# Chapter 5
# Complexity and Information Theory

What do we mean when by saying that a given system shows "complex behavior", can we provide precise measures for the degree of complexity? This chapter offers an account of several common measures of complexity and the relation of complexity to predictability and emergence.

The chapter starts with a self-contained introduction to information theory and statistics. We will learn about probability distribution functions, the law of large numbers and the central limiting theorem. We will then discuss the Shannon entropy and the mutual information, which play central roles both in the context of time series analysis and as starting points for the formulation of quantitative measures of complexity. This chapter then concludes with a short overview over generative approaches to complexity.

## 5.1 Probability Distribution Functions

Statistics is ubiquitous in everyday life and we are used to chat, e.g., about the probability that our child will have blue or brown eyes, the chances to win a lottery or those of a candidate to win the presidential elections. Statistics is also ubiquitous in all realms of the sciences and basic statistical concepts are used throughout these lecture notes.[1]

**Variables and Symbols** Probability distribution functions may be defined for continuous or discrete variables as well as for sets of symbols,

$$x \in [0, \infty], \qquad x_i \in \{1, 2, 3, 4, 5, 6\}, \qquad \alpha \in \{\text{blue}, \text{brown}, \text{green}\} \ .$$

For example we may define with $p(x)$ the probability distribution of human life expectancy $x$, with $p(x_i)$ the chances to obtain $x_i$ when throwing a dice or

[1] In some areas, like the neurosciences or artificial intelligence, the term "Bayesian" is used for approaches using statistical methods, in particular in the context of hypothesis building, when estimates of probability distribution functions are derived from observations.

with $p(\alpha)$ the probability to meet somebody having eyes of color $\alpha$. Probabilities are in any case positive definite and the respective distribution functions normalized,

$$p(x), p(x_i), p(\alpha) \geq 0, \qquad \int_0^\infty p(x)\, dx = 1 = \sum_\alpha p(\alpha), \qquad \ldots \ .$$

The notation used for a given variable will indicate in the following its nature, i.e. whether it is a continuous or discrete variable, or denoting a symbol. For continuous variables the distribution $\rho(x)$ represents a probability density function (PDF).

**Continuous vs. Discrete Stochastic Variables** When discretizing a stochastic variable, e.g. when approximating an integral by a Riemann sum,

$$\int_0^\infty p(x)\, dx \approx \sum_{i=0}^\infty p(x_i)\, \Delta x, \qquad x_i \ = \ \Delta x\, (0.5 + i) \ , \tag{5.1}$$

the resulting discrete distribution function $p(x_i)$ is not any more normalized; the properly normalized discrete distribution function is $p(x_i)\Delta x$. Note, that both notations $p_i$ and $p(x_i)$ are used for discrete distribution functions.[2]

**Mean, Median and Standard Deviation** The average $\langle x \rangle$, denoted also by $\bar{x}$, and the standard deviation $\sigma$ are given by

$$\langle x \rangle = \int x\, p(x)\, dx, \qquad \sigma^2 = \int (x - \bar{x})^2\, p(x)\, dx \ . \tag{5.2}$$

One also calls $\bar{x}$ the expectation value or just the mean, and $\sigma^2$ the variance.[3]
For everyday life situations the median $\tilde{x}$,

$$\int_{x<\tilde{x}} p(x)\, dx = \frac{1}{2} = \int_{x>\tilde{x}} p(x)\, dx \ , \tag{5.3}$$

is somewhat more intuitive than the mean. We have a 50 % chance to meet somebody being smaller/taller than the median height.

**Exponential Distribution** Let us consider, as an illustration, the exponential distribution, which describes, e.g. the distribution of waiting times for radioactive decay,

[2] The expression $p(x_i)$ is therefore context specific and can denote both a properly normalized discrete distribution function as well as the value of a continuous probability distribution function.

[3] In formal texts on statistics and information theory the notation $\mu = E(X)$ is often used for the mean $\mu$, the expectation value $E(X)$ and a random variable $X$, where $X$ represents the abstract random variable, whereas $x$ denotes its particular value and $p_X(x)$ the probability distribution.

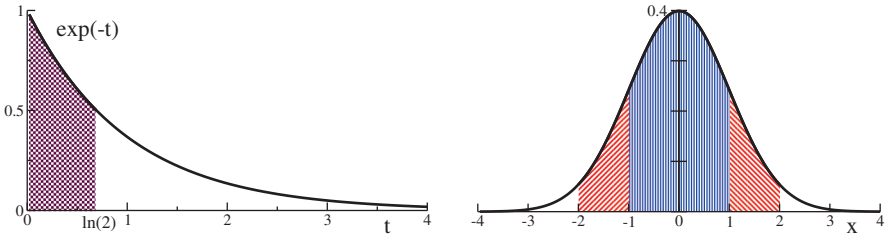


**Fig. 5.1** *Left*: The exponential distribution $\exp(-t/T)/T$, for an average waiting time $T = 1$. The *shaded area*, $t \in [0, \ln(2)]$, is $1/2$, where $\ln(2)$ is the median. *Right*: The normal distribution $\exp(-x^2/2)/\sqrt{2\pi}$ having a standard deviation $\sigma = 1$. The probability to draw a result within one/two standard deviations of the mean ($x \in [-1, 1]$ and $x \in [-2, 2]$ respectively, *shaded regions*), is 68 and 95 %

$$p(t) = \frac{1}{T}\,\mathrm{e}^{-t/T}, \qquad \int_0^\infty p(t)\,dt = 1\ , \tag{5.4}$$

with the mean waiting time

$$\langle t\rangle = \frac{1}{T}\int_0^\infty t\,\mathrm{e}^{-t/T}\,dt = -t\,\mathrm{e}^{-t/T}\big|_0^\infty + \int_0^\infty \mathrm{e}^{-t/T}\,dt = T\ .$$

The median $\tilde{t}$ and the standard deviation $\sigma$ are evaluated readily as

$$\tilde{t} = T\ln(2), \qquad \sigma = T\ .$$

In 50 % of times we have to wait less than $\tilde{t} \approx 0.69\,T$, which is smaller than our average waiting time $T$, compare Fig. 5.1.

**Standard Deviation and Bell Curve** The standard deviation $\sigma$ measures the size of the fluctuations around the mean. The standard deviation is especially intuitive for the "Gaussian distribution"

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \langle x\rangle = \mu, \qquad \langle (x-\bar{x})^2\rangle = \sigma^2\ , \tag{5.5}$$

also denoted "Bell curve", or "normal distribution". Bell curves are ubiquitous in daily life, characterizing cumulative processes (see Sect. 5.1.1).

The Gaussian falls off rapidly with distance from the mean $\mu$, compare Fig. 5.1. The probability to draw a value within $n$ standard deviation of the mean, viz the probability that $x \in [\mu - n\sigma, \mu + n\sigma]$, is 68, 95, 99.7 % for $n = 1$, 2, 3. Note, that these numbers are valid only for the Gaussian, not for a general PDF.

**Probability Generating Functions** We recall the basic properties of the generating function

$$G_0(x) = \sum_k p_k\,x^k\ , \tag{5.6}$$

introduced in Sect. **??**, for the probability distribution $p_k$ of a discrete variable $k = 0, 1, 2, \ldots$, namely

$$G_0(1) = \sum_k p_k = 1, \qquad G_0'(1) = \sum_k k\, p_k = \langle k \rangle \equiv \bar{k} \tag{5.7}$$

for the normalization and the mean $\langle k \rangle$ respectively. The second moment $\langle k^2 \rangle$

$$\langle k^2 \rangle = \sum_k k^2\, p_k\, x^k \Big|_{x=1} = x \frac{d}{dx} \left( x\, G_0'(x) \right) \Big|_{x=1} \tag{5.8}$$

allows to express the standard deviation $\sigma$ as

$$\begin{aligned} \sigma^2 = \langle (k - \bar{k})^2 \rangle = \langle k^2 \rangle \, - \, \bar{k}^2 &= \frac{d}{dx} \left( x\, G_0'(x) \right) \Big|_{x=1} \, - \, \left( G_0'(1) \right)^2 \\ &= G_0''(1) \, + \, G_0'(1) - \left( G_0'(1) \right)^2 \ . \end{aligned} \tag{5.9}$$

The importance of probability generating functions lies in the fact that the distribution for the sum $k = \sum_i k_i$ of independent stochastic variables $k_i$ is generated by the product of the generating functions $G_0^{(i)}(x)$ of the respective individual processes $p_{k_i}^{(i)}$, viz

$$G_0(x) = \sum_k p_k x^k = \prod_i G_0^{(i)}(x), \qquad G_0^{(i)}(x) = \sum_{k_i} p_{k_i}^{(i)} x^{k_i} \ ,$$

see Sect. **??** for further details and examples.

### *5.1.1 The Law of Large Numbers*

Throwing a dice many times and adding up the results obtained, the resulting average will be close to $3.5\, N$, where $N$ is the number of throws. This is the typical outcome for cumulative stochastic processes.[4]

> Law of Large Numbers. Repeating $N$ times a stochastic process with mean $\bar{x}$ and standard deviation $\sigma$, the mean and the standard deviation of the cumulative result will approach $\bar{x}\, N$ and $\sigma\sqrt{N}$ respectively in the thermodynamic limit $N \to \infty$.

The law of large numbers implies, that one obtains $\bar{x}$ as an averaged result, with a standard deviation $\sigma/\sqrt{N}$ for the averaged process. One needs to increase the number of trials by a factor of four in order to improve accuracy by a factor of 2.

**Proof** For a proof of the law of large numbers we consider a discrete process $p_k$ described by the generating functional $G_0(x)$. This is not really a

[4] Please take note of the difference between a cumulative stochastic process, when adding the results of individual trials, and the "cumulative PDF" $F(x)$ defined by $F(x) = \int_{-\infty}^{x} p(x')\, dx'$.

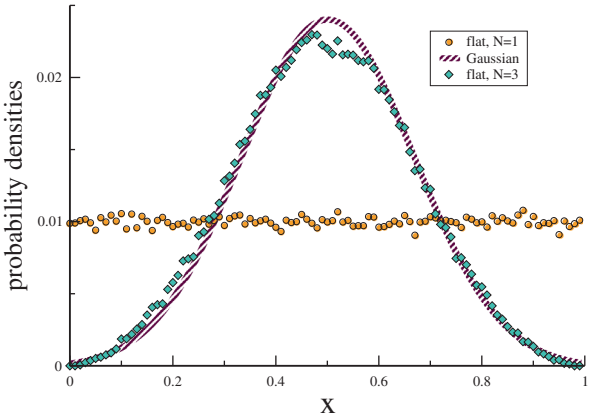


**Fig. 5.2** The flat distribution, which has a variance of $\sigma^2 = 1/12$ is shown together with the probability density of the sum of $N = 3$ flat distribution, which approximates already very well the limiting Gaussian with $\sigma = 1/6$, compare Eq. (5.11), in accordance with the central limiting theorem. 100 bins and a sampling of $10^5$ have been used

restriction, since PDFs of continuous variables can be discretized with arbitrary accuracy. The cumulative stochastic process is then characterized by a generating functional

$$G_0^N(x), \qquad \bar{k}^{(N)} = \frac{d}{dx} G_0^N(x)\Big|_{x=1} = N\, G_0^{N-1}(x)\, G_0'(x)\Big|_{x=1} = N\,\bar{k}$$

and the mean $\bar{k}^{(N)} = N\bar{k}$ respectively. For the standard deviation $\sigma^{(N)}$ of the cumulative process we use Eq. (5.9),

$$\begin{aligned}
\left(\sigma^{(N)}\right)^2 &= \frac{d}{dx}\left(x\,\frac{d}{dx} G_0^N(x)\right)\Big|_{x=1} - \left(N\bar{k}\right)^2 \\
&= \frac{d}{dx}\left(x\,N\,G_0^{N-1}(x)\,G_0'(x)\right)\Big|_{x=1} - N^2\,(G_0'(1))^2 \\
&= NG_0'(1) + N(N-1)\,(G_0'(1))^2 + NG_0''(1) - N^2\,(G_0'(1))^2 \\
&= N\left(G_0''(1) + G_0'(1) - (G_0'(1))^2\right) \equiv N\,\sigma^2\ , \qquad (5.10)
\end{aligned}$$

and obtain the law of large numbers.

**Central Limiting Theorem** The law of large numbers tells us, that the variance $\sigma^2$ is additive for cumulative processes, not the standard deviation $\sigma$. The "central limiting theorem" then tells us, that the limiting distribution function is a Gaussian.

Central Limiting Theorem. Given $i = 1, \ldots, N$ independent random variables $x_i$, distributed with mean $\mu_i$ and standard deviations $\sigma_i$. The cumulative distribution

$x = \sum_i x_i$ is then described, for $N \to \infty$, by a Gaussian with mean $\mu = \sum_i \mu_i$ and variance $\sigma^2 = \sum_i \sigma_i^2$.

In most cases one is not interested in the cumulative result, but in the averaged one, compare Fig. 5.2, which is obtained by rescaling of variables

$$y = x/N, \qquad \bar{\mu} = \mu/N, \qquad \bar{\sigma} = \sigma/N, \qquad p(y) = \frac{1}{\bar{\sigma}\sqrt{2\pi}}\, \mathrm{e}^{-\frac{(y-\bar{\mu})^2}{2\bar{\sigma}^2}} \ .$$

The rescaled standard deviation scales with $1/\sqrt{N}$. To see this, just consider identical processes with $\sigma_i \equiv \sigma_0$,

$$\bar{\sigma} = \frac{1}{N}\sqrt{\sum_i \sigma_i^2} = \frac{\sigma_0}{\sqrt{N}} \ , \tag{5.11}$$

in accordance with the law of large numbers.

**Is Everything Boring Then?** One might be tempted to draw the conclusion that systems containing a large number of variables are boring, since everything seems to average out. This is actually not the case, the law of large numbers holds only for statistically independent processes. Subsystems of distributed complex systems are however dynamically dependent and these dynamical correlations may lead to highly non-trivial properties in the thermodynamic limit.

### *5.1.2 Bayesian Statistics*

The notions of statistics considered so far can be easily generalized to the case of more than one random variable. Whenever a certain subset of the set of random variables is considered to be the causing event for the complementary subset of variables one speaks of inference, a domain of the Bayesian approach.

**Bayesian Theorem** Events and processes may have dependencies upon each other. A physician will typically have to know, to give an example, the probability that a patient has a certain illness, given that the patient shows a specific symptom.

Conditional Probability. The probability that an event $x$ occurs, given that an event $y$ has happened, is denoted "conditional probability" $p(x|y)$.

Throwing a dice twice, the probability that the first throw resulted in a 1, given that the total result was $4 = 1+3 = 2+2 = 3+1$, is $1/3$. Obviously,

$$p(x) = \int p(x|y)\, p(y)\, dy \tag{5.12}$$

holds. The probability of finding $x$ is given by the probability of finding $x$ given $y$, $p(x|y)$, integrated over the probability of finding $y$ in the first place.

The probability distribution of throwing $x$ in the first throw and $y$ in the second throw is determined, on the other hand, by the joint distribution $p(x, y)$.

Joint Probability Distribution. The probability of events $x$ and $y$ occurring is given by the "joint probability" $p(x, y)$.

Note, that $\int p(x, y)\, dx\, dy = 1$. The self-evident relations

$$p(x, y) = p(x|y)\, p(y) = p(y|x)\, p(x)$$

lead to

$$p(y|x) = \frac{p(x|y)\, p(y)}{p(x)} = \frac{p(x|y)\, p(y)}{\int p(x|y) p(y)\, dy} \ , \tag{5.13}$$

where we have used Eq. (5.12) in the second step. Equation (5.13) is denoted "Bayes' theorem". The conditional probability $p(x|y)$ of $x$ happing given that $y$ had occurred is denoted "likelihood".

**Bayesian Statistics** As an exemplary application of Bayes' theorem (5.13) consider a medical test.

- The probability of being ill/healthy is given by $p(y)$, $y =$ ill/health.
- The likelihood of passing the test is $p(x|y)$, with $x =$ positive/negative.

Let's consider an epidemic outbreak with, on the average, 1 % of the population being infected. We assume that the medical test has an accuracy of 99 %, $p(\text{positive}|\text{ill}) = 0.99$ with a low rate $p(\text{positive}|\text{healty}) = 0.02$ of false positives. The probability of a positively tested person of being infected is then actually just 33

$$\begin{aligned} p(\text{ill}|\text{pos}) &= \frac{p(\text{pos}|\text{ill}) p(\text{ill})}{p(\text{pos}|\text{ill}) p(\text{ill}) + p(\text{pos}|\text{healthy}) p(\text{healthy})} \\ &= \frac{0.99 \cdot 0.01}{0.99 * 0.01 + 0.02 * 0.99} = \frac{1}{3} \ , \end{aligned}$$

where we have used Bayes' theorem (5.13). A second follow-up test is hence necessary.

**Statistical Inference** We consider again a medical test, but in a slightly different situation. A series of test is performed in a city where an outbreak has occurred in order to estimate the percentage of people being infected.

We can then use expressing (5.12) for the marginal probability $p(\text{positive})$ for obtaining positive test results,

$$p(\text{positive}) = 0.99\, p(\text{ill}) \ + \ 0.02\, (1 - p(\text{ill})) \tag{5.14}$$

and solve for our estimate $p(\text{ill})$ of infections. In addition one needs to estimate the confidence of the obtained result, viz the expected fluctuations due to the limited number of tests actually carried out.

**Bayesian Inference** We start by noting that both sides of Bayes' theorem (5.13) are properly normalized,

$$\int p(y|x)\,dy = 1 = \frac{\int p(x|y)p(y)\,dy}{p(x)}\ .$$

For a given $x$ the probability that any $y$ happens is unity, and vice versa. For a given $x$ we may hence interprete the left-hand side as the probability that $y$ is true,

$$p_1(y) \equiv \frac{p(x|y)p_0(y)}{\int p(x|y)p_0(y)\,dy}\ . \tag{5.15}$$

One denotes

- $p_1(y) = p(y|x)$ the "posterior" distribution,
- $p(x|y)$ the likelihood and with
- $p_0(y)$ the "prior".

Equation (5.15) constitutes the basis of Bayesian inference. In this setting one is not interested in finding a self-consistent solution $p_0(y) = p_1(y) = p(y)$. Instead it is premised that one disposes of prior information, viz knowledge and expectations, about the status of the world, $p_0(y)$. Performing an experiment a new result $x$ is obtained which is then used to improve the expectations of the world status through $p_1(y)$, using Eq. (5.15).

**Bayesian Learning** The most common application of Bayesian inference is the situation when inference from a given set of experimental data needs to be drawn, using (5.15) a single time.

Alternatively one can consider Eq. (5.15) as the basis of cognitive learning processes, updating the knowledge about the world iteratively with any further observation $x_1, x_2, \ldots, x_n$,

$$p_i(y) \propto p(x_i|y)\, p_{i-1}(y), \qquad \forall y\ .$$

This update procedure of the knowledge $p_i(y)$ about the world is independent of the grouping of observations $x_i$, viz

$$p_0 \to p_1 \to \cdots \to p_n \qquad \text{and} \qquad p_0 \to p_n$$

yield the same result, due to the multiplicative nature of the likelihood $p(x|y)$, viz when considering in the last relation all consecutive observations $\{x_1, \ldots, x_n\}$ as a single event.

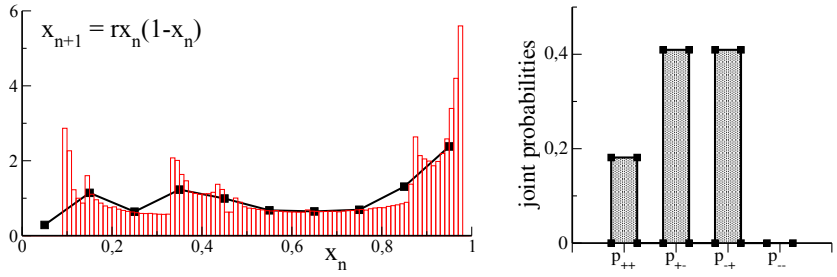


**Fig. 5.3** For the logistic map with $r = 3.9$ and $x_0 = 0.6$, two statistical analyses of the time series $x_n$, $n = 0, \ldots, N$, with $N = 10^6$. *Left*: The distribution $p(x)$ of the $x_n$. Plotted is $N_{bin}p(x)/N$, for $N_{bin} = 10/100$ bins (*curve* with *square symbols* and *open vertical bars* respectively). The data is plotted at the midpoints of the respective bins. *Right*: The joint probabilities $p_{\pm\pm}$, as defined by Eq. (5.18), of consecutive increases/decreases of the $x_n$. The probability $p_{--}$ that the data decreases consecutively twice vanishes

### *5.1.3 Time Series Characterization*

In many cases one is interested in estimating the probability distribution functions for data generated by some known or unknown process, like the temperature measurements of a weather station. It is important, when doing so, to keep a few caveats in mind.

**Binning of Variables** Here we will be dealing mainly with the time series of data generated by dynamical systems. As an example we consider the logistic map, compare Sect. **??**,

$$x_{n+1} = f(x_n) \equiv r\, x_n\, (1 - x_n), \qquad x_n \in [0, 1], \qquad r \in [0, 4]\ . \quad (5.16)$$

The dynamical variable is continuous and in order to estimate the probability distribution of the $x_n$ we need to bin the data. In Fig. 5.3 the statistics of a time series in the chaotic regime, for $r = 3.9$, is given.

One needs to select the number of bins $N_{bin}$ and, in general, also the positions and the widths of the bins. When the data is not uniformly distributed one may place more bins in the region of interest, generalizing the relation (5.1) through $\Delta x \to \Delta x_i$, with the $\Delta x_i$ being the width of the individual bins.

For our illustrative example see Fig. 5.3, we have selected $N_{bin} = 10/100$ equidistant bins. The data is distributed over more bins, when $N_{bin}$ increases. In order to make the distribution functions for different number of bins comparable one needs to rescale them with $N_{bin}$, as it has been done for the data shown in Fig. 5.3.

The selection of the binning procedure is in general a difficult choice. Fine structure will be lost when $N_{bin}$ is too low, but statistical noise will dominate for a too large number of bins.

**Symbolization** One denotes by "symbolization" the construction of a finite number of symbols suitable for the statistical characterization of a given time series.[5] The binning procedure discussed above is a commonly used symbolization procedure.

For a further example of a symbolization procedure we denote with $\delta_t = \pm 1$,

$$\delta_t = \text{sign}(x_t - x_{t-1}) = \begin{cases} 1 & x_t > x_{t-1} \\ -1 & x_t < x_{t-1} \end{cases} \tag{5.17}$$

the direction of the time development. The consecutive development of the $\delta_t$ may then be encoded in higher-level symbolic stochastic variables. For example one might be interested in the joint probabilities

$$\begin{array}{ll} p_{++} = \langle p(\delta_t = 1, \delta_{t-1} = 1)\rangle_t & p_{+-} = \langle p(\delta_t = 1, \delta_{t-1} = -1)\rangle_t \\ p_{-+} = \langle p(\delta_t = -1, \delta_{t-1} = 1)\rangle_t & p_{--} = \langle p(\delta_t = -1, \delta_{t-1} = -1)\rangle_t \end{array} , \tag{5.18}$$

where $p_{++}$ gives the probability that the data increases at least twice consecutively, etc., and where $\langle \dots \rangle_t$ denotes the time average. In Fig. 5.3 the values for the joint probabilities $p_{\pm\pm}$ are given for a selected time series of the logistic map in the chaotic regime. The data never decreases twice consecutively, $p_{--} = 0$, a somewhat unexpected result.

There are many possible symbolization procedures and the procedure used to analyze a given time series determines the kind of information one may hope to extract, as evident from the results illustrated in Fig. 5.3. The selection of the symbolization procedures needs to be given attention, and will be discussed further in Sect. 5.2.1.

**Self Averaging** A time series produced by a dynamical system depends on the initial condition and so will generally also the statistical properties of the time series. As an example we consider the XOR series[6]

$$\sigma_{t+1} = \text{XOR}(\sigma_t, \sigma_{t-1}), \qquad \sigma_t = 0,\, 1\ . \tag{5.19}$$

The four initial conditions 00, 01, 10 and 11 give rise to the respective time series

$$\begin{array}{ll} \dots 000000\underline{00} & \dots 101101\underline{01} \\ \dots 110110\underline{10} & \dots 011011\underline{11} \end{array} \tag{5.20}$$

where time runs from right to left and where we have underlined the initial conditions $\sigma_1$ and $\sigma_0$. The typical time series, occurring for 75 % of the initial conditions, is $\dots 011011011011\dots$, with $p(0) = 1/3$ and $p(1) = 2/3$ for the probability to find a 0/1. When averaging over all four initial conditions, we have on the other hand $(2/3)(3/4) = 1/2$ for the probability to find a 1. Then

[5] For continuous-time data, as for an electrocardiogram, an additional symbolization step is necessary, the discretization of time. Here we consider however only discrete-time series.

[6] Remember, that $\text{XOR}(0,0) = 0 = \text{XOR}(1,1)$ and $\text{XOR}(0,1) = 1 = \text{XOR}(1,0)$.

$$p(1) = \begin{cases} 2/3 & \text{typical} \\ 1/2 & \text{average} \end{cases} .$$

When observing a single time series we are likely to obtain the typical probability, analyzing many time series will result on the other hand in the average probability.

> Self Averaging. When the statistical properties of a time series generated by a dynamical process are independent of the respective initial conditions, one says the time series is "self averaging".

The XOR series is not self averaging and one can generally not assume self averaging to occur. An inconvenient situation whenever only a single time series is available, as it is the case for most historical data, e.g. of past climatic conditions.

**XOR Series with Noise** Most real-world processes involve a certain degree of noise and one may be tempted to assume, that noise could effectively restart the dynamics, leading to an implicitly averaging over initial conditions. This assumption is not generally valid but works out for XOR process with noise,

$$\sigma_{t+1} = \begin{cases} \text{XOR}(\sigma_t, \sigma_{t-1}) & \text{probability} 1-\xi \\ \neg\,\text{XOR}(\sigma_t, \sigma_{t-1}) & \text{probability} \xi \end{cases} \qquad 0 \leq \xi \ll 1 \; . \tag{5.21}$$

For low level of noise, $\xi \to 0$, the time series

$$\ldots 000000\underline{0}01101101\underline{0}10110110110\underline{1}11011011\underline{0}1\underline{1}00000000\ldots$$

has stretches of regular behavior interseeded by four types of noise induced dynamics (underlined, time running from right to left). Denoting with $p_{000}$ and $p_{011}$ the probability of finding regular dynamics of type $\ldots 000000000\ldots$ and $\ldots 011011$
$011\ldots$ respectively, we find the master equation

$$\dot{p}_{011} = \xi p_{000} \; - \; \xi p_{011}/3 = -\dot{p}_{000} \tag{5.22}$$

for the noise-induced transition probabilities. In the stationary case $p_{000} = p_{011}/3$ for the XOR process with noise, the same ratio one would obtain for the deterministic XOR series averaged over the initial conditions.

The introduction of noise generally introduces a complex dynamics akin to the master Eq. (5.22) and it is generally to be expected that the resulting time series is self-averaging. This is also the case for the OR time series, for which the small noise limit does however not coincide with the time series obtained in the absence of noise.

**Time Series Analysis and Cognition** Time series analysis is a tricky business whenever the fundamentals of the generative process are unknown, e.g. whether noise is important or not. This is however the setting in which

cognitive systems, see Chap. **??**, are operative. Our sensory organs, eyes and ears, provide us with a continuous time series encoding environmental information. Performing an informative and fast time series analysis is paramount for surviving.

> Online vs. offline analysis. If one performs an analysis of a previously recorded time series one speaks of "offline" analysis. An analysis performed on-the-fly during recording is denoted "online".

Animals need to perform online analysis of their sensory data input streams, otherwise they would not survive long enough to react.

**Trailing Averages** Online characterization of a time series in terms of its basic statistical properties, like mean and standard deviation, is quite straightforward.

We consider a continuous time input stream $x(t)$ and define with

$$\mu_t = \frac{1}{T}\int_0^\infty d\tau\, x(t-\tau)\, e^{-\tau/T} \tag{5.23}$$

$$\sigma_t^2 = \frac{1}{T}\int_0^\infty d\tau\, \big(x(t-\tau)-\mu_t\big)^2\, e^{-\tau/T} \tag{5.24}$$

the "trailing average" $\mu_t$ and the trailing variance $\sigma_t^2$. The trailing average exponentially discounts older data, the first two moments of the PDF of the input stream $x(t)$ are recovered in the limit $T \to \infty$. The factor $1/T$ in (5.23) and (5.24) normalizes the respective trailing averages. For the case of a constant, time independent input $x(t) \equiv \bar{x}$ we obtain correctly

$$\mu_t \to \frac{1}{T}\int_0^\infty d\tau\, \bar{x}\, e^{-\tau/T} = \bar{x}\ .$$

The trailing average can be evaluated by a simple online update rule, there is no need to store all past data $x(t-\tau)$. To see this we calculate the time dependence

$$\dot{\mu}_t = \frac{1}{T}\int_0^\infty d\tau\, e^{-\tau/T}\frac{d}{dt}x(t-\tau) = \frac{-1}{T}\int_0^\infty d\tau\, e^{-\tau/T}\frac{d}{d\tau}x(t-\tau)\ .$$

The last expression can be evaluated by a simple partial integration. One obtains

$$\dot{\mu}_t = \frac{x(t)-\mu_t}{T}\ , \tag{5.25}$$

and an analogous update rule for the variance $\sigma_t^2$ by substituting $x \to (x-\mu)^2$. Expression (5.25) is an archetypical example of an online updating rule for a time averaged quantity, here the trailing average $\mu_t$.

## 5.2 Entropy and Information

Entropy is a venerable concept from physics encoding the amount of disorder present in a thermodynamic system at a given temperature. The "Second Law of Thermodynamics" states, that entropy can only increase in an isolated (closed) system. The second law has far reaching consequences, e.g. determining the maximal efficiency of engines and power plants, and philosophical implications for our understanding of the fundamentals underpinning the nature of life as such.

**Entropy and Life** Living organisms have a body and such create ordered structures from basic chemical constituents. Living beings therefore decrease entropy locally, in their bodies, seemingly in violation of the second law. In reality, the local entropy depressions are created on the expense of corresponding entropy increases in the environment, in agreement with the second law of thermodynamics. All living beings need to be capable of manipulating entropy.

**Information Entropy and Predictability** Entropy is also a central concept in information theory, where it is commonly denoted "Shannon entropy" or "information entropy". In this context one is interested in the amount of information encoded by a sequence of symbols

$$\dots \sigma_{t+2}, \sigma_{t+1}, \sigma_t, \sigma_{t-1}, \sigma_{t-2}, \dots \ ,$$

e.g. when transmitting a message. Typically, in everyday computers, the $\sigma_t$ are words of bits. Let us consider two time series of bits, e.g.

$$\dots 101010101010 \dots , \qquad \dots 1100010101100 \dots \ . \tag{5.26}$$

The first example is predictable, from the perspective of a time-series, and ordered, from the perspective of an one-dimensional alignment of bits. The second example is unpredictable and disordered respectively.

Information can be transmitted through a time series of symbols only when this time series is not predictable. Talking to a friend, to illustrate this statement, we will not learn anything new when capable of predicting his next joke. We have therefore the following two perspectives,

$$\text{high entropy} \mathrel{\hat{=}} \begin{cases} \text{large disorder} & \text{physics} \\ \text{high information content} & \text{information theory} \end{cases} ,$$

and vice versa. Only seemingly disordered sequences of symbols are unpredictable and thus potential carriers of information. Note, that the predictability of a given time series, or its degree of disorder, may not necessarily be as self evident as in above example, Eq. (5.26), depending generally on the analysis procedure used, see Sect. 5.2.1.

**Extensive Information** In complex system theory, as well as in physics, we are often interested in properties of systems composed of many subsystems.

Extensive and Intensive Properties. For systems composed of $N$ subsystems a property is denoted "extensive" if it scales as $O(N^1)$ and "intensive" when it scales with $O(N^0)$.

A typical extensive property is the mass, a typical intensive property the density. When lumping together two chunks of clay, their mass adds, but the density does not change.

One demands, both in physics and in information theory, that the entropy should be an extensive quantity. The information content of two independent transmission channels should be just the sum of the information carried by the two individual channels.

**Shannon Entropy** The Shannon entropy $H[p]$ is defined by

$$H[p] = -\sum_{x_i} p(x_i)\, \log_b(p(x_i)) = -\langle\, \log_b(p)\,\rangle, \qquad H[p] \geq 0\ , \tag{5.27}$$

where $p(x_i)$ is a normalized discrete probability distribution function and where the brackets in $H[p]$ denote the functional dependence.[7] Note, that $-p\log(p) \geq 0$ for $0 \leq p \leq 1$, see Fig. 5.4, the entropy is therefore strictly positive.

$b$ is the base of the logarithm used in Eq. (5.27). Common values of $b$ are 2, Euler's number $e$ and 10. The corresponding units of entropy are then termed "bit" for $b = 2$, "nat" for $b = e$ and "digit" for $b = 10$. In physics the natural logarithm is always used and there is an additional constant (the Boltzmann constant $k_B$) in front of the definition of the entropy. Here we will use $b = 2$ and drop in the following the index $b$.

**Extensiveness of the Shannon Entropy** The log-dependence in the definition of the information entropy in Eq. (5.27) is necessary for obtaining an extensive quantity. To see this, let us consider a system composed of two independent subsystems. The joint probability distribution is multiplicative,

$$p(x_i, y_j) = p_X(x_i)p_Y(y_j), \quad \log(p(x_i, y_j)) = \log(p_X(x_i)) + \log(p_Y(y_j))\ .$$

The logarithm is the only function which maps a multiplicative input onto an additive output. Consequently,

$$H[p] = -\sum_{x_i, y_j} p(x_i, y_j) \log(p(x_i, y_j))$$

[7] A function $f(x)$ is a function of a variable $x$; a functional $F[f]$ is, on the other hand, functionally dependent on a function $f(x)$. In formal texts on information theory the notation $H(X)$ is often used for the Shannon entropy and a random variable $X$ with probability distribution $p_X(x)$.

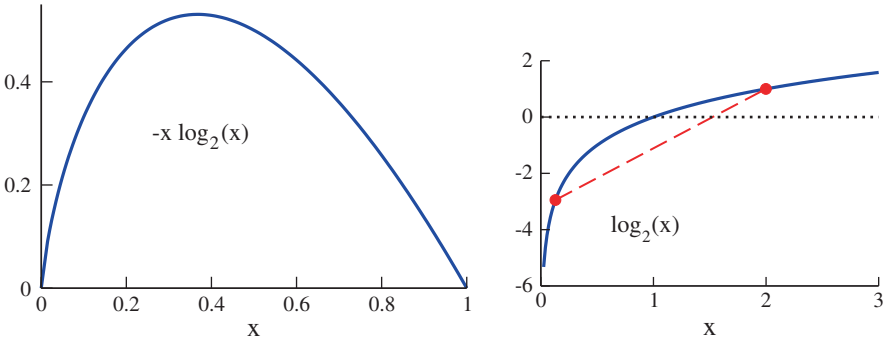


**Fig. 5.4** *Left*: Plot of $-x \log_2(x)$. *Right*: The logarithm $\log_2(x)$ (*full line*) is concave, every cord (*dashed line*) lies below the graph

$$
\begin{aligned}
&= -\sum_{x_i,y_j} p_X(x_i) p_Y(y_j) \Big[ \log(p_X(x_i)) + \log(p_Y(y_j)) \Big] \\
&= -\sum_{x_i} p_X(x_i) \sum_{y_j} p_Y(y_j) \log(p_Y(y_j)) \\
&\quad -\sum_{y_j} p_Y(y_j) \sum_{x_i} p_X(x_i) \log(p_X(x_i)) \\
&= H[p_Y] + H[p_X] \ ,
\end{aligned}
$$

as necessary for the extensiveness of $H[p]$. Hence the log-dependence in Eq. (5.27).

**Degrees of Freedom** We consider a discrete system with $x_i \in [1, \ldots, n]$, having $n$ "degrees of freedom" in physics' slang. If the probability of finding any value is equally likely, as it is the case for a thermodynamic system at infinite temperatures, the entropy is

$$
H = -\sum_{x_i} p(x_i) \log(p(x_i)) = -n \frac{1}{n} \log(1/n) = \log(n) \ , \tag{5.28}
$$

a celebrated result. The entropy grows logarithmically with the number of degrees of freedom.

**Shannon's Source Coding Theorem** So far we have shown, that Eq. (5.27) is the only possible definition, modulo renormalizing factors, for an extensive quantity depending exclusively on the probability distribution. The operative significance of the entropy $H[p]$ in terms of informational content is given by Shannon's theorem.

Source Coding Theorem. Given a random variable $x$ with a PDF $p(x)$ and entropy $H[p]$. The cumulative entropy $NH[p]$ is then, for $N \to \infty$, a lower bound for the number of bits necessary when trying to compress $N$ independent processes drawn from $p(x)$.

If we compress more, we will lose information, the entropy $H[p]$ is therefore a measure of information content.

**Entropy and Compression** Let's make an example. Consider we have words made out of the four letter alphabet $A$, $B$, $C$ and $D$. Suppose, that these four letters would not occur with the same probability, the relative frequencies being

$$p(A) = \frac{1}{2}, \qquad p(B) = \frac{1}{4}, \qquad p(C) = \frac{1}{8} = p(D) \ .$$

When transmitting a long series of words using this alphabet we will have the entropy

$$\begin{aligned} H[p] &= \frac{-1}{2}\log(1/2) \ - \ \frac{1}{4}\log(1/4) \ - \ \frac{1}{8}\log(1/8) \ - \ \frac{1}{8}\log(1/8) \\ &= \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8} = 1.75 \ , \end{aligned} \tag{5.29}$$

since we are using the logarithm with base $b = 2$. The most naive bit encoding,

$$A \to 00, \qquad B \to 01, \qquad C \to 10, \qquad D \to 11 \ ,$$

would use exactly 2 bit, which is larger than the Shannon entropy. An optimal encoding would be, on the other hand,

$$A \to 1, \qquad B \to 01, \qquad C \to 001, \qquad D \to 000 \ , \tag{5.30}$$

leading to an average length of words transmitted of

$$p(A) \ + \ 2p(B) \ + \ 3p(C) \ + \ 3p(D) = \frac{1}{2} \ + \ \frac{2}{4} \ + \ \frac{3}{8} \ + \ \frac{3}{8} = 1.75 \ , \tag{5.31}$$

which is the same as the information entropy $H[p]$. The encoding given in Eq. (5.30) is actually "prefix-free". When we read the words from left to right, we know where a new word starts and stops,

$$110000010101 \qquad \longleftrightarrow \qquad AADCBB \ ,$$

without ambiguity. Fast algorithms for optimal, or close to optimal encoding are clearly of importance in the computer sciences and for the compression of audio and video data.

**Discrete vs. Continuous Variables** When defining the entropy we have considered hitherto discrete variables. The information entropy can also be defined for continuous variables. We should be careful though, being aware that the transition from continuous to discrete stochastic variables, and vice versa, is slightly non-trivial, compare Eq. (5.1):

$$
\begin{aligned}
H[p]\Big|_{\text{con}} &= -\int p(x)\log(p(x))\,dx \approx -\sum_i p(x_i)\log(p(x_i))\,\Delta x \\
&= -\sum_i p_i \log(p_i/\Delta x) = -\sum_i p_i \log(p_i) \;+\; \sum_i p_i \log(\Delta x) \\
&= H[p]\Big|_{\text{dis}} \;+\; \log(\Delta x)\;, \qquad (5.32)
\end{aligned}
$$

where $p_i = p(x_i)\Delta x$ is here the properly normalized discretized PDF, compare Eq. (5.1). The difference $\log(\Delta x)$ between the continuous-variable entropy $H[p]\big|_{\text{con}}$ and the discretized version $H[p]\big|_{\text{dis}}$ diverges for $\Delta x \to 0$, the transition is discontinuous.

**Entropy of a Continuous PDF** From Eq. (5.32) it follows, that the Shannon entropy $H[p]\big|_{\text{con}}$ can be negative for a continuous probability distribution function. As an example consider the flat distribution

$$
p(x) = \begin{cases} 1/\epsilon & \text{for} x \in [0,\epsilon] \\ 0 & \text{otherwise} \end{cases}\;, \qquad \int_0^\epsilon p(x)\,dx \;=\; 1
$$

in the small interval $[0,\epsilon]$, with the entropy

$$
H[p]\Big|_{\text{con}} = -\int_0^\epsilon \frac{1}{\epsilon}\log(1/\epsilon)\,dx = \log(\epsilon) < 0, \qquad \text{for} \quad \epsilon \;<\; 1\;.
$$

The absolute value of the entropy is hence not meaningful for continuous PDFs, only entropy differences. $H[p]\big|_{\text{con}}$ is therefore also referred-to as "differential entropy".

**Maximal Entropy Distributions** Which kind of distributions maximize entropy, viz information content? Remembering that

$$
\lim_{p\to 0,1} p\log(p) = 0, \qquad \log(1) = 0\;,
$$

see Fig. 5.4, it is intuitive that a flat distribution might be optimal. This is indeed correct in the absence of any constraint other than the normalization condition $\int p(x)dx = 1$.

**Variational Calculus** We consider generically the task to maximize the functional

$$
H[p] = \int f(p(x))\,dx, \qquad f(p) = -p\log(p)\;, \qquad (5.33)
$$

where the notation used will turn out useful later on. Maximizing a functional like $H[p]$ is a typical task of variational calculus. One considers with

$$
p(x) = p_{opt}(x) \;+\; \delta p(x), \qquad \delta p(x)\text{arbitrary}
$$

a general variation of $\delta p(x)$ around the optimal function $p_{opt}(x)$. At optimality, the dependence of $H[p]$ on the variation $\delta p$ should be stationary,

$$0 \equiv \delta H[p] = \int f'(p)\, \delta p\, dx, \qquad 0 = f'(p)\ , \tag{5.34}$$

where $f'(p) = 0$ follows from the fact that $\delta p$ is an arbitrary function.

For the entropy functional $f(p) = -p \log(p)$ we find then with

$$f'(p) = -\log(p) - 1 = 0, \qquad p(x) = \text{const.} \tag{5.35}$$

the expected flat distribution.

**Maximal Entropy Distributions with Constraints** We consider now the entropy maximization under the constraint of a fixed average $\mu$,

$$\mu = \int x\, p(x)\, dx\ . \tag{5.36}$$

This condition can be enforced by a Lagrange parameter $\lambda$ via

$$f(p) = -p \log(p)\ -\ \lambda x p\ .$$

The stationary condition $f'(p) = 0$ then leads to

$$f'(p) = -\log(p) - 1 - \lambda x = 0, \qquad p(x) \propto 2^{-\lambda x} \sim \mathrm{e}^{-x/\mu} \tag{5.37}$$

the exponential distribution, see Eq. (5.4), with mean $\mu$. The Lagrange parameter $\lambda$ needs to be determined such that the condition of fixed mean, Eq. (5.36), is satisfied. For a support $x \in [0, \infty]$, as assumed above, we have $\lambda \log_e(2) = 1/\mu$.

One can generalize this procedure and consider distribution maximizing the entropy under the constraint of a given mean $\mu$ and variance $\sigma^2$,

$$\mu = \int x\, p(x)\, dx, \qquad \sigma^2 = \int (x-\mu)^2\, p(x)\, dx\ . \tag{5.38}$$

Generalizing the derivation leading to (5.37) one sees that the maximal entropy distribution constrained by (5.38) is a Gaussian, as given by expression (5.5).

**Pairwise Constraints** We consider a joint distribution function $p(x_1, \ldots, x_n)$ for $n$ variables $x_i$ with pairwise correlations

$$\langle x_i x_j \rangle = \int dx^n x_i x_j p(x_1, \ldots, x_n)\ . \tag{5.39}$$

Pair correlations can be measured in many instances experimentally and can be hence considered as constraints for modelling. One can adjust in the max-

imal entropy distribution

$$p(x_1, \ldots, x_n) = \frac{\mathrm{e}^{-H}}{N}, \qquad H = \sum_{ij} J_{ij} x_i x_j + \sum_i \lambda_i x_i \tag{5.40}$$

the $n(n-1)/2$ variational parameters $J_{ij}$, in order to reproduce given $n(n-1)/2$ pairwise correlations (5.39), and the Lagrange multiplier $\lambda_i$ for regulating the respective individual averages $\langle x_i \rangle$.

The maximal entropy distribution (5.40) has the form of a Boltzman factor of statistical mechanics with $H$ representing the Hamiltonian, the energy function, and with the coupling constants $J_{ij}$ encoding the strength of pairwise interactions.

### 5.2.1 Information Content of a Real-World Time Series

The Shannon entropy is a very powerful concept in information theory. The encoding rules are typically known in information theory, there is no ambiguity regarding the symbolization procedure (see Sect. 5.1.3) to employ when receiving a message via some technical communication channel. This is however not any more the case, when we are interested in determining the information content of real-world processes, e.g. the time series of certain financial data or the data stream produced by our sensory organs.

**Symbolization and Information Content** The result obtained for the information content of a real-world time series $\{\sigma_t\}$ depends in general on the symbolization procedure used. Let us consider, as an example, the first time series of Eq. (5.26),

$$\ldots 101010101010 \ldots \ . \tag{5.41}$$

When using a 1-bit symbolization procedure, we have

$$p(0) = \frac{1}{2} = p(1), \qquad H[p] = -2\,\frac{1}{2}\log(1/2) = 1 \ ,$$

as expected. If, on the other hand, we use a 2-bit symbolization, we find

$$p(00) = p(11) = p(01) = 0, \qquad p(10) = 1, \qquad H[p] = -\log(1) = 0 \ .$$

When 2-bit encoding is presumed, the time series is predictable and carries no information. This seems intuitively the correct result and the question is: Can we formulate a general guiding principle which tells us which symbolization procedure would yield the more accurate result for the information content of a given time series?

**The Minimal Entropy Principle** The Shannon entropy constitutes a lower bound for the number of bits, per symbol, necessary when compressing the data without loss of information. Trying various symbolization procedures, the symbolization procedure yielding the lowest information entropy then allows us to represent, without loss of information, a given time series with the least number of bits.

> Minimal Entropy Principle. The information content of a time series with unknown encoding is given by the minimum (actually the infimum) of the Shannon entropy over all possible symbolization procedures.

The minimal entropy principle then gives us a definite answer with respect to the information content of the time series given in Eq. (5.41). We have seen that at least one symbolization procedure yields a vanishing entropy and one cannot get a lower value, since $H[p] \geq 0$. This is the expected result, since $\ldots 01010101 \ldots$ is predictable.

**Information Content of a Predictable Time Series** Note, that a vanishing information content $H[p] = 0$ only implies that the time series is strictly predictable, not that it is constant. One therefore needs only a finite amount of information to encode the full time series, viz for arbitrary lengths $N \to \infty$. When the time series is predictable, the information necessary to encode the series is intensive and not extensive.

**Symbolization and Time Horizons** The minimal entropy principle is rather abstract. In practice one may not be able than to try out more than a handful of different symbolization procedures. It is therefore important to gain an understanding of the time series at hand.

An important aspect of many time series is the intrinsic time horizon $\tau$. Most dynamical processes have certain characteristic time scales and memories of past states are effectively lost for times exceeding these intrinsic time scales. The symbolization procedure used should therefore match the time horizon $\tau$.

This is what happened when analyzing the time series given in Eq. (5.41), for which $\tau = 2$. A 1-bit symbolization procedure implicitly presumes that $\sigma_t$ and $\sigma_{t+1}$ are statistically independent and such missed the intrinsic time scale $\tau = 2$, in contrast to the 2-bit symbolization procedure.

### *5.2.2 Mutual Information*

We have been considering so far the statistical properties of individual stochastic processes as well as the properties of cumulative processes generated by the sum of stochastically independent random variables. In order to understand complex systems we need to develop tools for the description of a large number of interdependent processes. As a first step towards this

direction we consider in the following the case of two stochastic processes, which may now be statistically correlated.

**Two Channels – Markov Process** We start by considering an illustrative example of two correlated channels $\sigma_t$ and $\tau_t$, with

$$\sigma_{t+1} = XOR(\sigma_t, \tau_t), \qquad \tau_{t+1} = \begin{cases} XOR(\sigma_t, \tau_t) & \text{probability} 1-\xi \\ \neg XOR(\sigma_t, \tau_t) & \text{probability} \xi \end{cases} . \tag{5.42}$$

This dynamics has the "Markov property", the value for the state $\{\sigma_{t+1}, \tau_{t+1}\}$ depends only on the state at the previous time step, viz on $\{\sigma_t, \tau_t\}$.

> Markov Process. A discrete-time memory-less dynamical process is denoted a "Markov process". The likelihood of future states depends only on the present state, and not on any past states.

When the state space is finite, as in our example, the term "Markov chain" is also used. We will not adhere here to the distinction which is sometimes made between discrete and continuous time, with Markov processes being formulated for discrete time and "master equations" describing stochastic processes for continuous time.

**Joint Probabilities** A typical time series of the Markov chain specified in Eq. (5.42) looks like

$$\begin{array}{ll} \ldots \sigma_{t+1}\sigma_t \ldots & : 0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0 \ldots \\ \ldots \tau_{t+1}\tau_t \ldots & : 0\,0\,0\,1\,\underline{1}\,0\,0\,0\,0\,0\,0\,1\,\underline{1}\,1\,1\,1 \ldots \end{array} ,$$

where we have underlined instances of noise-induced transitions. For $\xi = 0$ the stationary state is $\{\sigma_t, \tau_t\} = \{0, 0\}$ and therefore fully correlated. We now calculate the joint probabilities $p(\sigma, \tau)$ for general values of noise $\xi$, using the transition probabilities

$$\begin{array}{ll} p_{t+1}(0,0) = (1-\xi)\,[p_t(1,1) + p_t(0,0)] & p_{t+1}(1,0) = \xi\,[p_t(0,1) + p_t(1,0)] \\ p_{t+1}(1,1) = (1-\xi)\,[p_t(1,0) + p_t(0,1)] , & p_{t+1}(0,1) = \xi\,[p_t(0,0) + p_t(1,1)] , \end{array}$$

for the ensemble averaged joint probability distributions $p_t(\sigma, \tau) = \langle p(\sigma_t, \tau_t) \rangle_{ens}$, where the average $\langle .. \rangle_{ens}$ denotes the average over an ensemble of time series. For the solution in the stationary case $p_{t+1}(\sigma, \tau) = p_t(\sigma, \tau) \equiv p(\sigma, \tau)$ we use the normalization

$$p(1,1) + p(0,0) + p(1,0) + p(0,1) = 1 .$$

We find

$$p(1,1) + p(0,0) = 1 - \xi, \qquad p(1,0) + p(0,1) = \xi ,$$

by adding the terms $\propto (1-\xi)$ and $\propto \xi$ respectively. It then follows immediately

$$\begin{array}{ll} p(0,0) = (1-\xi)^2 & p(1,0) = \xi^2 \\ p(1,1) = (1-\xi)\xi\,, & p(0,1) = \xi(1-\xi) \end{array}\,. \tag{5.43}$$

For $\xi = 1/2$ the two channels become 100 % uncorrelated, as the $\tau$-channel is then fully random. The dynamics of the Markov process given in Eq. (5.42) is self averaging and it is illustrative to verify the result for the joint distribution function, Eq. (5.43), by a straightforward numerical simulation.

**Entropies** Using the notation

$$p_\sigma(\sigma') = \sum_{\tau'} p(\sigma', \tau'), \qquad p_\tau(\tau') = \sum_{\sigma'} p(\sigma', \tau')$$

for the "marginal distributions" $p_\sigma$ and $p_\tau$, we find from Eq. (5.43)

$$\begin{array}{ll} p_\sigma(0) = 1-\xi & p_\tau(0) = 1-2\xi(1-\xi) \\ p_\sigma(1) = \xi\,, & p_\tau(1) = 2\xi(1-\xi) \end{array} \tag{5.44}$$

for the distributions of the two individual channels. We may now evaluate both the entropies of the individual channels, $H[p_\sigma]$ and $H[p_\tau]$, the "marginal entropies", viz

$$H[p_\sigma] = -\langle \log(p_\sigma) \rangle, \qquad H[p_\tau] = -\langle \log(p_\tau) \rangle\,, \tag{5.45}$$

as well as the entropy of the combined process, termed "joint entropy",

$$H[p] = -\sum_{\sigma',\tau'} p(\sigma', \tau') \log(p(\sigma', \tau'))\,. \tag{5.46}$$

In Fig. 5.5 the respective entropies are plotted as a function of noise strength $\xi$. Some observations:

- In the absence of noise, $\xi = 0$, both the individual channels as well as the combined process are predictable and all three entropies, $H[p]$, $H[p_\sigma]$ and $H[p_\tau]$, vanish consequently.
- For maximal noise $\xi = 0.5$, the information content of both individual chains is 1 bit and of the combined process 2 bits, implying statistical independence.
- For general noise strengths $0 < \xi < 0.5$, the two channels are statistically correlated. The information content of the combined process $H[p]$ is consequently smaller than the sum of the information contents of the individual channels, $H[p_\sigma] + H[p_\tau]$.

**Mutual Information** The degree of statistical dependency of two channels can be measured by comparing the joint entropy with the respective marginal entropies.

Mutual Information. For two stochastic processes $\sigma_t$ and $\tau_t$ the difference

$$I(\sigma, \tau) = H[p_\sigma] + H[p_\tau] - H[p] \tag{5.47}$$

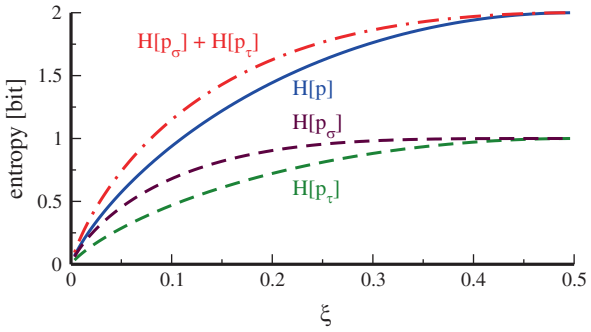


**Fig. 5.5** For the two-channel XOR-Markov chain $\{\sigma_t, \tau_t\}$ with noise $\xi$, see Eq. (5.42), the entropy $H[p]$ of the combined process (*full line*, Eq. (5.46)), of the individual channels (*dashed lines*, Eq. (5.45)), $H[p_\sigma]$ and $H[p_\tau]$, and of the sum of the joint entropies (*dot-dashed line*). Note the positiveness of the mutual information, $I(\sigma, \tau) = H[p_\sigma] + H[p_\tau] - H[p] > 0$

between the sum of the marginal entropies $H[p_\sigma] + H[p_\tau]$ and the joint entropy $H[p]$ is denoted "mutual information" $I(\sigma, \tau)$.

When two dynamical processes become correlated, information is lost and this information loss is given by the mutual information. Note, that $I(\sigma, \tau) = I[p]$ is a functional of the joint probability distribution $p$ only, the marginal distribution functions $p_\sigma$ and $p_\tau$ being themselves functionals of $p$.

**Positiveness** We will now discuss some properties of the mutual information, considering the general case of two stochastic processes described by the joint PDF $p(x, y)$ and the respective marginal PDFs $p_X(x) = \int p(x, y)\, dy$, $p_Y(y) = \int p(x, y)\, dx$.

The mutual information

$$I(X, Y) = \langle \log(p) \rangle \; - \; \langle \log(p_X) \rangle \; - \; \langle \log(p_Y) \rangle \qquad I(X, Y) \geq 0 \; , \tag{5.48}$$

is strictly positive. Rewriting the mutual information as

$$\begin{aligned} I(X, Y) &= \int p(x, y) \Big[ \log(p(x, y)) - \log(p_X(x)) - \log(p_Y(y)) \Big] \, dx \, dy \qquad (5.49) \\ &= \int p(x, y) \log \left( \frac{p(x, y)}{p_X(x) p_Y(y)} \right) dx \, dy = - \int p \log \left( \frac{p_X p_Y}{p} \right) dx \, dy \; , \end{aligned}$$

we can easily show that $I(X, Y) \geq 0$ follows from the concaveness of the logarithm, see Fig. 5.4,

$$\log(p_1 x_1 + p_2 x_2) \geq p_1 \log(x_1) + p_2 \log(x_2), \qquad \forall x_1, x_2 \in [0, \infty] \; , \tag{5.50}$$

and $p_1, p_2 \in [0, 1]$, with $p_1 + p_2 = 1$; any cord of a concave function lies below the graph. We can regard $p_1$ and $p_2$ as the coefficients of a distribution

function and generalize,

$$p_1\delta(x-x_1)+p_2\delta(x-x_2) \quad \longrightarrow \quad p(x)\ ,$$

where $p(x)$ is now a generic, properly normalized PDF. The concaveness condition, Eq. (5.50), then reads

$$\log\left(\int p(x)\,x\,dx\right) \geq \int p(x)\log(x)\,dx\ , \qquad \varphi\left(\langle x\rangle\right) \geq \langle\,\varphi(x)\,\rangle\ , \tag{5.51}$$

the "Jensen inequality", which holds for any concave function $\varphi(x)$. This inequality remains valid when substituting $x \to p_X p_Y/p$ for the argument of the logarithm.[8] We then obtain for the mutual information, Eq. (5.49),

$$\begin{aligned} I(X,Y) &= -\int p\log\left(\frac{p_X p_Y}{p}\right)dx\,dy \geq -\log\left(\int p\,\frac{p_X p_Y}{p}\,dx\,dy\right) \\ &= -\log\left(\int p_X(x)\,dx\int p_Y(y)\,dy\right) = -\log(1) = 0\ , \end{aligned}$$

viz $I(X,Y)$ is non-negative. Information can only be lost when correlating two previously independent processes.

**Conditional Entropy** There are various ways to rewrite the mutual information, using Bayes theorem $p(x,y) = p(x|y)p_Y(y)$ between the joint PDF $p(x,y)$, the conditional probability distribution $p(x|y)$ and the marginal PDF $p_Y(y)$, e.g.

$$\begin{aligned} I(X,Y) &= \left\langle \log\left(\frac{p}{p_X p_Y}\right)\right\rangle = \int p(x,y)\,\log\left(\frac{p(x|y)}{p_X(x)}\right)dx\,dy \\ &\equiv H(X)\ -\ H(X|Y)\ , \end{aligned} \tag{5.52}$$

where we have used the notation $H(X) = H[p_X]$ for the marginal entropy and defined the "conditional entropy"

$$H(X|Y) = -\int p(x,y)\log(p(x|y))\,dx\ dy\ . \tag{5.53}$$

The conditional entropy is positive for discrete processes, since

$$-p(x_i,y_j)\log(p(x_i|y_j)) = -p(x_i|y_j)p_Y(y_j)\log(p(x_i|y_j))$$

is positive, as $-p\log(p) \geq 0$ in the interval $p \in [0,1]$, compare Fig. 5.4 and Eq. (5.32) for the change-over from continuous to discrete variables. Several

[8] For a proof consider the generic substitution $x \to q(x)$ and a transformation of variables $x \to q$ via $dx = dq/q'$, with $q' = dq(x)/dx$, for the integration in Eq. (5.51).

variants of the conditional entropy may be used to extend the statistical complexity measures discussed in Sect. 5.3.1.

**Causal Dependencies** For independent processes one has $p(x,y) = p(x)p(y) = p(x|y)p(y)$ and hence

$$p(x|y) \;=\; p(x), \qquad\qquad H(X|Y) \;\to\; H(X)\;.$$

The opposite extreme is realized when the first channel is just a function of the second channel, viz when

$$x_i = f(y_i), \qquad p(x_i|y_i) = \delta_{x_i,f(y_i)}, \qquad p(x_i,y_i) = \delta_{x_i,f(y_i)}\,p(y_i)\;.$$

The conditional entropy (5.53) then vanishes,

$$H(X|Y) \;=\; -\sum_{x_i,y_j} \delta_{x_i,f(y_j)}\,p_Y(y_j)\log\left(\delta_{x_i,f(y_j)}\right) \;=\; 0\;,$$

since $\delta_{x_i,f(y_j)}$ is either unity, in which case $\log(\delta) = \log(1) = 0$, or zero, in which case $0\log(0)$ vanishes as a limiting process. The conditional entropy $H(X|Y)$ measures hence the amount of information, present in the stochastic process $X$, which is not causaly related to the process $Y$.

The mutual entropy reduces to the marginal entropy, as a corollary,

$$I(X,Y) \to H(X)\;,$$

for the case that $X$ is fully determined by $Y$. Compare Eq. (5.52).

### *5.2.3 Kullback-Leibler Divergence*

One is often interested in comparing two distribution functions $p(x)$ and $q(x)$ with respect to their similarity. When trying to construct a measure for the degree of similarity one is facing the dilemma that probability distributions are positive definite and one can hence not define a scalar product as for vectors; two PDFs cannot be orthogonal. It is however possible to define with the "Kullback-Leibler divergence" a positive definite measure.

Kullback-Leibler Divergence. Given two probability distribution functions $p(x)$ and $q(x)$ the functional

$$K[p;q] = \int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx \geq 0 \tag{5.54}$$

is a non-symmetric measure for the difference between $p(x)$ and $q(x)$.

The Kullback-Leibler divergence $K[p;q]$ is also denoted "relative entropy" and the proof for $K[p;q] \geq 0$ is analogous to the one for the mutual informa-

tion given in Sect. 5.2.2. The Kullback-Leibler divergence vanishes for identical PDFs, viz when $p(x) \equiv q(x)$.

**Relation to the $\chi^2$ test** We consider the case that the two distribution functions $p$ and $q$ are nearly identical,

$$q(x) = p(x) + \delta p(x), \qquad \delta p(x) \ll 1 \ ,$$

and expand $K[p;q]$ in powers of $\delta p(x)$, using

$$\log(q) = \log(p + \delta p) \approx \log(p) + \frac{\delta p}{p} - \left(\frac{\delta p}{p}\right)^2 + \dots$$

and obtaining

$$\begin{aligned} K[p;q] &\approx \int dx\, p \left[\log(p) - \log(p) - \frac{\delta p}{p} + \left(\frac{\delta p}{p}\right)^2\right] \\ &= \int dx\, \frac{(\delta p)^2}{p} = \int dx\, \frac{(p-q)^2}{p} \ , \end{aligned} \tag{5.55}$$

since $\int \delta p\, dx = 0$, as a consequence of the normalization conditions $\int p\, dx = 1 = \int q\, dx$. This measure for the similarity of two distribution functions is termed "$\chi^2$ test". It is actually symmetric under exchanging $q \leftrightarrow p$, up to order $(\delta p)^2$.

**Example** As a simple example we consider two distributions, $p(\sigma)$ and $q(\sigma)$, for a binary variable $\sigma = 0, 1$,

$$p(0) \ = \ 1/2 \ = \ p(1), \qquad q(0) \ = \ \alpha, \qquad q(1) \ = \ 1 - \alpha \ , \tag{5.56}$$

with $p(\sigma)$ being flat and $\alpha \in [0,1]$. The Kullback-Leibler divergence,

$$\begin{aligned} K[p;q] &= \sum_{\sigma=0,1} p(\sigma) \log\left(\frac{p(\sigma)}{q(\sigma)}\right) = \frac{-1}{2}\log(2\alpha) \ - \ \frac{1}{2}\log(2(1-\alpha)) \\ &= -\log(4(1-\alpha)\alpha)\,/\,2 \geq 0 \ , \end{aligned}$$

is unbounded, since $\lim_{\alpha\to 0,1} K[p;q] \to \infty$. Interchanging $p \leftrightarrow q$ we find

$$\begin{aligned} K[q;p] &= \alpha \log(2\alpha) \ + \ (1-\alpha)\log(2(1-\alpha)) \\ &= \log(2) \ + \ \alpha\log(\alpha) \ + \ (1-\alpha)\log(1-\alpha) \geq 0 \ , \end{aligned}$$

which is now finite in the limit $\lim_{\alpha\to 0,1}$. The Kullback-Leibler divergence is highly asymmetric, compare Fig. 5.6.

**Kullback-Leibler Divergence vs. Mutual Information** The mutual information, Eq. (5.49), is a special case of the Kullback-Leibler Divergence. We first write (5.54) for the case that $p$ and $q$ depend on two variables $x$ and

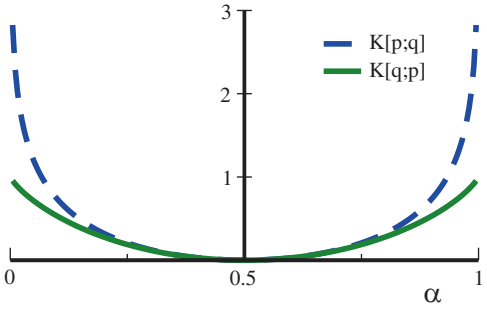


**Fig. 5.6** For the two PDFs $p$ and $q$ parametrized by $\alpha$, see Eq. (5.56), the respective Kullback-Leibler divergences $K[p;q]$ (*dashed line*) and $K[q;p]$ (*full line*). Note the maximal asymmetry for $\alpha \to 0, 1$, where $\lim_{\alpha\to 0,1} K[p;q] = \infty$

$y$,

$$K[p;q] = \int p(x,y) \log\left(\frac{p(x,y)}{q(x,y)}\right) dx\, dy \ . \tag{5.57}$$

This expression is identical to the mutual information (5.49) when considering for $q(x,y)$ the product of the two marginal distributions of $p(x,y)$,

$$q(x,y) = p(x)p(y), \quad p(x) = \int p(x,y)\, dy, \quad p(y) = \int p(x,y)\, dx \ .$$

Two independent processes are described by the product of their PDFs. The mutual information hence measures the distance between a joint distribution $p(x,y)$ and the product of its marginals, viz the distance between correlated and independent processes.

**Fisher Information** The Fisher information $F(\theta)$ measures the sensitivity of a distribution function $p(y,\theta)$ with respect to a given parametric dependence $\theta$,

$$F(\theta) = \int \left(\frac{\partial}{\partial\theta} \ln(p(y,\theta))\right)^2 p(y,\theta)\, dy \ . \tag{5.58}$$

In typical applications the parameter $\theta$ is a hidden observable one may be interested to estimate.

**Kullback-Leibler Divergence vs. Fisher Information** We consider the infinitesimal Kullback-Leibler divergence between $p(y,\theta)$ and $p(y,\theta+\delta\theta)$,

$$\begin{aligned} K &= \int dy\, p(y,\theta) \log\left(\frac{p(y,\theta)}{p(y,\theta+\delta\theta)}\right) \approx -\int dy\, p \log\left(\frac{p+p'\delta\theta}{p}\right) \\ &= -\int dy\, \frac{\partial p(y,\theta)}{\partial\theta}\delta\theta \ + \ \frac{(\delta\theta)^2}{2}\int dy\, \frac{1}{p(y,\theta)}\left(\frac{\partial p(y,\theta)}{\partial\theta}\right)^2 \end{aligned} \tag{5.59}$$

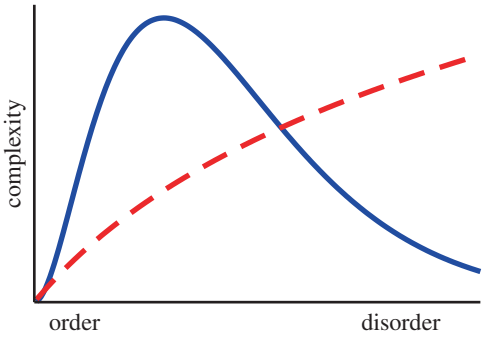


**Fig. 5.7** The degree of complexity (*full line*) should be minimal both in the fully ordered and the fully disordered regime. For some applications it may however be meaningful to consider complexity measures maximal for random states (*dashed line*)

with $p = p(y, \theta)$ and $p' = \partial p(y, \theta)/\partial\theta$. The first term in (5.59) can be written as

$$(-\delta\theta)\frac{\partial}{\partial\theta}\int dy\, p(y,\theta) = (-\delta\theta)\frac{\partial}{\partial\theta}1 \equiv 0 \ ,$$

and vanishes. The second term in (5.59) contains the Fisher information (5.58) and hence

$$K\big[p(y,\theta); p(y,\theta+\delta\theta)\big] = F(\theta)\frac{(\delta\theta)^2}{2} \ , \tag{5.60}$$

which establishes the role of the Fisher information as a metric.

## 5.3 Complexity Measures

Can we provide a single measure, or a small number of measures, suitable for characterizing the "degree of complexity" of any dynamical system at hand? This rather philosophical question has fascinated researchers for decades and no definitive answer is known.

The quest of complexity measures touches many interesting topics in dynamical system theory and has led to a number of powerful tools suitable for studying dynamical systems, the original goal of developing a one-size-fit-all measure for complexity seems however not anymore a scientifically valid target. Complex dynamical systems can show a huge variety of qualitatively different behaviors, one of the reasons why complex system theory is so fascinating, and it is not appropriate to shove all complex systems into a single basket for the purpose of measuring their degree of complexity with a single yardstick.

**Intuitive Complexity** The task of developing a mathematically well defined measure for complexity is handicapped by the lack of a precisely defined goal. In the following we will discuss some selected prerequisites and constraints one may postulate for a valid complexity measure. In the end it is, however, up to our intuition for deciding whether these requirements are appropriate or not.

An example of a process one may intuitively attribute a high degree of complexity are the intricate spatio-temporal patterns generated by the forest fire model discussed in Sect. **??**, and illustrated in Fig. **??**, with perpetually changing fronts of fires burning through a continuously regrowing forest.

**Complexity vs. Randomness** A popular proposal for a complexity measure is the information entropy $H[p]$, see Eq. (5.27). It vanishes when the system is regular, which agrees with our intuitive presumption that complexity is low when nothing happens. The entropy is however maximal for random dynamics, as shown in Fig. 5.5.

It is a question of viewpoints to which extend one should consider random systems as complex, compare Fig. 5.7. For some considerations, e.g. when dealing with "algorithmic complexity" (see Sect. 5.3.2) it makes sense to attribute maximal complexity degrees to completely random sets of objects. In general, however, complexity measures should be concave and minimal for regular behavior as well as for purely random sequences.

**Complexity of Multi-component Systems** Complexity should be a positive quantity, like entropy. Should it be, however, extensive or intensive? This is a difficult and highly non-trivial question to ponder.

Intuitively one may demand complexity to be intensive, as one would not expect to gain complexity when considering the behavior of a set of $N$ independent and identical dynamical systems. On the other side we cannot rule out that $N$ strongly interacting dynamical systems could show more and more complex behavior with an increasing number of subsystems, e.g. we consider intuitively the global brain dynamics to be orders of magnitude more complex than the firing patterns of the individual neurons.

There is no simple way out of this quandary when searching for a single one-size-fits-all complexity measure. Both intensive and extensive complexity measures have their areas of validity.

**Complexity and Behavior** The search for complexity measures is not just an abstract academic quest. As an example consider how bored we are when our environment is repetitive, having low complexity, and how stressed when the complexity of our sensory inputs is too large. There are indeed indications that a valid behavioral strategy for highly developed cognitive systems may consist in optimizing the degree of complexity. Well defined complexity measures are necessary in order to quantify this intuitive statement mathematically.

### 5.3.1 Complexity and Predictability

Interesting complexity measures can be constructed using statistical tools, generalizing concepts like information entropy and mutual information. We will consider here time series generated from a finite set of symbols. One may, however, interchange the time label with a space label in the following, whenever one is concerned with studying the complexity of spatial structures.

**Stationary Dynamical Processes** As a prerequisite we need stationary dynamical processes, viz dynamical processes which do not change their behavior and their statistical properties qualitatively over time. In practice this implies that the time series considered, as generated by some dynamical system, has a finite time horizon $\tau$. The system might have several time scales $\tau_i \leq \tau$, but for large times $t \gg \tau$ all correlation functions need to fall off exponentially, like the autocorrelation function defined in Sect. **??**. Note, that this assumption may break down for critical dynamical systems, which are characterized, as discussed in Chap. **??**, by dynamical and statistical correlations decaying only slowly, with an inverse power of time.

**Measuring Joint Probabilities** For times $t_0$, $t_1$, ..., a set of symbols $X$, and a time series containing $n$ elements,

$$x_n, x_{n-1}, \ldots, x_2, x_1, \qquad x_i = x(t_i), \qquad x_i \in X \tag{5.61}$$

we may define the joint probability distribution

$$p_n : \qquad p(x_n, \ldots, x_1) \ . \tag{5.62}$$

The joint probability $p(x_n, \ldots, x_1)$ is not given a priori. It needs to be measured from an ensemble of time series. This is a very demanding task as $p(x_n, \ldots, x_1)$ has $(N_s)^n$ components, with $N_s$ being the number of symbols in $X$.

It clearly makes no sense to consider joint probabilities $p_n$ for time differences $t_n \gg \tau$, the evaluation of joint probabilities exceeding the intrinsic time horizon $\tau$ is a waste of effort. In practice finite values of $n$ are considered, taking subsets of length $n$ of a complete time series containing normally a vastly larger number of elements. This is an admissible procedure for stationary dynamical processes.

**Entropy Density** We recall the definition of the Shannon entropy

$$H[p_n] = - \sum_{x_n, \ldots, x_1 \in X} p(x_n, \ldots, x_1) \, \log(p(x_n, \ldots, x_1)) \equiv - \langle \log(p_n) \rangle_{p_n} \ , \tag{5.63}$$

which needs to be measured for an ensemble of time series of length $n$ or greater. Of interest is the entropy density in the limit of large times,

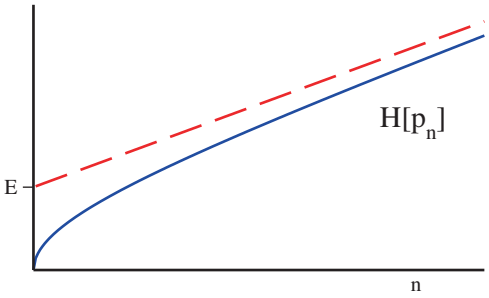


**Fig. 5.8** The entropy (*full line*) $H[p_n]$ of a time series of length $n$ increases monotonically, with the limiting slope (*dashed line*) $h_\infty$. For large $n \to \infty$ the entropy $H[p_n] \approx E + h_\infty n$, with the excess entropy $E$ given by the intercept of asymptote with the $y$-axis

$$h_\infty = \lim_{n\to\infty} \frac{1}{n} H[p_n] \ , \tag{5.64}$$

which exists for stationary dynamical processes with finite time horizons. The entropy density is the mean number of bits per time step needed for encoding the time series statistically.

**Excess Entropy** We define the "excess entropy" $E$ as

$$E = \lim_{n\to\infty} \big( H[p_n] \ - \ n\, h_\infty \big) \geq 0 \ . \tag{5.65}$$

The excess entropy is just the non-extensive part of the entropy, it is the coefficient of the term $\propto n^0$ when expanding the entropy in powers of $1/n$,

$$H[p_n] = n\, h_\infty + E + O(1/n), \qquad n \ \to \ \infty \ , \tag{5.66}$$

compare Fig. 5.8. The excess entropy $E$ is positive as long as $H[p_n]$ is concave as a function of $n$ (we leave the proof of this statement as an exercise to the reader), which is the case for stationary dynamical processes. For practical purposes one may approximate the excess entropy using

$$h_\infty = \lim_{n\to\infty} h_n, \qquad h_n = H[p_{n+1}] \ - \ H[p_n] \ , \tag{5.67}$$

since $h_\infty$ corresponds to the asymptotic slope of $H[p_n]$, compare Fig. 5.8.

- One may also use Eqs. (5.67) and (5.53) for rewriting the entropy density $h_n$ in terms of an appropriately generalized conditional entropy.
- Using Eq. (5.66) we may rewrite the excess entropy as

$$\sum_n \left[ \frac{H[p_n]}{n} - h_\infty \right] \ .$$

  In this form the excess entropy is known as the "effective measure complexity" (EMC) or "Grassberger entropy".

**Excess Entropy and Predictability** The excess entropy vanishes both for a random and for an ordered system. For a random system

$$H[p_n] = n\, H[p_X] \equiv n\, h_\infty \ ,$$

where $p_X$ is the marginal probability. The excess entropy, Eq. (5.65) vanishes consequently. For an example of a system with ordered states we consider the dynamics

$$\ldots 000000000000000 \ldots , \qquad \ldots 111111111111111 \ldots \ ,$$

for a binary variable, occurring with probabilities $\alpha$ and $1-\alpha$ respectively. This kind of dynamics is the natural output of logical AND or OR rules. The joint probability distribution then has only two non-zero components,

$$p(0,\ldots,0) = \alpha, \qquad p(1,\ldots,1) = 1-\alpha, \qquad \forall n \ ,$$

all other $p(x_n,\ldots,x_1)$ vanish and

$$H[p_n] \equiv -\alpha \log(\alpha) \ - \ (1-\alpha)\log(1-\alpha), \qquad \forall n \ .$$

The entropy density $h_\infty$ vanishes and the excess entropy $E$ becomes $H[p_n]$; it vanishes for $\alpha \to 0, 1$, viz in the deterministic limit.

The excess entropy therefore fulfills the concaveness criteria illustrated in Fig. 5.7, vanishing both in the absence of predictability (random states) and for the case of strong predictability (i.e. for deterministic systems). The excess entropy does however not vanish in above example for $0 < \alpha < 1$, when two predictable states are superimposed statistically in an ensemble of time series. Whether this behavior is compatible with our intuitive notion of complexity is, to a certain extent, a matter of taste.

**Discussion** The excess entropy is a nice tool for time series analysis, satisfying several basic criteria for complexity measures, and there is a plethora of routes for further developments, e.g. for systems showing structured dynamical activity both in the time as well as in the spatial domain. The excess entropy is however exceedingly difficult to evaluate numerically and its scope of applications therefore limited to theoretical studies.

### *5.3.2 Algorithmic and Generative Complexity*

We have discussed so far descriptive approaches using statistical methods for the construction of complexity measures. One may, on the other hand, be interested in modelling the generative process. The question is then: which is the simplest model able to explain the observed data?

**Individual Objects** For the statistical analysis of a time series we have been concerned with ensembles of time series, as generated by the identical underlying dynamical system, and with the limit of infinitely long times. In this section we will be dealing with individual objects composed of a finite number of $n$ symbols, like

$$00000000000000000000, \qquad 00100000111010010111001 \ .$$

The question is then: which dynamical model can generate the given string of symbols? One is interested, in particular, in strings of bits and in computer codes capable of reproducing them.

**Turing Machine** The reference computer codes in theoretical informatics is the set of instructions needed for a "Turing machine" to carry out a given computation. The exact definition for a Turing machine is not of relevance here, it is essentially a finite-state machine working on a set of instructions called code. The Turing machine plays a central role in the theory of computability, e.g. when one is interested in examining how hard it is to find the solution to a given set of problems.

**Algorithmic Complexity** The notion of algorithmic complexity tries to find an answer to the question of how hard it is to reproduce a given time series in the absence of prior knowledge.

> Algorithmic Complexity. The "algorithmic complexity" of a string of bits is the length of the shortest program that prints the given string of bits and then halts.

The algorithmic complexity is also called "Kolmogorov complexity". Note, that the involved computer or Turing machine is supposed to start with a blank memory, viz with no prior knowledge.

**Algorithmic Complexity and Randomness** Algorithmic complexity is a very powerful concept for theoretical considerations in the context of optimal computability. It has, however, two drawbacks, being not computable and attributing maximal complexity to random sequences.

A random number generator can only be approximated by any finite state machine like the Turing machine and would need an infinite code length to be perfect. That is the reason why real-world codes for random number generators are producing only "pseudo random numbers", with the degree of randomness to be tested by various statistical measures. Algorithmic complexity therefore conflicts with the common postulate for complexity measures to vanish for random states, compare Fig. 5.7.

**Deterministic Complexity** There is a vast line of research trying to understand the generative mechanism of complex behavior not algorithmically but from the perspective of dynamical system theory, in particular for deterministic systems. The question is then: in the absence of noise, which are the features needed to produce interesting and complex trajectories?

Of interest are in this context the sensitivity to initial condition for systems having a transition between chaotic and regular states in phase space, see Chap. **??**, the effect of bifurcations and non-trivial attractors like strange attractors, see Chap. **??**, and the consequences of feedback and tendencies toward synchronization, see Chap. **??**. This line of research is embedded in the general quest of understanding the properties and the generative causes of complex and adaptive dynamical systems.

**Complexity and Emergence** Intuitively, we attribute a high degree of complexity to ever changing structure emerging from possibly simple underlying rules, an example being the forest fires burning their way through the forest along self-organized fire fronts, compare Fig. **??** for an illustration. This link between complexity and "emergence" is, however, not easy to mathematize, as no precise measure for emergence has been proposed to date.

**Weak and Strong Emergence** On a final note one needs to mention that a vigorous distinction is being made in philosophy between the concept of *weak emergence*, which we treated here, and the scientifically irrelevant notion of *strong emergence*. Properties of a complex system generated via weak emergence result from the underlying microscopic laws, whereas strong emergence leads to top-level properties which are strictly novel in the sense, that they cannot, like magic, linked causally to the underlying microscopic laws of nature.