolution-dependent optimal perforated Markov model for the generalized Hénon map for delay $K = 20, 25$ and 30 under the constraint of (hopefully reasonably) a priori omitted past components.

6.4.3 Mackey-Glass dynamics

A special time-continuous deterministic dynamics with explicit memory is the Mackey-Glass dynamics given by

$$\dot{x}(t) = \frac{ax(t-\tau)}{1 + [x(t-\tau)]^c} - bx(t)\ . \tag{6.11}$$

The state at time t depends explicitly on the state at time $t - \tau$. Mackey-Glass dynamics is a representative of the class of delay differential equations, a subset of the set of infinite-dimensional dynamical systems. It serves as a model for the regeneration of white blood cells for patients with leucemia. Discretized, the equation of motion reads

$$x_{n+1} = (1 - b\Delta t)x_n + \frac{ax_{n-k}}{1 + x_{n-k}^c}\Delta t \tag{6.12}$$

with the delay

$$k = \frac{\tau}{\Delta t} \in \mathbb{N}\ . \tag{6.13}$$

Typical parameter values ([35], [21]) are

$$a = 0.2\,, \qquad b = 0.1\,, \qquad c = 10\ . \tag{6.14}$$

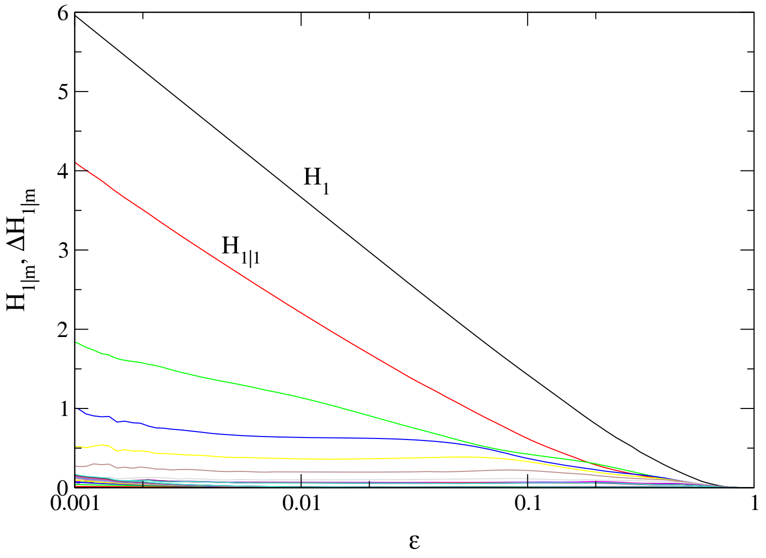


Figure 6.10: Conditional entropies of the Mackey-Glass dynamics with effective delay= 6

As an example taking $\Delta t = 0.01$ time units, a delay of e.g. $k = 1800$ time steps leads to a time delay of $\tau = 18$ time units. For $\tau > 16.8$ time units it is known that the dynamics is essentially chaotic. Using every 300th time step in the dataset to analyze, an effective delay of 6 time steps is on hand. For the following analysis, data sets of 12000 effective data points are used. Because of the underlying rule of multiple mutual information shifts as close to the presence as possible, in the plot of usual (nonperforated) conditional entropies (fig. 6.10) the delay is invisible. The reason for this in comparison with the visibility of the delay in the example of the generalized Hénon map could be found in the smoothness of flows opposed to maps.

Even though in fig. 6.10 the delay is not visible, the entropic-statistical criterion (6.3) selects it. This is seen in fig. 6.11, where a whole series of optimal perforated Markov models for different effective delays is shown. The right part of the panels is again subject to coarse graining effects. For higher resolution more structure is visible. The most important point to stress is that all panels have in common that there is an interval of resolutions, where the optimal perforated Markov model contains omissions behind the first step of conditioning and the first following time step taken into account is exactly the time step corresponding to the effective delay of the dynamics. The presence is always securely selected, because of the x_n-term in eq. (6.12).

Knowing the delays behind the data sets, I would claim that it is possible to see the effects in fig. 6.11. Otherwise having an arbitrary dataset, I would not claim to be able to make reliable statements on the properties behind the dataset under those 'effective' conditions. Further optimization and fine tuning will finally be necessary for real practical application of the algorithm.

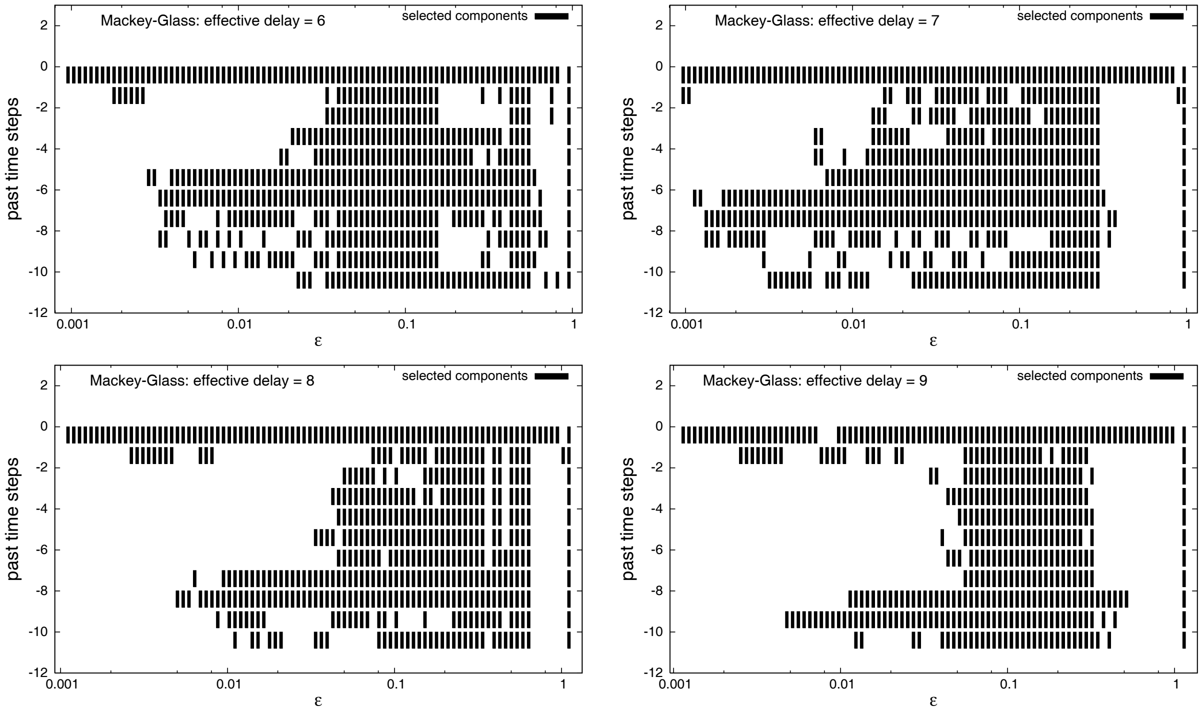


Figure 6.11: Resolution-dependent optimal perforated Markov model for the Mackey-Glass dynamics for different effective delay.

6.5 Application for different parameter values in the perforated case

6.5.1 Variable future lead time

In the framework of prediction it is questionable if for prediction of the further future it is useful to take into account the further past or if it is best to predict only with the knowledge close to the presence, i.e., in the language of time series analysis it is asked for the necessity of larger embedding windows for the prediction of the further future. The idea comes from the fact that most information for prediction of a quantity at some time scale in the future could rather depend on quantities in the past being a time distance of the same time scale apart. In the framework of wind speed prediction in [43] we proposed "Preliminary studies suggest to increase the embedding window m when increasing the prediction horizon (lead time) h." without proving it.

For an analysis of the question stated it is used an AR(3) process with chosen parameters $a_i = 0.3; i = 0, 1, 2$ for the generation of a dataset. Varying the time into the future it is now asked which components in the past fulfil best the chosen criterion and contain therefore most information about the respective time step of the future under statistical demands.

Taking first one time step into the future, fig. 6.12 shows that the model behind the dataset is found for a broad range of resolutions. The fluctuations for high resolution are caused by the larger fluctuations in entropy estimation which influence the result, if the criterion with $H_{\{f\}|\mathbb{Z}_0^-}(\epsilon)$ set to zero is used. This is necessary for the loop with restricted

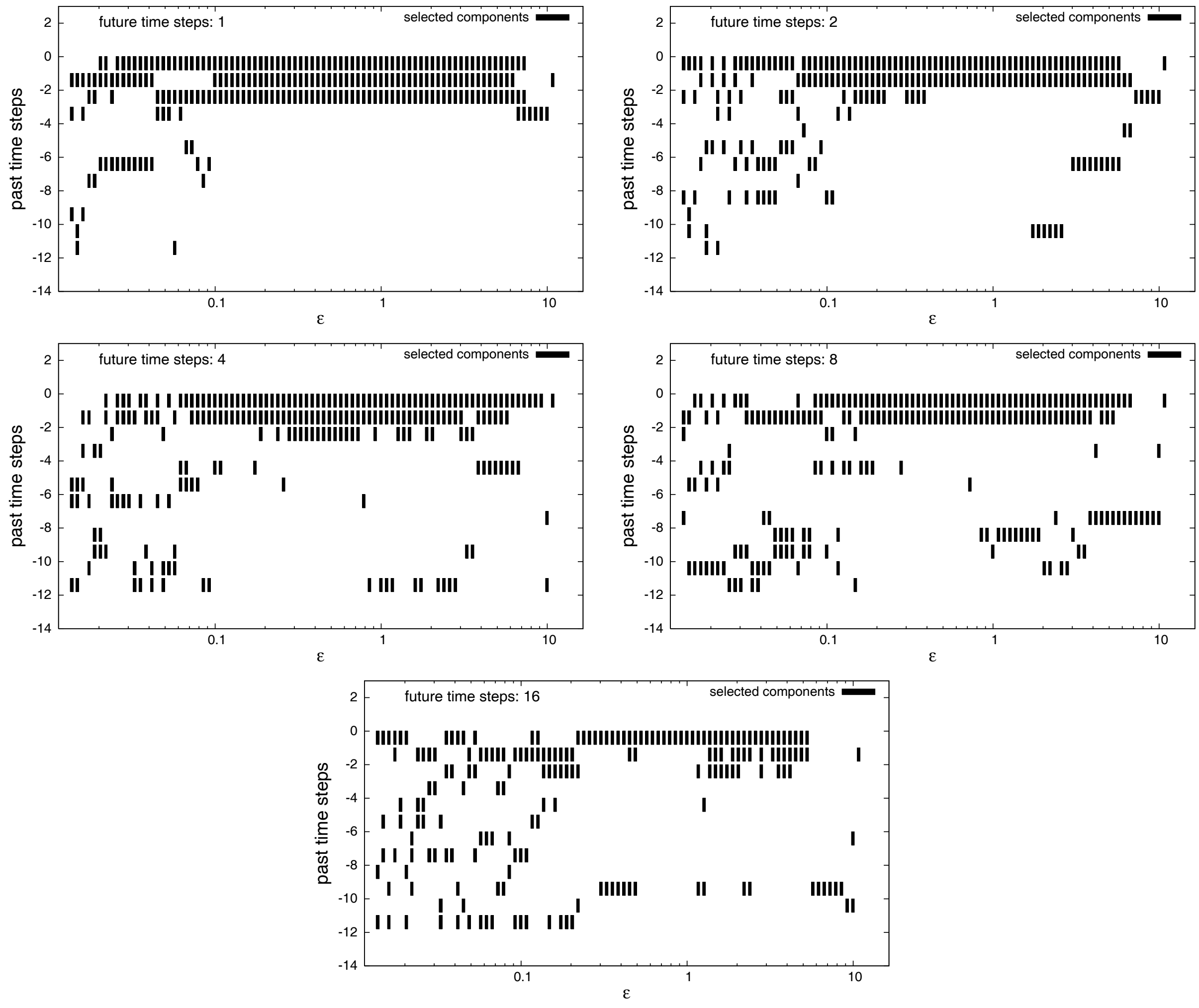


Figure 6.12: Change of the optimal perforated Markov model of an AR(3) process ($a_0 = a_1 = a_2 = 0.3$), if the future time is increased from 1 to 16 time steps.

number of conditioning time steps in the perforated Markov model.

Retrieving the model more and more breaks down the farther in the future the time step lies, for which uncertainty reduction is looked for. The numerical results tend to validate from information theory the proposal of [43] given at the beginning of this section, at least for stochastic dynamics. Especially the panel with 16 time steps ahead gives this hint.

6.5.2 Dataset length dependence of optimal perforated Markov model

From fig. 6.13 one can conclude that in the case of longer data sets the time steps of memory in the used dynamics can be retrieved on a broader interval of resolutions with higher reliability. For shorter data sets the influence of the statistical error in the criterion (6.3) increases and the domain of dominance of the information term is shifted to coarser resolutions seen in the selection of fewer components for the optimal model, where the statistical error term is dominant. The structure of conditioning of the underlying dynamics becomes blurred, if the domain of dominance of the statistical error starts to touch the domain of coarse graining effects for sufficiently short data sets. In the bottom right panel of fig. 6.13 this case is almost reached.

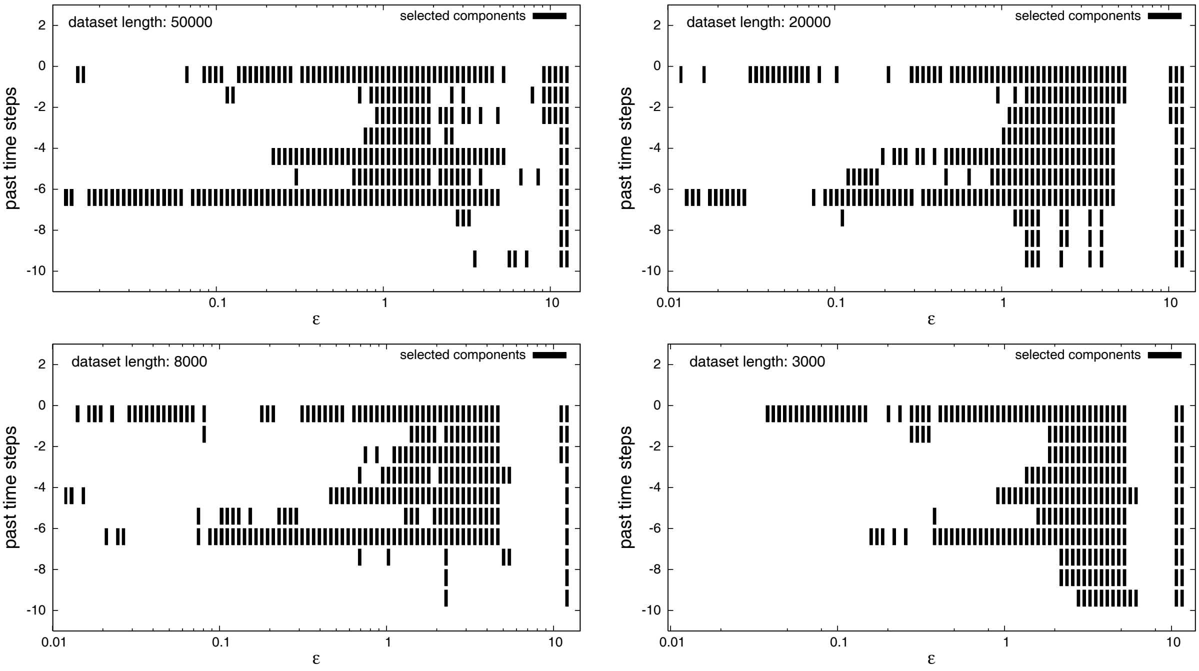


Figure 6.13: Resolution-dependent optimal perforated Markov model of an AR(7) with coefficients $a_0 = a_4 = a_6 = 0.3$ under changed length of data sets. The results are obtained from a full loop over all possible conditionings restricted only by the maximal number of 10 past time steps.

6.5.3 Optimal perforated Markov model in hypothetical case of infinite length of the dataset

In the case of infinite dataset length the statistical error becomes zero for all resolutions. If memory ranges infinitely far into the past, then a Markov approximation is always accompanied by a loss of information. According to the criterion (6.3) a Markov approximation of finite order can thus not be selected as optimal.

If the range of memory is finite into the past, a Markov approximation is possible where no information is found in the further past, but it would not be necessary, because components of the past without information about the future nevertheless kept do not diminish the quality of the model with respect to the first part of criterion (6.3) in the case of infinite data sets. The second part of the criterion given in eq. (6.5) decides for the shortest conditioning in the set of degenerated selected perforated Markov models.

6.6 Iterative procedure for the construction of the optimal conditioning and improved estimation of the KS entropy

In the methods of complete scanning of variants of conditioning in former sections the needed computational time increases exponentially as 2^m with the number m of past time steps taken into account. Also the improvements by restricting the maximal cardinality of the conditioning set $\mathbf{J}$ or the a priori omission of certain past time steps do not improve the problem of computational time qualitatively. The need for an alternative arises and one suggestion is presented in this section. Two main changes compared with formerly suggested methods are carried out. First, an iterative character of the search for single time steps instead of full scanning of all variants of conditioning is chosen, which is the central reason for the reduction of numerical costs. Second, integrating entropies for all available resolutions, what is only possible because of the iterative character, yields one resolution-independent optimal perforated Markov model.

6.6.1 Presentation of the method

This method aims for iteratively searching for the single component in the past, which reduces most the uncertainty in the chosen future time step, given the uncertainty reduction already obtained from the previously selected components of the past, i.e., it is searched for a component which maximizes the difference of conditional entropies under additional conditioning.

Since it is decided to evaluate the difference of the conditional entropies integrated over all available resolutions, the statistical error behind the entropy estimates varies, a fact already known from formerly introduced methods. The possibility of strong underestimation of entropies for higher resolution is the reason, why a suitable weight function involving the statistical error has to be created, which penalizes contributions from entropy estimates with a larger error. The problem is illustrated in fig. 6.14. It is possible to see that under conditioning on steps 0,2,5 the estimated entropy indicates wrongly a largely reduced uncertainty and would be preferred for non-normalized integrated entropy-differences. In general for smaller differences of entropies in the situation of higher conditioning the fluctuations of entropy estimations for smaller ϵ are a really disturbing effect concerning the selection of the best information containing components.

From the previous demands as a selection criterion for the determination of the new best component m_{new} it is suggested[4]:

$$m_{new} = m_{try} \quad \text{if} \quad \sum_{\epsilon} \max(0, \frac{H_{\{f\}|\mathbf{J}_{opt}}(\epsilon) - H_{\{f\}|[\mathbf{J}_{opt} \cup \{m_{try}\}]}(\epsilon)}{c + (1 + a \cdot \Delta H_{\{f\}|[\mathbf{J}_{opt} \cup \{m_{try}\}]}(\epsilon))^b - 1}) = \max \ . \tag{6.15}$$

Having found the suitable component, it is fixed for all next steps. Empirically chosen parameters are

$$a = 4 \ , \quad b = 4 \quad \text{and} \quad c = 0.2 \ . \tag{6.16}$$

---

[4] The sum is taken over a suitable set of ϵ-values, which are logarithmically equidistant.

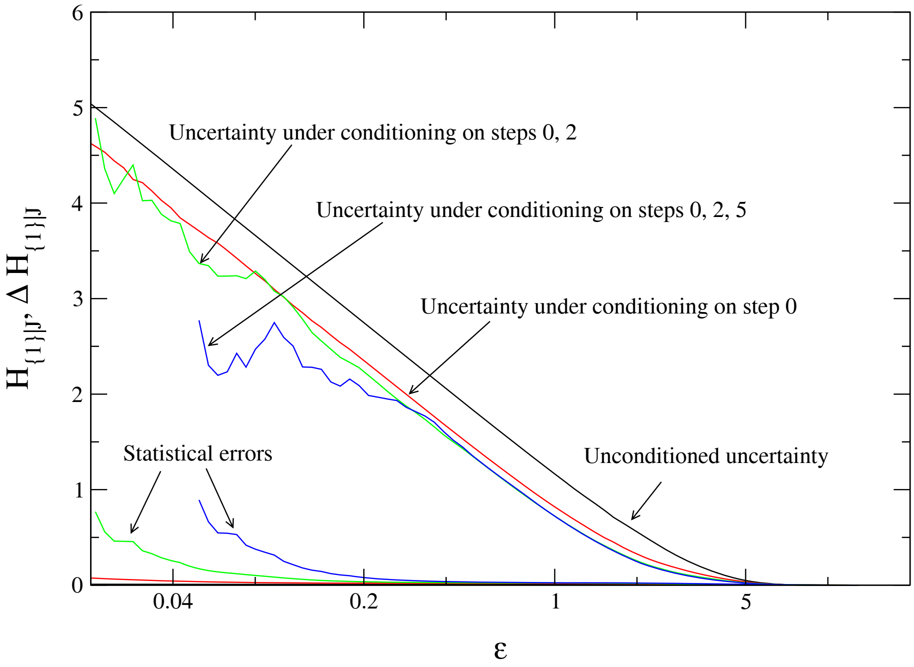


Figure 6.14: Illustration of the problem of fluctuations in entropy estimation for the case of a dataset of length 2000 for AR(3) dynamics with parameters $a_0 = a_1 = a_2 = 0.3$.

By the choice in eq. (6.16) it is tried to adjust the parameters a and b to suppress the disturbing effect of the fluctuations. Since the statistical error of the entropies ΔH tends to zero for the largest ϵ it is necessary to introduce an offset with $c > 0$ for the denominator to prevent the entropies for really coarse resolution to dominate the result. Theoretically the numerator with the entropy difference can only be larger than or equal to zero. Hence the maximum-function is included to eliminate parts of the fluctuation effects. If the estimation of the Renyi entropy with largest conditioning diverges because of vanishing correlation sum for certain resolutions, the contribution to the sum in (6.15) is a priori excluded.

Discussing in terms of mutual information, for the first selection of a component according to the numerator of the given criterion it is searched among all components of the past between the presence and a chosen maximal past time for the one which has the largest double mutual information with the future time step of interest. The next component is found by searching for the maximum of the difference of the double mutual information of the future time step with a variable past time step and the triple mutual information of the future time step, the formerly fixed past time step and again the variable past time step. In other words: It is searched for the maximum of the conditional mutual information of the future time step and the variable past time step conditioned on the formerly fixed past time step. In the third step it is searched for the maximum of the double mutual information of the future time step with a new variable past time step minus the sum of both triple mutual informations of the future time step, the variable past time step and a formerly fixed component of the past plus the quadruple mutual information of the future time step, both fixed components of the past and the variable past time step. Continuing in this way with every fixed component of the past the order of the multiplicity of involved

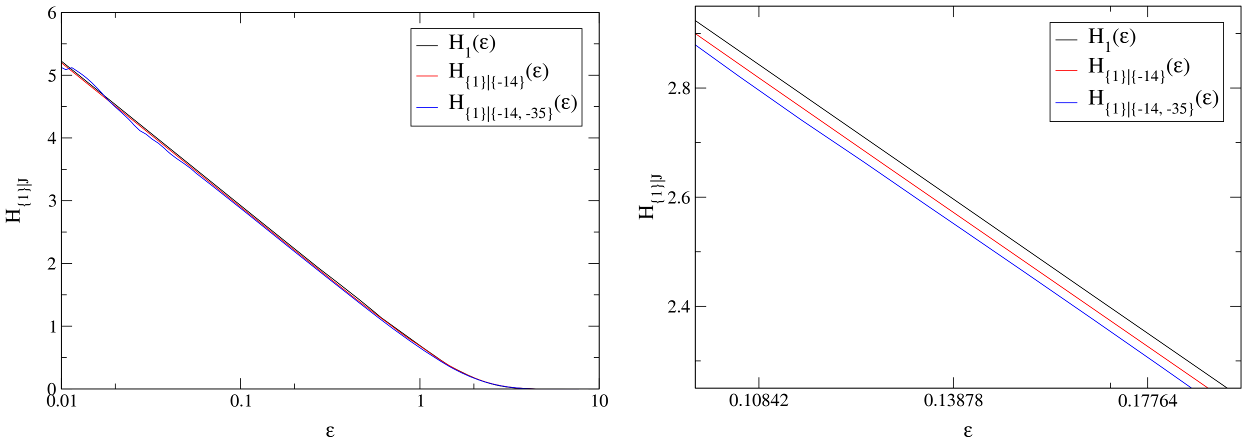


Figure 6.15: Perforated conditional entropies after successive conditioning on the most important found components in a dataset of a perforated AR process. The right panel is a zoom of the left panel.

mutual informations increases.

Knowing that the further contributed uncertainty reduction is a decreasing function, it is possible to formulate a criterion for which this procedure can be finished, ending up with the components of the past containing most information about the future.

The parameters needed by the algorithm are the maximal past time steps (could be almost data file length) and the parameters in the criterion function. The maximal allowed length of conditioning could be given.

6.6.2 Examples

6.6.2.1 AR process

As an example a dataset of length 24000 from an AR(35) process with parameters $a_{14} = a_{35} = 0.2$ and $a_i = 0$ otherwise is treated. It is important to notice that the parameters are taken comparatively small.

The suggested algorithm finds the relevant components and gives the entropic output in fig. 6.15. Even though the entropy differences and henceforth the memory effects are rather small, the algorithm was able to detect the memory containing components, noticing that the entropy estimations of the other components underly also some fluctuations.

6.6.2.2 Generalized Hénon map

The second example treated is the generalized Hénon map of eq. (6.8) with two different values of the delay ($K = 10$ and $K = 20$ time steps). The chosen maximal number of past time steps is 40 and the length of the dataset is 50000. The conditional entropies of the selected components are shown in fig. 6.16. As a result it is seen that as expected the conditioning components according to the delay and their weight in the map are selected. Furthermore multiples of the delay are chosen for uncertainty reduction. A similarity to the results of sec. 6.4.2.1 should be noticed. For longer conditioning and high resolution the

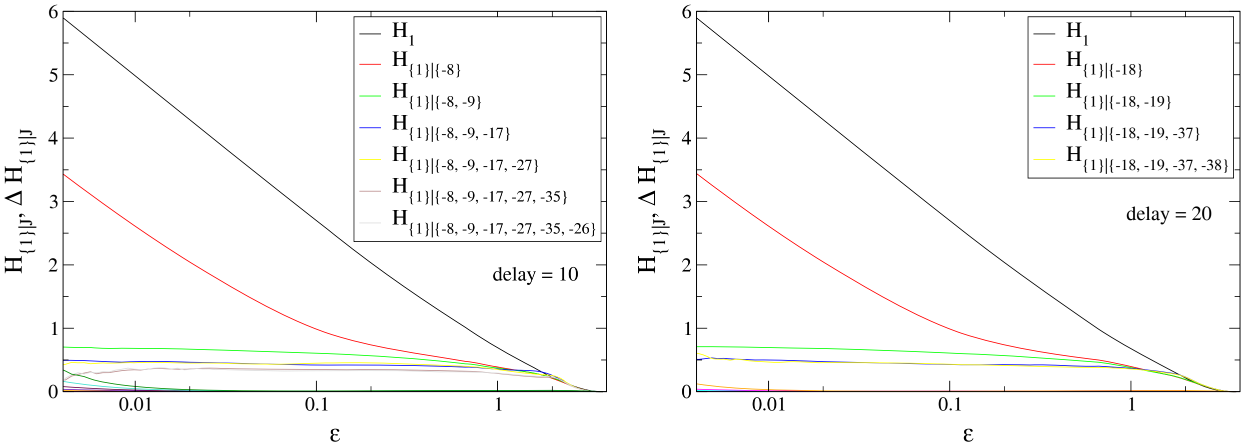


Figure 6.16: Perforated conditional entropies after successive conditioning on the most uncertainty reducing components in data sets of generalized Hénon dynamics with different delay. In the left panel for delay = 10 the succession of selected past components is -8, -9, -17, -27, -35, -26. In the right panel for delay = 20 the succession of selected components is -18, -19, -37, -38. (It is important to notice that the time step in the immediate future is denoted as '1' and the delay of e.g. 10 corresponds to time step '-9'.)

statistical error starts to become relevant even though the best perforated Markov model is chosen.

6.6.3 Discussion of the method

The advantage of this idea consists in the fact that in principle with finite effort of time every memory depth is detectable and the possibility of missing far memory is excluded for every finite dataset. This is the central reason why this method gives an improved reliability in the estimation of $H_{1|\infty}$ ($H_{\{f\}|\mathbb{Z}_0^-}$ for variable future) or of the Kolmogorov-Sinai entropy in the case of dynamical systems.

Previous methods for estimating $H_{1|\infty}$ used nonperforated entropies and there was no possibility to estimate entropies for conditioning on more than the last few time steps and in this sense there was no possibility to be sure to have found the final value of $H_{1|\infty}$, because information in the further past about the outcome of a random variable of the future was not detectable with this method. With the method suggested in sec. 6.6 this changes and for all(!) finite data sets with the help of arbitrary perforation there now exists a tool for excluding the possibility of missing the existence of reasonable information for prediction in the further past and for making really reliable estimations of the Kolmogorov-Sinai entropy from entropy-estimation for arbitrary dynamics behind data sets.

The essential components of the past carrying memory are detected first and a much faster speed of convergence to the limit case is reached if compared to the nonperforated case. In comparison with the full loops for search of optimal perforated Markov models the computational time resources increase only linearly with the number of past time steps taken into account. The integration over all possible resolutions in the presented method causes more robustness against fluctuations in the estimation procedures compared to the resolution-dependent case.

One of the disadvantages of this method is the fact that having selected a single component it will be kept in all next steps even though it would perhaps not have appeared, if having searched collectively from the beginning. Another disadvantage is that the truncation criterion for finishing the algorithm is purely empirical until now. Furthermore this method is not anymore space (ϵ) resolving as compared to the methods presented before. However, with respect to prediction this has not necessarily to be seen as a disadvantage.

6.6.4 Delay determination and search for feedback time

As a special case this program can be used as a delay finder in the sense of information theory for arbitrary delay dynamics. This serves as an alternative to what is presented in [26], where singular value decomposition is used for delay detection. The cited paper supports the notion of a 'smart embedding' instead of 'generalized embedding'.

The arguments in the paper [73] are quite similar from the point of view of information theory to the above presented method of delay finding, but since the method in this work is resolution integrating, it should be more robust against fluctuations in entropy estimation.

6.7 Comparison with the literature

In the following sections a short excursion on statistical model selection is undertaken. One of the central problems is the avoidance of overfitting. The following criteria yield ideas how to prevent it and act somehow in the same sense as the criteria based on information theory and statistics introduced at the beginning of this chapter. Hence they are illuminated here. The criteria are looking for a compromise between the best goodness of fit and the minimal model complexity. They follow Occam's razor to prefer the shortest hypothesis and obey the principle of parsimony, not a physical principle in the narrow sense.

6.7.1 Information criteria

6.7.1.1 An information criterion (AIC)

The question of model selection was adressed first by Akaike in [2]. The criterion was formulated via penalization of the best fit via the maximum likelihood method by the number k of parameters of the model:

$$AIC = -2 \ln p_{\{\hat{a}_j\}}(\{s_n\}) + 2k = \min \ . \tag{6.17}$$

Given a time series $\{s_n\}$ the parameters $\{\hat{a}_j\}$ and their values are inferred as a maximum-likelihood-estimation, i.e., the parameters $\{\hat{a}_j\}$ are chosen such that $p_{\{\hat{a}_j\}}(\{s_n\})$ is maximal. This is equivalent to a best fit of the probability distribution to the given data in the sense of properties of maximum likelihood estimators. From the monotony of the logarithm the best fit of a distribution to the data makes the first term minimal. Very important to understand is that the first term becomes the smaller the bigger $|\{a_j\}| = k$ because then p increases. This justifies the need for some penalization term.

As a first ad hoc trial AIC may be treated as the naive historical starting point of model selection, which nowadays became improved by other criteria. It is commonly agreed that model complexity is not penalized enough by AIC such that models with too many parameters are deduced to be optimal. AIC is not fully accepted by professional statisticians, even though widespread in practical use.

6.7.1.2 Minimum description length (MDL) and Bayesian information criterion (BIC)

An improvement of the AIC criterion is given by the MDL criterion founded by Rissanen [63]. The idea behind MDL, originating in algorithmic coding theory, is that regularity can be used to compress data, an idea analogous to the complexity measure 'algorithmic complexity'. A special case of this criterion is the Bayesian information criterion (Schwarz):

$$BIC = -2 \ln p_{\{\hat{a}_j\}}(\{s_n\}) + k \ln N = \min \quad . \tag{6.18}$$

Here again k is the number of parameters of the model and N is the number of data points. From multiplication of the number of parameters k with the logarithm of the dataset length it is inferred that the BIC criterion penalizes the model complexity stronger than AIC does, what is in better agreement with example dynamics. Another advantage of BIC against AIC is that other estimator-properties than consistency usually are not proven for the criteria, but BIC is proven to be optimal in the Bayesian sense of minimizing the error probability.

6.7.2 Comparison of AIC and MDL with the newly introduced criterion

AIC does not care for the dataset length which is not acceptable. AIC as well as BIC are criteria proposed in a framework of linear processes. Their applicability to nonlinear processes for probably good reasons is not discussed in the literature. The new criterion presented at the beginning of this chapter works quite well for nonlinear processes.

Furthermore the newly introduced criterion gives a resolution-dependent optimal model, whereas AIC and BIC do not allow for any resolution dependence.

The BIC criterion has a penalizing term increasing with the number of parameters and increasing with the length of the dataset. In criterion (6.3) increasing the number of parameters increases the penalization implicitly by an increase of the statistical error, but in contrast to BIC increasing the length of the dataset leads to lower penalization.

To conclude, similar ideas lying behind the introduced model selection criteria in the sense of avoiding overfitting were already applied in former contexts. The entropic contribution to the newly introduced criterion has some analogy to the maximum likelihood term in the sense that a maximum amount of information taken along corresponds to an optimal fit or stating it the other way around: Ignoring some memory is somehow analogous to a decrease of the maximum likelihood term (including the negative sign) by changing from optimal parameter values to non-optimal ones. As a penalization term the statistical error of the redundancy acts as a counterpart to the penalization terms with the number of modelling parameters (with above mentioned differences). Even though the general idea behind the model selection is somehow similar, the precise realization of the criterion is quite different.

The penalization terms in AIC and BIC are introduced in an ad hoc-sense, whereas the penalization in the newly introduced criterion arises via the statistical error of entropies in the author's opinion in a more natural sense and the central penalization objects k and N are included implicitly in the new approach.

6.7.3 Comparison of methods based on functional independence with the newly introduced criterion

The paper [57] is concerned with the selection of components for constructing a best structure of generalized delay vectors with the intention of optimal reconstruction. The authors Pecora et al. develop a unified approach of joint optimization for time delays and the embedding dimension in a framework of multivariate data (not present in this work), large time delays and embedding dimensions as well as multiple time scales. It is complained about heuristical and empirical arguments in the usual method of time delayed embedding for attractor reconstruction, which is in principle in the same spirit as the ideas presented in this chapter. The difference of [57] and this work is that on the one hand a strategy for optimal attractor reconstruction whereas on the other hand a strategy for finding an optimal Markov approximation is developed. In order to bring those tasks closer to each other it is pointed out that attractor reconstruction is not restricted to the deterministic whereas in this work generalized Markov approximations are not restricted to the stochastic case. Attractor reconstruction for dynamics with stochasticity also gets the quality of an approximation. Both tasks look for an optimal structure of generalized delay vectors, i.e. an optimal generalized embedding. Thus a comparison could be justified.

The abstract main criterion in [57] for the inclusion of new candidate components in an iterative scheme is that they have to be functionally independent of the previous components (functional independence is also the basis for the statement of Takens [71]) and the concrete implementation is realized by a continuity statistic. This work wants to suggest the idea that the functional independence could be a necessary and not a sufficient condition for the inclusion of new components into the optimal structure of generalized delay vectors. The iterative procedure for component selection suggested in sec. 6.6 on the other hand should be called 'sufficient' for selection. However, it is not clear, how far apart 'sufficient' from 'necessary' could be in examples.

The approach suggested by Pecora et al. needs two different statistics for the component selection (continuity statistic) and for the stopping rule (undersampling statistic). In contrast, the method used in this work only needs a single statistic for both tasks and hence could seem to be more 'unified' or natural. It could possibly be claimed that an approach based purely on information theory is more intuitive for component selection of the structure of generalized delay vectors than continuity statistic mixed with an undersampling statistic, but this of course doesn't say anything about the performance and efficiency of the method.

A general disadvantage of an 'inclusion one after the other'-strategy as e.g. also in hierarchical clustering or in the iterative procedure presented in sec. 6.6 is always that once a component is selected, one cannot get rid of it anymore in later stages, even though an a priori 'combined strategy' possibly would not have selected it at all.

Besides the approach of Pecora et al. in a sense of joint optimization, different variants in the spirit of functional independence restricted to the detection of only an optimal time delay, were already tried for attractor reconstruction. Examples are the search for the minimum of mutual information or the determination of the first zero-crossing of the autocorrelation function. This indicates that there is no naturally given fixed cost function to minimize for optimality of reconstruction methods, seen already within the class of methods using functional independence. Thus from until now non-widespread knowledge of the multiple mutual information of sec. 3.1.6 of this work there could exist potential for a starting point of a new spirit of reconstruction methods in an average sense not based on functional independence. If the method suggested in sec. 6.6 would be accepted, it could finally be varied and extended by reduction of uncertainties of random variables correponding to different sets of time steps in the future, because the reduction of uncertainty in the immediate future time step does not necessarily need to be the best choice.

6.8 Conclusion

Criteria for usual and in the sense of perforation generalized Markov approximations based on information theory and statistics were suggested. In the perforated case for sufficiently large dataset length and for sufficiently small ϵ (smaller than the coarse graining regime) the dependence structure of the dynamics behind the dataset is in principle correctly retrieved for arbitrary given dynamics. This indicates the functional capability of the implemented tool and serves as a posteriori justification of the introduced ad hoc criterion. The suggested algorithm yields a suitable truncation concerning memory for processes with infinite memory. Especially for higher resolutions the results confirm the intuition that the memory-structure behind the dynamical law loses optimality and it becomes adaequate not to keep the full memory indicated from pure information theory.

An iterative resolution-integrating procedure for the selection of the optimal conditioning structure is suggested and tested with example dynamics. This tool can be used for an improved estimation of KS entropy rates from information theory.

Chapter 7

Prediction

7.1 Point prediction and prediction error

General point prediction reads

$$\hat{s}_{n+1} = F(\mathbf{s}_n) \tag{7.1}$$

with a suitably chosen function F. The mean quality of predictions can be evaluated by estimating an accuracy measure. There are a lot of accuracy measures available. For this work it is chosen to take a simple one, the root mean squared (rms) prediction error given by

$$\hat{e} = \sqrt{\overline{(s_{n+1} - \hat{s}_{n+1})^2}}\ . \tag{7.2}$$

As a consequence the *mean value* of the estimated distribution of S_{n+1}, the random variable corresponding to the measured value s_{n+1}, is the optimal F. This distribution is estimated by a selected set of s_{k+1}, which are obtained from those $\mathbf{s}_k$, which are in some sense suitably related to $\mathbf{s}_n$. A decision, what a 'suitable relation' should be, is not immediately given by eq. (7.2) and has to be made additionally. Another possible accuracy measure would be the mean absolute error, which would lead to an optimal F given by the median.

Point prediction is sometimes also called 'deterministic prediction' even though from the prediction error it is seen that probabilities are not absent. The prediction error depends on the lead time[1] f, the dataset length N, possibly on the resolution ϵ and the noise ξ in the dynamical modeling F.

7.2 Locally constant prediction with generalized delay vectors

7.2.1 Specialized point prediction

A special point prediction used in the following, which is locally constant (cmp. sec. 7.3) and perforated, reads

---

[1] The notion 'horizon' instead of 'lead time' is also in use, but cannot be supported. There is a common agreement that the notion 'horizon' indicates something like the unique final limit of any visibility and is wrong in use as a shortcut for arbitrary future time distance.

$$\boxed{\begin{aligned}\hat{s}_{n+1}(\epsilon) &= \frac{\sum_{k\neq n}\Theta(\epsilon - \|P_{\mathbf{J}}\mathbf{s}_n - P_{\mathbf{J}}\mathbf{s}_k\|)\cdot s_{k+1}}{\sum_{k\neq n}\Theta(\epsilon - \|P_{\mathbf{J}}\mathbf{s}_n - P_{\mathbf{J}}\mathbf{s}_k\|)} \\ &= \frac{1}{|\{P_{\mathbf{J}}\mathbf{s}_{k\neq n} \in \mathcal{U}(\epsilon, P_{\mathbf{J}}\mathbf{s}_n)\}|}\sum_{P_{\mathbf{J}}\mathbf{s}_{k\neq n}\in\mathcal{U}(\epsilon,P_{\mathbf{J}}\mathbf{s}_n)} s_{k+1}\ . \qquad (7.3)\end{aligned}}$$

$P_{\mathbf{J}}$ is the projection operator onto the perforation structure given by the set $\mathbf{J}$ already encountered in sec. 4.3 and $\mathcal{U}(\epsilon, P_{\mathbf{J}}\mathbf{s}_n)$ is the ϵ-neighborhood of the vector $P_{\mathbf{J}}\mathbf{s}_n$. A strict resolution dependence of the prediction error

$$\hat{e}(\epsilon) = \sqrt{\overline{(s_{n+1} - \hat{s}_{n+1}(\epsilon))^2}} \qquad (7.4)$$

is taken into account, which is not common in the literature. Without focus on perforatedness and resolution-dependence, (7.4) in combination with (7.3) is also known under the name 'leave-one-out cross-validation' [38].

7.2.2 Empirical minimization of the prediction error

In the following for different example dynamics with certain given memory structure in the dynamical law the resolution-dependent prediction error is calculated for point prediction under various conditionings according to perforation structures in the sense of generalized delay vectors. The main focus is on the resolution-dependent conditioning, which minimizes the prediction error, in comparison with the given memory structure of the dynamics.

7.2.2.1 Linear stochastic example: AR(3)

In order to choose suitable parameter values of the AR(3) with restriction $a_1 = 0$, in fig. 2.2 the stability area of the process for the remaining parameters a_0 and a_2 was shown. For certain parameter values of the AR(3) the prediction error under various conditionings is shown in fig. 7.1. A minimum in the prediction error with a weak rise towards higher and a strong rise towards coarser resolutions can be seen. Somewhere close to $\epsilon \sim 0.2$ always the model behind the data is found as optimal from the prediction error. In the coarse graining region optimality is dominated by long conditioning. On the other hand, for higher resolution the decreasing number of neighbors for longer conditioning causes an increasing prediction error and a lower significance of it. Above some threshold resolution it is not anymore possible to estimate the prediction error towards smaller ϵ.

Against intuition single conditioning does not necessarily minimize the prediction error towards higher resolution seen in the second row of panels in fig. 7.1. Furthermore the components of the conditioning of the map are not necessarily the best components reducing the prediction error, if single conditioning is demanded, seen in the third row.

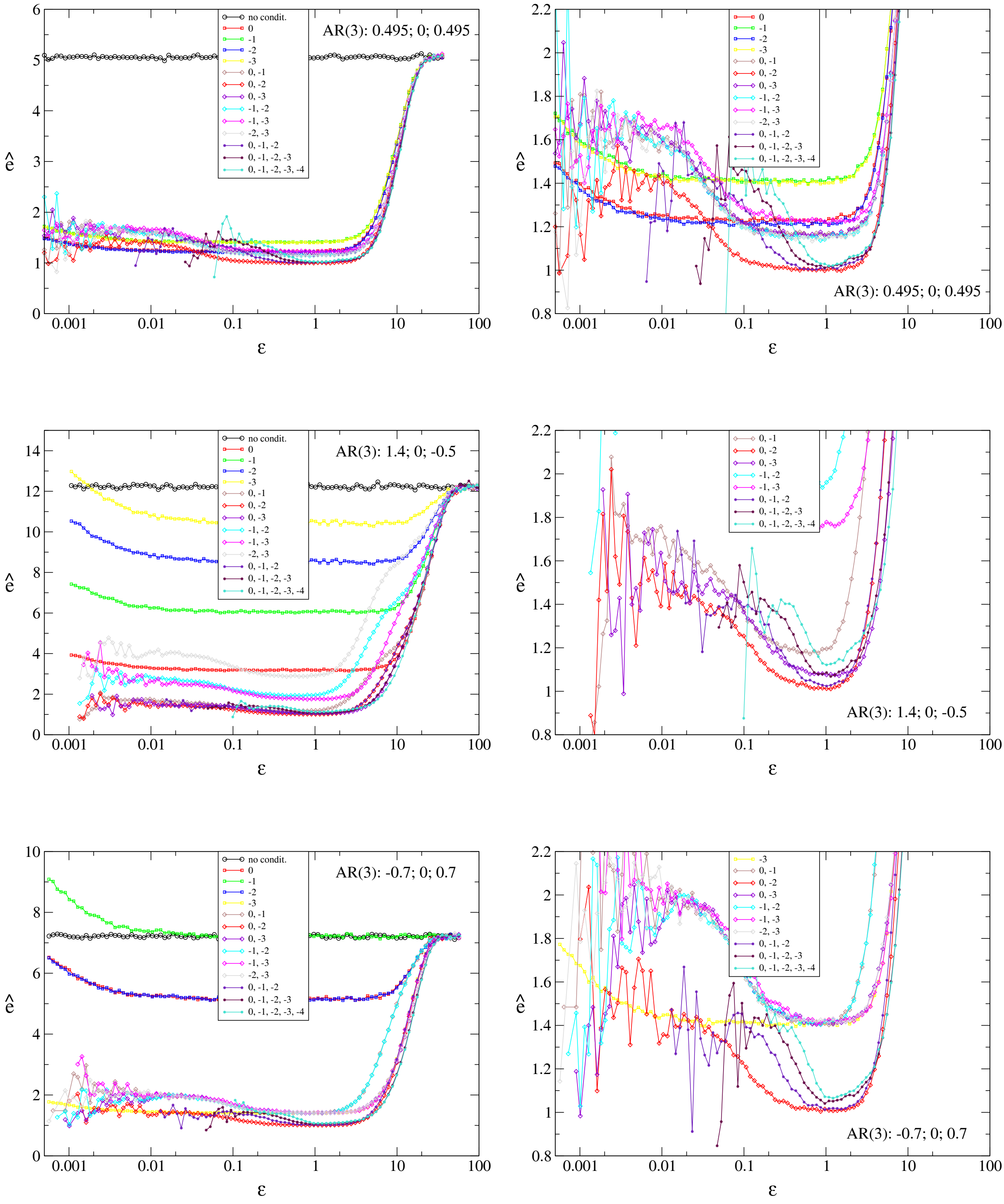


Figure 7.1: Resolution-dependent prediction error for data sets of 45000 data points of AR(3) dynamics with variable a_0, a_2 and $a_1 = 0$ fixed under various conditionings. In all cases the right panels are a zoom of the left panels.

7.2.2.2 Nonlinear deterministic example: Generalized Hénon map

In fig. 7.2 the prediction error for the generalized Hénon dynamics of eq. (6.8) is shown. Since the delay $K = 4$ is chosen, the conditioning time points in the past are '-2' and '-3'.

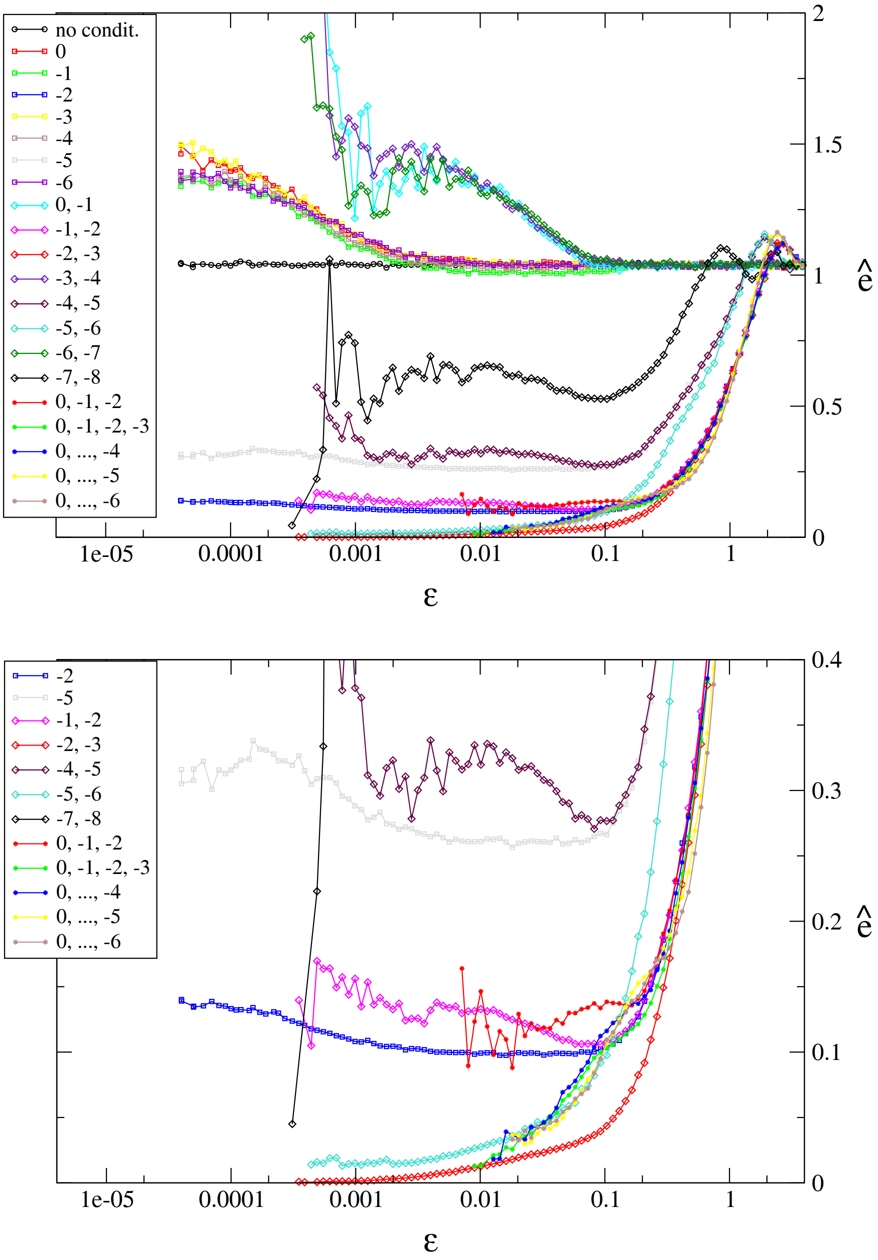


Figure 7.2: Resolution-dependent prediction error of dataset of 50000 data points of generalized Hénon dynamics ($a = 1.76; c = 0.025$) with delay of 4 time steps subjected to various conditionings. The lower panel is a zoom of the upper panel.

As in the stochastic case there is a central regime where conditioning according to the introduced model optimizes the prediction error (lower panel in fig. 7.2) and for coarser resolution longer conditioning optimizes the prediction error (upper panel in fig. 7.2). The longer conditionings become competitive again for higher resolutions (until finally no neighbors are found anymore), because the increasing fluctuations from fewer neighbors are restricted to a width increasing only linearly into the future starting from ϵ at the presence. At $(\epsilon = 0.1, e = 0.1)$ even a crossover from unrestricted increase of the prediction error in consequence of fewer neighbors to ϵ-bounded behaviour can be observed for the longer conditionings. In the framework of usual Markov approximations fig. 6.3 for the usual Hénon map shows a qualitative similarity with the behaviour found and described here empirically for the prediction error.

In the case of deterministic dynamics, the result for an optimal resolution for prediction as the absolute minimum of the prediction error is seen in the example to be the minimal available resolution for which still neighbors can be found. The phenomenon of a minimum in the prediction error for intermediate resolution as in the stochastic case is not seen. This means that performing predictions with only the closest neighbor, which in the single case of course carries the risk of making a rather large error, is the best in an average sense for deterministic dynamics.

For a longer nonperforated conditioning in the sense of Takens only a certain resolution threshold can be reached, above which no neighbors are available anymore for performing predictions. In case of perforated embedding, but still fully reconstructed dynamics, higher resolutions are reached and hence the absolute minimum of the prediction error needs perforatedness.

As in the stochastic case, fig. 7.2 shows that without conditioning the prediction error being independent of the resolution is identical to the standard deviation. The deviations of the prediction error from the standard deviation to higher values for higher resolution in the upper panel of fig. 7.2 are caused by the fact that in eq. (7.3) the mean value is not estimated anymore correctly.

In the case of incomplete reconstruction of the dynamics the ϵ-border characterizing predictions in the deterministic case (seen in the cases of long conditioning) doesn't exist and effectively predictions are performed stochastically with the possibility of increasing prediction error for smaller resolutions.

The deterministic tendency, not to prefer lowest number of conditioning components for higher resolution was already observed for the stochastic AR(3) process with parameters $a_0 = 1.4; a_1 = 0; a_2 = -0.5$ in fig. 7.1. The interpretation is that because of the large ratio of the variance of the process compared to the variance of the input noise $(\sigma = 1)$, the process is weakly stochastic and already dominated by determinism.

The results for the empirical access to the prediction error give a posteriori justification for the introduction of the concept of arbitrary perforation in the framework of delay vectors and Markov approximations.

7.2.3 Discussion on the balance of the introduced criterion

In sec. 6.3.1 a criterion for the optimal perforated Markov model was introduced with a free parameter adjusting the balance of the information term and a penalizing statistical error term. Knowing from the previous section empirically a resolution-dependent optimality of the conditioning, a comparison with the optimal perforated Markov model is expected to give hints on a suitable balance factor. Hence for purposes of comparison the resolution-dependent

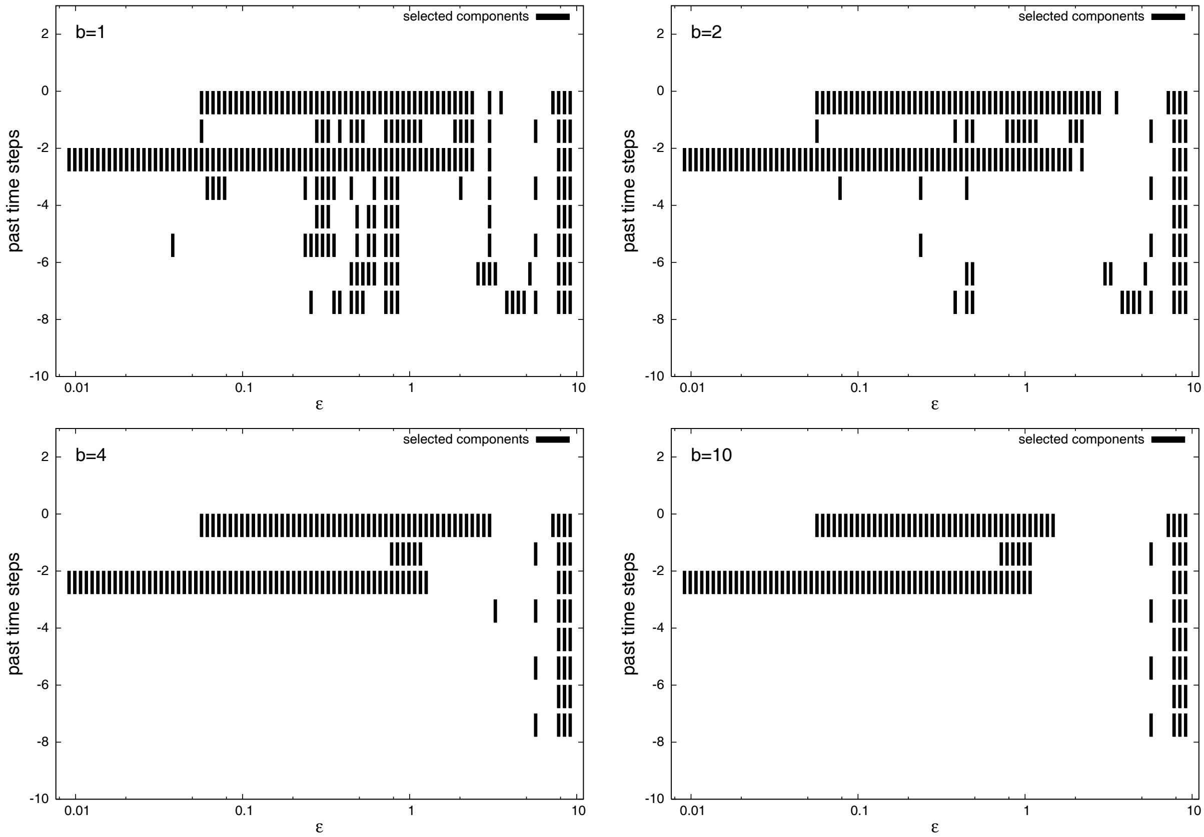


Figure 7.3: Change of the balance factor of eq. (6.3) between ignored memory and statistical error of the redundancy by $b = 1, 2, 4, 10$ for a dataset of length 45000 for AR(3) dynamics with $a_0 = a_2 = 0.3$ and $a_1 = 0$.

optimal perforated Markov model is determined for various balance factors shown in fig. 7.3.

A value for b between 4 and 10 yields concordance of the results concerning the optimal model for the chosen stochastic dynamics and for the chosen locally constant prediction scheme. There seems to be a rather robust behaviour of the optimal perforated Markov model for those b-values instead of a clear single value for one optimal b. On the other hand, from the upper right panel in fig. 7.1 there is an argument also for a smaller balance factor than $b = 4$ at least for a certain subset of the parameter space of the AR(3) processes with $a_1 = 0$.

A dependence of the optimal balance factor on the type of dynamics cannot be excluded. As an advantage of the variability of the suggested criterion (6.3), depending on the finally chosen prediction scheme a suitable value for the balance is always adjustable. Nevertheless, a maximal scientific consequence in the selection of the correct balance factor of the criterion is missing here. Rather intuitive arguments are prevailing.

7.2.4 Prediction from optimal perforated Markov model

After having found the optimal perforated Markov model from the criterion (6.3) with the suitable balance, the corresponding component structure can be used for the calculation of a point prediction according to eq. (7.3) and the rms prediction error according to eq. (7.4).

In the following the prediction error corresponding to the optimal perforated Markov model is compared with the minimum of the prediction error of usual standard embeddings (1 - 5 delay vector components in presence and past; delay of 1 - 5 time steps) and the relative position of the prediction error from the optimal perforated Markov model is given.

7.2.4.1 AR(3)

In this section an AR(3) process with $a_0 = 0$ is treated, since this process doesn't have the structure of a standard delay vector and thus a comparison of the prediction error from the optimal perforated Markov model and the minimal prediction error of standard embeddings makes sense. In fig. 2.2 the characteristic function of the stability area of this process class is shown.

In fig. 7.4 it is seen that for some resolutions the perforated Markov model can slightly reduce, i.e., improve the prediction error in comparison with standard embeddings. From noise variance $\sigma = 1$ the theoretical minimum of the prediction error is one and hence for re-

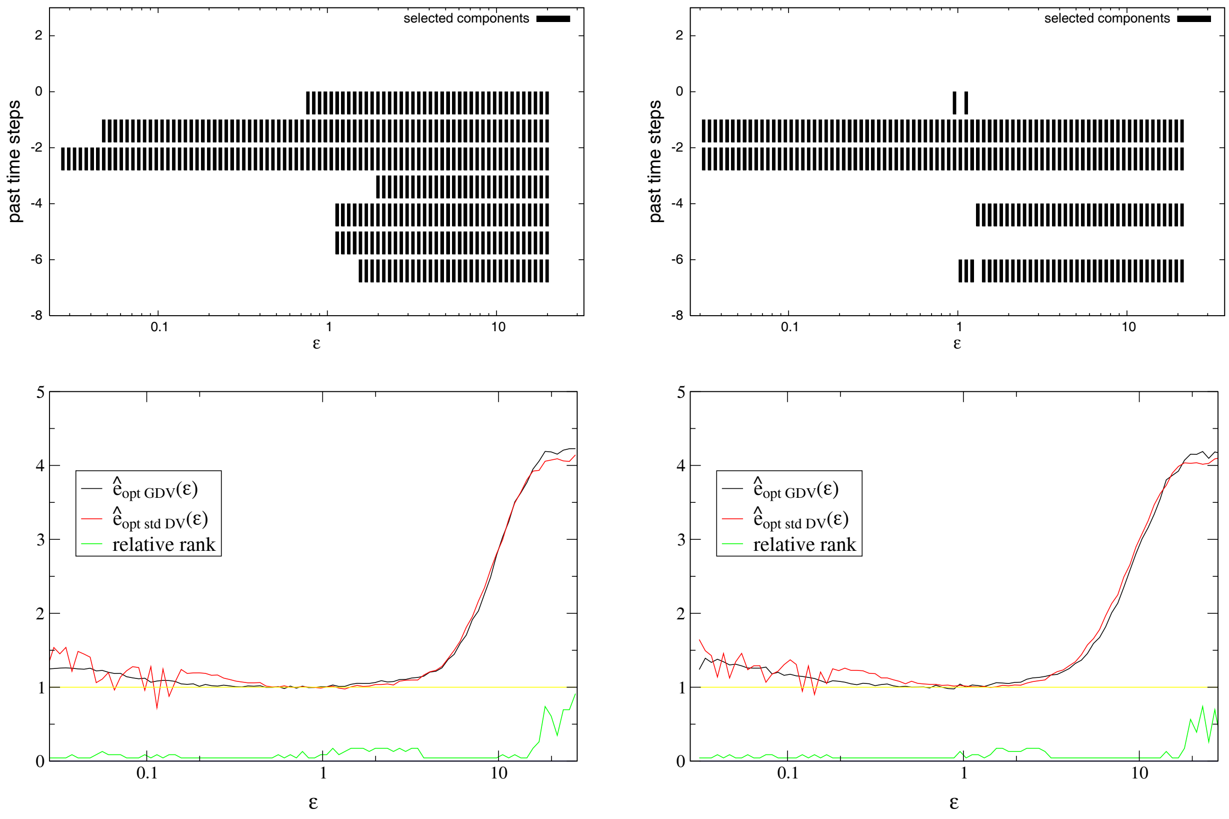


Figure 7.4: Left panels: AR(3) with coefficients $a_0 = 0$; $a_1 = 0.495$; $a_2 = 0.495$; 15000 data points. Right panels: AR(3) with coefficients $a_0 = 0$; $a_1 = -0.72$; $a_2 = 0.5$; 15000 data points. Upper panels: Optimal resolution-dependent perforated Markov model; Lower panels: Resolution-dependent prediction error from optimal perforated Markov model, minimal prediction error of standard delay vectors and relative rank of the prediction error from the optimal perforated Markov model.

solutions, for which this is already achieved by the optimum of conditionings in the sense of standard embeddings, a real improvement by optimized generalized conditionings cannot be expected. This is a weakness of the example.

For higher resolutions long standard delay vectors cause few neighbors concerning estimation of the prediction error. Therefore the minimal prediction error over standard embeddings is subject to fluctuations seen.

7.2.4.2 Generalized Hénon dynamics

In fig. 7.5 results for prediction from optimal perforated Markov models are shown for non-linear deterministic dynamics with memory for two different balance factors $b = 1$ and $b = 4$ of (6.3). Especially for $b = 4$ it is seen that the prediction error from the optimal perforated Markov model is smaller than the minimum of prediction errors for standard embeddings. This serves as another justification for the introduction of perforatedness into the framework of Markov approximations, if the *resolution-dependent* minimum of the prediction error is important.

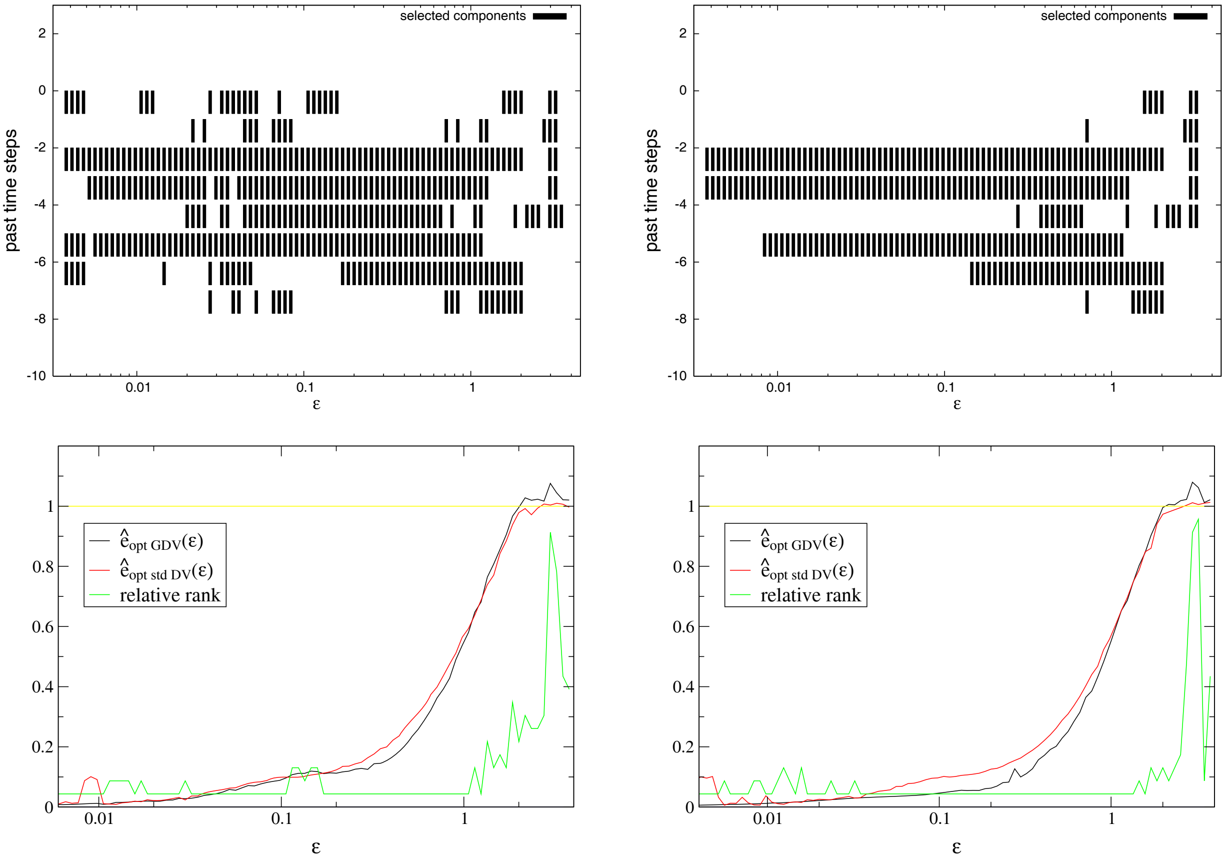


Figure 7.5: Generalized Hénon dynamics with delay $K = 4$, $a = 1.76$, $c = 0.1$, 10000 data points; Left panels: balance factor: $b = 1$; Right panels: balance factor: $b = 4$; Upper panels: Optimal resolution-dependent perforated Markov model; Lower panels: Resolution-dependent prediction error from optimal perforated Markov model, minimal prediction error of standard delay vectors and relative rank of the prediction error from the optimal perforated Markov model.

7.2.4.3 Conclusion concerning examples

Perforatedness in Markov modelling of higher order in principle can improve the resolution-dependent prediction error. On the other hand, in the given examples an improvement of the absolute prediction error cannot be seen.

7.2.5 Comparison of empirical prediction error optimization with entropic access to prediction error optimization

Practically the empirical method is faster, but after having determined the balance factor, the entropic method yields qualitatively the same results and a profound theoretical basis of what is found empirically. The empirical method doesn't yield any explanation for the selected optimal conditioning. From reasons of numerical costs the developed method finally could possibly not be very useful for practical purposes, even though it is truely prediction optimizing, but its real strength can be found in its value for theoretical explanation.

It has to be pointed out that a true comparison of both methods is only possible, if the balance factor b is fixed from the beginning (e.g. $b = 1$) or another than the empirical method could be used for the determination of the balance factor b. If this is not the case, then the empirical method is already used for a certain step in the entropic method and a true comparison of independent methods is not anymore on hand.

7.2.6 Detection of suitable resolution interval or even of optimal resolution for prediction

In contrast to the discussion of searching for optimal models, prediction finally doesn't need to be performed resolution-dependent since the resolution ϵ is an internal parameter in the algorithm addressing restrictions of the neighbor search. The presented resolution-dependent prediction error determined empirically in sec. 7.2.2 or from optimal perforated Markov models in sec. 7.2.4 shows an absolute minimum, which is suggested as the optimal resolution for performing the predictions. It is observed that local methods can be more or less skillful depending on the chosen degree of locality. Practically, seen from the very flat neighborhood of the minima of the prediction error, the optimal resolution can rather be treated as an optimal resolution interval.

Comparing stochastic and deterministic dynamics it has to be stressed again that the absolute minimum is found at qualitatively different resolutions. Deterministically the minimum of the prediction error is obtained for the minimal resolution for which in a fully (perforated) reconstructing dynamics any neighbors in the dataset are found, whereas stochastically the minimum of the prediction error is found at finite (larger) resolutions where statistical errors of the estimation start to become important. It is expected that in a continuous transition from stochastic to deterministic dynamics (e.g. reducing the variance of the input noise) the quality of the curves for the resolution-dependent prediction error approach each other.

7.3 An overview of prediction methods

With this section it is intended to present some classificational aspects of prediction methods, in order to have on hand a part of the range of the theoretical framework, and in order to locate the used prediction method in the general space of prediction methods. It has to be stressed that not 2^k prediction methods can be combined from the following collection of k dichotomic properties. The given list does not claim to be complete.

7.3.1 Classificational aspects

7.3.1.1 Univariate vs. multivariate forecasting

The distinction of univariate and multivariate forecasting is imposed by the length of the datavector at a certain time, where in the univariate case the vector shrinks to a single variable. Of course, if carefully treated, multivariate procedures will improve univariate procedures, but it is necessary to be aware of the fact that multivariate procedures seduce to overfit, what is one of the central problems concerning forecasting as stressed in this work. An example for the second case is multivariate regression.

7.3.1.2 Model-based vs. ad hoc forecasting

An example for model-based prediction methods is prediction via AR or ARMA models. Further examples of model-based prediction can be found below in the distinction of physical vs. statistical or of linear vs. nonlinear models.

An example for an univariate ad hoc method are the (exponential) smoothing methods. Those methods are techniques for extrapolating univariate time series. The main strategy of this method consists in decreasing the influence of an observation on prediction, the further in the past it was obtained. For modelling smoothing is irrelevant, because not even oscillations can be generated, but for forecasting it is said to be widespread [16]. Another example of an ad hoc method is 'trend extrapolation' being close to smoothing.

7.3.1.3 Physical model-based vs. statistical model-based forecasting

Physical model-based methods include assumed dynamical laws of the system in the phase space. In the numerical weather prediction method nowadays used this is partially realized.

An example for a statistical model-based method is the Box and Jenkins forecasting method (ARIMA model-based forecasting). An ARIMA model of order (p, d, q) is given by

$$\phi(B)(1-B)^d X_n = \theta_0 + \theta(B)\xi_n \tag{7.5}$$

with p, q polynomial order of ϕ, θ respectively.

In the first case, a physical interpretation of parameters is always possible, whereas in the second case this is not necessarily the case.

7.3.1.4 Density- and interval forecasting vs. point forecasting plus error

The variants opposed here are special cases of statistical, probabilistic (cmp. also sec. 7.3.1.11) forecasting. Finding the entire probability distribution of a future value usually is called density forecasting. In this case prediction skill evaluation is carried out by a so-called scoring rule called ignorance $-\log[p(x)]$ [13]. An example for density forecasting can be found in ([46], p.266). Percentile and quantile estimation of the conditional probability distribution are projections from density forecasting.

Prediction intervals are intervals in which the future value lies with prescribed probability [16]. It has to be taken care not to mix up prediction intervals concerning random variables and confidence intervals concerning model parameters.

Point forecasts are forecasts of single numbers (see sec. 7.1) and prediction errors are given by typical accuracy measures as e.g. the rms prediction error.

7.3.1.5 Ensemble vs. single trajectory forecasting

In contrast to the former section, where methods of statistical forecasting were discussed, this section concerns physical model-based forecasting.

In ensemble forecasting it is started with a finite sample from the probability distribution, which describes the uncertainty of the initial state. This sample is then iterated forward in time and estimates the evolution of the true probability density. The evolution of the single trajectories is usually guided by physical model-based methods. Ensemble forecasting is an element of the set of Monte-Carlo methods.

Single-trajectory forecasting is the physical model-based pendant to what in the framework of statistical methods is called point forecasting, asthonishingly also practically applied for weather forecasting nowadays.

7.3.1.6 Local vs. global methods of forecasting

In the framework of statistical model-based forecasting the local method is characterized by the selection of a certain neighborhood of a state represented by a time series segment from which the prediction is made. Technically, the selection of the neighborhood is carried out by the introduction of so-called kernel functions, which could be Gaussians or, as used, the Heaviside-function. In a global method no resolution dependence is treated and all data are taken into account for estimating the forecasting model. Local methods are usually computationally more expensive than global methods, since finally typically more parameters are involved.

7.3.1.7 Linear (and constant) vs. nonlinear methods of forecasting

In the framework of point forecasting methods, linear modeling for forecasting concerns parameter estimation of autoregressive models (under omission of the noise term) and is common locally as well as globally. Affin-linear modelling via

$$F(\mathbf{x}) = \mathbf{a}\mathbf{x} + b \tag{7.6}$$

is usually included, which contains the special case of *constant* modelling by fixing $\mathbf{a} = 0$. The locally constant and the locally linear methods are popular examples of local prediction methods. In the locally constant case the constant b depends on the actual state $\mathbf{s}_n$ and the degree of locality ϵ similar to eq. (7.3). The exact functional form for optimal determination of b depends on the chosen type of prediction error.

Nonlinear modeling F for forecasting is usual, e.g., with polynomials or radial basis functions in the framework of global methods. Details of global nonlinear modeling can be found in [46]. Local nonlinear modeling is in principle also possible for huge data sets, but it is not common practice.

7.3.1.8 Time-translation invariant vs. updated forecasting

In time-translation invariant forecasting a model is estimated once and this is kept then for all further times. In updated forecasting the most recent data are always also taken into account for the determination of the forecasting model. Neural networks, which are a subclass of machine learning, are global parameter updating methods.

7.3.1.9 In-sample vs. out-of-sample forecasting

In-sample forecasting means that training set and test set are not distinguished whereas out-of-sample forecasting is performed on a different dataset than the training set. True forecasting is out-of-sample forecasting.

7.3.1.10 Objective vs. subjective forecasting

Whereas objective forecasting is based scientifically, the subjective forecasting rather relies on experience. It is mentioned here, because it is still common practice concerning very complex issues and it becomes scientifically more justified from a posteriori objective validation by, e.g., reliability methods.

7.3.1.11 Comments on deterministic vs. probabilistic (stochastic) forecasting

This terminology is in use; especially also the first ([77], p.233). In the author's opinion those properties should be assigned to dynamics instead of to prediction methods and should not be used for classifying prediction methods, because every prediction method can be used for deterministc as well as for stochastic data. Furthermore every serious prediction method is somewhere handling with probabilities, which are sometimes hidden behind some neglected error (and this doesn't change if the underlying dynamics is deterministic!)

Seemingly 'deterministic forecasting' is better called point forecasting and stochastic forecasting either appears as ensemble forecasting if physical laws are involved, or in the statistical case as the variants of sec. 7.3.1.4. Otherwise one would have to be aware of the fact that stochastic dynamics could be deterministically forecasted.

7.3.2 Arrangement of the prediction method based on perforated Markov models into the general classification of prediction methods

The prediction methods based on perforated Markov models from minimization of an entropic-statistical criterion fix one important aspect of the statistical model underlying prediction, namely the relevance of components in the past as optimal inputs for the prediction algorithm. Until now, theoretically this is treated only as truncation for the nonperforated case with standard information criteria (compare sec. 6.7) or concentrating only on the embedding dimension with the method of 'false nearest neighbors' [47], which yields the appropriate minimized embedding dimension in comparison with the results of the statement of Takens [71]. Otherwise an ad hoc treatment of this problem rather often occurs without true theoretical justification.

In advance, it has to be mentioned that the suggested perforated component selection method is not usefully applicable in combination with few of the prediction methods given in sec. 7.3.1, e.g., smoothing methods as examples for ad hoc methods do not ask for perforated component selection. It is also expected that an ensemble method, being physical model-based, could not profit from it.

In a univariate framework, based on perforated component selection, a statistical (almost) non-model-based, locally constant, first moment point prediction method, which should be called objective, was implemented and analyzed.

The question arises, if the presented component selection method performs better in combination with more advanced forecasting methods than with the locally constant forecasting scheme. A numerical analysis on this question was not carried out. The suggested entropic-statistical method seems to intrinsically prefer the locally constant prediction scheme. For example in the deterministic case the optimal resolution for the locally constant scheme was the minimal resolution with the suggestion to take only one neighbor. This could be enough for an optimal (in an average sense) point prediction for the locally constant scheme, but probably not enough for a parameter fit procedure for a finite number of parameters. On the other hand, it could be argued that the optimal resolution for other prediction schemes than the locally constant method differs and is, e.g., reached in the deterministic case for the locally linear model at a coarser resolution than for the locally constant model. It is expected that the locally constant prediction scheme would win this minimization competition, because it starts initially from a smaller resolution, but this would finally have to be proven in order to become a serious scientific statement, which could depend on the underlying dynamics.

Chapter 8

Fluids and wind

8.1 Theoretical basis of hydrodynamics

8.1.1 Navier-Stokes equation

In the compressible case the basic equations are the *continuity equation*

$$\frac{\partial \rho}{\partial t} + \nabla(\rho \mathbf{u}) = 0 \tag{8.1}$$

and the *momentum equations*

$$\frac{\partial u_i}{\partial t} + (u_j \nabla_j) u_i = \frac{1}{\rho} \frac{\partial \sigma_{ij}}{\partial x_j} + F_i \; . \tag{8.2}$$

Here ρ is the density, $\mathbf{u}$, u_i are velocities and F_i is the specific external force. The shear stress σ_{ij} is given by

$$\sigma_{ij} = -p\delta_{ij} + 2\mu(e_{ij} - \frac{1}{3}\delta_{ij}\frac{\partial u_k}{\partial x_k}) \quad , \qquad [\sigma_{ij}] = \frac{M}{LT^2} \; . \tag{8.3}$$

p is the pressure, μ is the dynamic viscosity ($[\mu] = \frac{M}{LT}$) and e_{ij} is the deformation rate

$$e_{ij} = \frac{1}{2}(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}) \quad , \qquad [e_{ij}] = \frac{1}{T} \; . \tag{8.4}$$

Often the dynamic viscosity μ is replaced by the kinematic viscosity

$$\nu := \frac{\mu}{\rho} \quad , \qquad [\nu] = \frac{L^2}{T} \; . \tag{8.5}$$

In the incompressible case the density ρ is independent of time and space. The continuity equation becomes

$$\nabla \mathbf{u} = 0 \tag{8.6}$$

and the momentum equation becomes

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\nabla)\mathbf{u} = -\frac{1}{\rho}\nabla p + \nu \nabla^2 \mathbf{u} + \mathbf{F} \; . \tag{8.7}$$

The continuity equation together with the momentum equation is called the Navier-Stokes equation.

8.1.1.1 Discussion of terms in incompressible case

$\frac{\partial \mathbf{u}}{\partial t}$ is called the inertia term. The advection term reads

$$(\mathbf{u} \cdot \mathrm{grad})\mathbf{u} = \mathrm{grad}\left(\frac{1}{2}\mathbf{u}^2\right) - \mathbf{u} \times \mathrm{rot}\ \mathbf{u}\ . \tag{8.8}$$

This term exists explicitly only in Eulerian coordinates and disappears in (suitably accelerated) Lagrangian coordinates moving with the fluid. If the flow is laminar with constant velocity, then the derivative of the velocity field by space yields zero, and the advection term vanishes. The term $-\frac{1}{\rho}\nabla p$ is the pressure gradient term, which is the origin of nonlocality. The consequences of the viscous or diffusion term $\nu\nabla^2\mathbf{u}$ are the shear stress, friction forces and 'damping'. External force terms (specific force) of different physical type are frequently discussed, e.g. gravity, heating and buoyancy or magnetic forces in magnetohydrodynamics, etc.

8.1.1.2 Difficulties of solution

The first problem of the Navier-Stokes equation is the fact that the advection term is nonlinear in the velocity. The second major problem of the (incompressible) Navier-Stokes equation is its intrinsic nonlocality in the solution for the pressure generated dynamically by the velocity. This is seen by applying the ∇-operator on the Navier-Stokes equation, whereby a Poisson equation for the pressure results. A formal inversion of this equation reveals the nonlocality. The pressure is nonlocal due to the nonlocality of the inverse of the operator ∇^2.

8.1.2 Turbulence

Typically fluid motion is divided into the states of laminar and turbulent motion. Whereas the laminar state is characterized by smooth fluid motion, the turbulent state shows much more complex behaviour. The states of fluid dynamics are usually characterized by dimensionless control parameters, which for shear flows is the Reynolds number $Re = \frac{uL}{\nu}$ (L is the characteristic length scale). Above a certain threshold of the Reynolds number the turbulent state is reached. Definitions of turbulence in the literature are not unique. Hence some are collected in the following:

- Turbulence is defined as a three-dimensional random motion that is unpredictable in detail, of an incompressible, uniform density, viscous fluid that satisfies the incompressible Navier-Stokes equations and that is either the decay of an initial velocity distribution or is driven by a deterministic or random force field [61].

- Turbulence may be regarded as a manifestation of flow instability occuring randomly in space and time.

- Turbulence is the name given to imperfectly understood class of chaotic solutions to the Navier-Stokes equation in which many degrees of freedom are excited.

- Turbulence is a time-dependent motion in which vortex stretching causes velocity fluctuations to spread to all wavelengths between a minimum determined by viscous forces and a maximum determined by the boundary conditions of the flow. It is the usual state of fluid motion except at low Reynolds numbers.

Some of those above and further different definitions of turbulence can be found in appendix A of [74].

8.1.2.1 Properties of turbulence

At large scales energy is injected, which is transferred to the small scales, where it is dissipated. The transfer of energy from large to small scales is called 'energy cascade' and was introduced by Richardson. Any turbulent flow is maintained by an external source of energy produced by possibly various mechanisms. It occurs at rather high Reynolds numbers.

Turbulence arises under various conditions and lots of different external forces, but the intrinsic properties of turbulent flows are already included in the simplest case of homogeneous and isotropic turbulence. Furthermore turbulence has the following properties. It is:

- highly dissipative
- strongly mixing
- irregular and (spatio-temporal) random
- non-smooth
- predictable only in statistical sense.
- statistically stable
- endowed with high levels of fluctuating vorticity
- stretching vortices
- producing enstrophy $\Omega := \langle \frac{1}{2} |\nabla \wedge \mathbf{u}|^2 \rangle$
- strain amplifying
- characterized by a large number of strongly interacting degrees of freedom [74] and an extremely wide range of interacting scales (resulting from nonlinearity)

'Turbulence is a natural response to instabilities in the flow - a response that tends to reduce the instability' ([76], p.378).

8.1.2.2 Fully developed turbulence

Turbulence at very high Reynolds numbers is known as fully developed turbulence, if all or some of the possible symmetries are restored in a statistical sense ([28], p.11). It is too complex for being describable by low-dimensional models with few degrees of freedom.

8.1.3 Kolmogorov theory

Though the basic equations of fluid dynamics are deterministic, the intrinsic randomness caused by sufficiently large Reynolds numbers ([74], p.23) makes an a priori statistical approach like the Kolmogorov theories more adequate.

8.1.3.1 Preparing definitions

The longitudinal velocity increment [28] is defined as

$$\delta u_{\parallel}(\mathbf{r},\mathbf{l}) := [\mathbf{u}(\mathbf{r}+\mathbf{l}) - \mathbf{u}(\mathbf{r})] \cdot \frac{\mathbf{l}}{l} \, . \tag{8.9}$$

Here $l = |\mathbf{l}|$. With the assumption of homogeneity the explicit dependence on $\mathbf{r}$ vanishes and isotropy turns the dependence on $\mathbf{l}$ to a dependence on l, i.e., taken any axis ($\mathbf{l}$ along this axis), $\delta u_{\parallel}(\mathbf{r},\mathbf{l})$ can be replaced by

$$\delta u_{\parallel}(l) = [\mathbf{u}(l) - \mathbf{u}(0)] \cdot \frac{\mathbf{l}}{l} \, . \tag{8.10}$$

In order to write down the right expression formally correct, $\mathbf{l}$ is unavoidable. Nevertheless the expression does not depend on it as indicated on the left side of the equation. *Structure functions* are defined as

$$S_n(l) := \langle (\delta u_{\parallel}(l))^n \rangle \, . \tag{8.11}$$

8.1.3.2 Hypotheses

l_0 is defined as the scale of the largest eddies and η is defined as the scale of dissipation.

- Kolmogorov's hypothesis of local isotropy: At sufficiently high Reynolds number, the small-scale turbulent motions at scales $l \ll l_0$ are statistically isotropic.

- Kolmogorov's first similarity hypothesis: In every turbulent flow at sufficiently high Reynolds number, the statistics of the small-scale motions ($l \ll l_0$) have a universal form that is uniquely determined by the kinematic viscosity ν and the rate of dissipation ϵ ($[\epsilon] = \frac{ML^2}{T^3}$).

- Kolmogorov's second similarity hypothesis: In every turbulent flow at sufficiently high Reynolds number, the statistics of the motions at scales l in the range $\eta \ll l \ll l_0$ have a universal form that is uniquely determined by ϵ, but independent of ν.

Details can be found in ([60], p.184 ff) and in ([28], p.74 ff), where the representations of the hypotheses are not identical.

8.1.3.3 Consequences

Two-thirds law ($\frac{2}{3}$ is exponent) [28]:

$$S_2(l) \equiv \langle (\delta u_{\parallel}(l))^2 \rangle \sim l^{\frac{2}{3}} \; . \tag{8.12}$$

Four-fifths law ($\frac{4}{5}$ is prefactor) [28]:

$$S_3(l) \equiv \langle (\delta u_{\parallel}(l))^3 \rangle = -\frac{4}{5}\epsilon l \; . \tag{8.13}$$

In the general case it holds

$$S_n(l) = C_n \cdot (\epsilon l)^{\zeta_n} \tag{8.14}$$

with C_n as a universal constant, what means, that it is independent of the particular flow under consideration. The Kolmogorov theory (K41) yields

$$\zeta_n = \frac{n}{3} \; . \tag{8.15}$$

The Kolmogorov scaling of the energy spectrum reads

$$E(k) = C\epsilon^{\frac{2}{3}} k^{-\frac{5}{3}} \; . \tag{8.16}$$

8.1.3.4 Generalizations

Nowadays discrepancies between Kolmogorov scaling and experimental measurements are observed and intermittency is used for explaining those effects [28]. Generalized functional relations for ζ_n for all used models always have the properties of $\zeta_3 = 1$, which is a consequence of the Navier-Stokes equation, and being concave as a function of n.

8.1.4 Boundary layer

The boundary layer is the whole region, where at least some effects of a wall influence the flow.

8.1.4.1 Preparation: Some definitions of wall quantities

Given $\bar{u}(y)$ as the mean horizontal velocity with vertical dependence. The mean wall shear stress is defined by

$$\sigma_w := \mu \left(\frac{d\bar{u}}{dy} \right)_w \; . \tag{8.17}$$

Then the friction velocity can be defined by

$$u_* := \sqrt{\frac{\sigma_w}{\rho}} \; . \tag{8.18}$$

Dimensionless wall quantities are then defined for the velocity

$$u^+ := \frac{\bar{u}}{u_*} \tag{8.19}$$

and for the distance

$$y^+ := \frac{y u_*}{\nu} \; . \tag{8.20}$$

8.1.4.2 Classification of regions and physics of turbulent boundary layers

The tripel 'viscous sublayer, buffer layer, log-law region' from close to further apart from the wall is a decomposition of the boundary layer into disjoint parts. In the viscous sublayer ($y^+ < \ 5$) the Reynolds shear stress $\tau_{ij} = -\rho\langle u_i u_j \rangle$ is negligible compared with the viscous stress, which causes laminar motion. The buffer layer ($5 < y^+ < 70$) is dynamically most active ([51], p.106). Here the gradient of the streamwise velocity profile is largest and a shearing instability is the mechanical origin for turbulence. A Kelvin-Helmholtz breakdown of shear flow ([76], p.381) takes place. The log-law region ($70 < y^+$) obtains its name from the validity of the following law:

$$u^+(y) = \kappa^{-1} \log y^+ + A \quad ; \quad \text{von Karman constant } \kappa \approx 0.40 \pm 0.01; \; A \approx 5.1 \; . \tag{8.21}$$

Further notions for the classification of regions in the boundary layer are the viscous wall region versus the overlap region and the inner versus the outer layer, which are visualized in ([60], p.275/276).

Coherent structures in the boundary layer like hairpin vortices [1], low-speed streaks, ejections, sweeps, bursts [51], etc., are discussed in the literature. A detailed analysis of the dynamical mechanisms, in which such structures are involved, is beyond the scope of this work.

Heating of boundary layers causes a thermal origin of turbulence from buoyancy, which is especially (but not only) of importance in the atmospheric boundary layer. In order to relate energetic influences to the flow of fluids from shear and from buoyancy the flux Richardson number is introduced as [9]

$$R_f = -\frac{\text{buoyant production rate of turbulent kinetic energy}}{\text{shear production rate of turbulent kinetic energy}} \; . \tag{8.22}$$

Furthermore the gradient Richardson number, which is easier measurable, can be obtained from

$$Ri = Pr \cdot R_f \; , \tag{8.23}$$

where the Prandtl number Pr measures the relative magnitude of momentum and heat diffusivity. The absolute value of the Richardson numbers always increases with height ([9], p.57). For convective flows the value of the Richardson numbers allow for a height-dependent determination of laminarity or turbulence. Laminar flow becomes turbulent for $Ri < Ri_c = 0.25$. For the opposite transition hysteresis effects have to be taken into account.

8.2 Theoretical basis of wind dynamics

Aspects concerning humidity are ignored in this work.

8.2.1 Origin of wind

From thermal differences high-pressure areas and low-pressure areas emerge and cause a movement of air masses. Including the Coriolis forces geostrophic winds, i.e., air motion perpendicular to the pressure gradient, result and the rotation in the mean wind velocity causes the formation of cyclones and anti-cyclones. Caused by the Coriolis force with changing height over the ground there is a change in direction of the wind between 'free' pressure-driven geostrophic wind and the wind close to the ground, which is called the Ekman spiral effect.

The large scale winds described drive the air layers below. Turbulence production mechanisms especially at the boundary cause fluctuations of the wind speeds around the mean. Including orographic aspects and convective motion a framework of complicated dynamics in the atmospheric boundary layer close to the ground is opened.

8.2.2 Atmospheric boundary layer

The atmospheric boundary layer of 1-2 km height is mainly influenced by the strength of the geostrophic wind, Coriolis effects, the surface roughness and thermal effects [14].

8.2.2.1 Classification of sublayers

The following sublayers are distinguished (cmp. sec. 8.1.4.2):

- Laminar sublayer (viscous sublayer): Thickness of order milimeters
- Dynamic sublayer (buffer layer): Thickness of few meters
- Prandtl or surface layer (log-law region): 20 - 100 meters height; Coriolis force ignorable
- Ekman layer: 1 km height; Coriolis force important

8.2.2.2 Monin-Obukhov (MO) similarity theory and wind profile

The MO theory is based only on similarity and dimensional arguments. In MO theory it is assumed that all meteorological relationships between dimensionless local variables in the atmospheric surface layer are assumed to be functions of $\frac{z}{L}$. Here z is the height above ground and the MO length L is given by

$$L = \frac{-\bar{\rho}_0 u_*^3 c_p}{\kappa \alpha g q} , \tag{8.24}$$

with the surface mean density $\bar{\rho}_0$, the friction velocity u_*, the specific heat under constant pressure c_p, the von Karman constant κ, the thermal expansion coefficient α, the gravitational acceleration g and the vertical heat flux q (positive upward). The MO length gives a measure of the relative contribution of energy supplied to the turbulence by buoyancy forces to that supplied by heat generated from friction; for large values of L the vertical heat flux has a small influence on the structure of the atmospheric boundary layer near the surface ([51], p.10). The notion 'length' is questionable however, because L can become negative. Especially the non-dimensional wind and temperature-profiles (derived by simple integration) are universal functions of z/L:

$$\frac{\partial u}{\partial z}\frac{\kappa z}{u_*} = \phi_m(z/L) \ , \tag{8.25}$$

$$\frac{\partial \theta}{\partial z}\frac{\kappa z}{T_*} = \phi_h(z/L) \ . \tag{8.26}$$

Here θ is the potential temperature[1], $T_* = \frac{-\overline{w'\theta'}}{u_*} = \frac{H}{\rho_0 c_p u_*}$ is a scaling temperature and H is the turbulent flux of sensible heat, which is an alternative notion for enthalpy.

If ϕ_m (and ϕ_h) can be treated approximately as constants, a logarithmic profile results:

$$u(z) = \frac{u_*}{\kappa}\ln(\frac{z}{z_0}) \ . \tag{8.27}$$

z_0 can be interpreted as a property of the surface and is called the aerodynamic roughness length. From the logarithmic law shear and thus instabilities, especially Kelvin-Helmholtz-type instabilities can be inferred. This supports the onset of turbulence and hence the atmospheric boundary layer becomes an interesting example for prediction.

The validity of the MO theory is restricted to $z < |L|$, i.e., for $\zeta := \frac{z}{L}$ this means $|\zeta| < 1$. Hence the validity essentially concerns the neutral stratification (cmp. sec. 8.2.2.3) and the lowest parts of the boundary layer (small z) in non-extreme cases of stable and unstable stratification.

Finally it has to be remarked that the Richardson number of eq. (8.23) and the MO length of eq. (8.24) are connected via

$$Ri = \frac{\phi_h}{\phi_m^2}\frac{z}{L} \ . \tag{8.28}$$

8.2.2.3 Thermal convection in the atmosphere; atmospheric stratification and stability of the boundary layer

Turbulence in the atmosphere is partly driven by shear, but partly also by buoyancy, because heat comes into play. Thermal effects can be classified into three categories[2]: Unstable, neutral and stable stratification [14].

---

[1] The potential temperature θ of an air parcel is defined as the temperature that the parcel of air would have if it were expanded or compressed adiabatically from its existing pressure and temperature to a standard pressure p_0 (generally taken as 1000 hPa)

[2] In ([70], p.95) the five categories convective, unstable, neutral, stable-continuous and stable-sporadic are distinguished, what is beyond the scope of this work.

Stratification-type	unstable	neutral	stable
Thermal equilibrium	no; surface warmer than air	yes	no; surface colder than air
Temperature of rising air compared with the environment	too hot	same temp.	too cold
Preferred occurence	daytime	transitional exceptional case	nighttime
Dominant turbulence generation mechanism	buoyancy; stress insignificant		shear production friction with ground
Buoyancy	important	unaffected by buoyancy ([78],p.208)	buoyancy effects oppose vertical motion ([54],p.127)
Turb. eddy size scale	large	intermediate	small
Boundary layer thickness	thick	intermediate	thin
Vertical mixing and transfer of momentum and heat	large	'sufficient' [14]	strongly suppressed
Compexity of dynamics	turbulent, hence statistically simple		intermittent, meandering motions
Wave development	suppressed		supported
Scaling (lower z)	free convect. scaling	MO-scaling	local scaling [54]
Scaling (higher z)	mixed layer scaling		'z-less' scaling
Obukhov stabil. param. ([30],p.50)	$\zeta := z/L \rightarrow -\infty$	$\lvert\zeta\rvert \ll 1$	$\zeta \rightarrow \infty$
Universal fct. ϕ_m [41]	$\phi_m(\zeta) = (1 - a\zeta)^{-1/4}$	$\phi_m(\zeta) := 1$	$\phi_m(\zeta) - 1 \propto \zeta$
Universal fct. ϕ_h [41]	$\phi_h(\zeta) = b(1 - c\zeta)^{-1/2}$	$\phi_h(\zeta) = \text{const}$	$\phi_h(\zeta) - 1 \propto \zeta$
Windspeed- and Temperature-profile	power-law ([41],p.222)	logarithmic law	log-linear law
turb. Coriolis forces	neglectable	structurally important	neglectable
Gradient of mean wind speed with height near boundary	small	intermediate	large
Wind-turbine-problems	sudden gusts possible		signif. asymm. loadings

The shown table makes clear that the neutral boundary layer fits in perfectly from the thermal point of view, but the discussion on it is usually performed in an asymmetric way in the literature if compared to the stable and unstable case, because of its role as a transitional state.

8.2.3 Aspects of the physics of gusts and extreme wind events

A comparatively low level of accumulation of gusts is produced by usual shear-driven turbulence. For convectively generated gusts (downbursts) there are two important processes: First, horizontal momentum is produced by pressure gradient forces, because convective downdraughts are blocked by the surface and distorted to spread out horizontally. Second, in the presence of vertical shear convective downdraughts transport horizontal momentum downward. Convective downdraught and strong geostrophic wind result in strong gusts near the surface [53].

The trigger mechanism for the deflection of air parcels from higher levels in the boundary layer to the surface can be attributed to vertical mixing by turbulent eddies in unstable stratification. The gust speed should be a function of the large-scale wind, the turbulence and the stability of the boundary layer [12].

A consequence of the Bernoulli theorem ([23], 40-6)

$$\frac{p}{\rho} + \frac{1}{2}v^2 + \phi = \text{const} \quad \text{(on a streamline)} \tag{8.29}$$

is that a fluid particle trapped into the depression (smaller p) will acquire a higher velocity than the one going into the pressure high. This is the reason why winds tend to be strong in troughs and weak in the highs ([52], p.28). The higher kinetic energy in the motion of air masses accounts for the higher frequency of gusts in throughs.

The idea of gusts as velocities $u(t)$ as a function of time having the shape of a so-called 'mexican hat' as sometimes possibly used in branches of engineering has to be rejected. The reason is that the mexican hat arises from a sound wave picture, what is an idea being too low-dimensional for the phenomenon to describe.

Tornadoes and hurricanes are further important phenomena in the class of strong wind events, which are not touched by this work.

8.3 Conclusion

Even though the basis is seemingly deterministic the occurence of turbulence causes irregularity and randomness in fluid dynamics. In boundary layers instabilities appear accumulated and hence the tendency for turbulence is amplified. Nevertheless, from the projection in the measurement process memory usually arises. Thus, wind in general seems to be a potential example for application of the suggested methods of chap.'s 6 and 7, where however, as will be seen in chap. 9, new types of problems will arise.

Turbulence, especially in convective circumstances, can furthermore cause strong gusts, for which methods like structure search in combination with a ROC plot in appendix B could be reasonable.

The combination of a priori deterministic structure with probabilistic aspects associated to turbulence in the boundary layer makes this to be a very interesting example for the task of prediction in either the average sense or with regard to extremes, especially because measurement of wind speeds is a rather simple task.

Chapter 9

Application of discussed concepts and proposed methods on wind data

9.1 Wind data

Example data sets of wind speed at 10 meters above ground measured with 8 Hz for whole days in flat terrain at the Lammefjord site in Denmark in the year 1987 downsampled to 1 Hz are shown in fig. 9.1.

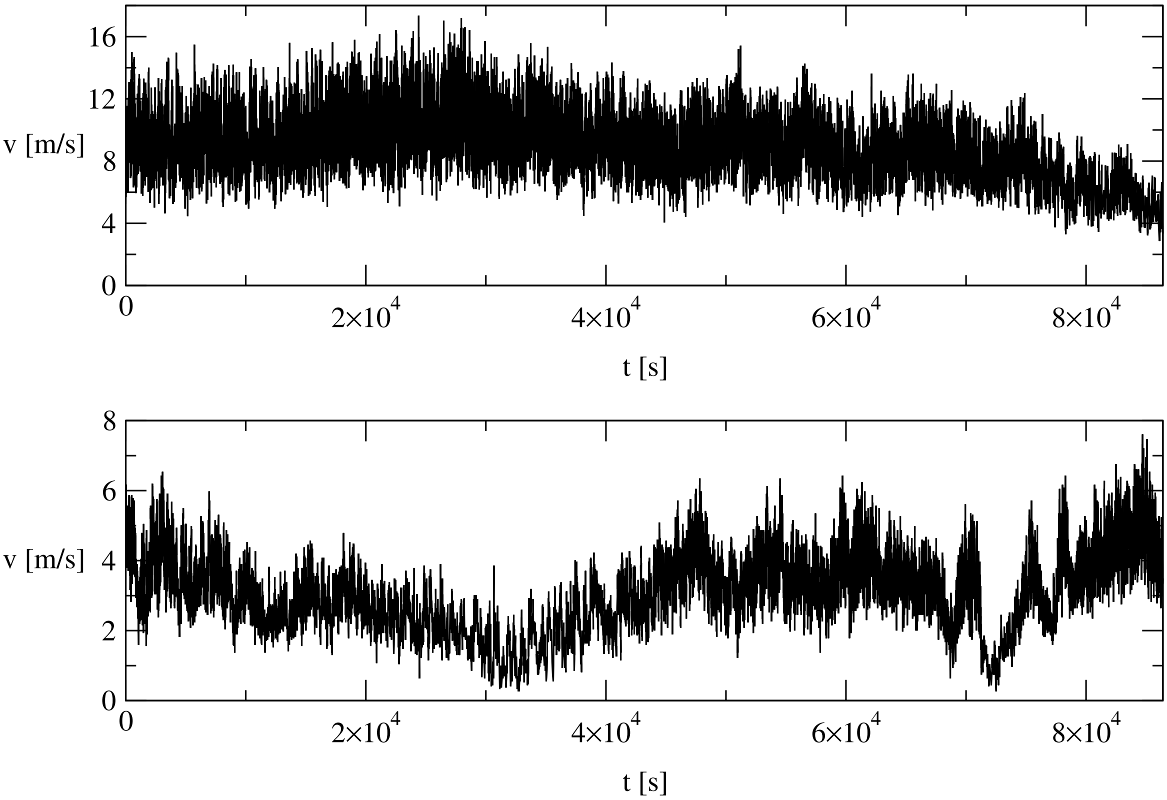


Figure 9.1: Examples of wind speed data sets. The upper plot corresponds to the dataset labeled with nr. 191 and the lower plot to the dataset labeled with nr. 192.

The upper plot of fig. 9.1 can be seen as comparatively stationary, at least in the first part. For the application of the developed methods it is tried to concentrate on the rather stationary parts of the data sets and possibly remaining nonstationarity is ignored even though this is questionable. In the lower plot nonstationarity is ubiquitous.

9.2 Stochastic modelling of wind data

Given σ_n as the width of the distribution of the velocity fluctuations around the mean value (mean taken over one minute) conditioned to the mean value $\bar{v}_n$ of the velocity $p(v_n - \bar{v}_n | \bar{v}_n)$, one obtains [44]

$$\sigma_n \approx 0.1 \bar{v}_n \ , \tag{9.1}$$

what qualitatively can be seen in fig. 9.1. This property is included in modelling of the dynamics behind the wind data by a term of multiplicative noise. A naive ansatz for the dynamics with suitable parameter values is henceforth

$$v_{n+1} = \alpha v_n + \beta v_n \xi_n \quad \text{with} \quad \xi_n \sim \mathcal{N}(0,1)\ ,\ \alpha = 0.95 \text{ and } \beta = 0.1\ . \tag{9.2}$$

Positivity of velocities as in fig. 9.1 may be reached by taking the root of the sum of squares of perpendicular velocities each obtained from (9.2). The iteration law (9.2) is a Markov process of first order. However, the entropy analysis of the following section will show that a Markov process of first order cannot be the final answer on the dynamics behind the data.

More advanced trials for modeling wind data involve a continuous time random walk (CTRW) in an upcoming PhD-thesis by D. Kleinhans or power-law truncated Levy flights by A. Chechkin.

9.3 Entropies of wind data in nonperforated case

In the left panel of fig. 9.2 the conditional entropies are plotted for stationary wind data. Starting from low resolution one observes first a typical stochastic nonperforated conditional entropy plot with almost all uncertainty reduction found in the first step of conditioning, but yet nonvanishing information in the further steps of conditioning, what will be especially important in the case of usual Markov approximations. The steps in the graphs for small ϵ appear due to a non-equidistant discretization present in our dataset as a result of an analog-digital converter. For higher resolution the plot seems to indicate increasing regularity with finally almost completely disappearing uncertainty, what has to be interpreted as an artefact of the discretization procedure, which influences the number of neighbors found in the correlation sum in a way that for given high resolution the number of neighbors found in the correlation sum decreases more slowly with increasing m reaching zero for much higher m than undiscretized. The existence of decrease is seen in non-zero $H_{1|m}$ for smallest ϵ.

On a rather nonstationary part of a wind dataset shown in the right panel of fig. 9.2 the artefact is diminished. This is in accordance with the explanation of slower decreasing number of neighbors in m due to dicretization in the stationary case. The stronger reduction of uncertainty in the first step of conditioning in the right panel of fig. 9.2 than in the left panel is not a consequence of the nonstationarity but instead a consequence of a lower mean velocity in the analyzed segment of the dataset.

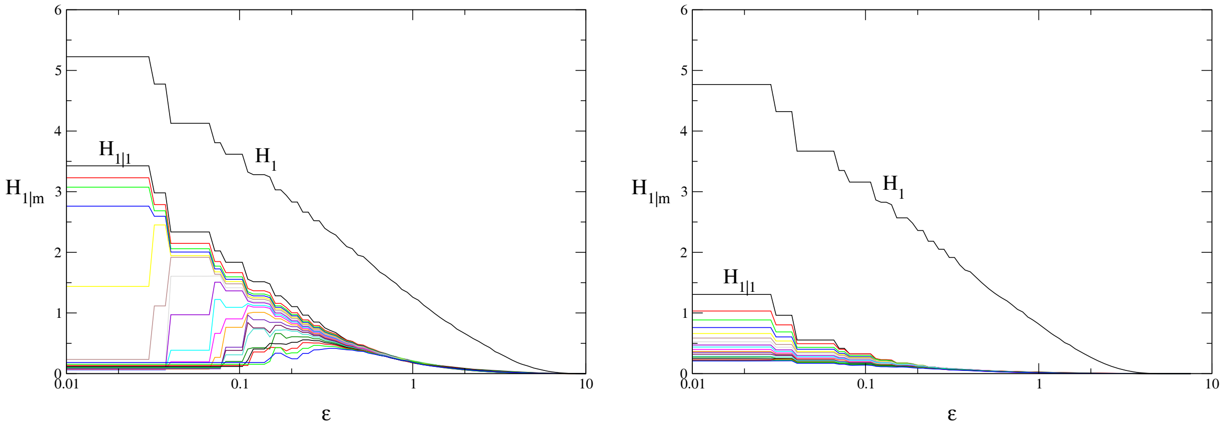


Figure 9.2: Estimation of conditional entropies of wind data with the TISEAN program d2 from segments of 100000 data points. From top to bottom the curves correspond to increasing m starting with $m = 0$ for the topmost curve. Left panel: Rather stationary dataset (nr. 191). Right panel: Rather nonstationary dataset (nr. 192).

The coarse graining regime with splitting of the entropies seen for autoregressive time series in chap. 6 is not seen for wind data.

Except for the huge amount of information about the future in the first step of the past, there is no special structure available giving a suggestion for a preferred Markov approximation to take along special components for prediction.

9.4 Entropies of wind data in perforated case

9.4.1 Variable future lead time

Comparison of the plots in fig. 9.3 among themselves and with the left plot in fig. 9.2 shows that uncertainty reduction of the future is less the larger the lead time into the future.

9.4.2 Regular perforation by downsampling

In the left panel of fig. 9.4 again nonperforated conditional entropies as in fig. 9.2 are shown, but only for a few conditioning indices in order to improve the comparability with the strongly downsampled case in the right panel of fig. 9.4. The first aspect in the comparison of the plots is the annihilation of information for prediction under the operation of downsampling.

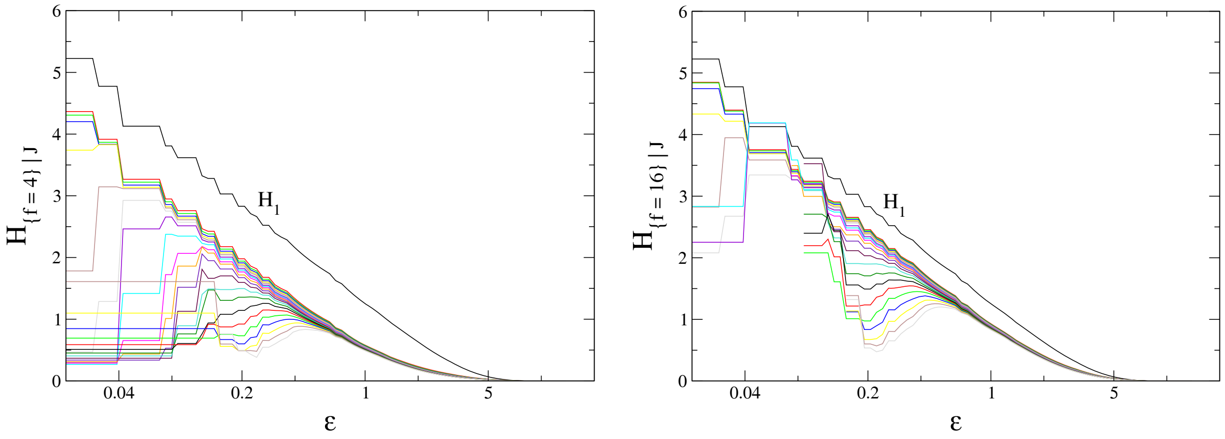


Figure 9.3: Conditional entropies for variable future time. Left panel: Lead time of 4 time steps. Right panel: Lead time of 16 time steps. In both cases the upper black curve corresponds to $\mathbf{J} = \phi$ and the further curves from top to bottom correspond to increasing number of conditioning indices $\mathbf{J} = \{0\}$, $\mathbf{J} = \{-1, 0\}$, $\mathbf{J} = \{-2, -1, 0\}$, etc., in this succession.

On the other hand, from the splitting of conditional entropies for intermediate resolutions it is possible to see also on wind data that downsampling generates memory with redistribution of information for prediction into the further past.

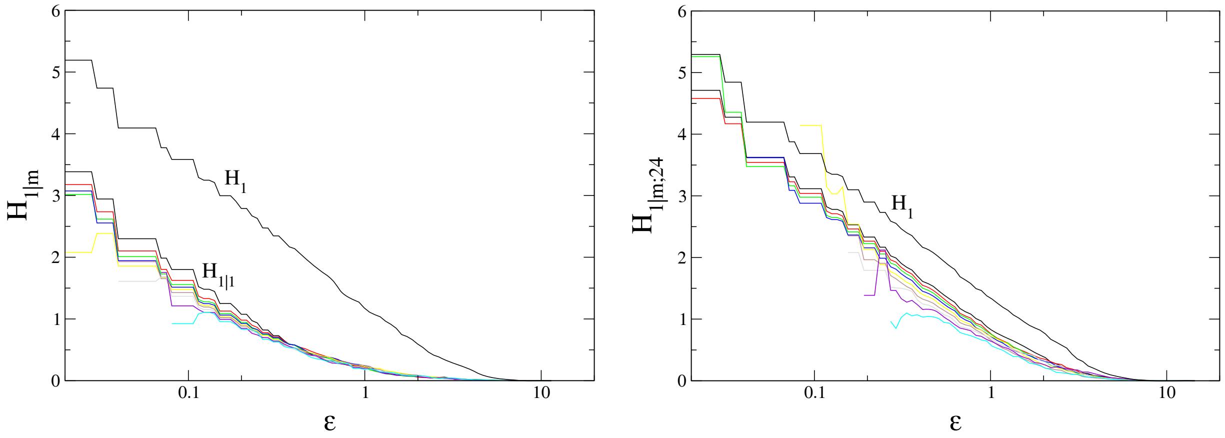


Figure 9.4: Conditional entropies for downsampling with the TISEAN program d2. Conditioning is in the same sense as for fig. 9.2. Left panel: Estimation of entropy of 30000 data points of a windspeed dataset. Right panel: Estimation of entropy under downsampled conditions taking only every 24th data point of the same basis wind dataset.

9.5 Finding optimal Markov approximation

9.5.1 Finding optimal usual Markov approximation for wind data

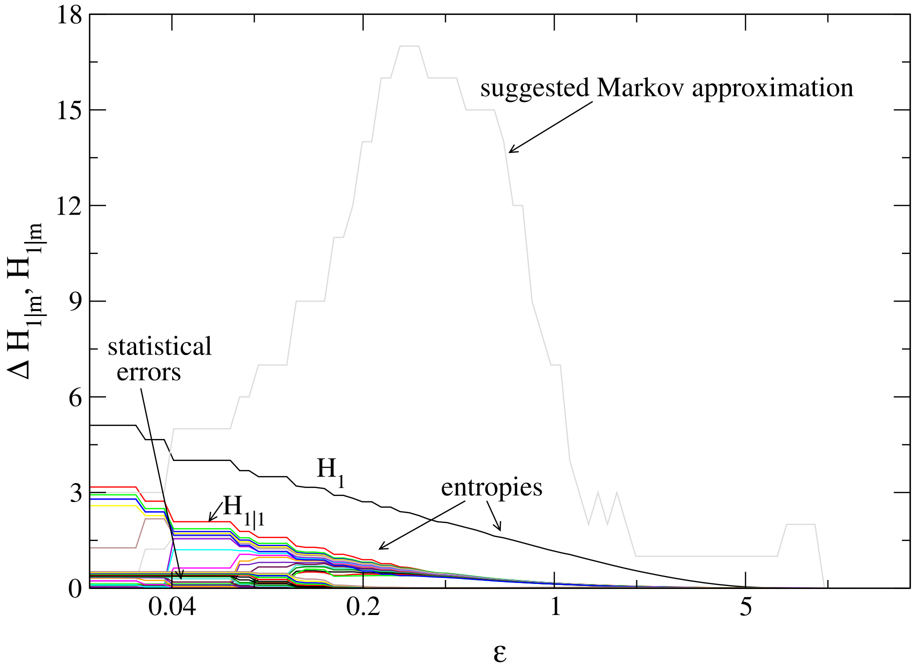


Figure 9.5: Suggested Markov approximation for wind data

In fig. 9.5 the Markov approximation according to the criterion proposed in sec. 6.2.1 is shown. Again the first result is that the suggested Markov approximation is resolution-dependent. For fine resolution the statistical error is large and this causes a Markov approximation of low order. Remarkable is the intermediate region where a Markov approximation of rather high order is suggested. Probably laminar phases in the flow cause that for intermediate resolutions the statistical error remains small and smallest information in the longer conditioning is still larger than the statistical error causing the high order.

It has to be stressed that in all of sec. 9.5 and hence especially in fig. 9.5 nothing is said about an optimal resolution for prediction.

9.5.2 Finding optimal perforated Markov approximation for wind data

9.5.2.1 Variable dataset length

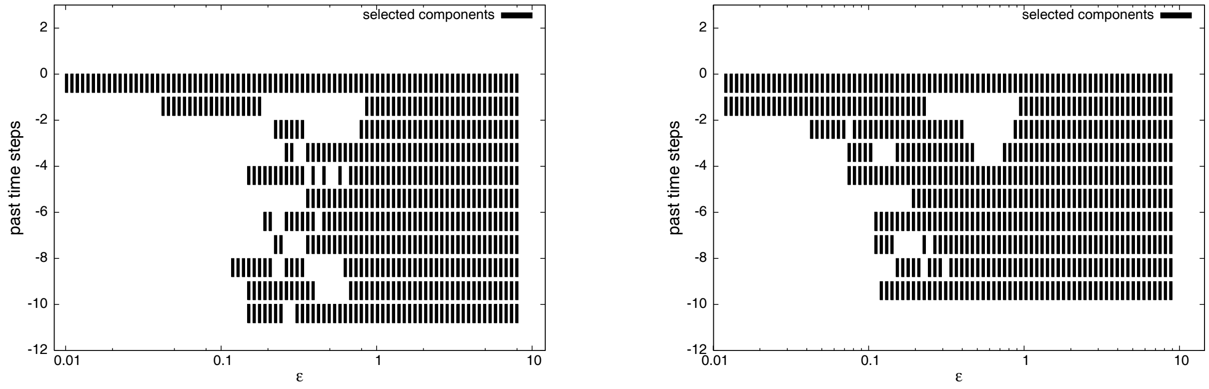


Figure 9.6: Nr. of day: 191; Height: 30m (column 3), balance-factor $b = 4$. Left panel: 5000 data points (lines 105000-110000). Right panel: 20000 data points (lines 100000-120000).

The result of fig. 9.6 is that the longer the dataset, the more conditioning indices are selected as optimal also for smaller resolutions.

9.5.2.2 Variable mean wind velocity

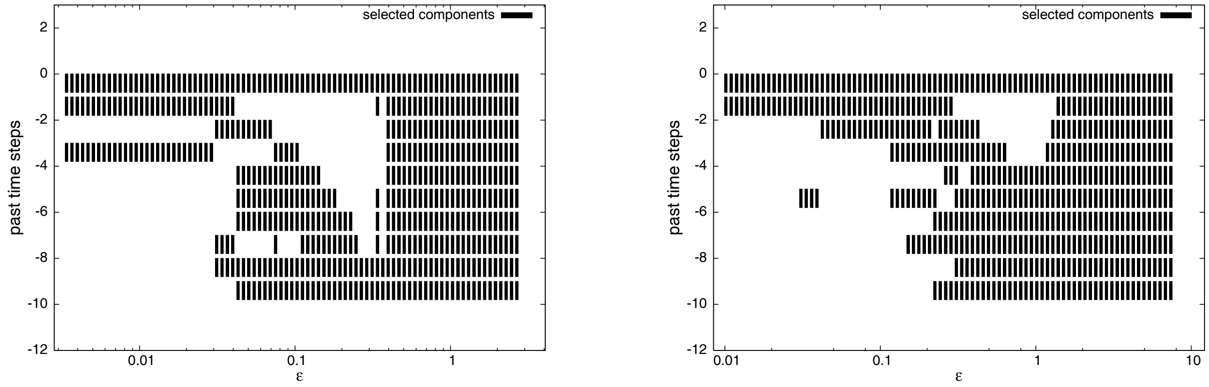


Figure 9.7: Height: 10m (column 1), 10000 data points, balance-factor $b = 4$, Left panel: Nr. of day: 186 (lines 270000-280000), $\bar{v} \approx 4.5m/s$. Right panel: Nr. of day: 191 (lines 50000-60000), $\bar{v} \approx 10m/s$.

As a result of fig. 9.7 it could be seen that a smaller mean velocity offers better chances for extraction of a statistically robust perforated structure for optimal conditioning. It is worth mentioning the robust exclusion of the second component in the past for more than a decade of high resolutions in the left panel of fig. 9.7.

9.5.2.3 Variable balance factor in the selection criterion

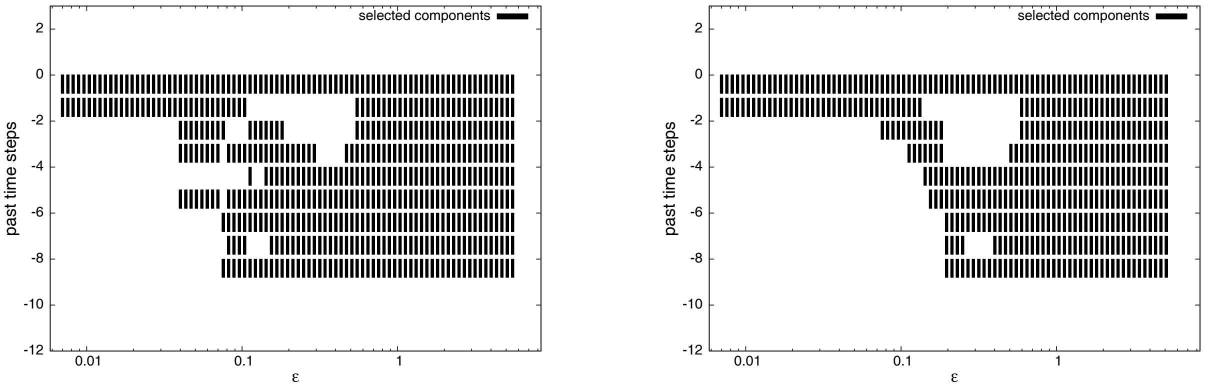


Figure 9.8: Nr. of day: 190, height 30m (column 3), 10000 data points (306000-316000). Left panel: Balance-factor $b = 4$. Right panel: Balance-factor $b = 16$.

Increasing the balance factor increases the importance of the term of statistical errors in the criterion and hence selection of too many conditioning indices is punished, what can be seen in fig. 9.8 also for wind data at intermediate resolutions.

9.5.2.4 General analysis

From the fig.'s 9.6 - 9.8 it is again seen that the chosen optimal set of conditioning indices is resolution-dependent. A clearly perforated structure does not seem to be extractable from the analyzed data sets. The resolution-dependent component selection algorithms did not exclude clearly particular components over a broad range of resolutions universally. The statistically averaged robust structure chosen was the following:

- The present observation is always part of the conditioning.

- A triangle-shape exclusion of components starting from the second component at intermediate resolutions in the entropic crossover regime is seen. A similarity in principle with the top right plot (and a bit with the bottom right plot) of fig. 7.1, where the long conditioning also reaches optimality in two different regimes of resolution, seen if zoomed into those plots, has to be pointed out.

- All components at coarse resolution are taken along

- The number of conditioning indices for higher resolution decreases (to usually two-dimensional embedding)

For the task of prediction the resolution-integrating strategy of finding the optimal perforated Markov model from sec. 6.6 needs inevitably a theoretically justified stopping criterion which is not available until now. For plotting the succession of conditional entropies the non-existence of such a stopping criterion is acceptable but this is not the case with the intention of prediction and hence skipped here.

9.6 Wind speed prediction

9.6.1 General remarks

One of the intentions of this work originally was a suggestion of a short time 'online' prediction scheme for strong damaging winds. Knowing or assuming the available time (probably smaller than 1s) between precursor and wind gust, taking into account the necessary reaction time and knowing the calculational speed of the available computational device it is calculable the *finite maximal number* of states described by a perforated embedding vector with which a comparison of the actual state is possible. This finite number serves as an input for the suggested component selection-algorithm applied to some usual time series parts of the given length at the location in real application. This should give in a statistical sense a preferred component selection which then is used for performing a point prediction at the optimal resolution on a selected data set of the given length with suitable comparison vectors, which would have to be selected really carefully. This point prediction would have to be transformed by a suitable threshold construction into adaequate actions. Criticism on this approach with respect to the prediction of extremes is outlined in appendix C.

9.6.2 Prediction from optimal perforated Markov model

From the criterion in eq. (6.3) the resolution-dependent optimal generalized Markov approximation is obtained, which is used in the projection operator of eq. (7.3) in order to perform a point prediction. The resolution-dependent prediction error is then calculated according to eq. (7.4). In the following figures $\hat{e}_{\text{opt GDV}}$ is the prediction error from conditionings according to structures of generalized delay vectors which are optimal in the sense of the criterion from information theory and statistics. $\min\{\hat{e}_{\text{std DV}}\}$ is the minimum of prediction errors from conditionings according to a certain set of structures of standard delay vectors (number of components from 1 to 5, number of delays from 1 to 5). The relative rank of $\hat{e}_{\text{opt GDV}}$ positions this prediction error in the set of all calculated prediction errors from conditionings in the sense of standard delay vectors.

9.6.2.1 Variable dataset length

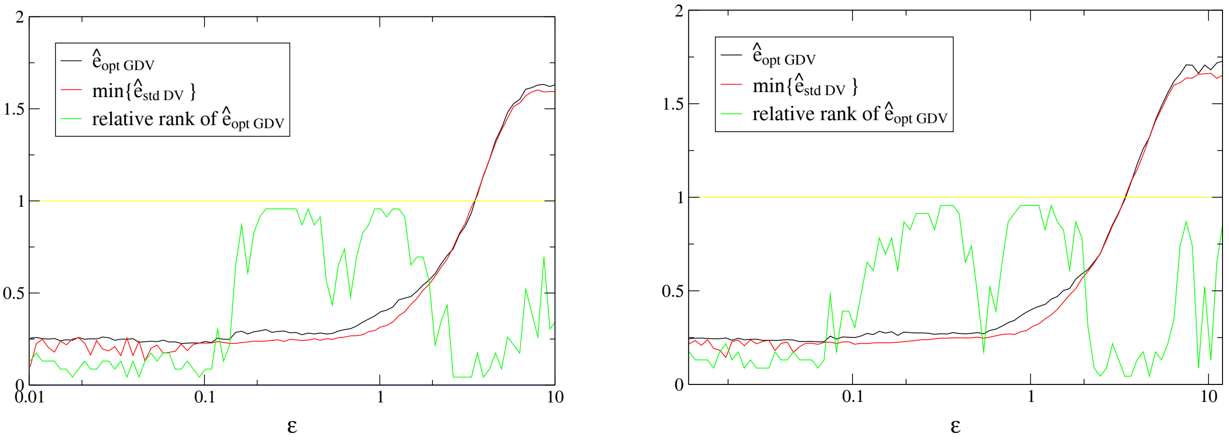


Figure 9.9: Variable dataset length. Left panel: Shorter dataset (5000 data points). Prediction error for wind data corresponding to left panel of fig. 9.6. Right panel: Longer dataset (20000 data points). Prediction error for wind data corresponding to right panel of fig. 9.6.

9.6.2.2 Variable mean wind velocity

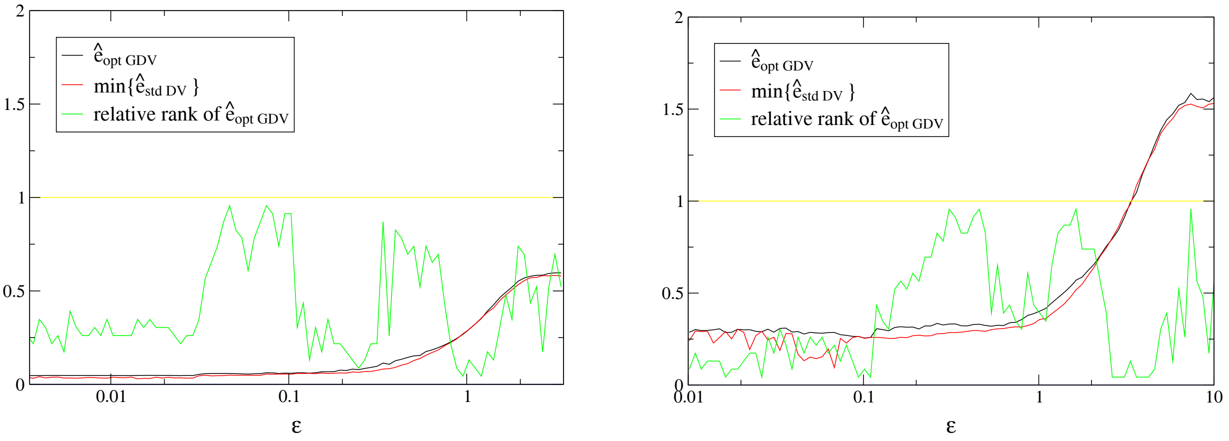


Figure 9.10: Variable mean wind velocity. Left panel: Lower mean wind velocity ($\bar{v} \approx 4.5m/s$). Prediction error for wind data corresponding to left panel of fig. 9.7. Right panel: Higher mean wind velocity ($\bar{v} \approx 10m/s$). Prediction error for wind data corresponding to right panel of fig. 9.7.

9.6.2.3 Variable balance factor in the selection criterion

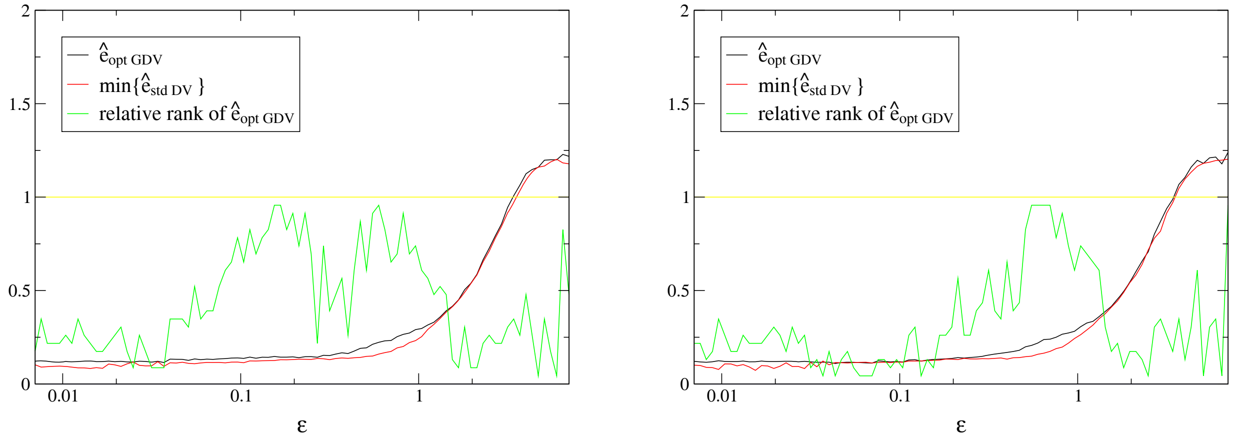


Figure 9.11: Variable balance factor. Left panel: Balance factor $b = 4$. Prediction error for wind data corresponding to left panel of fig. 9.8. Right panel: Balance factor $b = 16$. Prediction error for wind data corresponding to right panel of fig. 9.8.

9.6.2.4 General analysis

For coarsest resolution the prediction errors approximate the width of the dataset and for increasing resolution the prediction errors reach a plateau. Instead of an optimal resolution for prediction as a result it is found that all resolutions above some particular intermediate resolution lead to approximately predictions of the same quality in an average sense. The fact that neither under change of the set of conditioning indices (standard or generalized) nor under change of resolution above some threshold resolution the prediction error does not further decrease, suggests that wind speeds are a stochastic process with transition probabilities having some minimal width which cannot be reduced by more precise conditioning. The increase of the prediction error for smaller ϵ as seen for stochastic processes in fig. 7.1 is expected to appear for smaller ϵ than the highest resolution shown.

A very general result of fig.'s 9.9 - 9.11 is that higher cardinality of the set of conditioning indices from the generalized Markov approximation leads to increased prediction errors which is seen in the two maxima of the relative rank appearing very robust under change of parameters.

The closeness of the curves for the prediction error from generalized conditionings and the minimum of the prediction error from standard conditionings (especially in the left panel of fig. 9.11) in combination with the fluctuating curve for the sometimes very bad position of the prediction error from generalized conditioning shows that all prediction errors for intermediate resolutions are quite close independent of the type of conditioning as long as the first component of the past is taken into account. This shows that no strong effects in the optimization of the conditioning can be expected for the wind data with respect to prediction improvement in an average sense[1] .

---

[1] The question why this seemingly rather simple thing wasn't tested empirically in a rather early stage of the project must be answered with the fact that the notion of a resolution-dependent prediction error and

9.7 Discussion of the results for prediction with generalized delay vectors from the criterion of information theory and statistics in the context of wind data

The following discussion is not especially specific for the wind, but more general, and only triggered by results on wind data.

The model selection from criteria based on information theory and statistics and the consequences for prediction were successful on several example dynamics. In contrast, the application for wind data cannot be called very successful, because finally the prediction error corresponding to the structure of the optimal generalized delay vector did not improve the optimum of the prediction errors from conditionings in the sense of standard delay vectors. The punishment by the statistical error of additional components in the optimal generalized Markov approximation seems to be not strong enough. Already for the usual Markov approximation in sec. 9.5.1 it was seen that for intermediate resolutions the statistical error is rather small.

In the following a list of possible reasons for the problems of the combination of the criterion of eq. (6.3) and the task of reduction of the prediction error in point prediction of wind data are presented. The first two entries of this list could account for the smallness of the estimation of the statistical error.

- Problems with dynamics of laminarity and bursts:
  It seems that too many neighbors are found in the case of the wind. The reason could be that laminar phases from intermittency, a kind of nonlinear deterministic dynamics, (and unused so-called 'Theiler window') could support this fact relatively dramatically. Problems of estimation with the correlation sum under laminar conditions were also reported in [39].

  In our most turbulent dataset, the day 191 (cmp. fig. 9.1), laminar phases are at first sight not visible and nonetheless the problem with the prediction error is present. This could be a minor reason, that this 'laminarity'-argumentation is not maximal important. On the other hand a medium number of rather short laminar phases are already enough to introduce the effect and this wouldn't be visible on large scales.

  As a solution of this problem it could be suggested to use time-dependent optimal embeddings, i.e., working on segments of the time series, or a preselection according to certain restricted neighbor search areas of the phase space, what is a different sense of locality compared to the usual epsilon-dependence of entropies. On the other hand the effectiveness of information theory on 'short' segments or restricted phase space areas each for its own is rather questionable, because the effect of the needed increase of the statistical error in the estimation of entropies by decrease of the total length of the

---

the idea of generalized perforatedness didn't exist from the beginning, but only emerged in the development of this work and the focus of this work was primarily directed to a justification of Markov approximations. Furthermore a comparison of prediction errors with fixed number of neighbors and hence belonging to different resolutions for varying conditionings is rather questionable.

parts of the dataset is expected to be a weaker gain than the mean loss of information in the relative accumulation of laminar phases. Another trial for the improvement of the situation of shortening laminar phases by downsampling is of course only apparent, because then the improvement by the increasing statistical error is decremented by smearing out of the intersting dynamical parts. Also this should not be a promising ansatz. Here information theory seems to reach its limits and other methods have to be preferred.

- Possible correlatedness of statistical errors:
  The statistical errors of conditional entropies are estimated from the statistical errors of unconditioned joint entropies in different dimensions. From this construction it is possible that the errors are correlated, but uncorrelatedness is a condition for the certain success of error propagation. Thus a wrong estimate of the error and eventually an underestimation could be possible.

  Practically one of the statistical errors was dominant, namely the one connected to the largest support and the others from experience are lower by at least approximately a factor of two or three. Summing squares as in usual error propagtion this effect is increased. In this sense a correlatedness of the errors could probably cause only minor effects on the final sum. If this effect finally turns out as relevant, a solution would be a more advanced error propagation strategy.

- Minimization of prediction error and optimization of Markov approximation are different tasks:
  Since finding the optimal generalized Markov approximation and optimization of forecasting are different tasks, an optimal forecast from an optimal generalized Markov approximation might be expected, but does not have to be necessarily demanded. The reason is the fact that different cost functions are optimized for the different cases. This means that the method of point prediction or especially the corresponding accuracy measure do not necessarily match up with the method of component selection.

  Entropies were estimated with the correlation sum (preferentially $q = 2$), where all pairs of neighbors are treated in a symmetrical sense. Concerning the chosen accuracy measure (rms prediction error) also all pairs of neighbors are treated, however they are grouped before it is operated with them. So the spirit concerning the use of locality is different.

  On the other hand, if this issue really is a problem, then it should always be a problem. I have doubts if a presumption of a real problem with the different senses of locality and different senses of cost functions is in accordance with the rather suitable results found for the example dynamics of autoregressive processes and Hénon dynamics. Nevertheless, it has of course to be admitted that in those examples the relevant components of the dynamics were chosen as rather unsuitable for standard embeddings such that the generalized embedding was in a constructed advantage.

  Eventually the combination with other accuracy measures or prediction methods can reveal a better performance of the criterion even though I feel that a criterion based on information theory crunched together with prediction methods not being in the spirit of working on mean properties or a prediction method for extremes is even more playing a lottery concerning the hope for success of the method (cmp. appendix C).

Two defending arguments of the results of this chapter shouldn't be completely ignored:

- Partially unfair comparison:
  From the ranking of prediction errors the suggested criterion is indicated as rather bad. It is not that bad, because the comparison with standard delay vectors is only performed for delay vectors of rather short length. If optimality of the prediction error depends primarily on the smallness of the cardinality of the set of conditioning indices then according to the criterion optimized rather large sets of conditioning indices in the sense of generalized embeddings will automatically lead to worse prediction errors and to the last rank. A comparison with standard delay vectors of the same length could have revealed a better rank in the corresponding subclass.

- Optimal resolution:
  Concerning prediction it is not important to have an optimal prediction error for all resolutions but to have the minimal prediction error from generalized conditionings according to the criterion at the optimal resolution. Assuming that this absolute minimum occurs at rather small ϵ, what cannot be excluded from fig.'s 9.9 - 9.11 the result concerning the rank wouldn't be that bad for wind data anymore.

Nevertheless, application of methods of information theory seems to be more problematic in the context of wind data than in the context of the example dynamics studied in chap. 6. First, it is the a priori nonstationarity which prevents success. Second, the 'intermittency-like' structures with their laminar phases and turbulent bursts are expected to cause huge problems with respect to entropic strategies because the averaging involved in entropies is presumably inadequate for this type of dynamics.

Chapter 10

Conclusion

Starting from the task of performing suitable necessary Markov approximations in the framework of infinite memory from projected stochastic dynamics with the boundary condition of using information theory, statistical reasons and the possibility of multiscale dependences suggested or rather enforced the generalization to a perforated case, where omissions of time steps in the past can be allowed.

A generalized notational framework of information theory in time series analysis was developed, which allows for a *unified* description of variable future time steps ahead, joint prediction, downsampling (regular perforation) and arbitrary irregular perforation with the tools of information theory. This is the central and unavoidable basis for the work presented.

A non-widespread generalization of the concept of mutual information of two random variables to mutual information of many random variables, which points out the close connection of information theory to usual set theory via Venn diagrams, is supported. It yields access to understanding and explicitly calculating the redistribution of information for prediction under perforation.

Considerations on the generalized correlation sum yielded explicit formulas on the connection of entropies and dimensions, which in principle are known from [68] for deterministic systems. This work unifies their validity for the stochastic and deterministic case for various embedding dimensions and especially for the perforated case.

In order to contribute to structural understanding, the information theory on a time-discrete setting was generalized to the time-continuous case under finite time conditioning, a case very close to the paper [31], but not discussed there. Considerations of the Lorenz and Roessler dynamics as well as the Ornstein-Uhlenbeck process yield for time-continuous entropies as a result a new possibility of distinction of chaos and noise. In the deterministic case an upper threshold of the joint uncertainty in the limit of infinitely high sampling rate can be given. In a three-dimensional representation of the joint entropy $H(\epsilon, \tau, t)$ with τ as a parameter the dimension and the KS entropy rate of the dynamics can be read out from *one* single graph. In the time-continuous limit the slope of the joint entropy with respect to time at finite rather small time t yields a possibility for the selection of optimal observables with respect to prediction. The numerically found surprisingly non-monotonous entropy rate for Renyi order $q = 2$ of the z-component of the Roessler system has to be mentioned here.

The origin of memory in complex systems is clarified and explained in detail.

Taking into account the theoretical constraint of minimizing heuristic input in contrast to the case of, e.g., Akaike's model selection, novel criteria for usual as well as for perforated Markov approximations were introduced. The explicit calculation of the statistical error of entropies made accessible those criteria based only on entropies and its derived quantities. Since the method is based on information theory, it is applicable to a broad class of dynamics. This is especially useful for an analysis of data sets, where it is not allowed to assume nice properties like linearity.

Numerically on the basis of the entropic part of the TISEAN-program d2 generalizations were carried out. First, the estimation of entropies under omission of intermediate conditioning components and variable time steps ahead were allowed. Via introducing the estimate of the statistical error of entropies the criteria for performing usual and generalized Markov approximations were implemented. Strategies of reaching the further past for the Markov approximations under the demand of finite computational time were introduced.

The criteria were successfully tested for linear stochastic and nonlinear deterministic dynamics for various dataset lengthes and varying lead time. It was found that the optimal perforated Markov model is resolution-dependent. For certain intervals of intermediate resolution the memory structure of the dynamical law was retrieved by the suggested criterion indicating the functional capability to yield suitable Markov approximations. For small resolutions coarse graining effects are clearly seen and for fine resolutions from statistical reasons fewer conditioning components are selected.

A new method for improved estimation of the KS entropy rate is presented, which excludes the possibility of missing relevant information from the further past. This resolution-integrating method for Markov approximations offers a new spirit for attractor reconstruction methods in contrast to methods based on functional independence. As a byproduct the developed tool can act as a delay finder.

An explicitly resolution-dependent rms prediction error was introduced. It was calculated for various conditionings, and characteristic differences of stochastic and deterministic dynamics were determined. The criterion for perforated Markov approximations yields a theoretical justification of the empirically found results for the rms prediction error. For certain resolutions an improvement of the rms prediction error from the resolution-dependent optimal perforated Markov model in comparison with the rms prediction error from standard embeddings was seen. The existence of an optimal resolution for local prediction methods was pointed out and qualitatively detected.

The perforated framework is an improvement of the Takens framework, because it improves strongly the statistics in the problem. This is especially the case, if stochasticity is present. For determinism the nonperforated case is closer to optimality, but especially the highest resolutions cannot be reached, where the prediction error has its absolute minimum.

Some underlying physics behind wind phenomena were discussed.

It was not possible to transfer the success of the method of information theory and statistics from the example dynamics to the case of wind data. Reasons were found in the laminarity of the data and in the interplay of the correlation sum with intermittent dynamics. Furthermore, correlatedness of errors was seen as a potential problem. From this result, the universality of the proposed method must be reduced to the subclass of stable nonlinear stochastic processes, for which of course any nonstationarity is absent, but it has also to be demanded the absence of intermittent-like characteristics. Information theory turned out not to be a useful concept for predicting extreme and rare events.

Chapter 11

Outlook

In the following a list of potential refinements and extensions of the presented work is given:

- If the time-continuous entropies can be established, the formalism could be carried to the perforated case. It is the missing link of sec. 3.8 with the title of this work. The problem of redistribution of information under omission of intervals in the time-continuous case should then be solved.
- Even though the suggested criterion arises in a natural way, it is not clear if the criterion for the generalized Markov approximation could be improved by introducing further terms or even changing the given terms. This should be checked.
- A more advanced error propagation strategy for the statistical errors in the estimation of entropies could be implemented, which is more appropriate for the violations of the assumptions of the usual error propagation.
- It could be strived for a more efficient programming of entropy estimation such that truely far past, what in principle is accessible, is also accessible in the algorithms with acceptable calculational demands.
- If the given criterion for the generalized Markov approximation finally stays as introduced, then especially with respect to prediction a scientifically more serious strategy for the determination of the balance factor b between the informational and the statistical term would be desirable.
- The improvement of the prediction error by (asymmetrically) perforated instead of standard embeddings at the optimal resolution could be proven for *more advanced* dynamics than linear stochastic or nonlinear deterministic dynamics.
- The optimal generalized Markov approximation from the criterion based on information theory and statistics could be used as input for various advanced forecasting methods other than point forecasting with evaluation by the root mean squared prediction error.
- Instead of being restricted to wind data the method of generalized Markov approximations from information theory could possibly be able to show its power more in being applied to other problems.

Appendix A

Determinism versus stochasticity

It is well known that determinism is defined by the fact that, given initial conditions, the trajectory for all future time steps is fully fixed, i.e., the initial value problem has a unique solution. The only deviation of this behaviour can be found in the fact that an uncertainty in the initial state is amplified by the dynamical time development. This is stronger, if the dynamics is chaotic. On the other hand, stochastic processes are characterized by their intrinsic probabilistic nature.

Nevertheless, deterministic dynamics, even the nonlinear case, is via deltafunction-like transition probability densities in principle included in the framework of stochastic processes. The various elements in the set of all deterministic dynamics can be approached by limit processes from the larger space of all stochastic processes by reduction of the noise amplitude. On the other hand, the limit of largest positive Lyapunov exponent approaching infinity in the case of a bounded state space leads from chaos to a certain kind of noise.

This relationship of deterministic and stochastic cases causes the following possible changes for interpretations of the plots in fig. 3.2 of the usual conditional entropies. If the conditional entropy reaches a plateau for finer resolutions, then, no doubt, this is a clear indication of determinism. If there is only '$-\ln\epsilon$' - behaviour in the entropies this is until now always interpreted as noise. It could always also be interpreted as determinism with huge Lyapunov exponent. The same statement formulated in a different way reads: Given a finite finest resolution and assumed the dynamics always interpreted as deterministic, then plots of the conditional entropy give either the exact sum of positive Lyapunov exponents from the plateau or they give a possibly huge lower threshold of the sum of positive Lyapunov exponents, depending on the finest resolution. In this sense plots of the usual conditional entropy do not distinguish chaos from noise, but yield on the assumption of always determinism, a lower threshold of positive Lyapunov exponents.

The central difference of determinism and stochasticity in continuous time, where the limit of sampling interval $\tau \to 0$ is carried out, is that in almost all points the trajectories of stochastic processes are continuous, but non-differentiable, whereas in determinism almost all points are even differentiable. The result of sec. 3.8, where the uncertainty in finite time for deterministic systems is finite, but without extractable upper threshold for stochastic systems, is best understood, if the non-differentiability in almost all points in the stochastic case is traced back to a slope of the sample trajectory without upper threshold in almost all points, such that the apparently infinite uncertainty can be associated with a length without upper threshold of the sample trajectory. Chaotic determinism and noise can hence be made comparable by adjusting parameters, if noise is treated in discrete time such that the number of points of non-differentiability, the length of sample trajectories and the total uncertainty for the stochastic case all become explicitly finite as in the deterministic case.

Concerning the optimal resolution for prediction from minimizing the prediction error it was found in sec. 7 that the optimum in the deterministic case was found for highest available resolution and in the stochastic case it was found for intermediate resolutions. The reason for this behaviour can be found in the fact that for determinism in an average sense one neighbor in the dataset is optimal for prediction, whereas in the stochastic case a distribution is explicitly required for prediction.

Concluding on qualitative discrepancies of stochastic and deterministic dynamics in this work the following table can be given:

Example	stochasticity	determinism
Conditional entropy $H_{1\|m}(\epsilon)$	$\sim -\ln \epsilon$	constant
Time-continuous joint entropy $H(\tau = 0, t > 0)$	infinite	finite
Optimal resolution for prediction error	intermediate	smallest

Nonetheless as pointed out at the beginning of this appendix, determinism is the limit of stochasticity for noise amplitude approaching zero and not something far distant from the stochastic case. I would like to propose that the qualitative change of properties in the limit from the Ornstein-Uhlenbeck process to the Wiener process described in chap. 2 is quite similar to the situation encountered here[1]. E.g. the stationarity and correlatedness of increments of the Ornstein-Uhlenbeck process are qualitatively changed by $\alpha \to 0$ in eq. (2.27) to nonstationarity and uncorrelatedness of increments of the Wiener process. Even though there are qualitatively different properties, in the limit from the Ornstein-Uhlenbeck process to the Wiener process the transition is smooth, because with decreasing α in the Ornstein-Uhlenbeck process, e.g., the stationary state is reached later and later until in the limit not reached anymore.

The same question of smoothness of the transition is interesting for limit procedures between stochasticity and determinism. In the following this is discussed for the examples given above. In the transition of standard deviation $\sigma \to 0$ from stochastic to deterministic behaviour the curves of conditional entropy in the left panel of fig. 3.2 move to the left and converge to zero, the value of conditional entopies for a regular deterministic system for sufficiently high conditioning. Instead of being uniform the convergence is pointwise. Also the curves in fig. 3.16 in the example of time-continuous entropies move to higher resolutions in the transition of $\sigma \to 0$. Hence also in this example a smooth transition from stochastic to deterministic behaviour can be found. The optimal resolution for minimizing the prediction error moves smoothly from an intermediate resolution in the stochastic case (fig. 7.1) to the highest available resolution in the deterministic case (fig. 7.2) in the transition of $\sigma \to 0$. Whereas in all former examples a smooth transition from the stochastic case to the deterministic case is found, a non-smooth transition is only found for a priori asymptotic non-extensive (with respect to time) quantities as, e.g., the Kolmogorov-Sinai entropy rate.

---

[1] It has to be stressed that with this statement it is of course in no way claimed that the Wiener process is deterministic, but only the qualitative change of properties is concerned.

On the other hand, also this non-smoothness can be removed, if the limit $\sigma \to 0$ is carried out before the limit $\epsilon \to 0$ is taken, what leads to a discussion of consequences of the succession of limit procedures.

The given arguments lead to the conclusion that it is of course very important to distinguish the stochastic from the deterministic case, especially if the point of view is adopted to try to find some hidden determinism for prediction in seemingly unordered data, but because of the above shown limits connecting both cases under parameter changes, a more unified framework for the variants of dynamics in which both cases can be found in their relative position should be preferred in comparison with treating both cases as elements of disjoint and (far) distant boxes. The existence of an intermediate regime, where aspects of both dynamical variants are present, should be emphasized stronger.

Appendix B

Prediction via precursors: ROC plot from clustering analysis

B.1 Introductory remarks

At the basis of the following methods lies the assumption, that special signals called precursors precede extreme events. Without asking for the physics behind pre-structures, it is tried to identify reoccuring structures before wind gusts from data by a statistical method in order to improve prediction of extreme events. The method chosen in this case is cluster analysis. The following sections are rather understood as a presentation of the method than as the trial for finally optimized results. Again the wind dataset used for the analysis is taken at Lammefjord in Denmark in the year 1987.

B.2 Central tool: ROC plot

The main tool for characterizing the prediction skill of the method presented in the following is the receiver operating characteristic (ROC). In this method two binary classifications are compared with one another. On the one hand it is asked for the occurence (or not) of an event, e.g. wind gust, and on the other hand it is asked for a given alarm (or not) of an optional event. A clever combination of the respective frequencies allows for statements of predictive abilitities and is described in the following. An example ROC plot taken from [42] is given in fig. B.1.

The quantity written on the x-axis is the fraction of false alarms, i.e., the number of alarms, where finally nothing happens, divided by the total number of situations, where nothing happens (no gust). The quantity written on the y-axis contains the fraction of correctly predicted events, i.e., the fraction of events in an eventclass giving alarms divided by the total number of events in the class.

The dream of every prediction scheme is the following: No false alarms, but always an alarm if an event occurs. This would be the top left point of the diagram, but usually this is unaccessible.

Getting an alarm for all events forces in general to make the threshold for giving an alarm very low and this causes to get many false alarms. In the limit of lowest threshold the

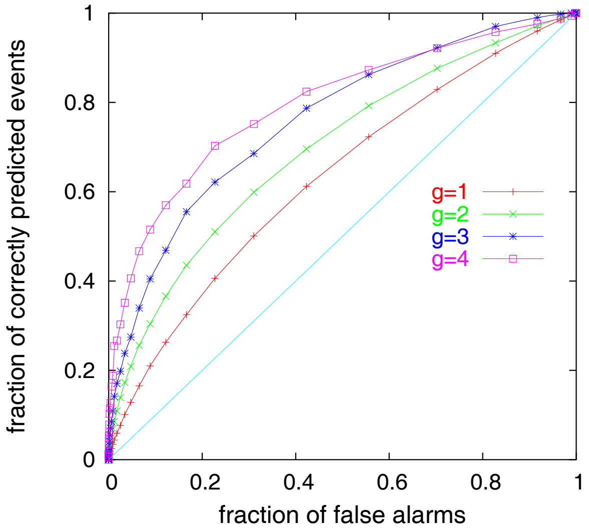


Figure B.1: ROC plot example

top right corner is reached. Decreasing the number of false alarms on the other hand forces a very high threshold and this causes to lose alarms if an event takes place. The limit of highest threshold corresponds to the bottom left corner.

If the threshold is parametrizable one gets a curve from the top right to the bottom left. The threshold parameter for giving an alarm decides for the point on the curves but doesn't appear explicitly.

If the alarms are given randomly the corresponding ROC curve is the diagonal which says that the prediction method is worthless. The interest is of course to raise the curve from the diagonal into a concave form as far to the top left corner as possible. For quantification of skills of the prediction scheme in a number it is possible to calculate the area below the curve. The smaller this area, the better the skills of the prediction scheme. Other summary indices, alternatives to the area above the ROC curve, are possible: The Kolmogorov-Smirnoff index is the largest distance from the diagonal [59] or derivative properties of the ROC curve are used [37].

The fraction of correctly predicted events depends on the chosen event class. This is the reason for different ROC curves for different event-sizes visible in fig. B.1.

B.3 Cluster analysis

In the following as a nonparametric clustering method it is used the *hierarchical clustering procedure* [20]. This method appears in the divisive (top-down) and agglomerative (bottom-up) mode, where the second is used. First, one has to decide about a measure of dissimilarity, i.e., an individual data distance measure. Examples are the usual euclidean or maximum norm, etc. From this it is possible to construct a distance matrix of every single event.

Second a cluster distance measure has to be defined. Examples for choices are the single linkage or the average linkage. With those ingredients the successive clustering procedure can be performed. For visualizing the clustering procedure it is possible to introduce a dendrogram, which shows in a tree structure the succession of clusterings. Quite similar to finding a strategy for an optimal Markov approximation also here a stopping rule for finishing the clustering has to be found.

An advantage of the hierarchical clustering procedure is that no assumptions about the underlying data distribution has to be made. Disadvantageous is the fact that it can never be repaired what was done in previous steps: Once two events are clustered, this cannot change anymore. Furthermore this method is quite memory-consuming.

B.4 Resulting prediction scheme

First, a simple criterion, what a gust should be, has to be defined. It is demanded for a wind gust that an increase of the wind speed should exceed a given threshold in a certain defined time window. After having collected all those gusts and shifted them in a way, such that all gusts start at the same time step, the mean value of the collected gusts at every time step is determined and shown in fig. B.2.

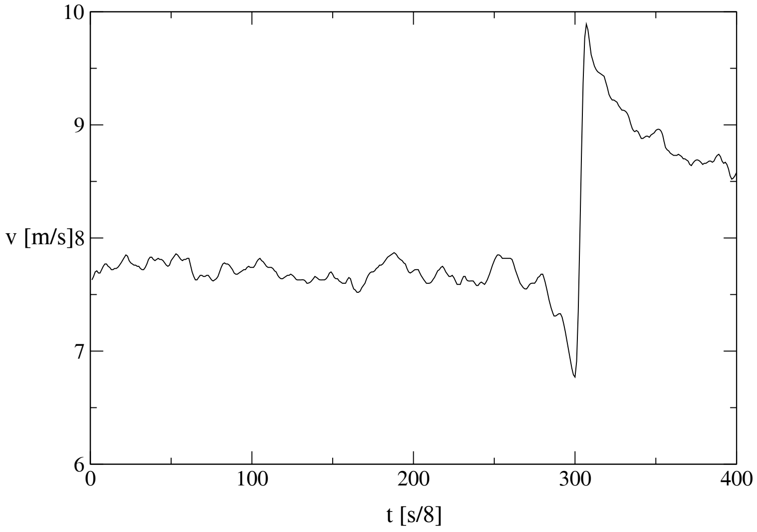


Figure B.2: Mean of gusts

It is possible to see that as some common feature there exists some decrease of the wind speed immediately in front of the wind gust or some wavy structure a bit further into the past which could be in some sense useful for prediction. On the other hand, the mean value as a statistic may be too coarse and not suitable for capturing more subtle precursor features. The idea is now to cluster similar pre-structures of wind gusts on a certain fixed analysis length (24 time steps chosen) in front of selected gusts. As a distance measure of single events it is chosen the euclidean norm and as the distance measure for existing clusters in the hierarchical clustering procedure the group average is chosen.

Because a coherent structure with higher velocity passes the measurement device faster, the data vector has to be horizontally rescaled as well as vertically shifted and possibly also rescaled in an adequate manner to make different gusts better comparable in the clustering procedure. It has to be admitted, that at this point a lot of unproved assumptions and heuristics enter the method and obscure the usefulness of it. The result of those transformations can be found in fig. B.3.

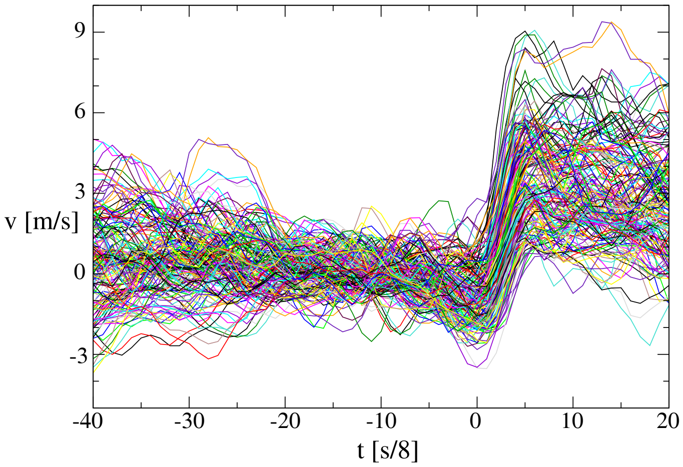


Figure B.3: Collection of adaequately transformed gusts

Fig. B.4 shows examples, which are obtained from the application of the clustering procedure.

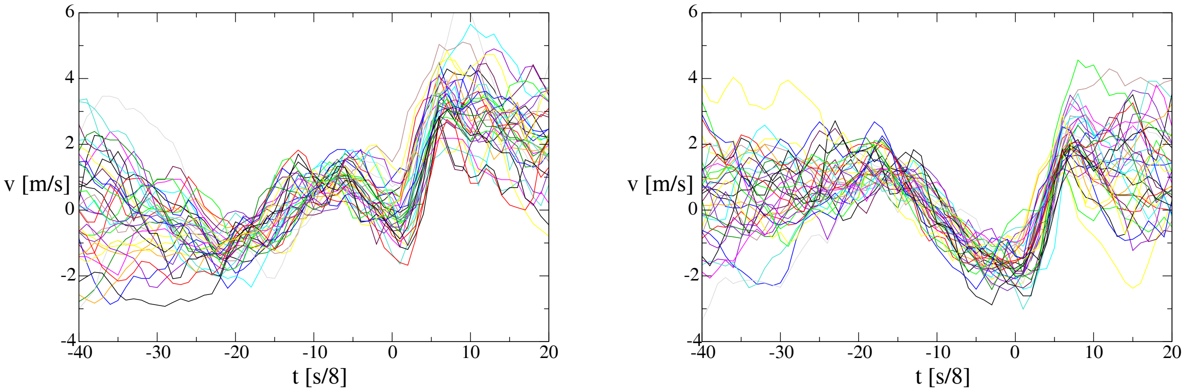


Figure B.4: Examples of clusters

After having finished the clustering procedure the mean structures of the clusters are calculated respectively.

The main idea of the implemented prediction scheme is now the following: In the same (in-sample) or another (out-of-sample) dataset from which the clusters were formed it is calculated for all data vectors the closest distance from one of the means of the clusters. A small distance gives a warning sooner than a big distance and now one is in the situation to construct a ROC plot, because one knows from the dataset if a gust is coming and from the distance one can give warnings or not.

B.5 Results

B.5.1 In-sample ROC as consistency check

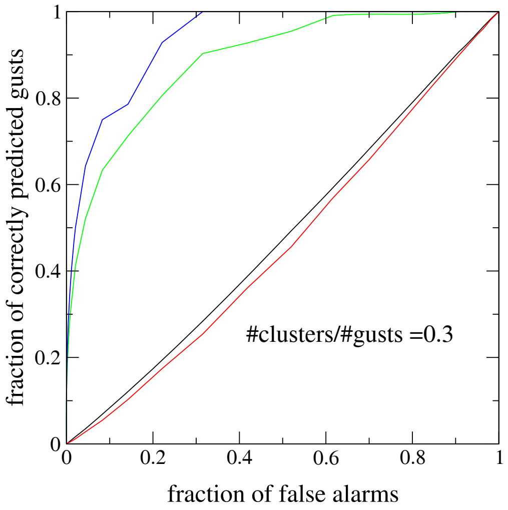


Figure B.5: In-sample ROC

In fig. B.5 the gust class corresponding to the green curve corresponds exactly to the selection criterion of gusts for the structure search and the gust class corresponding to the blue curve corresponds to even higher velocity increments (but also of course selected events). For small degree of clustering the clusters are more or less the events itself (clusters of one event) and in this sense very good ROC curves are trivially found in the in-sample case. In case of intermediate degree of clustering of 0.3 shown in fig. B.5, still rather good ROC curves can be found for the gust classes contributing to the pre-structure formation. However, a tendency of the ROC curves to depart from the top left point of the square is visible. The red and black curves correspond to gust classes of smaller gusts and their rather diagonal curves can be interpreted as a warning not to try to predict smaller gusts with the pre-structures of stronger gusts.

B.5.2 Out-of-sample prediction from clustered pre-structures

It should be mentioned that in the following the skill of the suggested prediction scheme is evaluated. The result gives a lower threshold on general predictability of wind gusts.

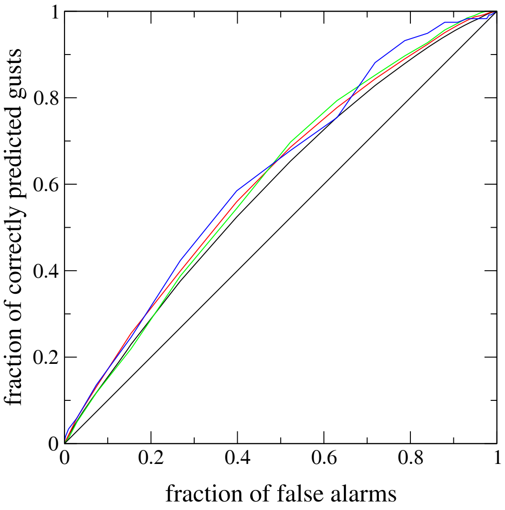


Figure B.6: Out-of-sample ROC

From the strategy of clustering from pre-structures some prediction skill for the prediction of wind gusts arises, but not much. Consequently some predictability must be inferred. No differences in predictability concerning different gust classes are seen.

B.6 Problems with this approach

- Because of its importance already mentioned in the explanation of the prediction scheme the transformations on the gusts for better comparability in the cluster analysis are absolutely essential, but the selection of the type of transformation is very questionable.
- The search for typical pre-structures under nonstationary conditions is also questionable
- Optimality in the break of the hierarchical clustering procedure is not clear
- Of course in principle it is possible to use the output of the scheme for finding the optimal perforated Markov model as the input model for the suggested clustering algorithm. Nevertheless it is quite questionable to use the optimal model found with concepts of information theory, i.e., with concepts in an 'average sense', as the input for an algorithm concerning extremes.

B.7 Research on precursors in general

The work in the spirit of the ROC analysis with focus on extreme events is continued more systematically and without clustering analysis mainly by colleague PhD-student S. Hallerberg for the models AR(1) and ARMA(p,q) and an easy precursor structure (one time step) analytically as well as numerically. Results are published in [37].

Appendix C

Is information theory a suitable starting point for predicting extremes, e.g., extreme wind events?

Information theory works out uncertainties in an average sense where all states are used and the result is mainly dictated by intermediate states. The extreme states have only a very tiny influence.

In contrast to information theory the ROC (cmp. appendix B), representatively used in the following for other potential scores for evaluating prediction schemes of extremes, receives its results from the explicit introduction of classes of certain events. Hence this predictive score (evaluation measure) is especially adaptable to extremes.

With respect to prediction of extremes there could be the idea of selecting relevant components of conditioning with the methods of information theory as carried out carefully in chap. 6, and using this as an input for precursory structures of the ROC. However, this was not carried out in this work due to several reasons, which will be presented in the following.

First, besides the time series of measurements of a physical quantity $\{x_i : i > 0\}$, e.g. wind velocities, a second binary time series $\{X_i : i > 0\}$ classifying event and non-event is introduced by

$$X_i = \begin{cases} 0 & \text{if no event at time } i \\ 1 & \text{if event at time } i \end{cases} . \tag{C.1}$$

The conditional probabilities $P(X_{n+1}|x_{n-m}, ..., x_n)$ of the two state random variable conditioned on a sequence of the physical quantity are used as the starting point for the method of information theory and statistics. The criterion given in eq. (6.3) should still work for the selection of optimal conditionings in case of this special random variable. Nevertheless, the selection of the optimal conditioning from information theory and statistics is dominated by non-events and is not adapted to the needs for a relative change of the relation of the hit rate over the false alarm rate. According to the normalizations, using information theory is only concerned with the x-axis of the ROC plot, i.e. the false alarm rate, but not with the y-axis, i.e. the hit rate, and hence also not with the ROC curve. The essential difference is that information theory carries out global averaging, whereas the ROC plot receives its quality from a clever representation of differently normalized quantities.

Thus, concerning prediction of extreme events, information theory as a basis for the selection of optimal conditioning should not even be tried out numerically and suggested as a promising alternative to a priori ROC plot optimization with a suitable summary index for obtaining the best conditioning. If anyway carried out the transition of the output of the method from information theory to the input of the statistical method of the ROC plot, it has to be clearly stated that the result is based on a *working hypothesis* and not obtained from a scientifically closed optimization procedure. Otherwise the use of the powerful concepts suggested pretends implicitly optimality, which is not present.

In order to use a method working in an average sense as input for a method with respect to extremes, the assumption must be made that the physics governing the extreme is the same as the physics governing the usual event. In some sense a scaling hypothesis including the largest scales is necessary. Is this generally justified or how large is the domain of validity of this assumption? Is it justified for wind gusts? For wind gusts one could argue with the Kolmogorov scaling, whereas it is not clear to what extent it can be assumed in the boundary layer. The result for increment statistics that the distribution of velocity increments is gaussian on large scales and becomes more and more intermittent (exponential) on smaller scales presented in ([10], [58]) is rather interpretable as a hint for qualitative discrepancies in the physics behind small and large scales. Without too much understanding nowadays about the physical details of the processes in the boundary layer, the validity of the posed hypothesis cannot generally be assumed. For arbitrary physical situations it is even more questionable if the assumption can be valid.

Another problem class of the combination of perforated Markov approximations based on the criterion (6.3) and extreme event prediction evaluated by the ROC in the case of wind speeds is the necessity of rescaling. A cluster analysis of prestructures of strong events according to appendix B, which serves for finding typical wind profiles before gusts, could be tried in order to know when to give an alarm for the ROC. This cluster analysis needs a rescaling of the velocity especially also in time direction in order to make prestructures for different mean velocities comparable, but exactly this necessary rescaling, which was already a finally somehow acceptable problem in the nonperforated case (cmp. appendix B), rescales also the selected perforation structure of the generalized Markov approximation. After rescaling, the selected components from information theory and statistics do not anymore belong to the same effective time distance into the past from the occurence of the extreme event. Finally the problem arises how to cluster objects with data at different effective times, i.e. with non-unified domains of definition. On the other hand, removing the rescaling makes the clustering extremely questionable because then the mean velocity dictates the clustering procedure. Rescaling of only the velocity and not the time in order to rescue the clustering procedure is scientifically unhonest. In order to solve this problem a lot of heurism is necessary.

It cannot be completely excluded that with some unavoidable suitable assumptions and under certain rather restricted circumstances there is a small chance that the output of information theory can be usefully used in a method for predicting extremes. However, the author is not a supporter of pushing together methods, if it is a priori clear that they do not belong together in a *natural* sense and which have a priori a small potential of being successfully applicable in a sufficient generality. Why study physics? Because of the intention of a very broad generality of the statements. We learn natural laws because of their huge invariance properties. Is it under such a point of view valuable to support an idea of using information theory for component selection in an average sense as input for a special variant of the ROC plot, where immediately examples could be constructed, in which such a procedure would not work?

In order to finally answer the posed question in the headline in the style of a proof, the arguments would have to be extended from the special prediction evaluation measure ROC to arbitrary evaluation measures. This is not carried out.

The serious problem on hand was solved in the following way: Instead of using information theory for the selection of optimal conditioning in combination with the ROC method of prediction evaluation in an unwanted sense with uncontrolled relative behaviour of hit rate and false alarm rate, it was decided to use a point prediction evaluated by the rms prediction error. From its 'average sense' this prediction method is close to the spirit of information theory, but not especially focussed on extremes. This is theoretically more smooth and honest than pretending a method suitable for prevention of damage from rotor blades. The price are less powerful results concerning prediction of the wind in chap. 9 than possible results for prediction of extreme wind events from a more suitable method as starting point. This is hopefully compensated by the much stronger results in the closer environment of information theory in chap.'s 3 (especially 3.1.6 and 3.8), 4 and 6 (especially 6.3ff and 6.6). As a consequence the ansatz for predicting extremes in appendix B stands rather unconnected with the rest of this work.

Concluding, with the intention of better access to the prediction of *extreme events*, a way based on information theory should not be recommended. This appendix is not written in order to tell that information theory is a priori a useless concept, but only that restrictions exist with respect to applicability.

Bibliography

[1] R.J. Adrian, C.D. Meinhart, and C.D. Tomkins. Vortex organization in the outer region of the turbulent boundary layer. *Journ. Fluid Mech.*, **422**, 1–54, 2000.

[2] H. Akaike. A new look on statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723, 1974.

[3] R. Badii and A. Politi. Statistical description of chaotic attractors: The dimension function. *Journ. Stat. Phys.*, **40**(5,6), 725–750, 1985.

[4] G. Baier and M. Klein. Maximum hyperchaos in generalized Hénon maps. *Phys. Lett. A*, **151**(6,7), 281–284, 1990.

[5] H. Bauer. *Probability Theory.* W. de Gruyter, Berlin, New York, 1996.

[6] C. Beck and F. Schlögl. *Thermodynamics of chaotic systems.* Cambridge University press, Cambridge, 1993.

[7] K. Behnen and G. Neuhaus. *Grundkurs Stochastik.* Teubner, Stuttgart, third edition, 1995.

[8] N.M. Blachman. Generalization of mutual information. *Proceedings of the Institute of Radio Engineers*, **49**(8), 1331–1332, 1961.

[9] A.K. Blackadar. *Turbulence and diffusion in the atmosphere.* Springer, Berlin, 1996.

[10] F. Böttcher, C. Renner, H.-P. Waldl, and J. Peinke. On the statistics of wind gusts. *Boundary-Layer Meteorology*, **108**, 163–173, 2003.

[11] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time Series Analysis: Forecasting and Control.* Prentice-Hall, New Jersey, third edition, 1994.

[12] O. Brasseur. Development and application of a physical approach to estimating wind gusts. *Monthly Weather Review*, **129**, 5–25, 2001.

[13] J. Bröcker. Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, **22**, 382–388, 2007.

[14] T. Burton, D. Sharpe, N. Jenkins, and E. Bossanyi. *Wind Energy Handbook.* Wiley, New York, 2001.

[15] M. Cencini, M. Falcioni, E. Olbrich, H. Kantz, and A. Vulpiani. Chaos or noise: Difficulties of a distinction. *Phys. Rev. E*, **62**(1), 427–437, 2000.

[16] C. Chatfield. *Time-series forecasting.* Chapman & Hall, Boca Raton, 2000.

[17] T.M. Cover and J.A. Thomas. *Elements of information theory.* Wiley, New York, 1991.

[18] J.P. Crutchfield and D.P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos*, **13**(1), 25–54, 2003.

[19] J.-P. Eckmann and D.Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, **57**, 617–656, 1985.

[20] B.S. Everitt, S. Landau, and M. Leese. *Cluster Analysis.* Arnold, London, fourth edition, 2001.

[21] J. Farmer. Chaotic attractors of an infinite-dimensional dynamical system. *Physica D*, **4**, 366–393, 1982.

[22] D. Feldman. Information theory, excess entropy and computational mechanics. Available from: http://hornacek.coa.edu/dave/Tutorial/index.html, 1998.

[23] R.P. Feynman. *The Feynman lectures on physics II.* Addison-Wesley, Reading, Massachusetts, 1963.

[24] E. Fick, M. Fick, and G. Hausmann. Logistic equation with memory. *Phys. Rev. A*, **44**, 2469–2473, 1991.

[25] M.E. Fisher. Renormalization group theory: Its basis and formulation in statistical physics. *Rev. Mod. Phys.*, **70**, 653–681, 1998.

[26] A.C. Fowler and G. Kember. Delay recognition in chaotic time series. *Phys. Lett. A*, **175**, 402–408, 1993.

[27] A.M. Fraser and H.L. Swinney. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, **33**(2), 1134–1140, 1986.

[28] U. Frisch. *Turbulence.* Cambridge University Press, Cambridge, 1995.

[29] C.W. Gardiner. *Handbook of Stochastic Methods: for Physics, Chemistry and the Natural Sciences.* Springer, Berlin, Heidelberg, third edition, 2003.

[30] J.R. Garratt. *The atmospheric boundary layer.* Cambridge University Press, Cambridge, 1992.

[31] P. Gaspard and X.-J. Wang. Noise, chaos and (ϵ, τ)-entropy per unit time. *Phys. Rep.*, **235**(6), 291–343, 1993.

[32] P. Grassberger. Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.*, **25**(9), 907–938, 1986.

[33] P. Grassberger. Finite sample corrections to entropy and dimension estimates. *Phys. Lett. A*, **128**(6-7), 369–373, 1988.

[34] P. Grassberger. Randomness, information and complexity. In R. Rechtman et al., editor, *Proceedings 5. Mexican School on Statistical Mechanics*, 1989.

[35] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D*, **9**, 189–208, 1983.

[36] P. Grassberger, T. Schreiber, and C. Schaffrath. Nonlinear time sequence analysis. *Int. J. of Bifurcations and Chaos*, **1**(3), 521–547, 1991.

[37] S. Hallerberg, E.G. Altmann, D. Holstein, and H. Kantz. Precursors of extreme increments. *Phys. Rev. E*, **75**(1), 016706, 2007.

[38] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer, New York, 2001.

[39] R. Hegger, H. Kantz, and E. Olbrich. Correlation dimension of intermittent signals. *Phys. Rev. E*, **56**(1), 199–203, 1997.

[40] R. Hegger, H. Kantz, and T. Schreiber. TISEAN 2.1, 2000. Available from: http://www.mpipks-dresden.mpg.de/∼tisean/TISEAN_2.1/index.html.

[41] U. Högström. Review of some basic characteristics of the atmospheric surface layer. *Boundary-Layer Meteorology*, **78**, 215–246, 1996.

[42] H. Kantz, D. Holstein, M. Ragwitz, and N.K. Vitanov. Extreme events in surface wind: Predicting turbulent gusts. In S. Bocccaletti et al., editor, *Experimental Chaos: Proceedings of the 8th Experimental Chaos Conference*, volume 742. American Institute of Physics, 2004.

[43] H. Kantz, D. Holstein, M. Ragwitz, and N.K. Vitanov. Markov chain model for turbulent wind speed data. *Physica A*, **342**, 315–321, 2004.

[44] H. Kantz, D. Holstein, M. Ragwitz, and N.K. Vitanov. Short time prediction of wind speeds from local measurements. In J. Peinke, P. Schaumann, and S. Barth, editors, *Proceedings of the EUROMECH Colloquium 464b Wind Energy.* Springer, 2005.

[45] H. Kantz and E. Olbrich. Coarse grained dynamical entropies: Investigation of high-entropic dynamical systems. *Physica A*, **280**, 34–48, 2000.

[46] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis.* Cambridge University Press, Cambridge, second edition, 2004.

[47] M.B. Kennel, R. Brown, and H.D.I. Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A*, **45**(6), 3403–3411, 1992.

[48] A. I. Khinchin. *Mathematical foundations of information theory.* Dover Publications, New York, 1957.

[49] P.E. Kloeden and E. Platen. *Numerical Solutions of Stochastic Differential Equations.* Springer, Berlin, Heidelberg, second edition, 1995.

[50] K. Kornwachs. *Information - new questions to a multidisciplinary concept.* Akademie Verlag, Berlin, 1996.

[51] M.T. Landahl and E. Mollo-Christensen. *Turbulence and random processes in fluid mechanics.* Cambridge University Press, New York, second edition, 1992.

[52] M. Lesieur. *Turbulence in fluids.* Kluwer Academic Publishers, Dordrecht, Boston, London, third edition, 1997.

[53] K. Nakamura, R. Kershaw, and N. Gait. Prediction of near-surface gusts generated by deep convection. *Meteorol. Appl.*, **3**, 157–167, 1996.

[54] F.T.M. Nieuwstadt and P.G. Duynkerke. Turbulence in the atmospheric boundary layer. *Atmospheric Research*, **40**, 111–142, 1996.

[55] E. Ott. *Chaos in dynamical systems.* Cambridge University press, Cambridge, second edition, 2002.

[56] H.C. Öttinger. *Stochastic processes in physics and chemistry.* Springer, Berlin, Heidelberg, 1996.

[57] L.M. Pecora, L. Moniz, J. Nichols, and T.L Carroll. A unified approach to attractor reconstruction. *Chaos*, **17**, 013110, 2007.

[58] J. Peinke, S. Barth, F. Böttcher, D. Heinemann, and B. Lange. Turbulence, a challenging problem for wind energy. *Physica A*, **338**, 187–193, 2004.

[59] M.S. Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, Oxford, 2003.

[60] S.B. Pope. *Turbulent flows.* Cambridge University Press, Cambridge, 2000.

[61] D.I. Pullin and P.G. Saffman. Vortex dynamics in turbulence. *Ann. Rev. Fluid Mech.*, **30**, 31–51, 1998.

[62] H. Risken. *The Fokker-Planck Equation.* Springer-Verlag, Berlin, second edition, 1989.

[63] J. Rissanen. Consistent order estimates of autoregressive processes by shortest description of data. In O. L. R. Jacobs et al., editor, *Analysis and Optimization of Stochastic Systems.* Academic Press, New York, 1980.

[64] T. Sauer and T. Yorke. How many delay coordinates do you need? *Int. J. Bifurcation and Chaos*, **3**, 737–744, 1993.

[65] T. Sauer, T. Yorke, and M. Casdagli. Embedology. *Journ. Stat. Phys.*, **65**, 579–616, 1991.

[66] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, **85**(2), 461–464, 2000.

[67] H.G. Schuster. *Complex Adaptive Systems*. Scator Verlag, Saarbrücken, 2003.

[68] H.G. Schuster and W. Just. *Deterministic Chaos: An introduction*. Wiley-VCH, Weinheim, fourth edition, 2005.

[69] C. Shalizi. Syllabus for Advanced Probability II, Stochastic Processes. lecture notes, 2006.

[70] Z. Sorbjan. *Structure of the Atmospheric Boundary Layer*. Prentic Hall, New Jersey, 1989.

[71] F. Takens. Detecting strange attractors in turbulence. In *Lecture Notes in Mathematics*. Springer, New York, 1981.

[72] J. Theiler. Estimating fractal dimension. *Journ. Opt. Soc. Am. A*, **7**, 1055–1073, 1990.

[73] Y.C. Tian and F. Gao. Extraction of delay information from chaotic time series based on information entropy. *Physica D*, **108**, 113–118, 1997.

[74] A. Tsinober. *An informal introduction to turbulence*. Kluwer Academic Publishers, Dordrecht, 2001.

[75] N.G. van Kampen. *Stochastic processes in physics and chemistry*. Elsevier Science Publishers B.V., Amsterdam, 1992.

[76] J.M. Wallace and P.V. Hobbs. *Atmospheric science*. Academic Press, San Diego, London, second edition, 2006.

[77] D.S. Wilks. *Statistical methods in the atmospheric sciences*. Academic Press, San Diego, London, second edition, 2006.

[78] J.C. Wyngaard. Atmospheric turbulence. *Ann. Rev. Fluid Mech.*, **24**, 205–233, 1992.

[79] R.W. Yeung. A new outlook on Shannon's information measures. *IEEE Transactions on information theory*, **37**(3), 466–474, 1991.

[80] L.-S. Young. Entropy in dynamical systems. In G. Keller and Warnecke, editors, *Entropy*, pages 313–328. Princeton Univ. Press, 2003.

[81] R. Zwanzig. *Nonequilibrium statistical mechanics*. Oxford University Press, Oxford, 2001.

Acknowledgements

First of all, I would like to thank Prof. Dr. Holger Kantz for the supervision of this work and for a lot of controversial and hence fruitful discussions.

Also I want to acknowledge all people involved in extensions of my working contract, which didn't lose the final confidence that this work could be finished properly, even though the time limit was strongly exceeded.

For careful corrections of or influential discussions about this work I would like to thank Holger Kantz, Markus Niemann, Thomas Laubrich, Rafael Vilela, Jochen Broecker, Sarah Hallerberg and Rainer Klages.

Sarah Hallerberg, Eduardo Goldani Altmann, Anja Riegert, Nikolay K. Vitanov, Mario Ragwitz and Holger Kantz I would like to thank for the collaboration concerning the scientific projects and publications.

The central advisor concerning my computational questions was Matthias Jurgk, who deserves special thanks.

My roommates Markus Porto, Giovanni Meacci and Rainer Bedrich enabled a pleasant working atmosphere during my stay at the MPIPKS.

Special thanks goes to my family Anke and Vincent Holstein for ... a lot of things.

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Die vorliegende Dissertation wurde am Max-Planck-Institut für Physik komplexer Systeme in Dresden unter der wissenschaftlichen Betreuung von Prof. Dr. Holger Kantz angefertigt.

Es haben keine früheren erfolglosen Promotionsverfahren von mir stattgefunden.

Hiermit erkenne ich die Promotionsordnung des Fachbereichs Mathematik und Naturwissenschaften der Bergischen Universität Wuppertal an.