1 Introduction.

1.1 Time series

In various situations information about a certain environment is represented as a sequence of measurements made consecutively in time. Such sequences are called time series. Data of this type play an important role in many geophysical sciences (meteorology, seismology, oceanology), in economical and financial applications, also in medical studies (electrocardiography, electroencephalography), etc.

The main goals of the analysis of time series are: modelling, forecasting, characterisation and classification. The aim of modelling is to build a model, i.e. a description of the mechanism (or the dynamics) that generates a given time series. In theory, this requires a thorough knowledge of the laws of physics, biology, etc., which govern the evolution of the environment related to the involved time series. However, in practice there is little or no prior knowledge of all the factors that influence the evolution. In that case one tries to build a model from the time series only, by studying its behaviour in the past. In general, a good model should capture essential features of the long-term behaviour of the system. On the other hand, the goal of forecasting, or predicting, is mainly to predict the future short-term evolution of the system. These two goals are not necessarily intersecting: a model that properly describes long-term governing mechanisms may fail in giving reliable short-term forecasts; and a model that is accurate in terms of short-term predictions may not give (and does not need to give) a good insight into the long-term properties of the system. In many cases, however, modelling and forecasting go hand in hand.

The purpose of characterisation is to determine the fundamental features of a system, such as the number of "degrees of freedom", the amount
Introduction.

of "randomness", etc., from the corresponding time series. This is closely related to classification, where the aim is to decide, on the basis of one or more characteristic features, whether a given time series belongs to a certain class. It can be the class of electrocardiograms of healthy persons, in which case nonassignment to this class may require further action or analysis, or, in a different context, the class of time series produced by some given mathematical model. For applications, the two most important objectives of time series analysis are classification and prediction.

In most practical cases discrete time series are considered, i.e. those series where time is indexed by integers (in contrast to reals, in which case we talk about continuous time series). We shall denote discrete time series by \( \{y_n\} \), the index \( n \in \mathbb{N} \) or \( n \in \mathbb{Z} \) denoting instants of time. A time series can be scalar \( (y_n \in \mathbb{R}) \) or multivariate \( (y_n \in \mathbb{R}^d, \ d > 1) \). Here we shall mainly concentrate on scalar time series.

One possible source of complex behaviour of time series is the presence of random factors, such as measurement errors, system noise, etc. The highest degree of randomness we can encounter would be a time series which represents a sequence of outcomes of independent random variables, with no further structure. This, however, is usually not of interest for time series analysis.

The crucial difference between time series, and situations usually considered in classical statistics, is that the measurements \( y_n \) and \( y_m, \ m \neq n \), will be stochastically dependent. Indeed, almost always a sequence of measurements made in equally spaced time instants is regarded as a realisation of a process which, in principle, has continuous sample paths \( \{y_t\}, t \in \mathbb{R} \). Then, as \( t \) and \( s \) get close, the variables \( y_t \) and \( y_s \) obviously be dependent. So giving up the assumption of independence is dictated by such a fundamental fact as (apparent) continuity of natural phenomena.

The opposite of pure randomness is considered by the theory of deterministic dynamical systems. Here the future evolution is uniquely determined by the initial state and the evolution law. The behaviour of a deterministic system is not necessarily simple. In fact, developments in the area of nonlinear dynamical systems show the existence of deterministic time series which display highly erratic behaviour and may look like a realisation of a random process. In this case one speaks of the so-called chaotic dynamics. We shall consider this in more detail in later sections.
Besides the question of the degree of determinism in a time series, it is important to model the dependence of future observations on the previous ones. Beginning with Yule's invention in 1927 of linear autoregression for the analysis of sunspot data, linear models have dominated time series analysis for about half a century. In these models the dependence of future observations on past ones is linear. To model complex behaviour by such a simple system, the presence of external random perturbations must be assumed. In traditional linear models driven by noise, such as the AR and ARMA models, a future observation is taken to be a linear combination of a certain number of previous observations and random, mostly Gaussian, disturbances, the so-called innovations. However, as we shall see in the following sections, there are simple examples of time series, for instance those related to chaotic dynamical systems, for which linear models are inadequate. This creates new problems: how to recognise such time series, and which methods to use for their modelling and prediction. These problems motivated many researches in the field of nonlinear time series analysis during the last decade. The present thesis is also an attempt to contribute to the theory of nonlinear time series.

The question of separating nonlinear time series from linear ones is rather complex and cannot be answered just by visual inspection. As an example, consider the four time series segments in Fig. 1.1-4 (A-D). Only one of them was obtained from a linear model driven by noise and the other three have strongly nonlinear features. We suggest to the reader to guess which one is linear.

One can see that it is difficult, even for an experienced eye, to discriminate nonlinear time series from linear ones just by visual inspection. However, there are ways (graphical as well as analytical) to detect nonlinear time series and time series arising from deterministic systems. An important recent development in the theory of dynamical systems, the so-called state space reconstruction, assures that it is possible to reconstruct characteristic features of the underlying dynamics from the observed output, i.e., the actual time series. We shall consider this more in detail in the following sections.

Other crucial developments of the last decade, which also encouraged wide-spread interest in nonlinear methods, are the increased availability of powerful computers and the emergence of the field of machine learning,
such as neural networks. This shifted modelling more towards data-driven methods and allowed for the exploration of a larger set of potential models, including nonlinear ones.

1.2 Nonlinear models

Let a time series \( \{y_n\}_{n \in \mathbb{N}} \) be obtained by the rule

\[
y_{n+1} = f(y_n),
\]

where \( f : \mathbb{R} \to \mathbb{R} \) is a nonlinear function and \( y_0 \in \mathbb{R} \) is an initial value. Even in this simple deterministic one-dimensional case the nonlinear function \( f \) can cause complex behaviour of the resulting time series such that
it appears random. Perhaps the best known example is the logistic map 
\( f : [0, 1] \rightarrow [0, 1] \) given by 
\( f(x) = \lambda x(1 - x) \), which is known to exhibit chaotic behaviour for certain parameter values, e.g. \( \lambda = 4 \). The time series \( A \) in Fig. 1.1 was generated by the logistic map

\[
y_{n+1} = 4y_n(1 - y_n).
\] (1.2)

Deterministic models of the type (1.1) can be generalised to multidimensional ones of the form

\[
X_{n+1} = F(X_n),
\] (1.3)

where \( X_n \in \mathbb{R}^k \), \( n \geq 0 \) and \( X_0 \) is an initial value. The function \( F : \mathbb{R}^k \rightarrow \mathbb{R}^k \) is vector-valued. The difference equation (1.3) describes a discrete dynamical system in \( \mathbb{R}^k \), with the evolution map \( F \).

In reality, observations rarely evolve according to the model (1.3): usually there is some observational or measurement noise, as well as other random disturbances may be present in the system. Moreover, a deterministic model will inevitably be inadequate for modelling of real data. Therefore, it is more realistic to replace (1.3) by a model

\[
X_{n+1} = F(X_n, e_{n+1}),
\] (1.4)

where \( F : \mathbb{R}^{2k} \rightarrow \mathbb{R}^k \), and \( \{e_n\}_{n \in \mathbb{N}} \) is a sequence of \( k \)-dimensional random vectors, such that \( e_n \) is independent of \( \{X_i\}_{i \leq n} \) (which, too, is quite an assumption). We will call (1.4) a (discrete) stochastic dynamical system. The sequence of random vectors \( \{e_n\} \) is called the dynamical noise (also the system, or the intrinsic noise). For convenience of analysis we shall further assume that the dynamic noise is additive, so that (1.4) reduces to the model with additive noise

\[
X_{n+1} = F(X_n) + e_{n+1},
\] (1.5)

here \( F : \mathbb{R}^k \rightarrow \mathbb{R}^k \). If we have the following vector representations:

\[
X_n = (y_n, y_{n-1}, \ldots, y_{n-k+1}),
\]
\[
F(X_n) = (f(X_n), y_{n-1}, \ldots, y_{n-k+2}),
\]
\[
e_n = (e_n, 0, \ldots, 0),
\]
then (1.5) together with (1.6) results into

\[
y_{n+1} = f(y_n, y_{n-1}, \ldots, y_{n-k+1}) + e_{n+1},
\] (1.7)
and the function $f$ is nonlinear if $F$ is. Relation (1.7) defines a \textit{nonlinear autoregression model of order $k$} for the time series $\{y_n\}$. Conversely, the model (1.7) for the time series $\{y_n\}$ can be written as a stochastic dynamical system (1.5) by vectorising $\{y_n\}$.

Using the example of the time series above we introduce a common exploratory tool for detecting the nonlinearity of a time series, the so-called \textit{delay-1 map}: $y_n \rightarrow y_{n+1}$. The delay-1 map will also be called return map. In practice one plots $y_{n+1}$ vs. $y_n$. For the time series A, due to the relation (1.2) between the present and the next value, the points of the return map (see Fig.1.5) fall on the parabola, revealing the deterministic structure of this time series.

In most real-life situations the observational noise or other disturbances
cause points to spread out, but in many cases the character of the underlying dynamics and nonlinearities can still be partially deduced from the structure of the return map. The return maps of the other three time series are shown in Fig. 1.6-8. The return map of the time series D (Fig. 1.8) immediately reveals its nonlinear structure. This is a time series coming from the so-called Henon dynamical system (we shall describe it later in this section).

Since the time series C is oversampled (i.e. the measurements are made with a very short time interval), we plotted not the delay-1, but delay-35 map, i.e. not \(y_{n+1}\) vs. \(y_n\), but \(y_{n+35}\) vs. \(y_n\) (35 was taken because it is approximately a quarter of the “pseudocycle” of this time series). The fact that the points on Fig. 1.7 fall along a clearly visible though erratic curve indicates the nonlinear structure of this time series. This time series comes from a laboratory experiment of pressure measurement in a fluidized bed, which is believed to be a good example of a low-dimensional chaotic dynamical system with a low level of noise (see v.d. Stappen [81]).

In contrast to the other three maps, the return map in Fig. 1.6 does not reveal any familiar structure, with points spreading over the whole graph. This time series is a realisation of a linear autoregression of order 1, driven by Gaussian noise.

The use of return maps for recognising nonlinear time series is closely related to the technique of the state space reconstruction. This technique is based on a remarkable result in the area of chaotic time series, the Takens reconstruction theorem [76]. We shall formulate it in Section 1.4, and here we try to give a flavour of the reconstruction technique.

When studying return maps for a time series, we are interested in the behaviour of pairs of observations \((y_{n-1}, y_n)\). In case of a nonlinear time series this can provide a better understanding of the dynamics that generated this time series. Intuitively, one may expect the behaviour of the vectors \((y_{n-1}, y_n)\) to provide more insight into the underlying dynamics than the behaviour of the scalar observations \(y_n\). Consequently, for more complicated cases more information about the dynamics can be gained by studying three-dimensional return maps, i.e. vectors \((y_{n-2}, y_{n-1}, y_n)\), or even \(k\)-dimensional vectors \((y_{n-k+1}, \ldots, y_n)\) for some higher value of \(k\). This is the basic idea of reconstruction: one replaces a time series \(\{y_n\}\) by a sequence of vectors \((y_{n-k+1}, \ldots, y_n)\), which are called reconstruction vectors.
The reconstruction vectors "live" in the space $\mathbb{R}^k$, also called the embedding space, and its dimension $k$ is the embedding dimension. Note that both the space $\mathbb{R}^k$, where the stochastic dynamical system (1.5) evolves, and its dimension are naturally related to the embedding space and the embedding dimension.

We illustrate this technique using the example of the Henon dynamical system. This system evolves in phase space $\mathbb{R}^2$ according to the law:

\[(x, y) \rightarrow (1 - ax^2 + y, bx). \quad (1.8)\]

For the values $a = 1.4$, $b = 0.3$ it is known to exhibit chaotic behaviour. In Fig. 1.9 a trajectory of this system in the original coordinates $(x, y)$ is shown. The corresponding sequence of first coordinates $\{x_n\}$, is shown in Fig. 1.4. In Fig. 1.8 we visualised the reconstruction vectors generated by this time series. The two-dimensional embedding space is in this case given by the delay coordinates $(x_n, x_{n+1})$. Comparing Fig 1.8 with Fig. 1.9, we see a very similar structure in the embedding space and in the original phase space.

Figure 1.9: A trajectory of the Henon dynamical system

In contrast to this example, in most real-life situations neither an original phase space nor a dynamical law are known. What is actually observed is a one-dimensional time series of measurements made on some physical system. If one tries to study the underlying system on the basis of the
observed time series, a fundamental problem arises: the physical system and the observed data live in different spaces. The original system evolves in a multidimensional phase space, its the coordinates are given by the fundamental dynamical variables that describe the evolution. In practice it is impossible to measure and record all variables that define such a multidimensional phase space. One has at most measurements of one or a few dynamical variables, or possibly of a function of them, i.e. some derived quantity. For example, in a fluidized bed the mechanism of formation and movement of bubbles needs to be described by various dynamical variables, such as position, size and velocity of all bubbles, etc. But we only observe the series of pressure measurements at some fixed location in a fluidized bed. It is presumed that the pressure changes are caused by a dynamical process, but the local pressure itself is hardly a fundamental variable.

The reconstruction theorem assures that a time series of only one scalar variable contains enough "information" to reconstruct the dynamics in the original multivariate phase space, without any prior knowledge of it. In particular, if the embedding dimension $k$ is sufficiently high, then the trajectory generated by the series of reconstruction vectors in the embedding space mimics the evolution of the system in the phase space. Here the question of how big $k$ should be to allow such reconstruction, is essential. If the original state space is $d$-dimensional, then the smallest value of the embedding dimension $k$ which assures successful reconstruction is given by $k = 2d + 1$. In general, to view a one-dimensional object unambiguously we need to live in $\mathbb{R}^3$. By unambiguity we mean roughly that there is a one-to-one map from that object to $\mathbb{R}^k$ which preserves differential information.

The reconstruction technique assures that essential features of the original dynamics can be deduced from the behaviour of the reconstruction vectors. In turn, this opens the way for classification of time series according to characteristics of the underlying dynamics. Two quantities, which are often used for classification, are the so-called correlation integral and the correlation dimension. We shall give precise definitions and discuss them in detail in Section 1.3 of Introduction and in Chapter 2. Here we explain them (rather informally) in the context of time series and suggest how they can be used for classification.

The correlation integral was originally defined by Grassberger and Proc-
cacia [34] for the classification of time series arising from deterministic dynamical systems. They defined it via the sample correlation integral, which is the fraction of pairs of reconstruction vectors within distance \( r \), in a finite segment of length \( N \) of the time series:

\[
C_N^{(k)}(r) = \frac{2}{N(N-1)} \# \{(i, j) : 1 \leq i < j \leq N, \| X_i - X_j \|_{\text{max}} \leq r \}.
\]

They then defined the correlation integral as the limit

\[
C^{(k)}(r) = \lim_{N \to \infty} C_N^{(k)}(r).
\]

We will return to the question of existence of this limit in Chapter 3.

Of course in practice one does not have an infinite time series, but only a finite segment. The correlation integral is then estimated by its sample analogue \( C_N^{(k)}(r) \). This quantity offers one possible way of classification: we estimate the correlation integrals and then compare the obtained estimates to distinguish different types of time series. Initially, however, correlation integrals were used to distinguish deterministic time series from stochastic ones (by a stochastic time series we mean here those driven by noise). In this context another quantity, the correlation dimension, is even more important.

If the correlation integral behaves as \( C^{(k)}(r) \sim \text{const} \cdot r^{\alpha^{(k)}} \) for \( r \) in some neighbourhood of 0, then the exponent \( \alpha^{(k)} \) is called the correlation dimension of the sequence of \( k \)-dimensional reconstruction vectors. The limit of this sequence

\[
\alpha(\infty) = \lim_{k \to \infty} \alpha^{(k)}
\]

is sometimes also called the correlation dimension of the time series, again, provided such limit exists (see, for instance, [54]). The correlation dimension \( \alpha^{(k)} \) can be estimated by various methods (see Chapter 2), often using the sample correlation integrals.

The correlation dimension is defined above in terms of a time series, or, more precisely, the sequence of reconstruction vectors generated by it. It turns out that if the time series arises from a deterministic dynamical system, then, by virtue of the reconstruction theorem, the correlation dimension of this time series can give an idea about the complexity of the underlying dynamics, i.e. the number of active degrees of freedom. The next section addresses this in more detail.
The basis for discrimination between time series is the study of the behaviour of the correlation dimension $\alpha(k)$ as a function of $k$. For a time series induced by a deterministic dynamical system, when $k$ is sufficiently large, the reconstruction vectors will concentrate on a subset of $\mathbb{R}^k$ of lower dimension, since the trajectories of the reconstruction vectors replicate those in the state space where the underlying dynamical system evolves. On the other hand, for a stochastic time series, as $k$ grows in some range, the reconstruction vectors will always live on the whole space $\mathbb{R}^k$ (or a $k$-dimensional subset of it). Thus, one expects that for deterministic time series the estimates for correlation dimensions at some point remain constant even when further increasing $k$, while for stochastic time series the estimates increase together with the embedding dimension $k$.

These rather intuitive arguments have been widely used for distinguishing deterministic time series and still remain a useful tools for their classification. However, there are situations when this method can fail. According to a result of Osborne and Provenzale [54], there are cases of stochastically generated time series for which the sequence of correlation dimension estimates $\hat{\alpha}(k)$ also converges to a finite limit. This can happen, for example, when there is too much dependence in a time series. To treat such cases a modification of the method above was suggested: first generate a linear time series driven by Gaussian noise, which has the same autocorrelations as the time series of interest, and then compare the behaviour of the correlation dimension estimates for these two series. A clear difference is an indication of determinism in the original time series.

For a numerical illustration of the phenomenon described above, let us again consider the time series D induced by the Henon dynamical system and the linear time series B, which is an example of a stochastic time series. We estimated the correlation dimensions in both cases for increasing values of the embedding dimension $k$. Plots of the obtained estimates vs. $k$ are shown in Fig. 1.10. Note that, for values of $k$ larger than 2 the correlation dimension estimates for the time series D (Henon) stabilise around the theoretical value $\alpha = 1.24$ (see, for instance, [79]). For the linear time series B the estimates keep on growing together with $k$.

This is just one of many methods of discrimination between deterministic and stochastic time series. There are other methods introduced in the literature, such as the BDS test [14], classification based on prediction [18],
Figure 1.10: Estimated correlation dimension vs. embedding dimension: series D and B

etc. For a good review on this subject see [75].

Knowledge of the presence of determinism and nonlinearities in a time series can be used to build better predictors. Many nonlinear methods of prediction have been introduced in the literature. Among these methods the local methods of prediction are of particular importance. The main aim is to capture the local dynamics of the time series, when the effects of nonlinearities and the amplification of noise are not yet that strong. On a small scale the nonlinear dynamics can successfully be approximated by a linear function. This is the basic idea behind the local linear predictors - perhaps the most well-known prediction technique for nonlinear and chaotic time series. Other methods are closely related with methods of computer learning, such as neural networks. Other methods of prediction based on the investigation of general mechanisms that generated the time series. These methods, however, require deep knowledge of the subject area where the time series comes from, rather than of statistical methods of analysis. Here we shall not consider this class of prediction methods.

Prediction can also be useful for discrimination between deterministic and stochastic time series. For a discussion on classification by prediction see [18], [75]. The main ideas behind these classification methods are often similar to the idea behind the methods based on the correlation dimension estimation. For example, one can study the quality of local linear predictors, applied to nonlinear time series, as a function of the order $k$ of the
1.2. Nonlinear models

autoregression (in (1.7)), i.e. the embedding dimension. The same reasoning as above tells us that if the time series comes from a deterministic dynamical system, then, if $k$ is sufficiently large, considering an autoregression of even higher order will not bring any additional information. Therefore, an increase of the embedding dimension does not essentially improve the forecasts. In fact, it can even lead to worse predictions due to the growing effect of nonlinearities. Other argument against further increase of the embedding dimension is increasing the number of unknown parameters in the model that have to be estimated. For stochastic time series, however, due to the presence of noise, models with higher values of $k$ may be required to obtain better predictions. (Here, however, the same argument against increasing $k$ as above holds.) Since in nonlinear stochastic time series the effects of noise and nonlinearities are combined, an autoregression with intermediate values of $k$ would provide the best forecasts. Another, quite elegant criterion was suggested by Casdagli [18] (the so-called DVS criterion). It is based on finding the optimal number of neighbours that is necessary to fit a local linear model. It exploits the ideas similar to those considered above. This test has been shown to work well not only for discriminating between deterministic and stochastic time series, but also for distinguishing nonlinear and linear stochastic models.

We conclude the first two sections by saying that the study of time series can have various aims: (1) it can be pleasant for the mind of a rationalist to elaborate on examples of mathematical structures capturing some qualitative aspects of systems evolving in time; (2) it can be necessary for practical purposes to try to understand the mechanism behind a concrete sequence of univariate or multivariate observations: such understanding may be useful for making predictions. This thesis is more in line with the rationalist perspective (1) than with the empirist perspective (2). Nevertheless, perspective (2) is not ignored. To the contrary, it plays an essential role in the discourse.

If a concrete time series has been generated by a deterministic dynamical system, then knowledge of the factual presence of such determinism may be useful in constructing better predictors. Examples of such time series can be found everywhere, in fields such as physiology, seismology, in other physical contexts. In practice, some additional noise may be involved in the realisation of the concrete time series, but this should not dominate our
thinking about it. Of course, there are other situations where the aspect of a deterministic dynamical system is dominated by other aspects, e.g. the essential nonstationarity which appears, for instance, if growth and development studies are made of individuals.

Restricting ourselves to situations where the aspects of a deterministic dynamical system is strong, the acceptance of the "presence of determinism" may be used to increase the understanding and to construct better procedures for making predictions.

An interesting difference between such approach and the more classical statistical ones is that the predictors based on nonlinear dynamical systems will be nonlinear whereas in classical statistical theories the attention is often restricted to linear ones. This restriction is not essential for classical statistics: if distributions are different from multivariate normal, then the classical statistician will worry about the restriction of linearity. An advantage, however, is that using this restriction, theory can be developed which, like in the Gauss-Markov theories, does not require restrictive assumptions about the functional form of distributions.

In whole, this thesis makes the emphasis on "stationary" time series (mainly) generated by some deterministic dynamical system and on the possibility to recover information about this system from the observed outcomes.

1.3 Chaotic dynamical systems

In this section we discuss some aspects of chaotic dynamical systems, define some general notions and give the main assumptions. We shall focus on systems with discrete time.

A *discrete time dynamical system* \((\mathcal{X}, T)\) is a pair consisting of the *state space* \(\mathcal{X}\) - the set of all possible values, and the time evolution map \(T : \mathcal{X} \to \mathcal{X}\) - the law according to which a state evolves to other states at later times. For an initial state \(x_0 \in \mathcal{X}\), the iterations of \(T\) give rise to a *trajectory*, or an *orbit* \(\{T^nx_0\}_{n \in \mathbb{N}}\) (or, if \(T\) is invertible, \(n \in \mathbb{Z}\)).

A dynamical system is related to a time series by means of the *read-out function*, or the *observable function* \(f : \mathcal{X} \to \mathbb{R}\), which assigns to each possible state in \(\mathcal{X}\) the recorded value when the system is in that state. (As was mentioned in the section above, this is a particular model of a
nonlinear time series.)

The problem of discrimination between deterministic and stochastic
time series does not make sense unless some conditions on the underlying
dynamical system are imposed. Without these conditions, as we shall see,
the discrimination is impossible in principle, since a dynamical system
which does not satisfy these conditions can generate any time series.

We shall assume that the state space is finite dimensional, i.e. that it
is a (subset of) a finite dimensional Euclidean space \( \mathcal{X} \subset \mathbb{R}^d \), or can be
embedded into such a vector space. Also we shall assume that all positive
orbits of the dynamical system (i.e. those indexed by \( \mathbb{N}^+ \)) \( \{T^n x_0\}_{n \geq 0} \) are bounded. Now we present two examples of dynamical systems which violate
one of these assumptions and can generate any time series.

First, let a state space be infinite dimensional: the vector space \( G \) of all
functions \( g : \mathbb{N} \rightarrow \mathbb{R} \). Let the dynamical law be the map \( T : G \rightarrow G \),
which assigns to the function \( g \in G \) the function \( Tg \) defined by \( (Tg)(n) = g(n+1) \). Let the read-out function \( f : G \rightarrow \mathbb{R} \) be given by \( f(g) = g(0) \).
For an initial state \( g_0 \in G \) the orbit is \( g_0, Tg_0, T^2 g_0, \ldots \), and the time series is \( \{f(T^n g_0)\}_{n \in \mathbb{N}} = \{(T^n g_0)(0)\}_{n \in \mathbb{N}} = \{g_0(n)\}_{n \in \mathbb{N}} \). Now this time series
can be made anything we want by choosing an appropriate initial state -
function \( g_0 \).

The second example is a dynamical system with unbounded positive
orbits. Let the state space be \( \mathbb{R} \) and let the map \( T : \mathbb{R} \rightarrow \mathbb{R} \) be defined
by \( Tx = x + 1 \). For \( x_0 = 0 \) the orbit is unbounded. Let the read-out
function be \( f : \mathbb{R} \rightarrow \mathbb{R} \). Here we see that the corresponding time series
\( \{f(T^n x_0)\}_{n \in \mathbb{N}} = \{f(n)\}_{n \in \mathbb{N}} \) depends completely on the choice of the read-out
function, and therefore can be chosen arbitrary as well.

These two conditions, together with the assumptions that both \( T \) and
\( f \) are differentiable, allow us to avoid some obvious exceptional cases, such
as those mentioned above.

A way to get an impression of the behaviour of a dynamical system is to
study what happens asymptotically as \( n \rightarrow \infty \). The simplest case occurs
when the trajectories converge to a single point or limit cycle, called the
attracting point or the attracting cycle, respectively. In the latter case the
limit behaviour is periodic. More complicated limit behaviour appears as
quasiperiodic motion, when the trajectories are attracted to a 2-dimensional
torus. In these cases the limiting set \( \mathcal{A} \subset \mathcal{X} \), which is called an attractor, is
a simple geometrical object. It is possible that the attractor is none of the simple objects mentioned above, for example, it can be some Cantor-like set with a non-integer dimension. In such cases one speaks of a strange attractor.

There is no unique definition of attractor in the literature on dynamical systems. Here we shall not give an entirely precise mathematical definition of it. For a thorough discussion on different notions and definitions of an attractor see [51]. For our purposes we can say that an attractor is the limiting set where the experimental orbits \( \{T^n x\}_{n \in \mathbb{N}} \) accumulate for large \( n \). An example of an attractor - the attractor of Henon dynamical system - was given in Fig. 1.9.

A more precise definition is that of an attracting set. The set \( A \) is called an attracting set with fundamental neighbourhood \( U \), if it satisfies the following properties:

1. **Attractivity**: for every open set \( V : A \subset V \) we have \( \{T^n x : x \in U\} \subset V \) for all sufficiently large \( n \);

2. **Invariance**: for all \( x \in A \) and all \( n \) we have \( T^n x \in A \).

Together with the two properties above one usually requires an attractor to be irreducible in some sense. In practice the attracting sets defined above are generally referred to as attractors. In what follows, when we talk about an evolution on the attractor, we actually mean "in a neighbourhood of the attractor".

The notion of invariant measure is associated with a dynamical system. A finite measure \( \mu \) on the Borel \( \sigma \)-field \( \mathcal{F} \) of \( X \) is called \( T \)-invariant if, for any set \( B \in \mathcal{F} \), \( \mu(T^{-1}B) = \mu(B) \). Since \( \mu \) is finite, we can assume without loss of generality that \( \mu \) is a probability measure, i.e. that \( \mu(X) = 1 \).

To indicate the link between the dynamical system and the corresponding invariant measure, we shall sometimes write \((X, T, \mu)\) for a dynamical system.

A transformation \( T \) (if it is a homeomorphism) always has at least one invariant measure associated with it (if the state space is compact) [47]; in fact it can have more. Typically, there are many invariant measures on an attractor. Of particular interest for us are the so-called ergodic measures.

A \( T \)-invariant measure \( \mu \) is called ergodic if all \( T \)-invariant sets in \( X \) (i.e. all \( A \in \mathcal{F} \) for which \( T^{-1}(A) = A \)) have measure \( \mu(A) \) either 0 or 1.

It is exceptional that an attractor carries only one ergodic invariant
measure. In typical cases there are uncountably many distinct ergodic measures. Intuitively, however, it seems that there is one natural, physical measure on an attractor produced by the evolution - the one that describes how much time a trajectory spends on average in various parts of the attractor. A candidate for such a physical measure is the so-called SRB measure (from Sinai, Ruelle, Bowen). It can be specified by taking a point $x_0$ at random with respect to the Lebesgue measure on $\mathcal{X}$, and, for $A \subset \mathcal{X}$, considering the time averages

$$\frac{1}{n} \sum_{k=0}^{n-1} \delta_{T^k x_0}(A),$$

where $\delta$ is the Dirac delta-measure. Then the SRB measure $\rho$ is defined for all $x_0$ in a set $B \subset \mathcal{X}$ with Lebesgue measure $m(B) > 0$ by

$$\rho(A) = \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \delta_{T^k x_0}(A). \ (1.9)$$

For certain dynamical systems (the so-called Axiom A systems) it has been shown (Ruelle [62]) that the limits in (1.9) exist for all $x_0$ in a set of positive Lebesgue measure, and provide a unique SRB measure with support on the attractor.

We shall not go into further details as to when such a measure exists on an attractor, or how it is related to other ergodic measures. For a complete discussion on these aspects see [61]. Our further investigation will concentrate more on the properties of the physical ergodic measures carried by attractors, which are naturally defined by a dynamical evolution or the corresponding time series.

An attractor and an invariant measure carried by it provide us a global description of the asymptotic behaviour of a dynamical system. The dynamic on the attractor itself does not need to be simple. For some dynamical systems the evolution on the neighbourhood of the attractor may depend sensitively on initial conditions, i.e. the trajectories starting in nearby initial points diverge from each other at an exponential rate and after some time can be found in totally different parts of the attractor. This property, called the sensitive dependence on initial conditions, is the characteristic feature of chaos.
We say that the evolution on the attractor \( \mathcal{A} \) exhibits **sensitive dependence on initial conditions** if there is a positive constant \( C \) such that for any \( \epsilon > 0 \) and \( x \in \mathcal{A} \), there are \( x' \in \mathcal{A} \) and \( N > 0 \) such that

\[
\begin{align*}
(1) & \quad \rho(x, x') < \epsilon \\
(2) & \quad \rho(T^N x, T^N x') > C,
\end{align*}
\]

where \( \rho(\cdot, \cdot) \) is some distance on \( \mathcal{A} \).

Sensitive dependence on initial conditions is related to the so-called **Lyapunov**, or **characteristic exponents**. It measures the mean exponential rate at which nearby trajectories diverge with time. Suppose that \( \{x_i\}_{i \in \mathbb{N}} \) is the orbit of a chaotic discrete dynamical system corresponding to the initial condition \( x_0 \). If we slightly change the position of the initial point: \( x_0 \rightarrow x_0 + \delta x_0 \), the point at time \( n \) will also be different. In general, one might expect that, if \( \delta x_0 \) is small, \( \delta x_n \) is also small. But, due to the sensitive dependence on the initial condition, when \( n \) becomes large, the small distance between the initial values grows exponentially fast: \( \delta x_n \sim \delta x_0 \exp(\lambda n) \), where the mean rate of divergence of the trajectories \( \lambda \) is the Lyapunov exponent.

A 1-dimensional dynamical system has exactly one Lyapunov exponent. It can be defined as

\[
\lambda = \mathbf{E}_{\mu}(\log |T'(X)|),
\]

where the expectation is taken with respect to an ergodic invariant probability measure, the existence of which is assumed. Another way is to consider

\[
\lambda = \lambda(x) = \lim_{n \to \infty} \frac{1}{n} \log \|T^n(x)\|,
\]

where \( (T^n)' \) is the derivative of \( T^n \) evaluated at \( x \). In the case of ergodic invariant measure \( \mu \), the ergodic theorem implies that the limit in (1.11) exists and is constant \( \mu \)-almost everywhere, and, moreover, the two definitions (1.10) and (1.11) of \( \lambda \) are equivalent.

A \( d \)-dimensional dynamical system for \( d > 1 \) has a spectrum of \( d \) Lyapunov exponents \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \), measuring the exponential rate of divergence or contraction in different directions. They are defined as logarithms of the eigenvalues of the matrix

\[
\Lambda_x = \lim_{n \to \infty} \left[ (D_x T^n)' D_x T^n \right]^{1/2n}
\]

where \( D_x T^n \) is the matrix of the partial derivatives of the components of \( T^n \) at \( x \). In the case of an ergodic invariant measure \( \mu \), The multiplicative
ergodic theorem of Oseledec [55] implies that the limit in (1.12) exists and is constant $\mu$-almost everywhere if $\mu$ is an ergodic measure.

For a one-dimensional dynamical system positivity of the Lyapunov exponent implies sensitive dependence on initial conditions. In higher dimensions the existence of at least one positive Lyapunov exponent is evidence for the sensitive dependence on initial conditions, i.e. chaotic behaviour. This happens because the behaviour of the system is essentially determined by the largest Lyapunov exponent $\lambda_1$. If it is positive, then there is at least one direction in which the expansion at an exponential rate takes place. Then the neighbouring orbits diverge, which results in chaotic behaviour. In general, the exponential rate of growth of distances in $X$ when iterating $T$ is given by $\lambda_1$, and the rate of growth of the $d$-dimensional volume element by $\lambda_1 + \lambda_2 + \cdots + \lambda_d$.

The exponential divergence of trajectories means that two orbits that are so close that they cannot be distinguished at time zero, become distinguishable after some time. In this case we may speak of the "creation of information". The entropy is another characteristic of a chaotic dynamical system. It describes the asymptotic rate of production of information when iterating $T$. The subject of entropy is outside the scope of this manuscript, and so, we shall not define it here.

As we mentioned above, positive orbits of deterministic dynamical systems can be attracted to low-dimensional subsets of the state space. Other quantities which give us an idea about the complexity of the system are a dimension of the attractor and of an invariant measure on the attractor. Loosely speaking, a dimension is a measure of the amount of information needed to specify points accurately. Precise definition of a dimension can be obtained in many ways. For example, the so-called box-counting dimension, or capacity of the attractor $A$, is defined via the minimum number of closed balls required to cover $A$. Let $\epsilon > 0$ be small and suppose that $A$ is an interval in $\mathbb{R}$. Then the required number of balls of diameter $\epsilon$ needed to cover $A$ is approximately the reciprocal of $\epsilon$. If $A$ is a rectangular plane segment in $\mathbb{R}^2$, then this number is inversely proportional to $\epsilon^2$, etc. This scaling behaviour motivates the following definition: if the number of closed balls of radius $\epsilon$ needed to cover a set $E$ scales as $(1/\epsilon)^{d_B(E)}$, then $d_B(E)$ is the box-counting dimension of $E$. Other dimensions studied in the literature are: the Hausdorff dimension, the information dimension, the correlation
dimension; in the context of strange attractors they are often referred to as “fractal dimension”, because non-integer values are possible. For a good review on the question of dimensions see [20].

For purpose of statistical estimation, the correlation dimension is most appropriate because it characterises the invariant measure on the attractor, and it is relatively easy to estimate. Above we defined the correlation dimension in terms of a time series. Here we shall define the correlation dimension of an invariant measure on the attractor. It is again defined via the correlation integral:

\[ C(r) = (\mu \times \mu) \{ (X, Y) : \| X - Y \| \leq r \} \]

for \( r > 0 \), where \( X, Y \) are independently chosen points on the attractor. If the correlation integral scales as \( C(r) \sim \text{const} \cdot r^\alpha \) as \( r \to 0 \), then the exponent \( \alpha \) is called the correlation dimension of the invariant measure \( \mu \).

Much of the essential information about the dynamical system \( (\mathcal{X}, T) \) is contained in characteristics such as the fractal dimension of the attractor, the entropy or the Lyapunov exponents. In practice the dynamical system itself is rarely known, and, even if it is known, it is often impossible to compute these quantities precisely. Then we face the problem of estimating these characteristics from a single orbit of the dynamical system: \( (x_0, Tx_0, T^2x_0, ...) \). As we have argued above, in reality the problem is even more complicated: what we actually observe is not the orbit in the original state space \( \mathcal{X} \) but a real-valued time series \( \{ y_n \}_{n \in \mathbb{N}} \) of measurements made on the orbit, and all estimates must be based on \( \{ y_n \} \). The first step toward estimation is always the reconstruction of the original dynamics by embedding one dimensional data into a higher-dimensional space.

### 1.4 State space reconstruction

The technique of state space reconstruction is based on the reconstruction theorem of Takens [76]. It enables us to reconstruct the dynamics of a system from an observed time series.

Let \( (\mathcal{X}, T) \) be a dynamical system with finite-dimensional state space \( \mathcal{X} \) and bounded positive orbits \( \{ T^n x \}_{n \geq 0} \). To relate a dynamical system to a time series, we introduced a read-out function \( f : \mathcal{X} \to \mathbb{R} \). Usually \( f(x) \) represents a real-valued measurement made on a point \( x \in \mathcal{X} \). If


\( (x_0, Tx_0, T^2x_0, \ldots) \) is an orbit of the dynamical system with initial state \( x_0 \), then the corresponding time series is obtained by applying \( f \) to each point of the orbit: \( (f(x_0), f(Tx_0), f(T^2x_0), \ldots) \).

Let both \( T \) and \( f \) be continuously differentiable (or \( T, f \in C^1 \)). Define the vector-valued reconstruction map \( \text{Rec}_k : \mathcal{X} \rightarrow \mathbb{R}^k \) by

\[
\text{Rec}_k(x) = (f(x), f(Tx), \ldots, f(T^{k-1}x)) \in \mathbb{R}^k. \tag{1.13}
\]

**Theorem** (Takens, [76]) *In the Cartesian product of the space of \( C^1 \)-mappings on \( \mathcal{X} \) and the space of \( C^1 \)-functions from \( \mathcal{X} \) to \( \mathbb{R} \) there exists an open and dense subset \( U \), such that if \( (T, f) \in U \), then the reconstruction map \( \text{Rec}_k \), defined in (1.13), is an embedding, whenever \( k > 2 \dim(\mathcal{X}) \).*

We give some remarks to clarify this statement and its consequences.

1. The interpretation of the condition \( (T, f) \in U \) is that the statement of the theorem holds for “almost all”, or “generic” pairs \( (T, f) \). For a more complete discussion on the notion of genericity (and on reconstruction problems in general) see [64]. The reason why these conditions have to be imposed, is to exclude exceptional cases, for which the reconstruction will obviously fail. These are cases such as the observable function \( f \) being constant, or the map \( T \) being the identity. In the former case, applying a constant function to the points of the orbit destroys all the information about the orbit, so it is impossible to make the reconstruction. In the later case, the second, the third, and all other consecutive points of the orbit do not contain any additional information, so the reconstruction also fails here.

2. The key word in the theorem is “embedding”. The transformation from \( \mathcal{X} \) to \( \mathbb{R}^k \) given by the reconstruction map \( \text{Rec}_k \) is an embedding if the mapping of \( \mathcal{X} \) into \( \text{Rec}_k(\mathcal{X}) \) is continuously differentiable with continuously differentiable inverse. In less technical terms it means that \( \mathcal{X} \) and its image under the reconstruction map \( \text{Rec}_k(\mathcal{X}) \) are the same up to a diffeomorphic transformation. Furthermore, in the presence of an attractor \( \mathcal{A} \subset \mathcal{X} \) the reconstruction maps transform \( \mathcal{A} \) into its image in \( \mathbb{R}^k \). Under this transformation the differential structure of the attractor is preserved. Also the invariant measure on the reconstructed attractor is an image of the invariant measure of the original transformation. Moreover, since a diffeo-
morphism restricted to a bounded set gives only the distortion of distances by a factor which is bounded and bounded away from zero, the correlation dimensions of the original and the reconstructed attractor are the same (see [76]).

Here we also mention that this theorem is a variation of the well-known Whitney embedding theorem, but restricted to mappings of special form, namely the reconstruction maps.

3. For practical purposes, i.e., for reconstructing the attractor from the observed real-valued time series \( \{y_n\}_{n \in \mathbb{N}} \), we define the reconstruction vectors by

\[
X_n = \text{Rec}(T^n x_0) = (y_n, y_{n+1}, \ldots, y_{n+k-1})
\]

for \( k \geq 2\dim(\mathcal{X})+1 \). Then the trajectory \( \{T^n x_0\}_{n \in \mathbb{N}} \) in \( \mathcal{X} \) is an image of the trajectory generated by the sequence of reconstruction vectors \( \{X_n\}_{n \in \mathbb{N}} \) in \( \mathbb{R}^k \) under the diffeomorphism. Moreover, the reconstruction vectors accumulate in \( \mathbb{R}^k \) in a neighbourhood of the limit set which is diffeomorphic to the attracting set of orbits of \( T \) in \( \mathcal{X} \). The reconstructed invariant measure on this limit set, the correlation dimension and some other characteristics that are defined by the sequence \( \{X_n\}_{n \in \mathbb{N}} \) do not depend on the read-out function \( f \) but describe intrinsic properties of the dynamics.

4. If it is not known that the time series is produced by a deterministic dynamical system, e.g., because significant noise is present, then it is harder to interpret the results of our reconstruction. Also one should not forget that the choice of the embedding dimension \( k \) is essential. The problem of selecting a sufficiently large \( k \) can be difficult since the dimension of \( \mathcal{X} \) is usually unknown, and, hence, the criteria \( k > 2\dim(\mathcal{X}) \) of the Takens theorem cannot be applied directly. There is a number of methods for choosing \( k \) introduced in the literature (see [19], [18]). We shall not concentrate on this problem here and simply assume that a sequence of reconstruction vectors \( \{X_n\}_{n \in \mathbb{N}} \) or \( z, X_i \in \mathbb{R}^k \) is obtained from the outcome of the actual time series \( \{y_n\} \). All further analysis and the estimates considered below will be based on this sequence. The question, for which dynamical systems the sequence of reconstruction vectors provides a stationary stochastic process with respect to some ergodic probability measure, is in general very complex. For certain types of systems, the so-called Axiom A systems, this has been shown by Ruelle [62]. We shall not address this problem here and assume that a given sequence \( \{X_n\} \) is the outcome of a stationary
1.5 Ergodicity and mixing

In this section some ideas and results from ergodic theory to be used in the sequel are gathered together.

Let \( (X, T) \) be a discrete dynamical system with the Borel \( \sigma \)-field \( \mathcal{F} \) and a finite \( T \)-invariant measure \( \mu \). Recall that a dynamical system together with an invariant measure is called ergodic if all \( T \)-invariant sets have the measure \( \mu \) of 0 or 1.

An ergodic dynamical system may possess some mixing property with respect to the invariant measure. Many different types of mixing conditions can be considered, such as weak or strong mixing condition, weak Bernoulli, etc. The essential meaning of mixing is that the future evolution becomes “almost independent” of the past, as time goes by. We shall concentrate here on weak Bernoulli mixing.

Let \( \mathcal{E} = \{E_1, \ldots, E_m\} \) and \( \mathcal{D} = \{D_1, \ldots, D_n\} \) be finite measurable partitions of \( X \). Consider the following partitions:

1. \( \mathcal{E} \cup \mathcal{D} = \{E \cap D : E \in \mathcal{E}, D \in \mathcal{D}\} \)
2. \( T^{-1}\mathcal{E} = \{T^{-1}(E) : E \in \mathcal{E}\} \)
3. \( \mathcal{E}_s^t = T^{-(s-1)}\mathcal{E} \cup T^{-(s-2)}\mathcal{E} \cup \cdots \cup T^{-(t+1)}\mathcal{E} \cup T^{-t}\mathcal{E}, \ s < t, s, t \in \mathbb{N} \)

Denote by \( \mathcal{F}_s \) the \( \sigma \)-field generated by \( \mathcal{E}_s \), and by \( \mathcal{F}_0^\infty \) the smallest \( \sigma \)-field which contains all \( \mathcal{F}_s \). A partition is called a generator if the completion of \( \mathcal{F}_0^\infty \) with respect to \( \mu \) is \( \mathcal{F} \).

A partition \( \mathcal{E} \) is called weak Bernoulli for the dynamical system if

\[
\beta_k = \frac{1}{2} \sup_{s,t \in \mathbb{N}} \sum_{A \in \mathcal{E}_0} \sum_{B \in \mathcal{F}_{s+t+1}^{s+t+1}} |\mu(A \cap B) - \mu(A)\mu(B)| \to 0 \tag{1.14}
\]

as \( k \to \infty \). Here the \( \beta_k \) is called mixing coefficient. Intuitively, \( \beta_k \) measures how far from independence two pieces of the trajectory are when they are taken \( k \) units of time apart. We call the dynamical system \((X, T, \mu)\) weak Bernoulli if it has a weak Bernoulli partition which is a generator.

The weak Bernoulli property corresponds to the mixing condition known in the stochastic processes literature as absolute regularity. It is described
later how a weak Bernoulli dynamical system is related to an absolutely regular stochastic process. First we shall define the absolute regularity condition.

Let \( \{X_n\}_{n \in \mathbb{N}} \) be a stochastic process defined on the probability space \((\Omega, \mathcal{M}, \mathbb{P})\) and taking values in \(\mathbb{R}^k\). Denote by \(\mathcal{M}_a^\infty\) and \(\mathcal{M}_{a+k}^\infty\) the \(\sigma\)-fields generated by \(\{X_i, 1 \leq i < a\}\) and \(\{X_i, i \geq a + k\}\) respectively. The mixing coefficients \(\beta_k\) are defined in this context as

\[
\beta_k = \sup_a \mathbb{E}\{ \sup_{A \in \mathcal{M}_{a+k}^\infty} |\mathbb{P}(A|\mathcal{M}_a^\infty) - \mathbb{P}(A)|\}. \tag{1.15}
\]

It will be convenient to define \(\beta_0 = 1\). The process \(\{X_n\}_{n \in \mathbb{N}}\) is called absolutely regular if \(\beta_k \rightarrow 0\) as \(k \rightarrow \infty\).

The notion of absolute regularity was proposed by Kolmogorov about 1950, and further studied in the dissertation of his PhD student Vinokurov [82] and in the later work of Vinokurov [83].

Let \(P_{0k}^\infty\) be the measure induced by the process \(\{X_i\}\) on the \(\sigma\)-algebra \(\mathcal{M}_0^\infty \cup \mathcal{M}_{a+k}^\infty\), and \(P_{1k}^\infty\) be the product measure, defined by

\[
P_{1k}^\infty(A \cap B) = P_{0k}^\infty(A)P_{0k}^\infty(B)
\]

for \(A \in \mathcal{M}_{a+k}^\infty\) and \(B \in \mathcal{M}_0^\infty\). Rozanov and Volkonskii [59] showed that

\[
\beta_k = \frac{1}{2} \sup_a \| P_{0k}^\infty - P_{1k}^\infty \|_{TV},
\]

where \(\| \cdot \|_{TV}\) denotes the total variation norm.

If (1.15) is replaced by

\[
\alpha_k = \sup_a \sup_{A \in \mathcal{M}_a^\infty, B \in \mathcal{M}_{a+k}^\infty} |\mathbb{P}(AB) - \mathbb{P}(A)\mathbb{P}(B)|
\]

and \(\alpha_k \rightarrow 0\) as \(k \rightarrow \infty\), then another condition on the dependence is obtained, called strong mixing. It was introduced by Rosenblatt [60]. Note that \(\alpha_k \leq \beta_k\), so absolute regularity implies strong mixing.

Convergence of the mixing coefficients \(\beta_k\) or \(\alpha_k\) to 0 as \(k \rightarrow \infty\) indicates the fact that, when taken far apart, the “past” and the “future” of the process become “almost independent”, and the sequence \(X_1, X_2, \ldots\) is expected to exhibit an asymptotic behaviour which is close to that of an independent sequence.
We can relate the dynamics of \((\mathcal{X}, T, \mu)\) to the time evolution of a stationary stochastic process, and, moreover, that this process is absolutely regular if the dynamical system has a weak Bernoulli partition. First, we give a simple example illustrating the situation.

**Example 1.1** Consider the map \(T : x \mapsto 2x \pmod{1}\) on the interval \([0,1)\). For an initial value \(x_0\), \(x_n = T^n x_0\) is the observed process. Define the sequence \(\{a_n\}\) by \(a_i = 0\) if \(T^i x_0 \in [0, \frac{1}{2})\), \(a_i = 1\) if \(T^i x_0 \in [\frac{1}{2}, 1)\). Note that \(0a_0a_1a_2a_3\ldots\) is a binary expansion of \(x_0\), i.e. \(x_0 = \sum_{i=0}^{\infty} a_i/2^i\). The interval \([0,1)\) together with the Borel \(\sigma\)-field and the Lebesgue measure is a probability space, and \(a_0, a_1, a_2, \ldots\) are independent identically distributed random variables with \(P(a_i = 0) = P(a_i = 1) = \frac{1}{2}\). On the other hand, \(T^n x_0 = \sum_{i=0}^{\infty} a_{n+i}/2^i\), and so, the fully deterministic process \(\{T^n x_0\}_{n \in \mathbb{N}}\) is a functional of the i.i.d. process \(\{a_n\}_{n \in \mathbb{N}}\). The independence of the \(a_i\)'s reflects good mixing properties of \(T\).

In general, the situation is more complicated. Suppose that the dynamical system possesses a weak Bernoulli partition \(E\). For an initial state \(x_0 \in \mathcal{X}\) define the label, or itinerary process \(\{a_n\}_{n \in \mathbb{N}}\) by:

\[ a_n = a_n(x_0) = i \quad \text{if} \quad T^n(x_0) \in E_i, \quad i = 1, \ldots, m, \quad (1.16) \]

so that \(S(x_0) = (a_0a_1a_2\ldots)\) is a sequence of labels of the elements \(E_i\) of the partition \(E\) visited by the trajectory. The sequence \(\{a_n\}_{n \in \mathbb{N}}\) is the outcome of a stationary stochastic process on the probability space \((\mathcal{X}, \mathcal{F}, \mu)\) taking values in \(\{1, 2, \ldots, m\}^\mathbb{N}\). From definitions (1.14), (1.15) and (1.16) it follows that the label process \(\{a_n\}\) is absolutely regular if the partition is weakly Bernoulli, and the mixing coefficients coincide up to a factor of \(1/2\).

The question remains as to whether there is a correspondence between the realisations of the label process and the points in the state space, i.e. whether the trajectory \(\{T^n x\}_{n \in \mathbb{N}}, x \in \mathcal{X}\), can be uniquely recovered from the realisation of the stochastic process \(\{a_n\}\). It was shown (see [56]) that this is possible if the original partition \(E\) is a generator.

In this case there exists a map (partially defined) \(f : \{1, 2, \ldots, m\}^\mathbb{N} \to \mathcal{X}\), such that \(T^n x = f(a_n, a_{n+1}, \ldots)\), if \(\{a_n\}\) is the label process of \(x\). So, for a weak Bernoulli dynamical system, \(T^n x\) is represented as a functional of an absolutely regular stochastic process. For certain examples of piecewise...
expanding interval maps this can be checked (see [42]), and, moreover, for these maps the mixing coefficients decay exponentially fast. Furthermore, in this case $f$ is Lipschitz-continuous in the sense that there exists $\alpha \in (0, 1)$ such that

$$|f(a_0, a_1, a_2, \ldots) - f(a'_0, a'_1, a'_2, \ldots)| \leq \text{const} \cdot \alpha^t,$$

if $a_0 = a'_0, \ldots, a_t = a'_t$.

In the examples above, each point of the trajectory was uniquely associated with a one-sided infinite sequence of its itineraries. Note that in that case the single action of the map $T$ on all points of the trajectory $\{T^n x\}_{n \geq 0}$ is equivalent to shifting the sequence of itineraries $(a_0 a_1 a_2 \ldots)$ one unit to the left, i.e. the shift map on the space of all sequences $\{1, \ldots, m\}^\mathbb{N}$ is the image of the map $T$ acting on the original state space $X$. If $T$ is invertible, then a pair of sequences can be associated with each point $x$ of the orbit. This is the case for many dynamical systems, for instance, for hyperbolic toral automorphisms. One sequence then gives the itinerary of the forward orbit of the point and the other describes the backward orbit, i.e. the orbit obtained by iterating $T^{-1}$ in place of $T$, provided $T$ is invertable. These sequences are usually combined into one doubly-infinite sequence, i.e. we define the itinerary of $x$ by the rule

$$S(x) = (\ldots a_{-2} a_{-1}, a_0 a_1 a_2 \ldots)$$

where $a_n = i$ if and only if $T^n x \in E_i$ and $\mathcal{E} = \{E_i\}_{i=1,\ldots,m}$ in the case of hyperbolic toral automorphisms is a Markov partition. Thus, the case of doubly-infinite sequences of itineraries, is of particular importance for chaotic dynamical systems.

Even if the label process $\{a_n\}_{n \in \mathbb{Z}}$ is absolutely regular, the process $\{T^n x\}_{n \in \mathbb{Z}}$ typically is not, because it is a functional of the entire sample path. However, the condition (1.17) can still assure that the behaviour of the process $\{T^n x\}_{n \in \mathbb{Z}}$ is not unlike that of the absolutely regular process. But, in general, one cannot treat the data coming from a chaotic dynamical system as an absolutely regular process, and more investigation is required. We shall come back to this question in Chapter 5.
1.6 $U$-statistics

In this section we introduce the notion of $U$-statistics and mention some of the existing results on the asymptotic behaviour of $U$-statistics of dependent processes.

Let $\{X_n\}_{n \in \mathbb{N}}$ be a stationary sequence of random vectors with marginal distribution function $F$, taking values in $\mathbb{R}^k$, or a subset of it. $U$-statistics were initially introduced by Hoeffding [41] as estimators for functionals of the form

$$\theta(F) = \int_{\mathbb{R}^{km}} h(x_1, \ldots, x_m) dF(x_1) \cdots dF(x_m),$$

where the kernel function $h : \mathbb{R}^{k \times m} \rightarrow \mathbb{R}$ is symmetric in its arguments.

Hoeffding showed that, in case $X_1, X_2, \ldots$ are independent and the distribution $F$ is completely unknown, the minimum variance unbiased estimator of $\theta(F)$ is given by

$$U_n = \left( \frac{n}{m} \right)^{-1} \sum_{1 \leq i_1 < \cdots < i_m \leq n} h(X_{i_1}, \ldots, X_{i_m}).$$

$U_n$ is called the $U$-statistic of degree $m$ corresponding to the kernel $h$.

Examples of $U$-statistics are: the sample variance (with $k = 1$, $m = 2$, $h(x, y) = \frac{1}{2}(x - y)^2$), Gini’s mean difference ($k = 1$, $m = 2$, $h(x, y) = |x - y|$), Wilcoxon’s signed rank statistics ($k = 1$, $m = 2$, $h(x, y) = \mathbf{1}\{x + y > 0\}$), the sample correlation integral ($m = 2$, $h(x, y) = \mathbf{1}\{\|x - y\| \leq r\}$).

In recent years there has been considerable interest in $U$-statistics of stationary dependent processes. This has been motivated by numerous applications, and one of them is the estimation of the dimension of a chaotic attractor. Since a stochastic process arising from a dynamical system is, in general, not independent, the asymptotic behaviour of the estimators for the correlation integral and the correlation dimension can be studied via the corresponding properties of the $U$-statistics of dependent observations.

Strong consistency of the $U$-statistics of general stationary and ergodic sequences is established by the following theorem of Aaronson et al. [1].

**Theorem A1** (Aaronson et al.) Let $\{X_n\}_{n \in \mathbb{N}}$ be a stationary ergodic sequence with marginal distribution $F$, and let $h : \mathbb{R}^{km} \rightarrow \mathbb{R}$ be a measurable, bounded and $F^m$-a.e. continuous function. Then

$$U_n \xrightarrow{a.s.} \theta(F) \text{ as } n \rightarrow \infty.$$
In the case of absolutely regular sequences, a result similar to Theorem A1 was established under milder conditions on the kernel function - no continuity condition is required in this case.

**Theorem A2** (Aaronson et al.) Let \( \{X_n\}_{n \in \mathbb{N}} \) be a stationary absolutely regular sequence with marginal distribution \( F \), and let \( h : \mathbb{R}^{kn} \rightarrow \mathbb{R} \) be a measurable bounded function. Then

\[
U_n \xrightarrow{a.s.} \theta(F) \quad \text{as} \quad n \to \infty.
\]

To establish asymptotic normality of \( U \)-statistics stronger assumptions than ergodicity are required. The weakest mixing condition under which the central limit theorem for \( U \)-statistics was shown, is absolute regularity.

One of the main tools in studying \( U \)-statistics is Hoeffding’s decomposition, or projection method. It says that every \( U \)-statistic can be written as a finite weighted sum of degenerate \( U \)-statistics, i.e. those for which the expectation with respect to the product distribution \( F \times F \times \ldots \times F \) is zero. We illustrate this method on \( U \)-statistics of degree 2. A \( U \)-statistic of degree 2 can be decomposed as

\[
U_n = \theta(F) + \frac{2}{n} \sum_{i=1}^{n} [h_1(X_i - \theta(F))] + R_n, \tag{1.19}
\]

where

\[
h_1(x) = \int_{\mathbb{R}^k} h(x, y) dF(y).
\]

The term \( R_n \), called the remainder, is actually defined by the relation (1.19). For \( U \)-statistics of degree 2 it can be written as

\[
R_n = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} [h(X_i, X_j) - h_1(X_i) + h_1(X_j) + \theta(F)].
\]

If \( U_n \) itself is not degenerate, \( R_n \) is “small” comparing to the leading term \( \frac{2}{n} \sum_{i=1}^{n} [h_1(X_i - \theta(F))] \), which dominates the asymptotic behaviour and the variance of the \( U \)-statistics. This statement will be made precise in the following chapters.

Denote by

\[
\sigma_n^2 = \mathbb{E} \left[ \left( \frac{2}{n} \sum_{i=1}^{n} [h_1(X_i - \theta(F))] \right)^2 \right]
\]
the exact variance of the leading term in the above decomposition, and by
\[
\sigma^2 = \left[ \mathbb{E} h_1^2(X_1) - \theta^2(F) \right] + 2 \sum_{k=1}^{\infty} \left[ \mathbb{E} (h_1(X_1) h_1(X_{k+1})) - \theta^2(F) \right] \tag{1.20}
\]
its asymptotic variance, provided the sum converges absolutely. In that case
\[
4\sigma^2 = \lim_{n \to \infty} n\sigma^2_{\text{r}}.
\]
The following CLT for $U$-statistics was shown by Yoshihara [86].

**Theorem Y** (Yoshihara) \ Let \( \{X_n\}_{n \in \mathbb{N}} \) be a stationary absolutely regular process in \( \mathbb{R}^k \) with marginal distribution \( F \) and with mixing coefficients satisfying \( \beta_n = O(n^{-2+\delta'/\delta}) \) for some \( \delta' > 0 \). Assume that
\[
\max \left[ \mathbb{E}(h^{2+\delta'}), \mathbb{E}_{\mathbb{P} \times F}(h^{2+\delta'}) \right] \leq M
\]
for some \( \delta' > \delta, M < \infty \). Then
\[
\sqrt{n}(U_n - \theta(F)) \xrightarrow{D} \mathcal{N}(0, 4\sigma^2) \quad \text{as} \quad n \to \infty,
\]
provided \( \sigma^2 > 0 \), and the series (1.20) converge absolutely.

Denker and Keller [26] studied applications of $U$-statistics to data from chaotic dynamical systems and considered the case when \( \{X_n\}_{n \in \mathbb{N}} \) is not itself absolutely regular, but when \( X_n \) is a functional of some absolutely regular sequence. They proved the central limit theorem for $U$-statistics under some additional conditions on the functional and the kernel \( h \). This extension is of particular importance for data from dynamical systems. As was mentioned in Section 1.3, the sequence of observations on a dynamical system with a weak Bernoulli partition is not itself absolutely regular, but it is a functional of the absolutely regular label process. Thus, the rigorous approach to such data is that by Denker and Keller [26]. In Chapter 5 we shall extend their results to a broader class of functionals.

### 1.7 Outline

In this chapter we discussed some aspects of the analysis of time series and nonlinear modelling, and described some of the main concepts of the theory of chaotic dynamical systems, state space reconstruction, ergodic theory and $U$-statistics.
In Chapter 2 we shall review some of the most commonly used methods for estimating the fractal dimension of a strange attractor from a time series. Among them are the Grassberger-Proccacia method, based on estimation of the sample correlation integral, the Takens estimator of the correlation dimension and some other methods. It turns out that in most of these methods, estimators in the form of $U$-statistics play an essential role. Sometimes questions about the consistency and the asymptotic distribution of these estimators can be answered using existing results from the area of $U$-statistics. However, the specific of the time series setting requires asymptotic results for $U$-statistics of dependent observations, and in this area many questions were still open. Chapters 3, 5 and a part of Chapter 6 contain some new results on the asymptotic behaviour of $U$-statistics of dependent observations.

Chapter 3 is devoted to the question of the consistency of the Takens estimator for the correlation dimension. Some examples will illustrate that for some systems the Takens estimator is not consistent. We look for additional conditions under which the Takens estimator becomes consistent, and establish somewhat more general results concerning the consistency of $U$-statistics of stationary ergodic and of absolutely regular sequences, possibly with unbounded kernel.

As we discussed in Section 1.4, data from chaotic dynamical systems can be considered as a realisation of stationary stochastic process, which by itself does not necessarily satisfy any mixing condition. However, in many cases it is a functional of some mixing (e.g. absolutely regular) process. The known asymptotic results for functionals of mixing processes are scattered over the literature on related subjects, and do not always cover the cases encountered in the applications we want to consider. Moreover, it seems that there is no unifying approach for treating such processes. So in Chapter 4 we develop a general machinery for dealing with functionals of absolutely regular processes. We prove some moment inequalities, that are analogous to existing moment inequalities for $\alpha-$, $\beta-$, $\phi-$mixing sequences. With their help a number of limit theorems can be proved in a standard way. We show this using the central limit theorem, shown by Ibragimov and Linnik [45], presenting a new proof based on our technique. Then we give some new results, such as the functional central limit theorems for ordinary empirical processes and empirical processes indexed by functions. This
technique is extended in Chapters 5 and 6, that also deal with functionals of absolutely regular processes.

Chapter 5 is dedicated to $U$-statistics of dependent observations. This is mainly motivated by the problem of estimating the correlation integral and the correlation dimension from a time series. We prove the central limit theorem for $U$-statistics of functionals of absolutely regular processes, extending the result of Denker and Keller [26] to doubly-infinite functionals. Then we consider empirical processes of $U$-statistics structure of dependent observations. The motivation comes from the estimation of the correlation dimension by the Grasberger-Procaccia method. In this method the sample correlation integral has the form of a $U$-statistics that depends on a real parameter. Thus, the $U$-statistics estimator becomes a random function of a real parameter, rather than a random variable. It has the form of an empirical process of $U$-statistics structure. For this process we prove the functional central limit theorem for absolutely regular sequences and functionals of them. We apply these results to estimation problems for some dynamical systems.

Chapters 6 and 7 are oriented toward applications. Chapter 6 considers the important problem of estimating the variance of the sample correlation integral. This problem is essential for the classification of time series on the basis of values of the correlation integrals and correlation dimension. Methods for the variance estimation, such as Monte-Carlo, bootstrap, $U$-statistics decomposition, are presented together with some theoretical background, and tested on a simulated chaotic time series.

Chapter 7 concentrates on prediction methods for nonlinear time series. Some new nonlinear methods for prediction are suggested, and a number of well-known methods, such as neural networks, are considered. These methods are applied to the prediction of some real-life data, and their performances are compared with each other and with the performance of linear prediction methods.