Information Theory for Complex Systems

—

Lecture Notes

Kristian Lindgren

January-March 2008

Complex Systems Group
Department of Energy and Environment
Chalmers University of Technology
Göteborg, Sweden

Preface

These lecture notes have been developed during the past eight years in connection with the course Information Theory for Complex Systems in the International Masters Programme in Complex Adaptive Systems offered at Chalmers since year 2000. The basis for the lecture notes is research done together with my colleague Karl-Erik Eriksson during the 1980's[1]. The past four years, new efforts have been put into the reasearch, primarily dealing with information theory for pattern forming systems. This research has been supported by two European projects: PACE (Programmable Artificial Cell Evolution), a European Integrated Project in the EU FP6-IST-FET Complex Systems Initiative, and by EMBIO (Emergent Organisation in Complex Biomolecular Systems) under EU FP6.

June 30, 2008,

Kristian Lindgren

[1] The information-theoretic research was published in a book: *Structure, Context, Complexity, Organization* (Eriksson, Lindgren and Månsson; World Scientific, Singapore, 1987).

1 Introduction

The term “complex systems” has become a unifying concept for research and studies of large multi-component systems in many disciplines like physics, chemistry, biology, social science, computer science, economics and geography. The fact that systems composed of a large number of simple components can exhibit complex phenomena is exemplified in all these areas: the second law of thermodynamics (as a statistical result of large physical systems), self-organising systems (in the form of chemical reaction-diffusion systems), neural networks, evolution of cooperation, cellular automata (as an example of an abstract computational class of systems), economic systems of interacting trading agents, urban growth and traffic systems. During the past two decades the area of complex systems has grown tremendously, leading to a large number of scientific journals. An important factor has been the fast development of computers, allowing for cheap and powerful experimental laboratories of complex systems models.

A complex systems usually involves a large number of components. These components may be simple, both in terms of their internal characteristics and in the way they interact. Still, when the system is observed over longer time and length scales, there may be phenomena that are not easily understood in terms of the simple components and their interactions.

There is no universal definition of a complex system, but there are several features that researchers usually consider, like those mentioned above, when they say that a system is complex. One important scientific question is whether these and other characteristics of the systems can be quantified. This is one of the aims with this book – to provide a set of tools that can be used to give a quantitative description of a complex system for a variety of different types of systems.

In order to analyse systems composed of many components, statistical methods serve as important tools. Within physics, this approach is taken in statistical mechanics. Information theory is a branch of probability theory that has a mathematical basis that is equivalent to foundations of statistical mechanics, and information theory also provides concepts and methods that are useful in order to analyse structure, disorder, randomness, etc in models of complex systems.

Information-theoretic concepts can be applied on the macro-level of a system, for example, in order to describe the spatial structure formed in a chemical self-organising system. The connection between information theory and statistical mechanics makes it possible to relate such an analysis to the thermodynamical properties and limitations of the system.

Information theory may also be applied on the micro-level in physical (and other) systems, for example in spin system models and other more abstract models for statistical mechanics. In this lecture series we shall show and illustrate how information-theoretic quantities can be related to statistical mechanics and thermodynamics properties, and in this way we may illustrate, in an information-theoretic perspective, complex phenomena like the second law of thermodynamics.

The concept of information has many meanings. For a historical survey and a discussion of its use in different scientific disciplines, see (Capurro and Hjørland, 2003)[2]. During the 20th century information has often been associated with knowledge, or transmission of knowledge. The introduction of information as a quantity in science and engineering was associated with the development of communication technology. Harry Nyquist and Ralph Hartley were pioneers in the 1920's and they published papers in the context of telegraph speed and "information" transmission, offering the first quantitative definition of information. Their ideas were based on the observation that such a quantitative measure should be proportional to the logarithm of the number of choices or possibilities there is for the message to be sent.

This basic quantity was then generalised by Claude Shannon and Warren Weaver in 1947 when they presented the information theory that serves as the basis for the field today. This generalisation is based on a probabilistic or statistical description of the system under study. The basic concepts will be discussed in detail in Chapter 2.

It should be noted that the information concept that is presented in this book is a very specific one, related to a probabilistic description of the system that is studied or related to how such a description is changed when the knowledge about the system is changed. This means that what is quantified is related to the *description* of the system rather than the system itself. One can choose a description that hides a lot of detail and in that case may get a different information quantity compared to the fully described system.

Another point that needs to be stressed is that the information we consider has no direct connection to *meaning*. This is stated in Shannon's first paper presenting the theory (Shannon 1948):

> "... the fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages."

The book is organised as follows. In Chapter 2 the basic concepts in information theory is presented along with the maximum entropy principle. Chapter 3 presents the application of the concept to symbol sequences, or, more general, lattice systems in one or more dimensions. In Chapter 4 this is applied to the simple class of discrete dynamical system called Cellular Automata (CA). Information theory provides several tools in which the dynamics and the patterns generated can be studied. We will primarily focus on one-dimensional CA using the information theory for symbol sequences. In Chapter 5, some connections between the information-theoretic quantities and physics (statistical mechanics and thermodynamics) is presented. This relation makes it possible to connect some of the information-theoretic quantities describing a system with the thermodynamic constraints that put constraints on the dynamics of the system.

[2] Online material available at http://www.capurro.de/infoconcept.html.

To be able to analyse spatially extended patterns in continues space, we develop *geometric information theory* in Chapter 6. These concepts can, for example, be applied to images to identify at what length scales and at what positions information is located. This also includes applications to fractal patterns and the concept of dimension. In Chapter 7 the geometric information theory is applied to pattern formation in chemical system, where the system is described as a set of concentration profiles for the different chemicals involved. We derive a continuity equation for information that describes how free energy that drives the pattern formation process flows into the system and is aggregated at certain positions and length scale when information in the chemical pattern is built up. In Chapter 8 we give an information-theoretic perspective on chaotic dynamical systems. We show how the sensitivity to small fluctuations can be seen as an information flow from “micro” to “macro”.

Each Chapter contains a set of problems of varying degree of difficulty. Many of these have been taken from exam problems given during the history of the course that started in 1990.

2 Information theory

A quantitative measure of information was presented in *Bell System Technical Journal* already 1928 by Ralph Hartley, but even before that Harry Nyquist had brought up the issue in the same journal in 1924. Hartley showed, in his paper on signal transmission, that the information content in a message consisting of n characters, each of them chosen from an alphabet of N different symbols, should be proportional to $n \log N$. The generalisation of this was not done until 1948, when Shannon presented the information theory still used today. The concepts of information was after that also applied to other areas in science, including physics, see, e.g., (Brillouin, 1956) and (Jaynes, 1957), which will be discussed in a later Chapter.

2.1 Basic concepts

Information is usually quantified in units of *bits*, "binary units", which is the amount of information that can be stored in a single binary symbol. In information theory this quantity is defined in terms of probabilities. When a person makes an observation, for example reads the next character in a text or takes a look at the watch, information is received. How much information that is gained from the observation depends on how unexpected the event was. For an event that has an a priori probability p, the corresponding information $I(p)$ is defined

$$I(p) = \log\frac{1}{p} \tag{2.1}$$

This definition is a generalisation of Hartley's information quantity in that it distinguishes between events of different probability. If the event is an observation of one character drawn from an alphabet consisting of N symbols, where all are assumed to be equally probable, then the probability is $p = 1/N$ and the information is $I = \log N$. Here we have used the logarithm of base 2, which then results in the unit bit. Later on we will switch to base e, denoting the logarithm "ln", resulting in the natural information unit, usually called "nat".

Using the definition we see that an unexpected or less probable event (small p) corresponds to a high information value, but if we are told something that we already know ($p = 1$) the received information is zero.

When one does not have full information on the state of a certain system, one may associate a probability to each possible state. This means that the state of system is described by a probability distribution. This is one of the basic ideas in statistical mechanics, where one does not know the exact state (microstate) of a physical system but describes it as a probability distribution (macrostate) over the possible microstates. Such a probability distribution is denoted

$$P = \left\{p_i\right\}_{i=1}^{n} \tag{2.2}$$

where p_i denotes the probability for state i, and n is the number of possible states the system can take. The probabilities should be non-negative and normalised

$$p_i \geq 0 \tag{2.3}$$

$$\sum_{i=1}^{n} p_i = 1 \tag{2.4}$$

When one observes a system and learns about its exact state the amount of gained information depends on which state is observed, as is stated in Eq. (2.1). Therefore one can characterise the system by the *average* or *expected* information one gets when the system is observed. This expectation value is calculated on the basis of the probability distribution P that describes the system, and the expected information is called the *entropy* of the (unobserved) system:

$$S[P] = \left\langle \log \frac{1}{p_i} \right\rangle_i = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} \ . \tag{2.5}$$

(We define $0 \cdot \log 0 = 0$.) This entropy is usually called the *Shannon entropy*. It is clear that $0 \leq S[P] \leq \log n$. The entropy is the expected gain of information when we observe a system characterised by a probability distribution P over its possible states. One can also say that the entropy quantifies the *lack of knowledge* of the system (before the exact state of the system is observed). Sometimes this is also called the *disorder* of the system. This interpretation may become more clear when we discuss entropies in symbol sequences in a later chapter.

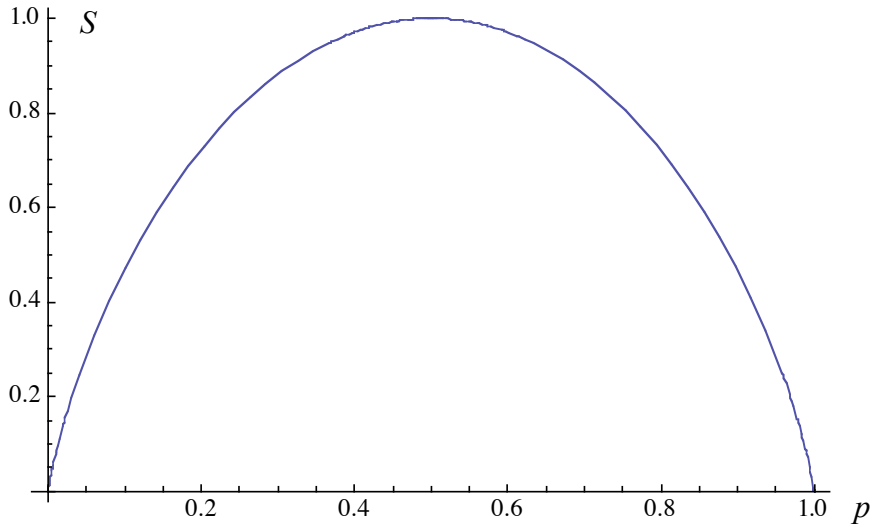


Figure 2.1. For a system with two possible states ($n = 2$), with the probabilities p and $1–p$, the entropy $S(p)$, or the lack of knowledge, is at maximum when both states are equally probable ($p = 1/2$). This means that we have no clue on which state we will find the system in when observing it. If we know that the system is in a certain state ($p = 0$ or $p = 1$), then the entropy is zero.

2.1.1 Entropy and coding – an extended example

Consider a stochastic process that generates randomly sequences of the symbols '`a`', '`b`', '`c`', and '`d`', for example '`bcaabada`…'. Suppose, to start with, that it is unknown with which probabilities the symbols are generated. Then the best guess is to assign probabilities 1/4 for each of the events, and that subsequent symbols are independent. The information gained in any of the possible observations of a single symbol is then log 4 = 2 (bits). This seems reasonable since we can simply code the four symbols with the binary code words '`00`', '`01`', '`10`', and '`11`', respectively.

Suppose again that symbols are generated independently of each other but that they are generated by the probabilities p(`a`) = 1/2, p(`b`) = 1/4, p(`c`) = 1/8, and , p(`d`) = 1/8, and that this is known a priori by the observer.

When observing a symbol, the amount of information one gets, according to Eq. (2.1), depends on the symbol. When observing an '`a`' we get 1 bit, but a '`d`' would give us 3 bits. The expectation value of the information we get from an observation is the weighted average of the information from the four possible events, which is the entropy of Eq. (2.5),

$$S = 1/2 \ + \ 1/4 \cdot 2 \ + \ 1/8 \cdot 3 \ + \ 1/8 \cdot 3 \ = \ 7/4.$$

We find that the entropy is now reduced, since we have some prior knowledge on the probabilities with which symbols occur. This also means that, by making a better code compared to the trivial one mentioned above, one could compress a message or a sequence of symbols in which the frequencies follow these probabilities. If we for example use the code words '`0`' for `a`, '`10`' for `b`, '`110`' for `c`, and '`111`' for `d`, we note that the average code word length decreases from 2 to S above. The knowledge of the frequencies of the four symbols has allowed us to compress the message from 2 bits per symbol to 1.75 bits per symbol. The trick lies in the fact that we have used a code word length (bits in the code word) that equals to the information gained if the corresponding symbol is observed. So, common symbols that carry little information should be given short code words and vice versa.

2.1.2 Entropy as an additive quantity

The entropy of a system composed by independent parts equals the sum of the entropies of the parts. The independency means that the probability for a certain microstate of the whole system equals the product of the probabilities of the corresponding microstates of the parts. Assume that the system consists of two subsystems, characterised by the probability distributions

$$Q = \{q_i\}_{i=1}^{n} \quad \text{and} \quad R = \{r_j\}_{j=1}^{m} \ , \tag{2.6}$$

with q_i and r_j representing probabilities for states i and j in the then the two subsystems, respectively. whole system is characterised by

$$P = \{q_i r_j\}_{i=1,j=1}^{n,m} \ . \tag{2.7}$$

The entropy can be written

$$S[P] = \sum_{i=1}^{n}\sum_{j=1}^{m} q_i\, r_j \log\frac{1}{q_i\, r_j} = \sum_{i=1}^{n} q_i \log\frac{1}{q_i} + \sum_{j=1}^{m} r_j \log\frac{1}{r_j} = S[Q] + S[R]. \tag{2.8}$$

Here we have used the normalisation $\Sigma_i\, q_i = \Sigma_j\, r_j = 1$.

2.1.3 Relative information, relative entropy, or Kullback information

Often when we make an observation we do not receive full information of the system. The exact microstate may not be revealed, but based on the observation, we may replace our original, *a priori*, distribution $P^{(0)}$ with a new one P. We here assume that whenever a microstate i is assumed to be impossible in the *a priori* situation, $p_i^{(0)} = 0$, then this state is also impossible in the new probability distribution, $p_i = 0$.

In order to quantify how much information we have gained from the observation, we calculate the decrease in our *lack of knowledge* of the system. Before the observation, we thought that $p_i^{(0)}$ described the probabilities, but when calculating the expectation value of that *lack of knowledge* before observation $S^{(0)}$, we should use the new probabilities for the weights,

$$S^{(0)} = \sum_{i=1}^{n} p_i \log\frac{1}{p_i^{(0)}}\ , \tag{2.9}$$

while the lack of knowledge after the observation is the ordinary entropy S, as in Eq. (2.5). The information gained in the observation, the *Kullback information* $K[P^{(0)}; P]$, when the *a priori* distribution $P^{(0)}$ is replaced by the new distribution P, is then

$$K\left[P^{(0)};P\right] = S^{(0)} - S = \sum_{i=1}^{n} p_i \log\frac{1}{p_i^{(0)}} - \sum_{i=1}^{n} p_i \log\frac{1}{p_i} = \sum_{i=1}^{n} p_i \log\frac{p_i}{p_i^{(0)}}. \tag{2.10}$$

(Here we define $0\cdot\log(0/0) = 0$.) This quantity is also called the *relative information* or the *relative entropy* between the distributions $P^{(0)}$ and P. This quantity fulfils the inequality

$$K\left[P^{(0)};P\right] \geq 0, \tag{2.11}$$

with equality only when the two distributions are identical.

Proof: Consider the function $g(x) = x - 1 - \ln x \geq 0$. By adding and subtracting 1, the expression for the contrast may be rewritten as an average over $g(p_i^{(0)} / p_i)$, after transforming from log of base 2 to base e,

$$K\left[P^{(0)};P\right] = \frac{1}{\ln 2}\sum_{i=1}^{n} p_i \ln\frac{p_i}{p_i^{(0)}} = \\ = \frac{1}{\ln 2}\sum_{i=1}^{n} p_i\left(\frac{p_i^{(0)}}{p_i} - 1 - \ln\frac{p_i^{(0)}}{p_i}\right) = \frac{1}{\ln 2}\sum_{i=1}^{n} p_i\, g\left(\frac{p_i^{(0)}}{p_i}\right) \geq 0, \tag{2.12}$$

which proves the inequality. It is also clear that equality requires that $p_i^{(0)} = p_i$ for all i.

2.1.4 The concavity of entropy

We shall now use inequality (2.12) to prove that the entropy is a concave function[3]. For the entropy function this means that, if P and Q are two probability distributions (both over n possible states), then the entropy of any weighted average of these is larger than the corresponding weighted average of their respective entropies,

$$S[a\cdot P + (1-a)\cdot Q] \geq a\cdot S[P] + (1-a)\cdot S[Q],$$

where a and $(1–a)$ are the weight factors ($0 \leq a \leq 1$). The probabilities in the distributions are denoted p_i and q_i, respectively. The proof is

$$S[a\cdot P + (1-a)\cdot Q] - \left(a\cdot S[P] + (1-a)\cdot S[Q]\right) = \\ = \sum_{i=1}^{n}(ap_i + (1-a)q_i)\log\frac{1}{ap_i + (1-a)q_i} - \sum_{i=1}^{n}\left(ap_i\log\frac{1}{ap_i} + (1-a)q_i\log\frac{1}{(1-a)q_i}\right) = \\ = a\sum_{i=1}^{n} p_i\log\frac{p_i}{ap_i + (1-a)q_i} + (1-a)\sum_{i=1}^{n} q_i\log\frac{q_i}{ap_i + (1-a)q_i} = \\ = aK[a\cdot P + (1-a)\cdot Q; P] + (1-a)K[a\cdot P + (1-a)\cdot Q; Q] \geq 0.$$

The non-negative property of the Kullback information will be used in several proofs for inequalities involving different entropy quantities in later chapters.

2.2 Maximum entropy formalism

Even if we do not know exactly the state (microstate) of a certain system, we may have some information on its state. We could, for example, have knowledge about the average energy or number of particles. Statistical physics is based on the idea that, with such limited information on the state of the system, we make an estimate of the probabilities for the possible microstates. Usually, there are, though, an infinite number of possible probability distributions that are consistent with the known properties of the system in study. The question is then: how should we choose the probability distribution describing our system?

Here it is reasonable to use the concept of entropy, since it can be interpreted as our lack of information on the state of the system. When assigning a probability distribution for the

[3] A function $f(x)$ is concave if $f(ax + (1–a)y) \geq a\,f(x) + (1–a)\,f(y)$, for all $0 \leq a \leq 1$.

system, we should not use a probability distribution that represent more knowledge than what we already have. For example, when assigning probabilities for the outcome of a through of a six-sided dice, we say that all probabilities are equal and 1/6. This corresponds to the situation of maximum entropy, or maximum lack of knowledge. In general we may have some knowledge about the system that implies that this maximum entropy level is not correct. There may be constraints that our knowledge implies which leads to limitations on to how the different probabilities can be varied.

Therefore, we choose, among the probability distributions that are consistent with the known system properties, the distribution that maximises the entropy. The probability distribution is in this way derived from a maximisation problem with constraints. This method assures that we do not include any more knowledge, in the description of the system, than we already know. This is the basic idea behind *the maximum entropy principle*, which also can be called *the principle of minimal bias*.

Let us assume that we shall derive a probability distribution $P = \{p_i\}$ describing a system for which we know that the following r averages (or expectation values) hold,

$$\langle f_k \rangle = \sum_{i=1}^{n} p_i f_k(i) = F_k \qquad (k = 1,...,r).$$

This means that we have r functions $f_k(i)$, $k = 1,\ldots, r$, of the microstates i, and that we know the expectation values of these, F_k. Such a function could, for example, give the energy of microstate i. The maximum entropy principle now states that we should choose the probability distribution that maximises the entropy under these conditions:

> Choose $P = \{p_i\}_{i=1}^{n}$, so that
> the entropy $S[P] = \sum_{i=1}^{n} p_i \ln \frac{1}{p_i}$ is maximised,
> subject to constraints $\langle f_k \rangle = F_k \qquad (k = 1,...,r)$,
> and the normalisation condition, $\sum_{i=1}^{n} p_i = 1$.

We solve this general problem by using the Lagrange formalism[4]. Therefore, we define the Lagrange function

$$L(p_1,\ldots,p_n,\lambda_1,\ldots,\lambda_r,\mu) = S[P] + \sum_{k=1}^{r}\lambda_k\left(F_k - \sum_{i=1}^{n} p_i f_k(i)\right) + (\mu-1)\left(1-\sum_{i=1}^{n} p_i\right), \tag{2.13}$$

where we have introduced Lagrange multipliers λ_k, for the r constraints, and $(\mu - 1)$ for the normalisation constraint. (Here we have chosen $\mu - 1$, instead of just μ, because it makes the expressions in the derivation a little simpler.) The optimisation problem is now solved by finding the P, λ_k, and μ for which partial derivatives with respect to L is 0. (Note that the constraints are equivalent to the conditions that the partial derivatives of L with respect to the Lagrange variables are 0.)

Derivation of L with respect to p_j results in

$$-\frac{\partial L}{\partial p_j} = \ln p_j + 1 + \sum_{k=1}^{r}\lambda_k f_k(j) + \mu - 1 , \tag{2.14}$$

which together with the optimisation requirement $\partial L / \partial p_j = 0$ gives

$$p_j = \exp\left(-\mu - \sum_{k=1}^{r}\lambda_k f_k(j)\right) = \exp\left(-\mu - \boldsymbol{\lambda}\cdot\mathbf{f}(j)\right) , \tag{2.15}$$

where we have introduced a vector notation for the Lagrange variables, $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_r)$, and $\mathbf{f}(j) = (f_1(j),\ldots,f_r(j))$. A probability distribution of this form is called a *Gibbs distribution*. The values on the Lagrangian variables can be derived from the constraints. Note that the distribution automatically fulfils $p_j \geq 0$. The normalisation condition $\Sigma_j p_j = 1$ determines μ as a function of $\boldsymbol{\lambda}$,

$$\mu(\boldsymbol{\lambda}) = \ln Z(\boldsymbol{\lambda}) , \tag{2.16}$$

where $Z(\boldsymbol{\lambda})$ denotes the *state sum*,

$$Z(\boldsymbol{\lambda}) = \sum_{j=1}^{n}\exp(-\boldsymbol{\lambda}\cdot\mathbf{f}(j)). \tag{2.17}$$

Finally, the Lagrangian variables $\boldsymbol{\lambda}$ are determined by the other constraints, which can be expressed as

[4] This means that we add, to the objective function, the constraints as separate terms (each being 0 when the constraint is fulfilled), each multiplied by a Lagrangian variable. The new objective function L is a function of all original variables and in addition one variable each for the different constraints. The maximum of the original constrained problem corresponds to the maximum of the unconstrained Lagrange function L, which is typically easier to solve.

$$\frac{\partial \mu(\lambda)}{\partial \lambda} = -\mathbf{F} \ , \tag{2.18}$$

where $\mathbf{F} = (F_1,\ldots, F_r)$. The maximum entropy value can be expressed in terms of the Lagrange variables as

$$S[P] = \mu + \lambda \cdot \mathbf{F} \ . \tag{2.19}$$

(This derivation is left as an exercise.) In the next Section, we illustrate the formalism with an example of a simple Gibbs distribution in physics.

2.2.1 The Bose-Einstein distribution

Suppose that we have a system composed of a number of particles, where the unknown microstate of the system is described by the exact number of particles n (n = 0, 1, 2,…). Let us assume that we gained the knowledge that the expected number of particles in this type of system is N, for example from measurements on a large number of such system. If this is all we know, the only constraint apart from normalisation we have is that the expected number of particles equals N. If p_n is the probability for the system being composed of n particles, then this constraint can be written

$$\sum_{n=0}^{\infty} p_n n = N . \tag{2.20}$$

Let λ be the Lagrange variable for this constraint and μ the variable for the normalisation. Then the state sum can be written

$$Z(\lambda) = \exp(\mu(\lambda)) = \sum_{n=0}^{\infty} \exp(-\lambda n) = \frac{1}{1-e^{-\lambda}} . \tag{2.21}$$

By using Eq. (2.18), we find that

$$\frac{\partial}{\partial \lambda} \ln \frac{1}{1-e^{-\lambda}} = -\frac{e^{-\lambda}}{1-e^{-\lambda}} = -N , \tag{2.22}$$

which results in

$$\lambda = \ln\left(1 + \frac{1}{N}\right) , \tag{2.23}$$

$$\mu = \ln(N+1) \ . \tag{2.24}$$

The probability distribution is then recognised as the Bose-Einstein distribution,

$$p_n = \frac{N^n}{(N+1)^{n+1}} \ . \tag{2.25}$$

2.3 Generalisation of entropies to a continuous state-space

The information-theoretic terminology can be generalised to a continuous state-space. Then we assume that we have a probability density $p(\mathbf{x})$ over this state space, and suppose that $\mathbf{x} = (x_1, \ldots, x_D)$ is a vector in a D-dimensional Euclidean space. This could, for example, mean that we do not know the exact position of a particle, but that we describe the position of the particle as a probability density over this space. (To be complete, we should also introduce a density of states function, but here we assume that to be identical to 1, and therefore we omit it in the expressions[5].) The probability to find the system in a certain volume V is then written

$$P(V) = \int_V d\mathbf{x}\, p(\mathbf{x}), \tag{2.26}$$

and the normalisation of the probability requires

$$\int_V d\mathbf{x}\, p(\mathbf{x}) = 1. \tag{2.27}$$

Furthermore,

$$0 \le p(\mathbf{x}) < \infty. \tag{2.28}$$

In analogy with Eq. (2.5), we can now define the entropy S as

$$S[p] = \int d\mathbf{x}\, p(\mathbf{x}) \ln \frac{1}{p(\mathbf{x})}. \tag{2.29}$$

Like in the discrete case, we can also define a *Kullback information* or *relative information*, assuming an *a priori* probability density $p^{(0)}(\mathbf{x})$,

$$K\left[p^{(0)}; p\right] = \int d\mathbf{x}\, p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{p^{(0)}(\mathbf{x})} \ . \tag{2.30}$$

It is assumed that $p(\mathbf{x}) = 0$ whenever $p^{(0)}(\mathbf{x}) = 0$, and we define $0 \cdot \ln(0/0) = 0$. One can show that this Kullback information is non-negative, as in the discrete case. This does not hold for the entropy in Eq. (2.29), though.

[5] The choice of the density of states function $v(\mathbf{x})$ affects the normalisation of the probability density, $\int d\mathbf{x}\, v(\mathbf{x}) p(\mathbf{x}) = 1$, and one needs to be careful if one makes variable transformations of the state-space variable.

In later Chapters, we shall use a continuous state-space in the analysis of, for example, spatial structure in chemical systems, in which we represent the spatial concentration distributions of chemical components by probability densities.

2.4 Exercises

2.1 Suppose you are to measure the voltage between two points in a circuit, and that you know the value to be anywhere in the range from 0 to 0.5 V (no value is more likely than any other). At your disposal is a Volt meter, showing two significant digits in that range. What is your Kullback information when you have read the result?

2.2 What is the Kullback information between two Gaussian distributions, with widths b_1 and b_2, respectively? Show that this quantity is always non-negative.

2.3 For a mechanical component to fit in a certain technical system, it is required that its length x fulfils $L - d < x < L + d$. In the production of these components the resulting lengths are normal (Gaussian) distributed, with an average of L and a standard deviation of $\sigma = d/2$. If we take a new (not yet tested) component and find that it does not fulfil the requirements, what is the Kullback information?

The normal distribution has the density function $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(x-L)^2}{2\sigma^2})$.

2.4 **A two-step observation.** The expected information gain you get when observing the result of throwing a six-sided dice, does not change if you make the observation in two steps. If you first learn whether the result is odd or even, and after that learn the exact outcome, the expectation value of the sum of these two information gains is the same as the expected information gain of the direct observation. Show that this results does **not** depend on the assumption that all outcomes are equally probable, i.e., that this holds for any probability distribution over the outcomes of the dice.

2.5 A system is described by a probability distribution over three states, characterised by the energies 0, 1, and 2, respectively. If the expectation value of the energy is 1, what probability distribution should we assign according to the maximum entropy principle?

2.6 Consider a radioactive atom with decay constant λ, so that the probability for the atom to remain after time t is $e^{-\lambda t}$. At time 0 the atom has not decayed. After a certain time $t = 1/\lambda$ the atom is observed again, and it is found to remain in its original state. What is the Kullback information from that observation?

Now we wait until $t = 2/\lambda$ and observe the atom again. This time it has decayed. What is our information gain in this observation?

2.7 Consider again, as in the previous problem, a radioactive atom with decay constant λ. How should we plan to time our measurements if we would like to maximise our expected information gain per observation?

2.8 A system is composed by n subsystems that each has two possible states, $k_j = 0, 1$ (j=1,…, n), which gives 2^n states for the whole system. Let $K = \sum_j k_j$, and assume that we know that the average of this function is $<K> = na$, where $0 < a < 1$. What is a microstate in this system? What probability distribution over microstates should we choose to describe the system?

2.9 What is the entropy of a Gaussian distribution with a width (standard deviation) b ? How can this result be interpreted?

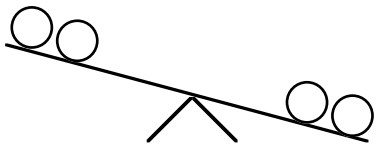
2.10 **Monty Hall Problem.** Behind 1 of 3 closed doors is a prize. You pick one of the doors. Monty opens one of the other doors behind which he knows there is no prize. You are given the choice of sticking with your choice or switching to the other unopened door. Should you switch? What are your chances of winning if you do?

Analyse this in information-theoretic terms. What is your initial uncertainty? How much information do you get when Monty opens a door? How does your probabilistic description change? What is the Kullback information between your a priori knowledge and the situation after Monty has opened the door?

2.11 **Balance information.** Suppose that you have four balls that all look the same, two of equal heavier weight and two of equal lighter weight. Assuming this knowledge, what is the uncertainty of the system. How many measurements using a balance would theoretically be needed in order to sort out which are the heavier and which are the lighter ones? Is there a procedure that accomplishes this?

2.12 **Weighing balls.** You have 12 balls, all have the same weight except one that deviates. By using a balance three times you should be able to find the deviating ball, and tell whether it is lighter or heavier. The task is to find this procedure.
To get started, you may consider the more easy problem, in which you have nine balls, of which you know that one is heavier but the others identical. Show how to find the heavy one in two measurements using a balance!

Example of balance; you may have any number of balls on each side, and you may assume the weighing to be perfect (equal weights on both sides gives a balanced result).



2.13 a) Use an information-theoretic approach to show that this is not always possible for 14 balls.
b) Show that this is not always possible for 13 balls.
c) **Tricky:** Construct a procedure that finds one ball out of 39 in three balance measurements

2.14 **Mathematical requirements for entropy.** The quantity S[P] that we have used for entropy of a distribution P = $\{p_1, p_2, \ldots, p_n\}$ over microstates (1, 2, …, n) is the only quantity fulfilling the following four conditions: (i) S is symmetric with respect to the probabilities, (ii) S is a continuous function of the probabilities, (iii) The information obtained when one gets to know the outcome of two equally probable events is 1 bit, and finally (iv):
The expected gain of information is the same for (I) an immediate observation of the microstate as for (II) a two step observation in which one distinguishes between, say, state 1 and state 2 only if a first observation rejects the other states.

Express the last condition (iv) in mathematical terms, i.e., express the entropy S as a sum of two entropies from two measurements as described in (II). Show that this expression holds for the Shannon definition of entropy: $S = \Sigma\, p_k \log(1/p_k)$.

2.15 **Information loss by aggregation.** Consider a picture composed of black and white dots, or pixels, in a two-dimensional square lattice of size NxN. At the finest resolution, we see exactly where are the pixels and we have "full" information. Assume that the resolution is made worse (by a factor of two in length scale) so that instead of seeing single pixels, we can only distinguish (N/2)x(N/2) cells, where each cell contains the aggregated information from 2x2 = 4 underlying pixels. At the aggregated level, the cells can take the (observed) "grey-scale" values 0, 1, 2, 3, or 4, corresponding to the number of black pixels that the cell was formed from.

This type of aggregation leads to a loss of information. The loss may vary (locally) depending on the local structure (local densities). Discuss briefly how this works.

If one assumes that the original picture is completely "random", with equal probabilities for black and white pixels, how much information (entropy) do we loose in average (per original pixel) when we aggregate?

2.16 **Maximum entropy of particle velocities.** Consider a system of uncorrelated particles moving on a one-dimensional lattice. Each particle can have velocity –2, –1, 0, +1, or +2. The only knowledge you have is that the average velocity is 0 and that the average square velocity is 1, i.e., $< v > = 0$ and $< v^2 > = 1$. Use the maximum entropy principle to determine the probability distribution over the different velocities.

3 Information theory for lattice systems

Information theory was developed for the application to signals or sequences of symbols. Suppose that we want to examine the information content in a symbol sequence, composed of the characters "0" and "1". The information carried per symbol *a priori* is 1 bit. This information quantity can be decomposed in two terms, an entropy term that quantifies the disorder of the system and another information quantity that quantifies the order in the system. The ordered information is usually called the *redundancy* (superfluous information) of the text or symbol sequence. Order here may, for example, mean that there is a higher chance to make a correct guess of the next character, if we may take into account preceding characters. This order depends on correlations between symbols in the sequence. The disordered information is the *entropy* of the symbol sequence, and it quantifies the uncertainty that remains (in average), when all correlations have been taken into account before observing the next character. Thus this is an average uncertainty (or "lack-of knowledge") measure per character in the sequence. It is in the entropy part of the text where we can transmit information between author and reader.

Analysis of written texts has for a long time been a popular application of information-theoretic concepts, not the least because of the possibilities to automatically generate texts. Randomly generated texts, based on correlation statistics, were present already in Shannon's original paper from 1948. As we have discussed above, the redundancy in a symbol sequence corresponds to the ordered part of the information content. Here, "order" means that certain characters are more common than others, or that some characters more often follow certain sequences of characters. This is especially the case when there are strong correlations in the text. The disordered information, the entropy, corresponds to the remaining uncertainty, when one is guessing the next character in the text, taking into account all correlations. A language with 32 characters (including "space" and some punctuation marks, but without using capital letters) has a maximum entropy value (when correlations are not included) per character that is log 32, or 5 bits, but the real entropy is usually much lower. In English about 4 bits are redundant, and only 1 bit is entropy.

How correlations contribute to the structure of the text can be illustrated by random generation of texts, using different correlation lengths. First one needs a large text (or a number of texts), from which statistics is collected on sequences of characters. The statistics from about a thousand lines of poetry may be sufficient to make the computer seem like an author. If the source for the statistics is only one text written by one author the generated text will not only show the characteristics of the language used, but also the characteristics of the author, as well as features of the story chosen.

Based on statistics from a source text, we form conditional probabilities, $p(x_n \mid x_1,\ldots,x_{n-1})$, expressing the probability for the next character given the $n-1$ preceding ones. Using these probabilities, we randomly generate characters, one by one. The longer correlation one decides to include (larger n), the longer preceding sequence is taken into account when calculating the probability. In the following six examples we go from a generated text using only density information to texts with correlations over block lengths two to six:

1: Tdory d neAeeeko,hs wieadad ittid eIa c i lodhign un a a svmb i ee' kwrdmn.
2: Le hoin. whan theoaromies out thengachilathedrid be we frergied ate k y wee ' e the sle! se at te thenegeplid whe tly titou hinyougea g l fo nd
3: 'Weed. Thed to dre you and a dennie. A le men eark yous, the sle nown ithe haved saindy. If - it to to it dre to gre. I wall much. 'Give th pal yould the it going, youldn't thave away, justove mouble so goink steace, 'If take we're do mennie.
4: I can light,' George tried in you and fire.' 'Nothen it and I want yourse, George some other ther. There's if his hand rolledad ther hisky, 'I little amonely we're we're with him the rain.
5: 'I...I'm not running.' The ranch, work on the time. Do you because you get somethings spready told you just him by heat to coloured rabbits. That's going grew it's like a whisky, place.
6: Million mice because it two men we'll sit by the future. We'll steal it. 'Aren't got it. 'About the fire slowly hand. 'I want, George,' he asked nervoulsly: 'That's fine. Say it too hard forget other.

Already in the case with correlation of two or three, it is quite clear what the language is. When correlation is increased, more of the words are correctly generated, and at the length of six, one may guess from what story the text is taken.

In the following sections, we present a formalism that can be used to analyse disorder and correlations in symbol sequences. We also extend the formalism to two dimensions. In later Chapters, we use this formalism to analyse states (in the form of symbol sequences) in the time evolution of discrete dynamical systems (cellular automata). In the Chapter on chaotic systems, we illustrate how the formalism can be applied to symbol sequences generated by dynamical systems in order to characterise the dynamics.

3.1 One-dimensional lattices

Like in the previous Chapter, our definitions on information quantities in symbol sequences depend on a probability description of the system. Let us assume that the systems under study consists of infinite sequences of symbols, in which each symbol is taken from some finite alphabet Λ. Actually, the system is a set of such sequences, an ensemble. We also assume that the system is *translation invariant*, so that the probability for finding a certain finite sequence of symbols at a certain position in the system does not depend on the position, but only on what other symbols that we may already have observed. In this type of system, we may form probability distributions over sets of finite sub-sequences of symbols, one for each length n (with $n = 1, 2, 3, \ldots$).

We consider symbol sequences that are generated by a *stationary* stochastic process, which implies the translation invariance of the sequence. The system can then be described by an infinite sequence of stochastic variables X_k,

$$X = (X_0, X_1, X_2, \ldots), \tag{3.1}$$

and since it is stationary, probabilities for certain sub-sequences do not depend on position,

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_{1+m} = x_1, X_{2+m} = x_2, \ldots, X_{n+m} = x_n), \tag{3.2}$$

for all n, m, and any symbols $x_k \in \Lambda$. Therefore we can characterise our system by a probability distributions P_n over symbol sequences of finite length n,

$$P_n = \{p_n(x_1,\ldots,x_n)\}_{x_1,\ldots,x_n \in \Lambda^n} \quad (n = 1, 2, \ldots), \tag{3.3}$$

or shorter

$$P_n = \{p_n(\sigma_n)\}_{\sigma_n \in \Lambda^n} \quad (n = 1, 2, \ldots). \tag{3.4}$$

(We are using x, y, z etc as variables for single symbols and Greek letters σ, α, β etc as variables for sequences of symbols.) The probabilities fulfil the general requirements of being non-negative and normalised, Eq. (2.3) and (2.4). Furthermore, there are conditions that relate probability distributions over lengths n and $n + 1$, based on the fact that the distribution over (n+1)-length sequences includes the distribution over n-length sequences. Summation over first or last variable in an (n+1)-length probability results in the corresponding n-length probability,

$$p_n(x_1,\ldots,x_n) = \sum_{x_{n+1} \in \Lambda} p_{n+1}(x_1,\ldots,x_n,x_{n+1}) \text{ , and} \tag{3.5}$$

$$p_n(x_1,\ldots,x_n) = \sum_{x_0 \in \Lambda} p_{n+1}(x_0,x_1,\ldots,x_n) \text{ .} \tag{3.6}$$

In the following, we will drop the subscript n of p and in short write $p(x_1\ldots x_n)$ or $p(\sigma_n)$, if it is clear which length that is considered.

We are usually assuming that the stochastic process that generates the ensemble of sequences is *ergodic*. This means that, almost always, we can get the right statistics from a single (infinitely long) sequence of symbols in the ensemble. In other words: the calculation of an average $<f>$ of a function f based on the probabilities $p(x_1\ldots x_n)$ results in the same as an average based on following the individual symbol sequence (s_1, s_2, s_3,…),

$$\sum_{x_1\ldots x_n \in \Lambda^n} p(x_1,\ldots,x_n) f(x_1,\ldots,x_n) = \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} f(s_k, s_{k+1},\ldots,s_{k+n}). \tag{3.7}$$

The implication of this is that we may consider one outcome of the stochastic process, a specific infinite symbols sequence, and that we may make an internal statistical analysis of it in order to determine a number of information-theoretic properties. Viewed as a physical system, a specific symbol sequence would correspond to a specific microstate. This would then open the possibility of discussing entropy and order of a single microstate. In statistical physics, this is not the common perspective since one associates the system with the macrostate, or the probability distribution over the possible microstates. We shall return to this view in the Chapter on physics and information theory.

Let us assume that we have an infinite sequence of symbols (or an ensemble of such sequences), characterised by probabilities for finite length sub-sequences. Our *a priori* knowledge of the system only reflects that each symbol belongs to a certain alphabet Λ, that contains $|\Lambda| = \nu$ different characters. Our initial uncertainty (or lack of knowledge) per symbol is then $S = \log \nu$. By successively adding probability distributions for sequences of increasing length P_n (n=1, 2, …), we may take correlations into account to reduce our uncertainty of the next symbol in the sequence. The entropy that still may remain when we include all lengths ($n \to \infty$) is the *Shannon entropy* of the symbol sequence, or for short the *entropy*. In some contexts this is called the *measure entropy* or the *entropy rate* of the stochastic process (to be discussed in the Chapter on chaotic systems).

Before we formalise this discussion, let us consider the following example involving two, possibly dependent, stochastic variables X_1 and X_2. Let us assume that the possible outcomes of these are characters belonging to a certain alphabet Λ. First, if we assume that they are independent and equally distributed, characterised by probabilities $p(x)$ each, then, according to the example under 2.1.2, the total entropy is

$$S[X_1, X_2] = S[X_1] + S[X_2]. \tag{3.8}$$

If there is a correlation (dependence) between X_1 and X_2, then their combined entropy $S[X_1,X_2]$ should be less than $S[X_1] + S[X_2]$. Therefore, we can use the following difference as a measure of the dependence between the variables — the *mutual information* $I[X_1; X_2]$,

$$\begin{aligned} I[X_1;X_2] &= S[X_1] + S[X_2] - S[X_1,X_2] = \\ &= \sum_{x_1} p(x_1)\log\frac{1}{p(x_1)} + \sum_{x_2} p(x_2)\log\frac{1}{p(x_2)} - \sum_{x_1x_2} p(x_1x_2)\log\frac{1}{p(x_1x_2)} = \\ &= \sum_{x_1x_2} p(x_1x_2)\log\frac{p(x_1x_2)}{p(x_1)p(x_2)} = K[P(X_1)P(X_2);P(X_1X_2)] \geq 0 \end{aligned} \tag{3.9}$$

Here we have made use of Eq. (3.5) and (3.6), when transforming the expression into a Kullback information. Thus, the mutual information is the information we get when we replace the separate distributions $P(X_1)$ and $P(X_2)$, as the description of the system, with the correct joint distribution $P(X_1\, X_2)$.

It is often useful to consider the conditional probability $p(x_2|x_1)$ which is the probability for a certain x_2 provided that a certain x_1 has been observed, defined by

$$p(x_2 \mid x_1) = \frac{p(x_1 \text{ and } x_2)}{p(x_1)} = \frac{p(x_1,x_2)}{p(x_1)}. \tag{3.10}$$

Using this, we can rewrite the mutual information

$$
\begin{aligned}
I[X_1;X_2] &= \sum_{x_1x_2} p(x_1)\frac{p(x_1x_2)}{p(x_1)}\log\frac{p(x_1x_2)}{p(x_1)p(x_2)} = \\
&= \sum_{x_1x_2} p(x_1)p(x_2\mid x_1)\log\frac{p(x_2\mid x_1)}{p(x_2)} = \\
&= \sum_{x_1} p(x_1)K[P(X_2);P(X_2\mid x_1)].
\end{aligned}
\tag{3.11}
$$

The Kullback information in the last row quantifies the information gained when we replace the probability description of X_2 in the form of the distribution $P(X_2)$ with the conditional probability distribution that includes the possible correlation that may exist when we have already observed a specific outcome x_1 of X_1. Then the average over the possible outcomes of X_1 is calculated. This is also an intuitively reasonable interpretation of the mutual information quantity.

The analysis of correlations in symbol sequences builds on how the entropy of n-length distributions varies with the length n. The *block entropy*,

$$
S_n = S[P_n] = \sum_{\sigma_n} p(\sigma_n)\log\frac{1}{p(\sigma_n)}\ , \tag{3.12}
$$

quantifies the disorder of n-length sub-sequences of the system. This quantity grows with n, but if there are correlations, the increase is less than the entropy of an isolated symbol. The larger n, the longer correlations can be taken into account, and thus the increase of S_n will be smaller for larger n. A qualitative picture of how S_n varies as a function of n is shown in Figure 3.1.

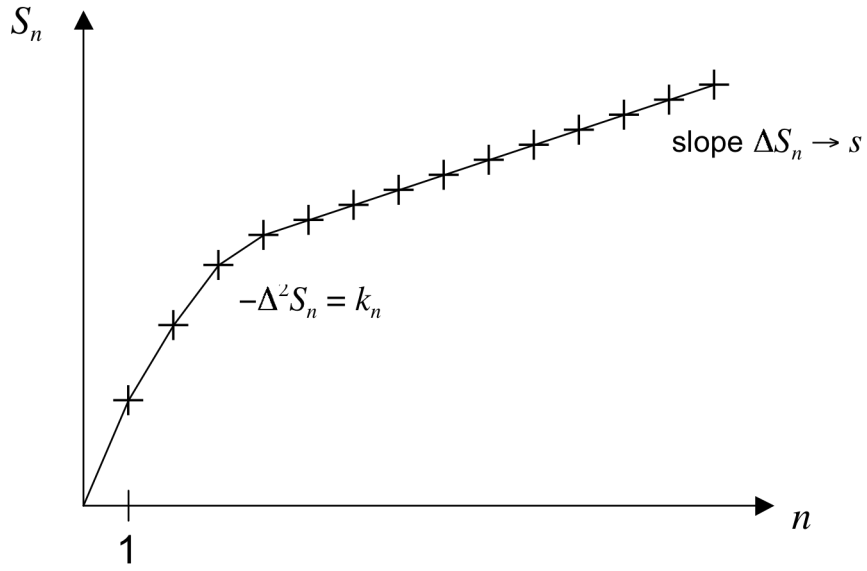


Figure 3.1. The block entropy S_n is an increasing function of length n. The increase is decreasing, ΔS_n decreases by n, implying $\Delta^2 S_n \leq 0$. This second difference in the block entropy can be interpreted as the information in correlations k_n over block length n.

In order to analyse the information contained in the correlations in the system, we introduce the conditional probability of a symbol, given that we have already observed the n–1 preceding symbols,

$$p(x_n \mid x_1 \ldots x_{n-1}) = \frac{p(x_1 \ldots x_{n-1} x_n)}{p(x_1 \ldots x_{n-1})} \ . \tag{3.13}$$

For each possible preceding sequence $(x_1 \ldots x_{n-1})$, this is a probability distribution $P(\bullet|x_1 \ldots x_{n-1})$ over the next symbol x_n. The entropy for this distribution is a measure of the difficulty in guessing the next symbol (uncertainty of the next symbol), based on the fact that we already know the preceding n–1 symbols. Of course, this entropy may vary with the exact preceding symbol sequence, and therefore we are interested in the average of this conditional entropy

$$\begin{aligned} \left\langle S[P(\bullet \mid x_1 \ldots x_{n-1})] \right\rangle &= \sum_{x_1 \ldots x_{n-1}} p(x_1 \ldots x_{n-1}) \sum_{x_n} p(x_n \mid x_1 \ldots x_{n-1}) \log \frac{1}{p(x_n \mid x_1 \ldots x_{n-1})} = \\ &= S_n - S_{n-1} = \Delta S_n \ . \end{aligned} \tag{3.14}$$

This shows that the average conditional entropy equals the slope of the block entropy S_n as a function of length n (see Figure 3.1). If we increase the length of the preceding symbol sequence in the conditional probability distribution, we may increase our chances to make a better estimate of the probability for the next symbol. This is due to the fact that increasing the length of the block includes more correlations in the system, and these correlations can be used when guessing the next symbol. The information content in such correlations can be derived as follows.

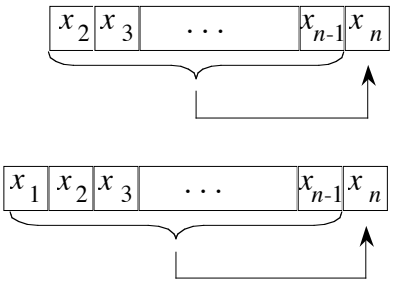


Figure 3.2. Extending the preceding block of symbols from $(x_2 \ldots x_{n-1})$ to $(x_1 x_2 \ldots x_{n-1})$ allows for longer correlations to be used in the guess of next symbol x_n.

To quantify the information in correlations of length n, suppose that we have an *a priori* conditional distribution $P^{(0)}(\bullet|x_2 \ldots x_{n-1})$ for a symbol x_n, given that a specific preceding sequence $(x_2 \ldots x_{n-1})$ is known. Then we are interested in the information we get when we observe the symbol x_1 and use that to change our probability description of the symbol x_n to $P(\bullet|x_1 x_2 \ldots x_{n-1})$, see Figure 3.2. The conditional probabilities in $P^{(0)}$ can not include any correlations of length n (stretching over a sequence of length n), but that is possible in the new distribution P. The Kullback information between $P^{(0)}$ and P is then a measure of the

correlation information of length n when a specific preceding sequence $(x_1x_2\ldots x_{n-1})$ is observed,

$$K[P^{(0)};P] = \sum_{x_n} p(x_n \mid x_1x_2\ldots x_{n-1}) \log \frac{p(x_n \mid x_1x_2\ldots x_{n-1})}{p(x_n \mid x_2\ldots x_{n-1})} \ . \tag{3.15}$$

If we now take the average over all possible preceding sequences $(x_1x_2\ldots x_{n-1})$, we get an expression for the average information content k_n in correlations of length n. This quantity can be rewritten in the form of a Kullback information, *the correlation information from length n*

$$\begin{aligned} k_n &= \sum_{x_1\ldots x_{n-1}} p(x_1\ldots x_{n-1}) K[P^{(0)};P] = \\ &= \sum_{x_1\ldots x_{n-1}} p(x_1\ldots x_{n-1}) \sum_{x_n} \frac{p(x_1\ldots x_{n-1}x_n)}{p(x_1\ldots x_{n-1})} \log \frac{p(x_1\ldots x_{n-1}x_n)p(x_2\ldots x_{n-1})}{p(x_1\ldots x_{n-1})p(x_2\ldots x_{n-1}x_n)} = \\ &= \sum_{x_1\ldots x_n} p(x_1\ldots x_n) \log \frac{p(x_1\ldots x_n)}{\tilde{p}(x_1\ldots x_n)} = K[\tilde{P}_n;P_n] \, . \end{aligned} \tag{3.16}$$

Here we have introduced an *a priori* distribution $\tilde{P}_n$ that is an estimate of P_n based only on $(n-1)$-length distributions, with probabilities

$$\tilde{p}(x_1\ldots x_n) = \frac{p(x_1\ldots x_{n-1})p(x_2\ldots x_n)}{p(x_2\ldots x_{n-1})} \ , \text{ for } n > 2, \text{ and} \tag{3.17}$$

$$\tilde{p}(x_1x_2) = p(x_1)p(x_2) \ , \text{ when } n = 2. \tag{3.18}$$

One can show, see exercise 4.1, that this is the maximum entropy probability distribution for n-length sequences that coincides with the $(n-1)$-length probabilities when summation over first or last symbol is applied, according to Eq. (3.5) and (3.6). The distributions $\tilde{P}_n$ and P_n differ in that the first one does not include any correlations of length n, which again is an argument for the interpretation of k_n as an measure of correlation information of length n. We can rewrite the correlation information by using (3.16) and decomposing the logarithmic term into a sum of four terms, which results in

$$k_n = -S_n + 2S_{n-1} - S_{n-2} = -\Delta S_n + \Delta S_{n-1} = -\Delta^2 S_n \quad (n = 2,3,\ldots) \ . \tag{3.19}$$

Here we define $S_0 = 0$, for Eq. (3.19) to be consistent with Eq. (3.16)-(3.18). Since k_n is a Kullback information, we know that $-\Delta^2 S_n \geq 0$, as was seen in Figure 3.1 showing the block entropy S_n as a function of block length n. Let us also introduce an information quantity that measures the difference in character frequency from a uniform distribution. As an *a priori* distribution we use the completely "uninformed" uniform distribution $P_1^{(0)}$ that assigns equal probabilities $p_1^{(0)}(x_1) = 1/\nu$, to all characters x_1 in Λ. The *density information* k_1 can then be written as a Kullback information between the *a priori* uniform distribution and the observed single character distribution P_1,

$$k_1 = K[P_1^{(0)};P_1] = \sum_{x_1} p(x_1)\log\frac{p(x_1)}{1/\nu} = \log\nu - S_1 \ . \tag{3.20}$$

We have thus defined a number of information quantities that captures the *ordered* information in the system — the density information k_1 and the series of correlation information contributions k_n (n = 2, 3, …). Let us combine all these into an information quantity, *the correlation information* k_{corr},

$$\begin{aligned} k_{\text{corr}} = \sum_{m=1}^{\infty} k_m = (\log\nu - S_1) + (-S_0 + 2S_1 - S_2) + \\ +(-S_1 + 2S_2 - S_3) + \\ +(-S_2 + 2S_3 - S_4) + \\ \dots \\ +(-S_{m-2} + 2S_{m-1} - S_m) + \\ \dots \quad = \\ = \log\nu - \lim_{m\to\infty}(S_{m+1} - S_m) = \log\nu - \Delta S_\infty \ . \end{aligned} \tag{3.21}$$

This is the ordered part, *the redundancy*, of the information in the system, expressed as an average per symbol. The entropy per symbol of the system, the Shannon entropy s, is a measure of the uncertainty that remains when all correlations, including deviation in symbol frequencies from uniformity, have been taken into account. One way to define this quantity, the remaining uncertainty, is to use the entropy of the conditional probability for the next character, given that we already have observed a preceding sequence of symbols. Then we take the average of this entropy over the possible preceding sequences, and we take the infinite limit of the length of the preceding sequence in order to include all possible correlations in the conditional entropy. The Shannon entropy s is then defined, according to Eq. (3.14),

$$s = \lim_{n\to\infty} \sum_{x_1\dots x_{n-1}} p(x_1\dots x_{n-1}) \sum_{x_n} p(x_n \mid x_1\dots x_{n-1}) \log\frac{1}{p(x_n \mid x_1\dots x_{n-1})} = \lim_{n\to\infty}\Delta S_n = \Delta S_\infty \tag{3.22}$$

From this we can conclude, that the total entropy per symbol of log ν, which is the maxmimum entropy per symbol, can be decomposed in two terms, the redundancy k_{corr} and the Shannon entropy s,

$$S_{\max} = \log\nu = (\log\nu - \Delta S_\infty) + \Delta S_\infty = k_{\text{corr}} + s \ . \tag{3.23}$$

Note also that the entropy can be expressed as the limit value of the infinite block entropy per symbol (proof left as an exercise),

$$s = \lim_{m\to\infty}\frac{1}{m} S_m \ . \tag{3.24}$$

The interpretation of these quantities in terms of coding theory is the following. If a text contains redundancy this can be removed in order to compress the text by using a coding procedure, replacing characters and sequences in an appropriate way. In the case of an optimal code, the compressed sequence may have a length of $Ls/\log \nu$ if the original length is L.

3.2 Markov processes and hidden Markov models

3.2.1 Markov processes and entropy

Some of the symbol sequences discussed in the course are the direct result of a simple Markov processes, i.e., processes in which the probability for the next state (or symbol) is fully determined by the preceding state (or symbol). This means that the conditional probability distribution $P(\bullet \mid z_1 \ldots z_{n-1})$ over the next symbol z_n, given a preceding sequence of symbols, converges already for $n = 2$. The entropy s of the process is then given already by ΔS_2, see Eq. (3.22),

$$s = \sum_{z_1} p(z_1) \sum_{z_2} p(z_2 \mid z_1) \log \frac{1}{p(z_2 \mid z_1)} \quad . \tag{3.25}$$

A Markov process can be described as a finite automaton with internal states z_i (with $i = 1, \ldots m$, and z_i belonging to the alphabet Λ) corresponding to the symbols generated. The process changes internal states according to transition probabilities, P_{ij}, denoting the probability to move to state j from state i. An example of such an automaton is given in the figure below. The internal states (the possible symbols) are a, b, and c. Let us assume in this example that, when there is a choice for the transition from a state, all possible transitions from that state are equally probable. Here this means that $P_{aa} = P_{ac} = P_{ba} = P_{bc} = ½$, $P_{cb} = 1$, and $P_{ab} = P_{ca} = P_{cc} = P_{bb} = 0$.

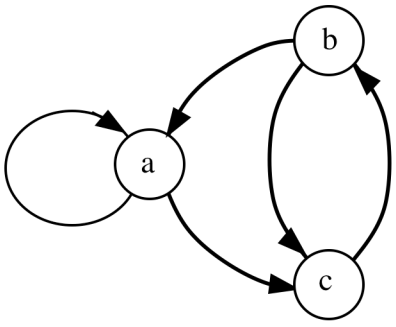


Figure 3.3. Example of finite state automaton representing a Markov process.

To calculate the entropy of the process (or the symbol sequence), we need $p(z)$, which is the probability distribution over the internal states (or, equivalently, the stationary probability distribution over the internal states). Since the (stationary) probability $p(z)$ to be in a certain state z must equal the sum over the probabilities of possible preceding states w weighted with the transition probabilities to state z, we can determine $p(z)$ by the equations

$$p(z) = \sum_w p(w)P_{wz} \quad \text{, for all } z \in \Lambda \tag{3.26}$$

together with the normalisation constraint $\Sigma_z\, p(z) = 1$. Note that the transition probabilities P_{zw} equals the conditional probabilities $p(w \mid z)$, which means that the entropy s can be written

$$s = \sum_z p(z) \sum_w P_{zw} \log \frac{1}{P_{zw}} \quad . \tag{3.27}$$

In the example above, the stationary distribution over the states is found by solving Eqs. (3.26), replacing, e.g., the last one with the normalisation constraint,

$$\begin{aligned} p(a) &= \frac{1}{2}p(a) + \frac{1}{2}p(b), \\ p(b) &= p(c), \\ p(c) &= 1 - p(a) - p(b), \end{aligned}$$

and one finds that p(a) = p(b) = p(c) = 1/3. The term $-P_{wz} \log P_{wz}$ only contributes to the entropy in Eq. (3.27) when w is a or b, so the resulting entropy then turns out to be s = 2/3 (bits). One way to view this is that the process needs one bit of "random information" in order to make the random choice of transition when leaving node a and node b. Since these nodes are visited with a fraction of 2/3 of the time and since this choice directly determines the symbols in the sequence, the average randomness, or the entropy, should be 2/3 bits.

3.2.2 An example of an optimal code exploiting correlations

Consider a stochastic process sending sequences of symbols from the alphabet Λ = {a, b, c, d}, like in example 2.1.1. Now we assume that there are correlations between neighbouring symbols. If the preceding character is observed, we have some additional knowledge about the probability for the following one.

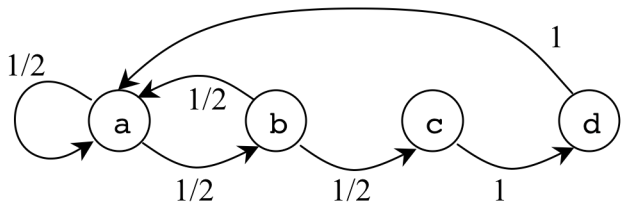


Figure 3.4. Markov process generating sequence of symbols from the alphabet {a, b, c, d}.

The correlations between the symbols are determined by the structure of the Markov process generating the symbol sequence. The finite state automaton in Figure 3.4 defines the Markov process. Solving for the stationary distribution, Eq. (3.26), results in the single symbol distribution P_1 = {1/2, 1/4, 1/8, 1/8}, as we also had in the example 2.1.1. The entropy of the process is then determined by Eq. (3.27). This implies that there is a single bit of contribution to the entropy each time the process is in any of the states a or b (since there is always 1/2 chance for choosing a specific transition), but no contribution when leaving states c or d. The

entropy is then 1/2 + 1/4 bits = 3/4 bits. This is significantly lower than the 7/4 bits that we found in example 2.1.1, and what we would also get here if we would disregard correlations, the entropy of the single symbol distribution P_1.

The implication of this is that we should be able to find an even more efficient coding, compressing the length of a binary coded message down to 3/4 bits in average per original symbol. One easy way to achieve this is to start with the code word for the initial symbol using the code of example 2.1.1. After that the code words are `0` or `1` only, for states `a` and `b`, with `0` meaning a transition to `a` and `1` the other transition. For states `c` and `d` no code word is needed as the next state is unique[6]. Then, in average for a long message, we only need one bit after state `a` and another bit after state `b`, which is only in 3/4 of all positions in the original sequence, resulting in an average code word length of 3/4.

3.2.3 Hidden Markov models and entropy

In a hidden Markov model, one does not observe the states z of the process (or of the finite state automaton), but one observes some function of the state $f(z)$. In the example above, if the function f is given by $f(\mathrm{a}) = 0$ and $f(\mathrm{b}) = f(\mathrm{c}) = 1$, then we have a process that generates a sequence of '0' and '1' symbols, which is not a Markov process but a *hidden Markov model*. The process can now be illustrated by the following automaton:

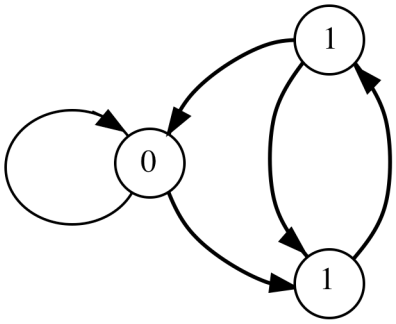


Figure 3.5. Example of finite state automaton representing a Hidden Markov model.

The process thus generates sequences of 0's and 1's in which the restriction is that blocks of 1's (separated by 0's) always are of even length. This example is discussed in the next section, but in that case we illustrate the process in a different (but equivalent) form, by associating the symbols generated by the process with the *transitions* rather than with the internal *states*. In that way one can get a slightly more compact description, see figure below.

[6] In order to make sure one knows where the end of the message is (since there is no code word following states `c` and `d`), one may encode the last symbol of the message using the reversed code words that were used for the initial symbol (so that one can read the last code word from right to left at the end).

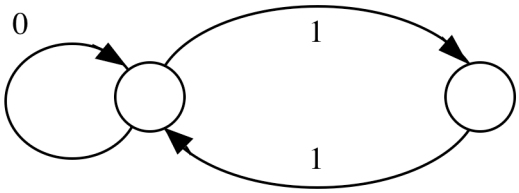


Figure 3.6. Automaton with symbols on the transition arcs gives a more compact automaton representation equivalent to the Hidden Markov model of the previous figure.

It turns out that the entropy of this process is still $s = 2/3$, even if other information-theoretic quantities is dramatically changed, to be discussed under example 3.3.4 below.

3.3 Some examples

3.3.1 Example: crystal

Consider a periodic symbol sequence of zeroes and ones,

…01010101010101010101010101010101…

The density information is zero, $k_1 = 0$, since the probabilities are equal for the two symbols, $p(0) = p(1) = 1/2$. The probabilities that are required for calculating the correlation information are

$$p(0) = p(1) = \frac{1}{2}, \quad p(0|1) = p(1|0) = 1, \quad \text{and } p(0|0) = p(1|1) = 0.$$

The average contrast form of the correlation information from length $n = 2$, Eq. (3.16), then gives

$$k_2 = \sum_{x_1} p(x_1) \sum_{x_2} p(x_2 \mid x_1) \log \frac{p(x_2 \mid x_1)}{p(x_2)} = \log 2 = 1 \text{ (bit)}.$$

Since this is the total information (per symbol) of the system, we can conclude that all other quantities are zero, $k_m = 0$ for $m \neq 2$, and $s = 0$. This is what one should expect. There is no entropy in this system — as soon as we see one symbol ("0" or "1"), we know the next symbol, and so on.

3.3.2 Example: Gas

Consider instead a symbol sequence generated by a completely random process, like coin tossing,

...11000011010001011001001101110 1...

The probability for next character to be "0" or "1" is 1/2 independent of how many preceding characters that we may observe. The entropy of the conditional probability is therefore always maximal, log 2 = 1 (bit),

$s = 1$,

and there are no contributions from the redundancy, $k_{\text{corr}} = 0$.

3.3.3 Example: Finite automaton generating short correlations

Let us now consider a symbol sequence generated by a stochastic process described by a finite automaton, see Figure 3.7.

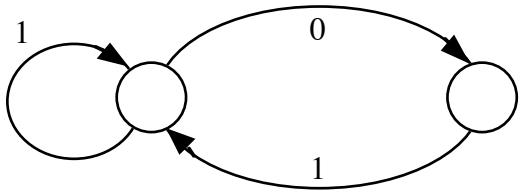


Figure 3.7. The finite automaton generates an infinite sequence of "0" and "1" by following the arcs between the nodes in the graph. When there are two arcs leaving a node, one is chosen randomly with equal probabilities for both choices.

The automaton in Figure 3.7 generates a symbol sequence where 0's cannot appear in pairs, e.g.,

```
...11010111110111101010111011011O...
```

In order to calculate the information-theoretic properties, we need to transform the characteristics of the automaton to a probabilistic description of the symbol sequence. First, we calculate the densities of 0's and 1's. Since the transition (arc) to the right node (R) always generates a "0" and is the only way a "0" can be generated the probability for being in R, $p(\text{R})$ equals the frequency of 0's, $p(0) = p(\text{R})$. Similarly, $p(1) = p(\text{L})$. The probabilities for the nodes are given by the stationary probability distribution over the nodes. Here, this can be expressed by the fact that the probability for being in the left node $p(\text{L})$ must be equal to the probability that we were in this node last step and generated a "1" plus the probability that we were in the right node last step,

$$p(\text{L}) = p(\text{L})\frac{1}{2} + p(\text{R}) \quad \Rightarrow \quad p(\text{L}) = \frac{2}{3} \text{ and } p(\text{R}) = \frac{1}{3} \ ,$$

where we have used the normalisation $p(\text{L}) + p(\text{R}) = 1$. This means that $p(0) = 1/3$ and $p(1) = 2/3$. Then the density information is

$$k_1 = \sum_{x_1} p(x_1) \log \frac{p(x_1)}{1/2} = \frac{5}{3} - \log 3 \approx 0.0817$$

All other redundant information is contained in correlations over block length $n = 2$, since if we observe one character we know which node in the automaton that is the starting point for generating the next character, and then we have the full knowledge about the true probabilities

for that character. Thus, $p(0|1) = p(1|1) = 1/2$, $p(1|0) = 1$, and $p(0|0) = 0$. The correlation information for length two is then

$$k_2 = \sum_{x_1} p(x_1) \sum_{x_2} p(x_2 \mid x_1) \log \frac{p(x_2 \mid x_1)}{p(x_2)} = \log 3 - \frac{4}{3} \approx 0.2516.$$

The Shannon entropy s is the entropy (uncertainty) that remains when we are guessing the next character in the sequence, based on our knowledge on all preceding characters. The preceding characters inform us on which node is used in generating the next character, and, actually, this information is in the last character alone. Formally, this can be expressed

$$p(1 \mid x_1 \ldots x_{n-1}1) = p(0 \mid x_1 \ldots x_{n-1}1) = \frac{1}{2}, \quad \text{for all possible } x_1 \ldots x_{n-1}1, \text{ and}$$
$$p(1 \mid x_1 \ldots x_{n-1}0) = 1, \quad \text{for all possible } x_1 \ldots x_{n-1}0 .$$

Equation (3.22) can now be used to calculate the entropy,

$$s = \lim_{m \to \infty} \Delta S_m = \Delta S_2 = \sum_{x_1} p(x_1) \sum_{x_2} p(x_2 \mid x_1) \log \frac{1}{p(x_2 \mid x_1)} =$$
$$= p(1) \cdot \log 2 + p(0) \cdot 0 = \frac{2}{3} \approx 0.6667$$

This is, of course, what we should expect, since we have already said that there is no more correlation information than what was calculated in k_1 and k_2, and then the rest of the 1 bit of information per symbol must be the entropy.

3.3.4 Example: Finite automaton generating long correlations

The finite automaton of Figure 3.7 is similar to the one of the previous example, with the difference that the arc leading from the right to the left node generates a "0" here. This means that the automaton generates sequences where 1's are separated by an even number of 0's. This is a hidden Markov model, as we discussed in the example of Section 3.2.3.

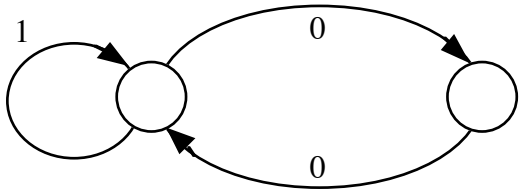


Figure 3.8. Finite automaton generating sequences of 0's and 1's in which 1's are always separated by an even number of 0's.

Suppose that we shall guess on the next character in the sequence, and that we may take into account an large (infinite) number of preceding characters, for example,

```
...0001000011100100000000110000000?
```

Then it is sufficient to go back to the closest preceding "1" and count how many 0's there are in between. If there is an even number (including zero 0's), we are in the left node, and if there is an odd number, we are in the right node. When we know which node we are in, we also have the best probability description of the next character. (Only if there are only 0's to the left, no matter how far we look, we will not be able to tell which is the node, but as the length of the preceding sequence tend to infinity the probability for this to happen tends to zero.)

Since the preceding sequence almost always determines (and corresponds to) the actual node, in the limit of infinite length, we can rewrite the entropy s as follows.

$$s = \lim_{n\to\infty} \sum_{x_1 \dots x_{n-1}} p(x_1 \dots x_{n-1}) \sum p(x_n \,|x_1 \dots x_{n-1}) \log \frac{1}{p(x_n \,|\, x_1 \dots x_{n-1})} =$$
$$= p(\mathrm{L}) \sum_x p(x \,|\mathrm{L}) \log \frac{1}{p(x \,|\mathrm{L})} + p(\mathrm{R}) \sum_x p(x \,|\mathrm{R}) \log \frac{1}{p(x \,|\mathrm{R})} =$$
$$= \frac{2}{3} \cdot \log 2 + \frac{1}{3} \cdot 0 = \frac{2}{3} \approx 0.6667$$

We have used the probabilities for the nodes L and R from the previous example, since they are the same. Expressed in this way, it is clear that the entropy of the symbol sequence comes from the random choice the automaton has to make in the left node. The right node does not generate any randomness or entropy. Therefore, we get the same entropy as in the previous example. Also the character frequencies are the same resulting in the same density information k_1, but when we get to the correlation information over blocks, we get a difference. In this case, we have correlation information in arbitrarily large blocks, since any number of consecutive 0's may occur, and then there is more information to be gained by observing one more character. To calculate the expressions for the different correlation information terms is now more complicated, and we leave it as a difficult exercise to show that

$$k_{2n-1} = \frac{1}{3 \cdot 2^n}(9 \log 3 - 14) \ , \ n = 1, 2, 3, \dots \text{, and}$$
$$k_{2n} = \frac{1}{3 \cdot 2^{n-1}}(5 - 3 \log 3), \ n = 1, 2, 3, \dots$$

So, even if the last two discussed examples are identical in terms of entropy and redundancy, this second example may be considered more complex, since the correlation information is spread out on larger distances. In a later section, we shall se that such a difference may be used as one way to characterise the complexity of symbol sequences.

3.4 Measuring complexity

What characterises a complex symbol sequence or a complex pattern? There are a large number of suggestions on how one should quantify complexity. What quantity to use depends on what one is looking for in the system under study. In this section we will focus on quantities related to how correlation information is distributed in the system. This approach

was suggested by Peter Grassberger in the 1980's, see (Grassberger, 1986). Consider two of the examples discussed above, the completely ordered ("crystal") sequence and the completely random ("gas") sequence:

```
...0101010101010101010101010101...
...1100001101000101100100110111101...
```

In the first case, the entropy is minimal (0 bit), and in the second case it is at maximum (1 bit). None of these are typically considered as complex, even though, in some contexts[7], the random one has been called complex. The precise configuration of the gas needs a lot of information to be specified, but the ensemble or the system that generate that type of sequence is simple: "toss a coin for ever…". The crystal has a simple description as it is: "repeat 01…".

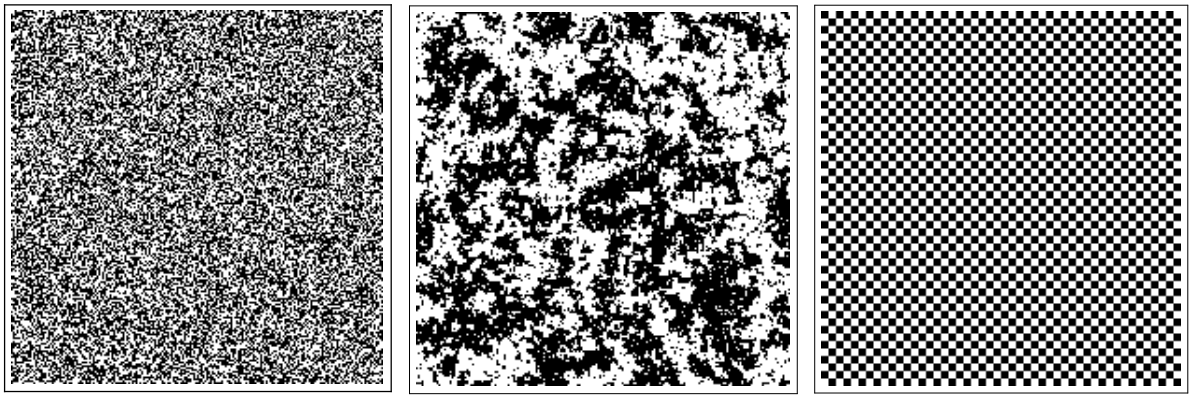


Figure 3.9. One is usually not considering a fully random pattern (left) as complex. Neither is a fully ordered structure exemplified by the checkerboard patter (right) considered as complex. The potential to exhibit complex characteristics usually requires patterns in between these extremes, patterns that may contain some order in correlations, but where correlations may reach over larger disctances, and where there is also some randomness as well. Here this is illustrated by a typical state in a physical spin system close to the critical temperature (middle).

If neither the most ordered nor the most random systems are considered complex, we should look for a measure of complexity that can take high values when the entropy is somewhere in between, see Figure ~~3.8~~. One information-theoretic characteristic that has been considered as an important component for a complex system is to what extent correlation information is spread out in the system. If there is a lot of information in long-range correlation, it may be more difficult to analyse and describe the system. Grassberger suggested a quantity, *the effective measure complexity*, which can be defined as a weighted sum of information contributions from different block lengths,

$$\eta = \sum_{m=1}^{\infty} (m-1)\, k_m \tag{3.28}$$

We call this *the correlation complexity*, and it can be rewritten (if $k_{\text{corr}} > 0$) as

[7] In algorithmic information theory, which we shall discuss in a later Chapter, the random sequence has the largest "algorithmic complexity".

$$\eta = k_{\text{corr}} \sum_{m=1}^{\infty} (m-1) \frac{k_m}{k_{\text{corr}}} = k_{\text{corr}} \, \overline{(m-1)} = k_{\text{corr}} \, d_{\text{corr}} \, . \tag{3.29}$$

Here we have introduced an average correlation distance d_{corr}. This is based on the average block length at which correlation information is found, but where correlation distance is defined to be one less than the block length (or the "distance" between first and last symbol of the block). This complexity measure is non-negative, but it is not unbounded. For the "crystal" example $\eta = 1$ and for the totally random sequence $\eta = 0$.

It also turns out that the correlation complexity can be interpreted as the average information contained in a semi-infinite symbol sequence $(\ldots x_{-2}, x_{-1}, x_0)$ has about its continuation $(x_1, x_2, x_3, \ldots)$. This quantity can be written as a Kullback information. Suppose first that the preceding sequence, that we may observe, has a finite length m and denote it $\sigma_m = (x_{-m+1}, \ldots, x_0)$. The continuation, has finite length n and we denote it $\tau_n = (x_1, \ldots, x_n)$. At the end we shall look at infinite limits of m and n. The *a priori* description of the continuation is given by probabilities $p(\tau_n)$, but after we have observed the preceding sequence σ_m, we can replace that with the conditional probabilities $p(\tau_n \,|\, \sigma_m)$. The information we gain by this is a Kullback information, and we take the average over all possible preceding sequences together with the limit of infinite lengths,

$$\eta = \lim_{m \to \infty} \lim_{n \to \infty} \sum_{\sigma_m} p(\sigma_m) \sum_{\tau_n} p(\tau_n \,|\, \sigma_m) \log \frac{p(\tau_n \,|\, \sigma_m)}{p(\tau_n)} \geq 0 \, . \tag{3.30}$$

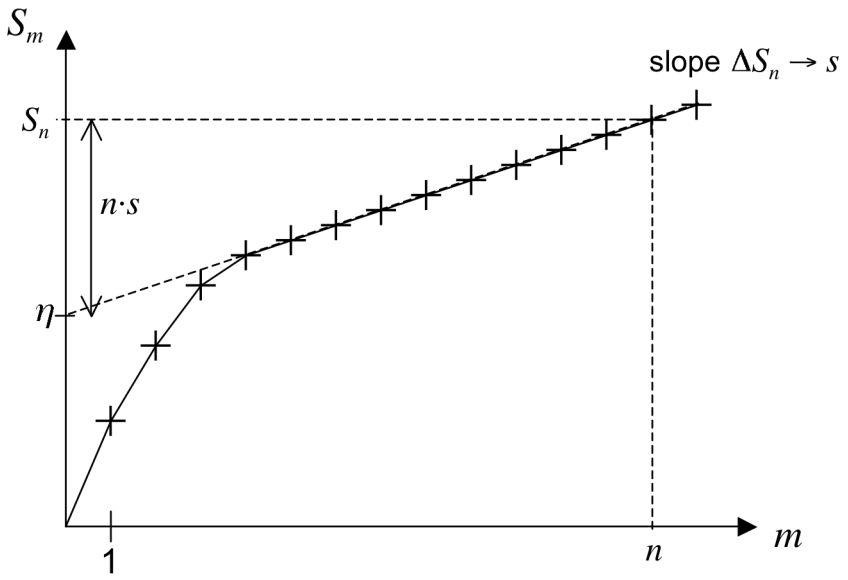


Figure 3.10. The block entropy S_n an asymptotic line with slope equal to the entropy s. Also the ratio S_n/n converges to to the entropy, but more slowly. Therefore, in the limit of infinite sequences, $S_n - n{\cdot}s$ is a measure of how fast S_n/n converges to to the entropy s. This is the correlation complexity η, and graphically, it is the intersection point on the vertical axis of the asymptotic line for the block entropy.

One can also relate the correlation complexity to the block entropies, see Figure 3.10,

$$\eta = \lim_{m\to\infty}(S_m - m\,s), \tag{3.31}$$

where s is the Shannon entropy of the symbol sequence (or more correctly of the stochastic process). The correlation complexity can thus be interpreted as the rate of convergence of S_m/m to s. The proof that these expressions are equal to the correlation complexity of Eq. (3.28) is left as an exercise.

3.4.1 Correlation complexity for Markov processes and hidden Markov models

For a Markov process the correlation complexity η is easy to calculate since there is no correlation information from blocks longer than 2, and therefore we get that $\eta = k_2$. For a hidden Markov model, the situation is different. As we saw in example 3.3.4, such a system may have correlation information in arbitrarily long blocks of symbols. In some situations there may be an easy way to calculate the correlation complexity using the form that expresses the information contained in the past about the future of an infinite symbol sequence, Eq. (3.30). Consider again the automaton describing the process of example 3.3.4, see Figure 3.11.

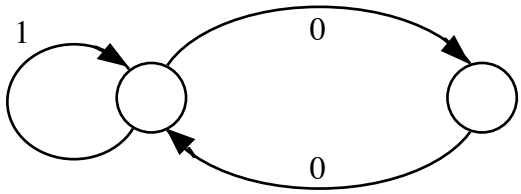


Figure 3.11. Example of hidden Markov model, represented by finite automaton generating sequences of 0's and 1's in which 1's are always separated by an even number of 0's. The probability for the arcs leaving the left node are both 1/2.

If almost all past sequences σ_m, in the infinite limit, determine whether we are in the left (L) or in the right (R) node, we can rewrite Eq. (3.30). The probability of the future sequence τ_n is given by the node we are in, so we can group together all σ_m leading to the left node L (and in the same way all leading to the right one, R). Then we have

$$\eta = \lim_{n\to\infty} \sum_{z\in\{L,R\}} p(z) \sum_{\tau_n} p(\tau_n \mid z) \log \frac{p(\tau_n \mid z)}{p(\tau_n)}, \tag{3.32}$$

where $p(L)$ and $p(R)$ are the stationary probabilities for the left and right node, respectively. We now use the definition of conditional probability $p(\tau_n \mid z)$ to rewrite the argument in the logarithm,

$$\frac{p(\tau_n \mid z)}{p(\tau_n)} = \frac{p(\tau_n, z)}{p(\tau_n)p(z)} = \frac{p(z \mid \tau_n)}{p(z)}.$$

We can then rewrite Eq. (3.32) as follows,

$$\begin{aligned}\eta &= \lim_{n\to\infty} \sum_{z\in\{L,R\}} p(z) \sum_{\tau_n} p(\tau_n \mid z) \log \frac{p(z \mid \tau_n)}{p(z)} = \\ &= \lim_{n\to\infty} \sum_{z\in\{L,R\}} p(z) \sum_{\tau_n} p(\tau_n \mid z) \left(\log \frac{1}{p(z)} + \log p(z \mid \tau_n) \right) = \\ &= \sum_{z\in\{L,R\}} p(z) \log \frac{1}{p(z)} - \lim_{n\to\infty} \sum_{\tau_n} p(\tau_n) \sum_{z\in\{L,R\}} p(z \mid \tau_n) \log \frac{1}{p(z \mid \tau_n)}.\end{aligned} \tag{3.33}$$

The last sum over nodes z is the entropy of which node we were in conditioned on observing the future sequence τ_n. But the starting node for the future sequence is almost always uniquely given by the sequence τ_n. Only when there are only zeroes in τ_n, we do not know, but that happens with probability 0 in the infinite limit. Therefore there is no uncertainty of z given the future τ_n, and the last entropy term is 0. Finally we get,

$$\eta = \sum_{z\in\{L,R\}} p(z) \log \frac{1}{p(z)}, \tag{3.34}$$

and we conclude that, for this situation, the correlation complexity equals the entropy of the stationary distribution of the states in the finite state automaton describing the hidden Markov model. Note, though, that this does not hold in general, but only for certain types of hidden Markov models.

3.5 Extensions to higher dimensions

The decomposition of information into entropy and contributions from different correlation lengths can be extended to lattice systems of any dimension. In this section we briefly indicate how the extension to the two-dimensional case is done, but it is easy to generalise the formalism also to higher dimensions.

Consider an infinite two-dimensional lattice in which each site is occupied by a symbol (0 or 1). (The generalisation to larger alphabets is straightforward.) Assume that the relative frequencies, with which finite configurations of symbols occur in the lattice, are well defined. Let $A_{M\times N}$ be a specific $M\times N$-block occurring with probability $p(A_{M\times N})$. Then the entropy, i.e., the average information per site, is

$$s = \lim_{M,N\to\infty} \frac{1}{MN} S_{M\times N}, \tag{3.35}$$

where the block entropy $S_{M\times N}$ is defined by

$$S_{M\times N} = \sum_{A_{M\times N}} p(A_{M\times N}) \log \frac{1}{p(A_{M\times N})}. \tag{3.36}$$

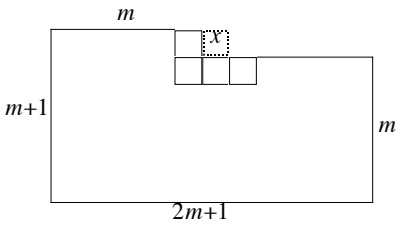


Figure 3.12. The configuration of cells B_m that is used in the conditional probability for the character x in the "next" cell, drawn with a broken line. By extending this configuration, we get better estimates of the unknown character x.

Let B_m be a certain configuration of symbols arranged as follows: m rows of symbols, each of length $2m+1$, are put on top of each other, and on the m first symbols of the top row a sequence of m symbols is placed, see Figure 3.12. We also introduce the notation $B_m x$ for the configuration that adds the symbol x to B_m after the mth symbol in the top row.

Then we can introduce the conditional probability for a certain character x given that we have already observed the characters in the configuration B_m,

$$p(x \mid B_m) = \frac{p(B_m x)}{p(B_m)}. \tag{3.37}$$

This can be interpreted as the conditional probability for the "next" character given that we have seen the "previous" $2m(m+1)$ characters. The average entropy is

$$H_m = \sum_{B_m} p(B_m) \sum_x p(x \mid B_m) \log \frac{1}{p(x \mid B_m)} . \tag{3.38}$$

For $m = 0$, we define $H_0 = S_{1\times 1}$, or the entropy of the single character distribution. One can prove, see Appendix (to be appended later), that in the limit $m \to \infty$, H_m is equal to the entropy (3.35),

$$s = \lim_{m\to\infty} H_m = H_\infty . \tag{3.39}$$

As in the one-dimensional case, the average information of ln 2, or equivalently 1 bit, per lattice site can be decomposed into a term quantifying the information in correlations from different lengths (including density information) and a term quantifying the internal randomness of the system,

$$1 = k_{\text{corr}} + s . \tag{3.40}$$

The density information k_1 does not depend on the dimensionality, so it should be as in the one-dimensional case,

$$k_1 = \sum_x p(x) \log \frac{p(x)}{1/2} = 1 - S_{1\times 1} = 1 - H_0 \ . \tag{3.41}$$

Then it is clear that if we define correlation information over length m by the difference between two consecutive estimates of the entropy s, i.e., $k_{m+1} = -H_m + H_{m-1}$, for $m > 0$, then the decomposition above is complete. In order to show that this definition leads to k_m being a non-negative quantity, we introduce an operator R that reduces a configuration B_m to a configuration $B_{m\text{-}1} = \mathrm{R}B_m$ by taking away the symbols from the leftmost and rightmost columns as well as from the bottom row. Then k_m can be written as the average Kullback information when the distribution for "next" character given a conditional configuration B_m replaces an *a priori* distribution with a smaller conditional configuration $B_{m\text{-}1} = \mathrm{R}B_m$,

$$k_{m+1} = -H_m + H_{m-1} = \sum_{B_m} p(B_m) \sum_x p(x \mid B_m) \log \frac{p(x \mid B_m)}{p(x \mid \mathrm{R}B_m)} \geq 0 \, . \tag{3.42}$$

Here we have used the fact that summation over the characters in the part that is being reduced in B_m connects the probabilities for B_m and $\mathrm{R}B_m$,

$$p(\mathrm{R}B_m) = \sum_{\substack{\text{all configurations} \\ \text{in the reduced part}}} p(B_m) . \tag{3.43}$$

The correlation information k_{corr} is then decomposed by

$$k_{\mathrm{corr}} = \sum_{m=1}^{\infty} k_m \, . \tag{3.44}$$

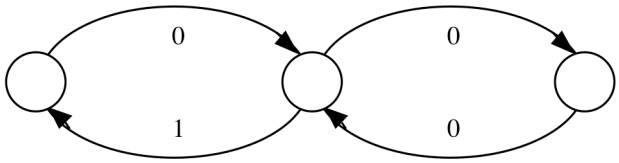
This procedure can be repeated in higher dimensions. Note also that the definition of correlation information contains a choice of direction and rotation: the B_m configuration can be chosen in eight different ways. In one dimension there are only two ways to choose, direction left or right, but the choice does not change the definitions. Furthermore, and this holds for both dimension one and higher, one may use other ways to decompose total redundancy which result in different type of correlation measures.

3.6 Excercises

3.1 How can the symbol sequences in Example 3.3.4 be coded so that one gets a maximally compressed coded sequence?

3.2 Suppose that we have a process that generates completely random binary sequences $(x_0, x_1, x_2, x_3, \ldots)$ with entropy 1 (bit). Construct a new binary sequence $(y_1, y_2, y_3, \ldots)$ by addition modulo 2 of pairs of symbols from the first sequence, $y_k = x_{k-1} + x_k \pmod 2$, $k = 1, 2, \ldots$. What is the Shannon entropy of the new sequence?

3.3 The finite automaton below represents a stochastic process generating binary sequences in which 1's are separated by an odd number of 0's. What is the Shannon entropy? How long correlations are there (from the information-theoretic point of view)? Assume that the arcs leaving the central node have equal probabilities.



3.4 Consider the class of stationary stochastic processes that generate binary symbol sequences in which pairs of 1's are forbidden. What is the largest entropy such a process could generate?

3.5 **Information loss in multiplication.** Consider two independent stochastic processes generating uncorrelated sequences of symbols 0, 1, and 2, with equal probabilities of the different symbols. Form the product process in which a symbol z_k is the product modulo three of the corresponding symbols from the original processes, $z_k = x_k * y_k \pmod 3$. Here, mod 3 means that 2 * 2 = 1, while all other multiplications work as usual. How much information is lost in average from the original pair of symbols (x_k, y_k) when the product z_k is formed?

3.6 Consider an infinite sequence of isolated 1's, separated by either one or two 0's (but not more), for example, "…100101010010010100101…". What probability distribution should we choose for describing the system, if we want the entropy s to be as large as possible? (The answer need not be explicit, but an equation that determines the parameter(s) is sufficient.

3.7 Show that (3.22) and (3.24) are equivalent as definitions of the Shannon entropy s.

3.8 Show that the distribution $\tilde{P}_n$ with the probabilities defined as in Eq. (3.17), can be derived by the maximum entropy formalism, using the constraints that the summation over first and last character results in the distribution P_{n-1} .

3.9 **Interpretation of a spatial average.** Consider a doubly infinite symbol sequence based on an alphabet of m different symbols, $\Gamma = \ldots\ \sigma_{i-1}\ \sigma_i\ \sigma_{i+1}\ \sigma_{i+2}\ \ldots$, where σ_i denotes the symbol in position i. Suppose that the sequence is generated by some ergodic, stationary stochastic process, so that the probability distributions for sub-sequences also can be

derived from internal statistics in the sequence Γ (as we usually assume). Consider the quantity

$$I_{i,n} = -\ln p(\sigma_i \mid \sigma_{i-n} \ldots \sigma_{i-2}\ \sigma_{i-1}) ,$$

where $p(\sigma_i \mid \sigma_{i-n} \ldots \sigma_{i-2}\ \sigma_{i-1})$ is the conditional probability defined by $p(\sigma_{i-n} \ldots \sigma_{i-2}\ \sigma_{i-1}\ \sigma_i)/p(\sigma_{i-n} \ldots \sigma_{i-2}\ \sigma_{i-1})$.

a) How can this quantity be interpreted?

b) What is its spatial average (for almost all Γ) in the limit $n \to \infty$,

$$\lim_{n\to\infty}\ \lim_{L\to\infty} \frac{1}{2L+1} \sum_{i=-L}^{L} I_{i,n} \quad ?$$

4 Cellular Automata

Cellular automata (CA) are simple models of dynamical systems that are discrete in space and time. The number of states per lattice site, or cell, is finite and usually small. The time development is governed by a local updating rule that is applied in parallel over the whole lattice. John von Neumann introduced cellular automata in the 1950's, and he wanted to use these models in his study of self-reproduction and noise-sensitivity of computation. One purpose was to demonstrate the existence of objects capable of complex behaviour combined with the capability of self-reproduction. This work lead to the design of a cellular automaton rule on a two-dimensional lattice with 29 states per cell. The designed object capable of making a copy of itself in this space also had the capability to simulate any computational process, usually termed a *computationally universal* system. In this way complex behaviour of the object was said to be guaranteed. This work was completed and published by Arthur Burks after von Neumann's death (von Neumann & Burks, 1966).

There are several possibilities to construct cellular automaton rules, even with only two states per cell, that demonstrate various examples of complex behaviour. One well-known is "Game of Life" that was introduced by John H. Conway in 1970. This rule has been studied extensively, mainly because of its capability of producing complex behaviour like propagating spatio-temporal structures from random initial states, exemplified by the simple "glider" in Fig. 4.1. The rule is based on a local configuration involving the cell and its 8 neighbours. Cells can be "alive" or "dead", represented by black and white in the figure. A "dead" cell becomes "alive" in the next time step if exactly 3 neighbours are "alive", while a cell that is "alive" remains so only if 2 or 3 neighbours are "alive".

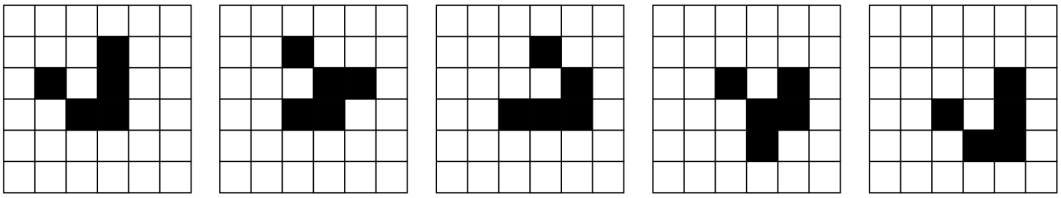


Figure 4.1. The cellular automaton "Game of Life" exemplified by five time steps of the "glider", a propagating object with an internal cycle of five time steps.

One can show that "Game of Life" also has the capability of universal computation. This can be done, for example, by constructing structures in the lattice that serves as wires that allow for propagating signals that may interact through structures serving as logical gates.

Already in the simplest class of CA, the CA rules in one dimension that have local interaction depending on nearest neighbours only, there are examples of various types of complex behaviour. In 1989 it was shown that with 7 states per cell, it is possible to construct CA rules that are capable of universal computation (Lindgren & Nordahl, 1989). A direct implication is that results from computation theory applies to this rule. For example, the halting problem, stating that there is no general procedure to determine whether a computer program will ever

halt, transforms into a theorem for computationally universal CA. This means that there are initial states for such a CA for which it is impossible to prove whether the CA will develop to a fixed point. Ten years ago Matthew Cook proved that the even more simple CA rule R110, depending on only two states per cell, is computationally universal (Cook, 2004). The space-time pattern of R110 is exemplified in Fig. 4.2.

Cellular automata as models for physical systems met with a renewed interest in the 1980's, partly because of the classic paper by Stephen Wolfram (1983), "Statistical mechanics of cellular automata." Among the physical applications of cellular automata, the "lattice gases" are the most well known. These CA simulate systems of particles with discrete (usually unit) velocities moving on a lattice. Despite the highly simplified microscopic dynamics, some of these systems approximate the Navier-Stokes equations for fluid dynamics, when averages are taken over large numbers of particles.

In this Chapter we shall demonstrate how the information-theoretic concepts presented in Chapter 3 can be used in order to analyse how order and disorder develop during the time evolution for different types of cellular automaton rules. We shall in particular study CAs that are reversible and discuss how apparently random patterns may be created in such systems. This will be related to the second law of thermodynamics to be discussed in a later chapter.

4.1 Elementary Cellular Automata

The simplest class of cellular automata is based on a one-dimensional lattice with two states per cell and a nearest neighbour rule for the dynamics. Such a rule is fully determined by specifying the next state of a cell for each of the eight possible local states that describes the present state of the local neighbourhood. An example of such a specification is shown in the following Table.

t	111	110	101	100	011	010	001	000
$t+1$	0	1	1	0	1	1	1	0

The time evolution of this rule, starting from a random sequence of 0's and 1's, is exemplified in Fig. 4.2. The top row of black (1) and white (0) dots represent the initial state, and the following rows represent the development in time when the rule in the table is applied in parallel over the whole row. In this CA class the rules can thus be described by the binary digits in the second row of the table, which means that eight binary symbols determine the rule. In the example, we have the rule number $(01101110)_2$ which in decimal form is 110 — the rule mentioned above being computationally universal. There are $2^8 = 256$ elementary CA rules, but several of these are equivalent (by symmetries, such as changing symbols and direction), which results in 88 different rules.

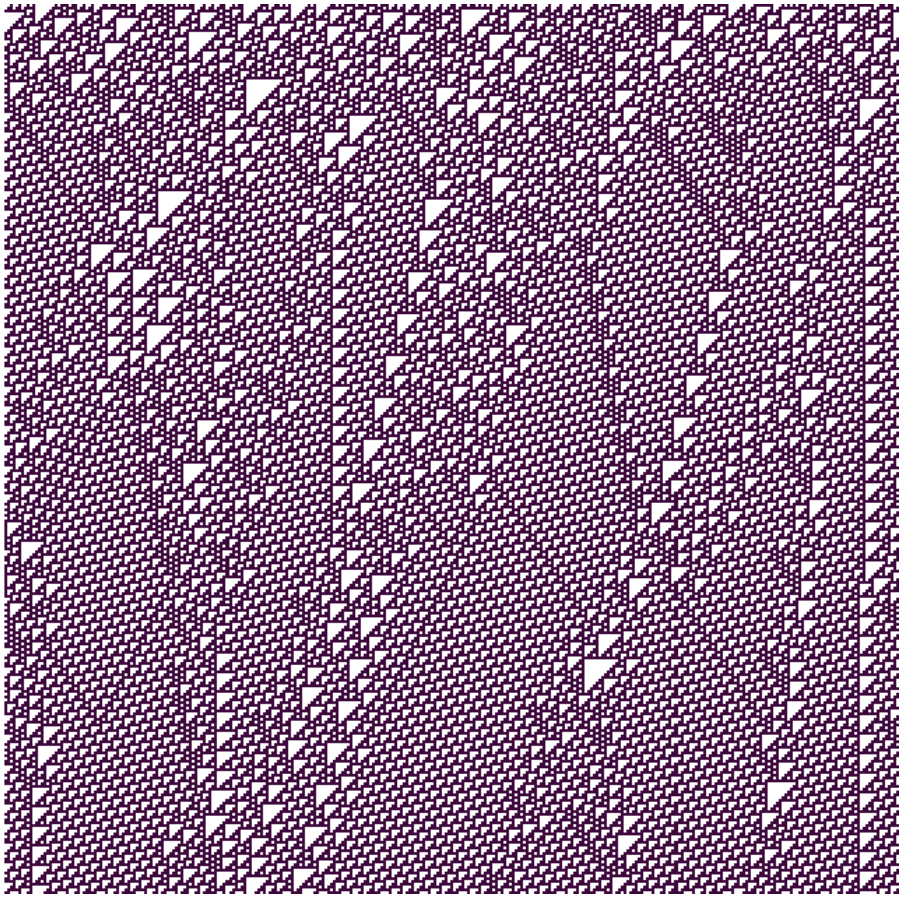


Figure 4.2. The time evolution of CA rule R110 starting with an "random" initial state (in the top row) shows how a periodic background pattern is built up at the same time as complex structures propagate and interact. This is a good example of the complexity that simple CA rules may exhibit.

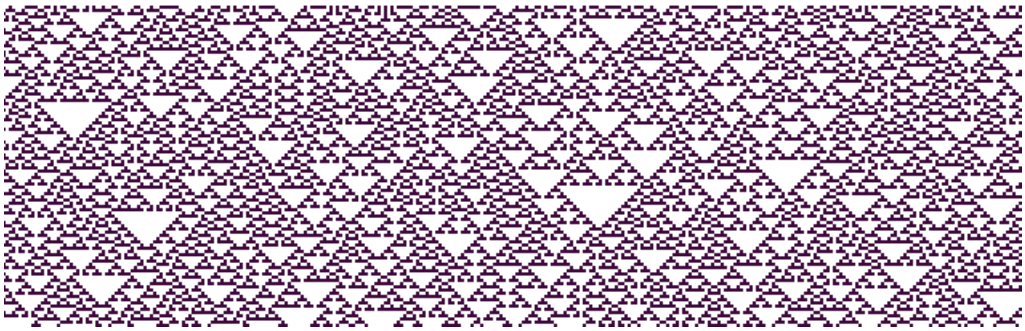
The dynamic behaviour of elementary CA can differ a lot from one rule to another. Wolfram suggested a classification with four types. The simplest CAs, class I, approach a homogenous fixed point as is exemplified by the rule in Fig. 4.3a. A class II rule develops into an inhomogenous fixed point or to a periodic and/or simple shift of the pattern like in Fig. 4.3b. The class III rules never seem to approach an ordered state, but their space-time patterns continue to look disordered, as illustrated in Fig. 4.3c. These rules are often called "chaotic". Then there is a class IV that is more vaguely defined as a border class between II and III with long complex transients, possibly mixed with a spatio-temporal periodic background pattern, see Fig. 4.3d. Also rule R110 in Fig. 4.2 belongs to this class.

An important characteristic of the class III, or the chaotic, rules is that they are sensitive to small perturbations, just like chaotic low-dimensional systems. If one follows the time evolution starting from two initial states, differing at one position only, the number of differing positions tends to increase linearly in time. This is illustrated in Figure 4.4, which shows the differing cells between the space-time patterns of two CAs following rule R22.

a)

b)

c)



d)

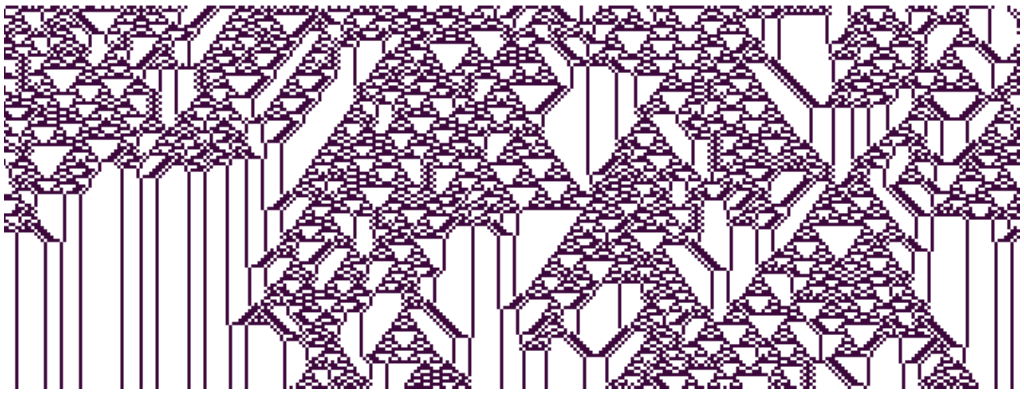


Figure 4.3. Cellular automata space-time patterns from the four classes of CA rules: a) Class I rules approach a fixed point (e.g., R160). b) Class II rules develop a periodic pattern in space and/or time (e.g., R213). c) Chaotic or class III rules are characterised by a continuous change of the patterns at the same time as a high disorder is kept (e.g., R22). d) Class IV is a border class involving features from both class II and class III with long complex transients. The last example is generated by a rule that depends on the number of living cells (1's) in a 5-cell neighbourhood, so that a cell survives if 1 or 3 neighbours are alive and a cell becomes living if 2 to 4 neighbours are alive.

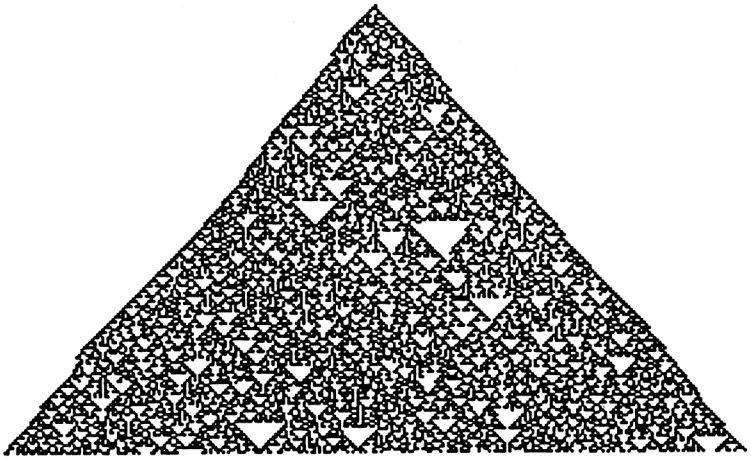


Figure 4.4. The difference pattern resulting from the time evolution of CA rule R22 starting from two different states, initially differing only by the state in one cell.

4.2 Information theory for Cellular Automata

What happens with the entropy that describes the internal disorder of the CA state in the time evolution? Initially we may have prepared the system as a maximally disordered sequence of symbols with an entropy of 1 bit per cell. Under what circumstances will this entropy decrease to create order in the system?

We shall use the information theory for symbol sequences presented in Chapter 3, in order to prove relations between the entropy of the CA at time t and the following time step $t + 1$. Suppose that we have a cellular automaton rule with range r, i.e., the local neighbourhood involved in the rule includes $2r + 1$ cells (r the left, r to the right, and the cell itself). We also assume that there are only two states per cell (0 and 1); the formalism is easily extended to more states. At each time step t, the state of the CA is described as a symbol sequence characterised by its entropy $s(t)$.

To begin with we shall assume that the rules are deterministic, which implies that the state at time t fully determines the state at the next time step. Let β_m denote a certain sub-sequence of symbols of length m at time $t + 1$. This sequence may have several possible predecessors at time t. Since the rule has range r, the predecessor sequence has length $m + 2r$, and we denote such a sequence α_{m+2r}. The probability distributions that describe the symbol sequences at different times may of course be different, and we use p for probabilities at time t and p' for probabilities at time $t + 1$. Then the probability for a sequence β_m at $t + 1$ can be expressed as the sum of the probabilities for all possible ancestors,

$$p'(\beta_m) = \sum_{\alpha_{m+2r} \to \beta_m} p(\alpha_{m+2r}). \tag{4.1}$$

This relation can be used to establish a connection between block entropies of length $m + 2r$ at time t with block entropies of length m at time $t + 1$,

$$S_m(t+1) = \sum_{\beta_m} p'(\beta_m) \log \frac{1}{p'(\beta_m)} = \sum_{\beta_M} \sum_{\alpha_{m+2r} \to \beta_M} p(\alpha_{m+2r}) \log \frac{1}{\sum_{\alpha_{m+2r} \to \beta_M} p(\alpha_{m+2r})} \leq$$
$$\leq \sum_{\beta_M} \sum_{\alpha_{m+2r} \to \beta_M} \left(p(\alpha_{m+2r}) \log \frac{1}{p(\alpha_{m+2r})} \right) = S_{m+2r}(t). \tag{4.2}$$

Here we have made use of the fact that $\log(1/x)$ is a decreasing function of x, when removing the summation in the logarithm[8]. This inequality can be used to derive the change in Shannon entropy between to successive time steps in the CA time evolution,

$$\Delta s(t+1) = s(t+1) - s(t) = \lim_{m \to \infty} \left(\frac{1}{m} S_m(t+1) - \frac{1}{m+2r} S_{m+2r}(t) \right) =$$
$$= \lim_{m \to \infty} \left(\frac{1}{m} (S_m(t+1) - S_{m+2r}(t)) + \left(\frac{1}{m} - \frac{1}{m+2r} \right) S_{m+2r}(t) \right). \tag{4.3}$$

By using Eq. (4.2) and the fact that the second term in Eq. (4.3) goes like $2r\, s(t)/m$, we can conclude that, for deterministic cellular automata rules, the entropy decreases (or stays constant) in the time evolution

$$\Delta s(t) \leq 0. \tag{4.4}$$

This irreversibility, when the inequality is strict, can be understood by the fact that deterministic CA rules reduce the number of possible states (symbol sequences) in the time evolution. This decrease in entropy is associated with an increase in total correlation information (including density information). This is clearly visible in the periodic pattern formed in the time evolution of rule 110, illustrated in Fig. 4.1. The initial state is a completely disordered sequence of zeroes and ones with en entropy $s = 1$ (bit), but as time goes on correlation information is built up and entropy decreases.

4.2.1 Almost reversible rules

An important property of the rules that govern a dynamical system is whether they can be considered reversible. In physical systems, the microscopic laws of motion usually are reversible, and therefore it is of interest to investigate the information-theoretic properties of cellular automata that can be considered reversible. Intuitively, we should expect the entropy of such a CA to be constant in time.

With a *reversible* CA rule we consider a CA in which each microstate (infinite symbol sequence) has a unique predecessor at the previous time step. We shall show, though, that there is a weaker form of reversibility that results in a constant entropy, i.e., equality in Eq. (4.4).

[8] We use the fact that $(x+y)\log(1/(x+y)) \leq x \log(1/x) + y \log(1/y)$.

Suppose that a cellular automaton rule R with range r can be written as

$$R(x_{-r},...,x_{r-1},x_r) = f(x_{-r},...,x_{r-1}) + x_r \mod 2, \tag{4.5}$$

where the summation is taken modulo 2. This means that the rule R is a one-to-one mapping with respect to its last argument — a flip of the state in the rightmost cell in the neighbourhood flips the result of the rule. (A corresponding type of rule can be constructed with respect to the leftmost argument.)

Note that this type of rule involves a certain type reversibility: Suppose that we know the semi-infinite microstate $y_1y_2y_3\ldots$ of the CA at time $t + 1$, and that we know the first $2r$ cells of the corresponding preceding microstate $x_1x_2\ldots x_{2r}$ at time t. Then we can use the rule (4.5), by entering $x_1x_2\ldots x_{2r}$ as an argument in f, to find the cell state x_{2r+1} at t that corresponds to the result y_{r+1}. By repeated use of this procedure, we may reproduce the complete semi-infinite microstate $x_1x_2x_3\ldots$ at time t. Only $2r$ bits of information are needed to reproduce the preceding microstate from the present one, and we call such a rule *almost reversible*. These rules differ from the *reversible* rules in that each microstate may have several (up to 2^{2r}) preceding microstates.

Then the probability for a sequence β_m at $t + 1$ can be expressed as the sum of the probabilities for all possible ancestors using a transfer function $T(\alpha_{m+2r},\beta_m)$. T takes the value 1 if $\alpha_{m+2r}\rightarrow\beta_m$ under the rule, otherwise 0.

$$p'(\beta_m) = \sum_{\alpha_{m+2r}} T(\alpha_{m+2r},\beta_m)p(\alpha_{m+2r}). \tag{4.6}$$

The transfer function has the properties

$$\sum_{\alpha_{m+2r}} T(\alpha_{m+2r},\beta_m) = 2^{2r}, \tag{4.7}$$

$$\sum_{\beta_m} T(\alpha_{m+2r},\beta_m) = 1. \tag{4.8}$$

To simplify the notation, we drop the length indices on α and β. The difference in block entropies between m-length blocks at $t + 1$ and $(m + 2r)$-length blocks at t can then be written

$$
\begin{aligned}
& S_m(t+1) - S_{m+2r}(t) = \\
& = \sum_\beta \left(\sum_\alpha T(\alpha,\beta)\, p(\alpha) \right) \log \frac{1}{\sum_\alpha T(\alpha,\beta)\, p(\alpha)} - \sum_\alpha p(\alpha) \log \frac{1}{p(\alpha)} = \\
& = \sum_\beta p'(\beta) \left(\sum_\alpha \frac{T(\alpha,\beta)\, p(\alpha)}{p'(\beta)} \log \frac{T(\alpha,\beta)\, p(\alpha) \,/\, p'(\beta)}{T(\alpha,\beta) \,/\, \sum_\alpha T(\alpha,\beta)} \right) - \log \sum_\alpha T(\alpha,\beta) = \\
& = \sum_\beta p'(\beta)\, K\left[T(\bullet,\beta) \Big/ \sum_\alpha T(\alpha,\beta) ; T(\bullet,\beta)\, p(\bullet) / p'(\beta) \right] - 2r \geq -2r \,.
\end{aligned}
\tag{4.9}
$$

Here we have used the fact that the Kullback information is non-negative. The symbol "•" in the position for α indicates that the constructed Kullback information is based on (normalised) probability distributions over α. This result then implies, cf. (4.3),

$$
\Delta s(t) \geq \lim_{m \to \infty} \left(-\frac{2r}{m} \right) = 0 . \tag{4.10}
$$

In combination with the law of non-increasing entropy for deterministic rules, Eq. (4.4), this results in a constant entropy for the time evolution of almost reversible rules.

$$
\Delta s(t) = 0 . \tag{4.11}
$$

For these cellular automata the initial distribution of information between entropy and redundancy is kept in the time evolution. Still, there may be non-trivial information-theoretic changes since the correlation length may change, provided that the system is prepared with an initial redundancy, for example in the form of density information. We shall look at illustrations of this in the section on examples.

4.2.2 Rules with noise

If noise, in the form of randomly flipped states, interfere with an otherwise deterministic rule, it seems reasonable that an increase in entropy is possible for some rules, provided that we start from an initial state with $s < 1$. Suppose that the deterministic rule R is applied as usual on the complete CA microstate, but that in the resulting microstate each cell state is flipped with a probability q. Then we denote such a *probabilistic rule* by the pair (R, q).

The change in entropy in one time step, can then be decomposed into two terms, one negative (or zero) term $\Delta_R s(t)$ associated with the deterministic change from rule R, cf. Eq. (4.4), and one positive term $\Delta_q s(t)$ due to the "added" entropy from the noise,

$$
\Delta s(t) = \Delta_R s(t) + \Delta_q s(t) \ . \tag{4.12}
$$

Let us now see how the noise increases the entropy s. Suppose that the probability distribution for m-length sequences of symbols are given by probabilities $p(\alpha_m)$, after the rule R has been applied. The noise then transforms the probabilities according to a transfer function $T_q(\alpha_m, \beta_m)$,

$$\hat{p}(\beta_m) = \sum_{\alpha_m} T_q(\alpha_m,\beta_m) p(\alpha_m) \ . \tag{4.13}$$

The transfer function depends on the Hamming distance $H(\alpha_m, \beta_m)$, i.e., the number of cells for which the disturbed sequence β_m differs from the original one α_m, which implies that

$$T_q(\alpha_m,\beta_m) = q^{H(\alpha_m,\beta_m)}(1-q)^{m-H(\alpha_m,\beta_m)}. \tag{4.14}$$

$$\sum_{\alpha_m} T_q(\alpha_m,\beta_m) = \sum_{\beta_m} T_q(\alpha_m,\beta_m) = 1. \tag{4.15}$$

The change in block entropy $\Delta_q S_m$ due to the noise can then be written (dropping the length index m on the sequence variables α and β)

$$\begin{aligned}
&\Delta_q S_m = S[\hat{p}] - S[p] = \\
&= \sum_{\beta}\sum_{\alpha} T_q(\alpha,\beta) p(\alpha) \log \frac{1}{\hat{p}(\beta)} - \sum_{\alpha} p(\alpha) \log \frac{1}{p(\alpha)} = \\
&= \sum_{\beta}\sum_{\alpha} T_q(\alpha,\beta) p(\alpha) \log \frac{T_q(\alpha,\beta) p(\alpha)/\hat{p}(\beta)}{T_q(\alpha,\beta)} = \\
&= \sum_{\beta} \hat{p}(\beta)\, K\left[T_q(\bullet,\beta); T_q(\bullet,\beta) p(\bullet)/\hat{p}(\beta)\right] \geq 0 \,,
\end{aligned}$$

since the Kullback information is non-negative. From this we can conclude that

$$\Delta_q s(t) = \lim_{m\to\infty} \frac{1}{m} \Delta_q S_m \geq 0 \,.$$

But it is also clear, from the fact that the Kullback information is zero only when both involved distributions are identical, that the added noise entropy is zero only when $p(\alpha_m) = \hat{p}(\beta_m)$ for all α_m and β_m. This only occurs when $p(\alpha_m) = \hat{p}(\beta_m) = 2^{-m}$, i.e., when the cellular automaton state is completely disordered with maximum entropy $s = 1$.

If the rule R involved is almost reversible, the only entropy change comes from the added noise. Starting from an ordered initial condition, $s < 1$, the entropy will increase until the system is completely disordered,

$$\Delta s(t) > 0, \quad \text{if } s(t-1) < 1, \tag{4.16}$$

$$\Delta s(t) = 0, \quad \text{if } s(t-1) = 1. \tag{4.17}$$

The following table summarises the results concerning the entropy change in deterministic and noisy cellular automata. For irreversible and noisy CA both increase and decrease in entropy is possible, depending on rule and the details of the current microstate.

	Deterministic	Noisy
Irreversible	$\Delta s(t) \leq 0$	–
Almost reversible	$\Delta s(t) = 0$	$\Delta s(t) \geq 0$

The results presented in this section are easily extended to two and more dimensions. One class of CAs that is of interest for this type of analysis is lattice gas models. As we will see in the next Chapter, the entropy analysed here is proportional to the thermodynamic entropy, and therefore the different relations shown, especially for (almost) reversible rules, has implications for the second law of thermodynamics, stating that physical entropy for a closed system cannot decrease.

4.3 Examples of information-theoretic properties in the evolution of simple CA

The irreversibility expressed by the law of non-increasing entropy for deterministic rules, Eq. (4.4), and illustrated by the time evolution of rule R110 in Fig. 4.2, can be quantified by calculating the increase in correlation information during the first time steps. In Fig. 4.5, we have plotted the contributions to the redundant information from density information k_1, as well as from correlation information from blocks of lengths up to 8, i.e., $k_2,...,k_8$, as functions of time. The initial state is completely "random" with an entropy of 1 and a total redundancy of 0. During the time evolution the entropy is transformed to density and correlation information.

If an almost reversible rule starts from a random initial state ($s = 1$), then the maximum entropy will be conserved in the time evolution, according to Eq. (4.11). For this type of system it is more interesting to study what is happening if one starts with a low entropy initial state, i.e., an initial configuration with $s < 1$. In the example shown in Fig 4.6, the time evolution of rule R60 is shown. This is an almost reversible rule given by $x_k' = R(x_{k-1}, x_k) = x_{k-1} + x_k \pmod 2$. This type of rule is also denoted *additive rule*. Note that the rule only depends on the cell itself and the left neighbour. The initial state is prepared with a 90% frequency of 1's, but without any correlations, resulting in $k_1 \approx 0.53$ and $s \approx 0.47$.

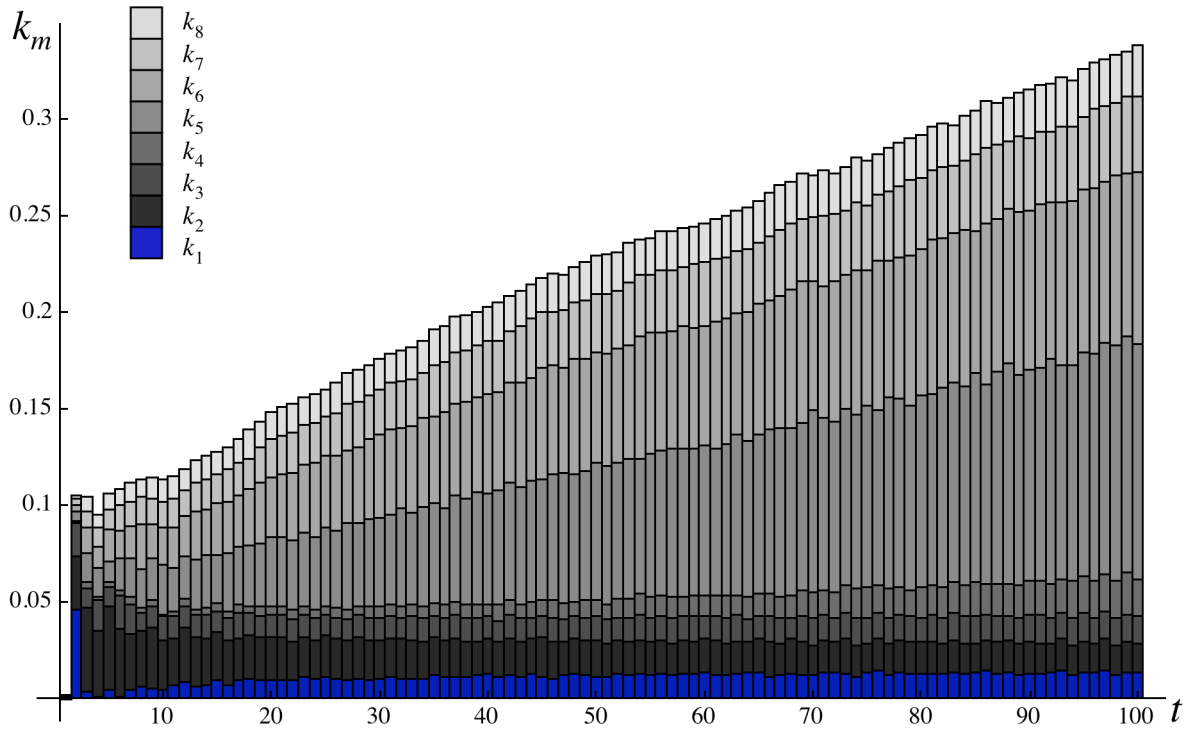


Figure 4.5. During the time evolution of CA rule 110, density and correlation information increases at the same time as entropy decreases by the same amount. In the diagram, the density information k_1 is represented by blue and contributions to correlation information from the first 7 terms, k_2 to k_8, represented by grayscale patterns from dark to light gray, are put on top of each other.

The time evolution of the CA in Figure 4.6 is characterised by the density information and the short length correlation information terms, see Figure 4.7. Even though the total redundancy is conserved, there are large changes in the specific k_m terms, showing that correlation information is changing significantly in length from one time step to another. Note also that at certain time steps, $t^* = 1, 2, 4, 8, 16$, etc, a large part of the redundancy is again gathered in the density information, k_0, and that these time steps are preceded by a series of steps in which correlation information is moving back from larger distances towards the density (or single cell) information term. Contributions from correlation information of length larger than 8 is not shown in the figure, but since we know that the total redundancy is conserved we also know that these contributions add up to the horizontal line at $k = 1 - s$. Of course, as time goes on, these events become more rare, which implies that in the long run most of the states (time steps) will not show any tracks of short length correlations. One can actually quantify this by using the correlation complexity η, and it can be proven (Lindgren & Nordahl, 1988) that the complexity increases linearly if $s < 1$,

$$\Delta\eta(t) = t\,s\,, \quad \text{if } s < 1\,, \tag{4.18}$$

otherwise $\Delta\eta(t) = 0$. In other words, the average correlation length increases linearly in time.

Therefore, even if the entropy is conserved for rule R60, the state of the CA will in the long run (at most time steps) appear more and more disordered, unless increasingly long correlations are taken into account when calculating the entropy. The low initial entropy may appear to increase, if the system is observed locally only.

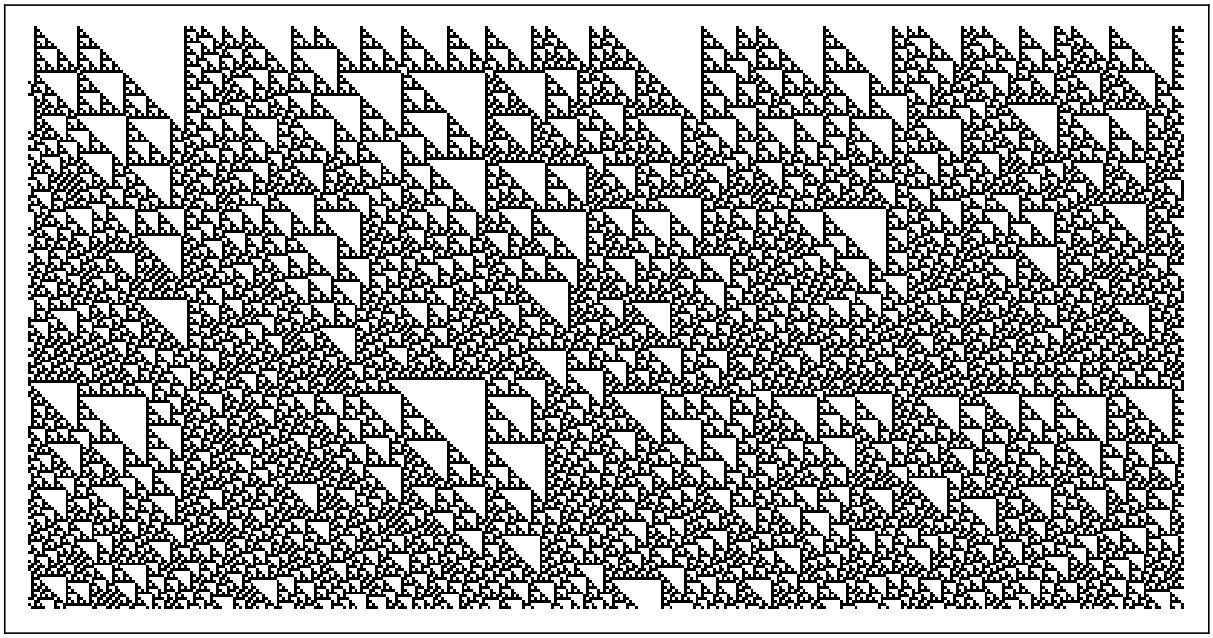


Figure 4.6. The space-time pattern of rule R60, an almost reversible rule, exhibits non-trivial behaviour when the entropy is less than 1 bit, $s < 1$. The initial state is dominated (90%) of 0's.

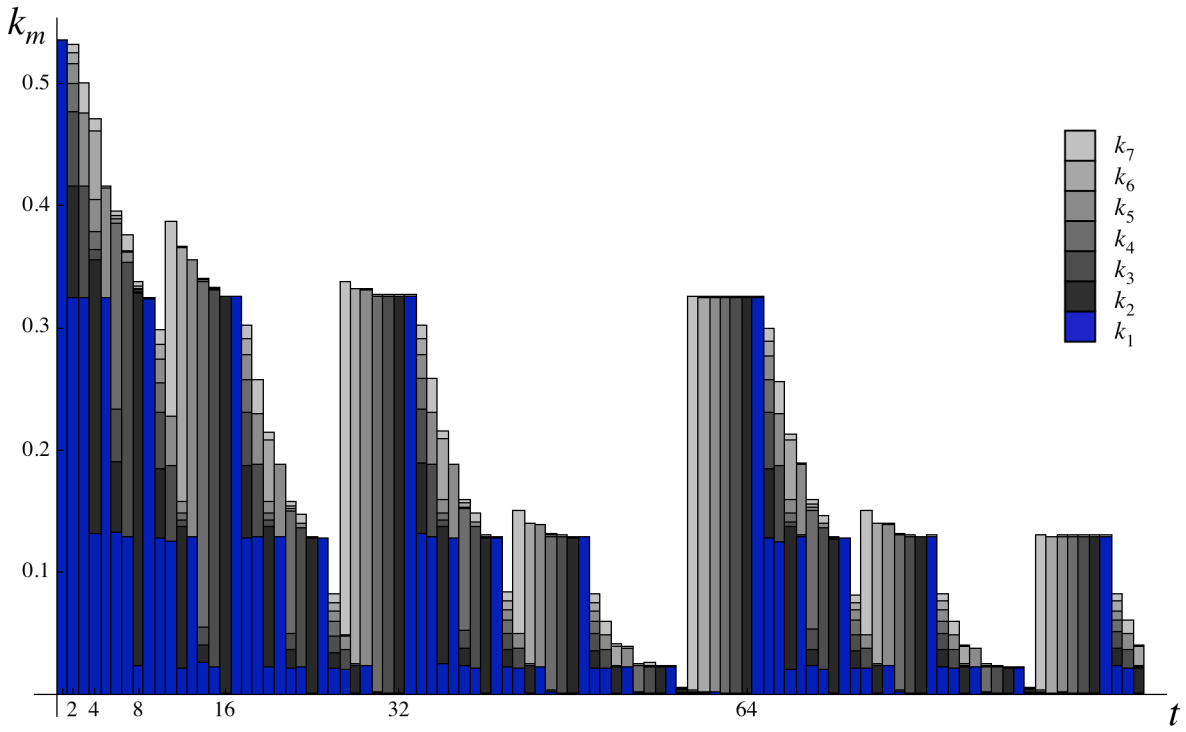


Figure 4.7. The diagram shows the contributions to the redundancy from density information (blue) and correlation information of lengths 2 up to 7 (grayscale from dark to light) put on top of each other, for the time evolution of the almost reversible rule R60. The initial state does not contain any correlations but there is density information due to the 90% dominance of the state 0. Therefore, the initial redundancy is larger than zero and the entropy $s < 1$. The reversibility implies that the entropy is conserved, and the figure then shows that information quickly is distributed over correlations of lengths larger than six.

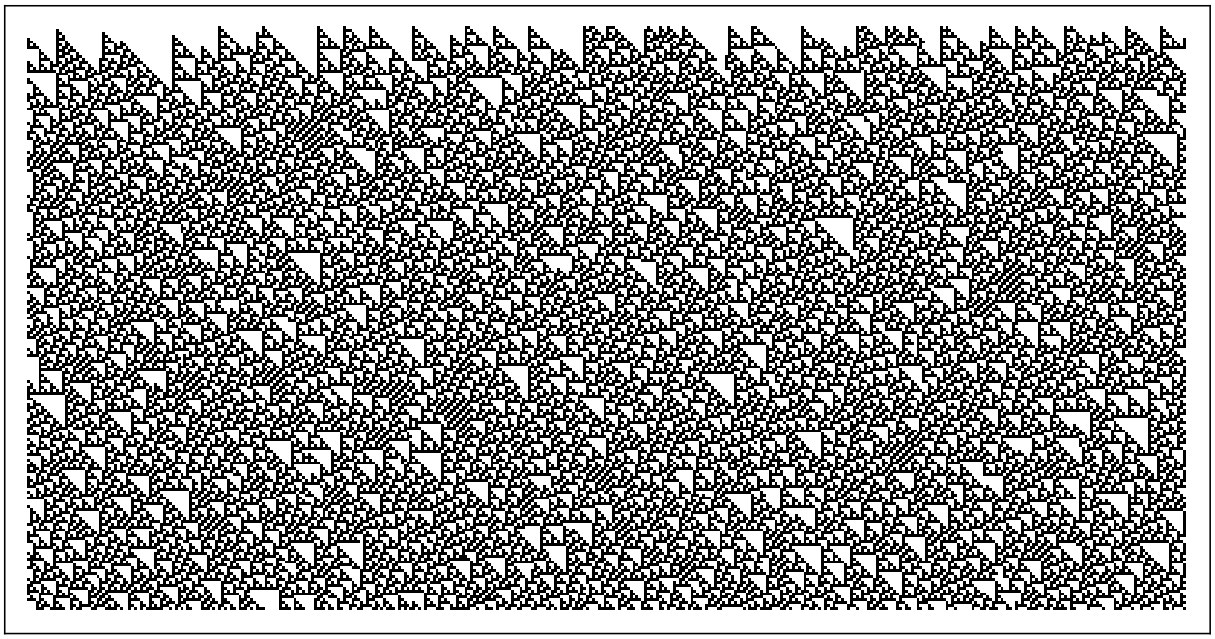


Figure 4.8. Noise at a level of 1% is added to rule R60, and the initial redundancy is efficiently destroyed in the time evolution, cf. Figure 4.6.

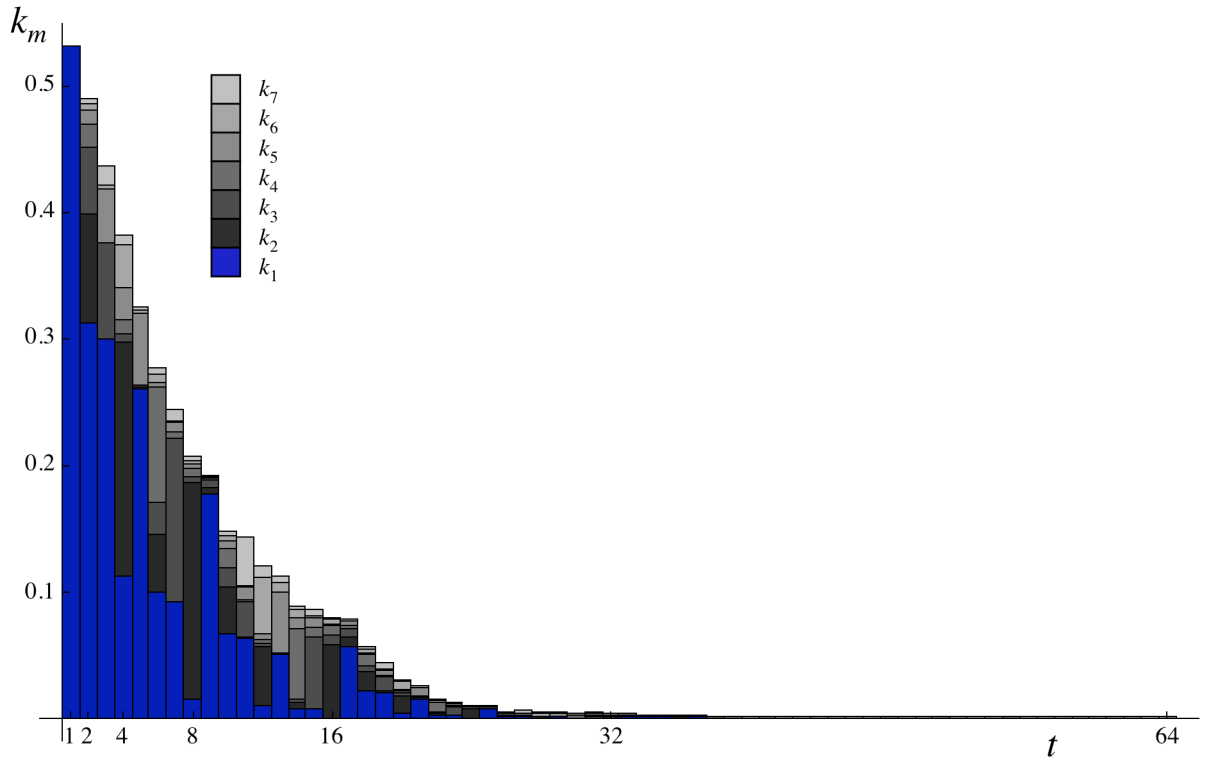


Figure 4.9. In the same way as was illustrated for rule R60 without noise, see Figure 4.7, the contributions to the redundancy, from density information (blue) and from correlation information over lengths 2 to 6 (dark to light grey), are put on top of each other. It is clear that the noise destroys the correlations, and the entropy increases.

If the time evolution of rule R60 is disturbed by random noise, the correlations that are built up from the initial density information are rapidly destroyed. In Figure 4.8, the space-time pattern is shown when a noise of 1% ($q = 0.01$) is added. It is clear, as also Figure 4.9

illustrates, that the time steps when a large part of the redundancy is recollected in the density information disappear. According to Eq. (4.16), the time evolution will lead to a completely disordered state with maximum entropy $s = 1$ (bit).

4.4 Analysis of CA time evolution using Hidden Markov models

The change of information-theoretic characteristics in the state of a one-dimensional CA, represented by the infinite sequence of symbols, from one time step to the next, can be analysed by inivestigating how the finite state automaton (FSA) representation of the sequence changes under the CA rule. Here we will assume that the state is given by a stochastic process, represented by a certain FSA. The FSA can correspond to a Markov process, which it does for the initial state if we, for example, start with uncorrelated symbols, like in the examples in previous sections. In general, though, also for such an initial condition, the CA rule transforms the automaton to one representing a hidden Markov model, as we shall ses in examples below. How the procedure works to transform one FSA to a new one under a CA rule will be illustrated by the following example.

Assume that we have, at time t, a certain description of the CA state in the form of the hidden Markov model shown in Fig. 4.10a, where we assume all transition probabilities to be ½ whenever ther eis a choice. First we rewrite that into a corresponding Markov model, Fig. 4.10b, in which all the states are represented by specific symbols.

(a) (b)

Figure 4.10. (a) FSA representing the hidden Markov model in which sequences of symbol 1 are separated by an odd number of the symbol 0. (b) FSA representing the corresponding Markov model in which the rightmost state has been given symbol 0' in order to distinguish between the two 0-states. The hidden Markov model in (a) is then given by the function $f(1)=1$, $f(0)=0$, and $f(0')=0$. All transition probabilities are ½ when there is a choice.

Next we construct a new Markov model out of the one in Fig 4.10b, by having each FSA state representing pairs of symbols form the sequence. These pairs are overlapping, so that a state can be seen as a two-symbol wide “window”, and each transitions in this automaton moves the window one step forward in the sequence of symbols, as illustrated in Fig. 4.11.

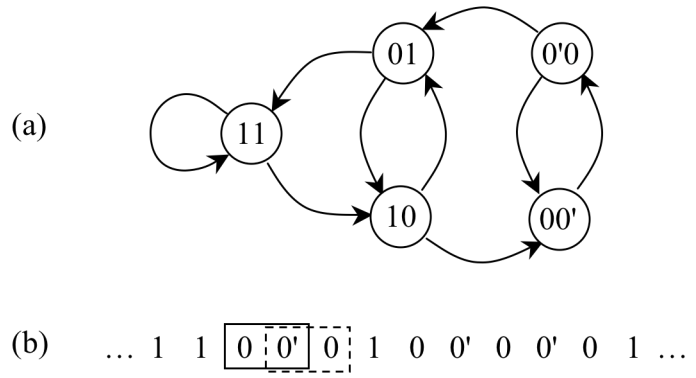


Figure 4.11. (a) FSA representing the same Markov model as in Fig. 4.10b, but where states corresponds to overlapping pairs of symbols, and where the transition move to the next pair overlapping with the previous one, as illustrated in (b). The transition probabilities from the original automaton remains, all being ½ when there is a choice.

The transitions in the FSA based on pairs of symbols, in Fig. 4.11, correspond to triplets of symbols (resulting from the two overlapping pairs of symbols of the states connected by the transition). Therefore, we can apply the elementary CA rule on the triplets to get the automaton representing the stochastic process that describes the next time step of the CA. In Fig. 4.12 this is illustrated with the resulting automaton shown for rule R86:

t	111	110	101	100	011	010	001	000
$t+1$	0	1	0	1	0	1	1	0

This rule is almost reversible (involving a 1-to-1 mapping between rightmost cell and resulting state). Thus we know that entropy s will not change under the CA rule, but correlation characteristics may do.

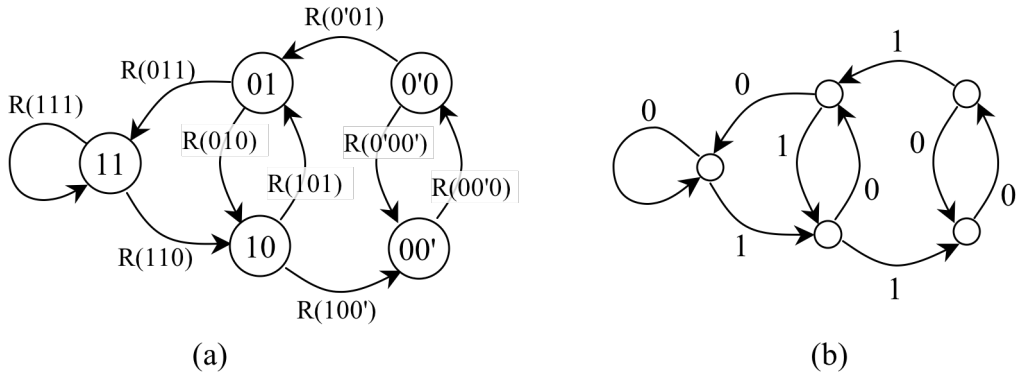


Figure 4.12. (a) The transitions on the FSA for the original time step t correspond to triplets of symbols and thus defines the local neighbourhood on which the CA rule is applied. In (b) the resulting FSA is shown where the states in the node are hidden, and this FSA represent the stochastic process describing the CA state at time $t + 1$.

It is now possible to use the FSA of Fig. 4.12b to calculate the properties of the CA at the new time step. One can simplify these calculations, though, by minimising the FSA exploiting the fact there there exist equivalent states in the automaton. The leftmost state (old 11) and the top middle state (old 01) both produce a 0 with probability ½ ending up in the leftmost state

and they both produce a 1 with the same probability ending up in the bottom middle state (old 10). Therefore they are equivalent in every aspect, and the states can be merged as shown in Fig. 4.13.

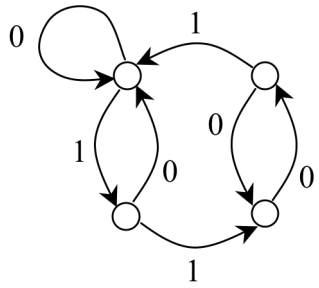


Figure 4.13. The most compact FSA representation of the CA state characteristics for time $t + 1$, when elementary CA rule R86 is applied to a CA state characterised by the FSA in Fig. 4.10a at time t.

For the FSA in Fig. 4.13, it is straightforward to calculate the stationary distribution over the nodes, which results in probability 2/5 for the upper left node and 1/5 for the others. One can check that the conditions for using the simplified expression of the entropy, as discussed in Example 3.3.4, is fulfilled here[9]. This means that the entropy equals the information required for the choices of transitions in the FSA. Each such choice involves 1 bit, as the probabilities are ½ whenever there are two outgoing transitions. This occurs in all nodes except the lower right one, with a total weight of 4/5. This results in the entropy $s = 4/5$ (bit), which one also finds for the original FSA in Fig. 4.10a. The correlation complexity η for the FSA at $t + 1$ in Fig. 4.13 can be calculated using the approach in Section 3.4.1 and Eq.(3.34) stating that η equals the entropy of the stationary distribution over the nodes, which gives $\eta_{t+1} = \log 5 - (2/5)\log 2 \approx 1.92$ (bits). For the original FSA at t the correlation complexity is similarly derived, $\eta_t = \log 5 - (4/5)\log 2 \approx 1.52$ (bits). This indicates an increase in average correlation length for this time step.

[9] For almost all preceeding sequences (generated by this FSA), in the infinite length limit, one can uniquely determine which node is the final one.

4.5 Local information detecting patterns in CA time evolution

In the time evolution of a cellular automaton rule it is often difficult to distinguish the normal behaviour from more rare events. For example, in a chatic rule like R18 the irregular space-time pattern contains local structures that are less common, but they are not easy to distinguish. Therefore it would be useful to be able to filter out the regular patterns to identify local configurations that deviate.

In information terms one would expect that when a local less common configuration is encountered at a certain position i in the sequence of cells, the conditional probability for that configuration given the n-length symbol sequence, for example, in the cells to the left of it will be relatively small. This implies a high local information of that conditional probability,

$$I_{i,n}^{(L)} = \log \frac{1}{p(s_i \mid s_{i-n} \ldots s_{i-1})}, \tag{4.19}$$

where s_i denotes the symbol at position i, see (Helvik *et al*, 2007). A corresponding local information $I_{i,n}^{(R)}$ conditioned on the n cells to the right of the position i is similarly defined. Such a local information quantity has a spatial average that can be written

$$\left\langle I_{i,n}^{(L)} \right\rangle = \lim_{N \to \infty, n \to \infty} \frac{1}{2N+1} \sum_{i=-N}^{N} \log \frac{1}{p(s_i \mid s_{i-n} \ldots s_{i-1})}. \tag{4.20}$$

By using the ergodicity theorem, Eq. (3.7), we find that

$$\left\langle I_{i,n}^{(L)} \right\rangle = \lim_{n \to \infty} \sum_{x_0 \ldots x_{n-1} x_n \in \Lambda^{n+1}} p(x_0 \ldots x_{n-1} x_n) \log \frac{1}{p(x_n \mid x_0 \ldots x_{n-1})} = \lim_{n \to \infty} \Delta S_{n+1} = s. \tag{4.21}$$

The same holds for the corresponding right-sided quantity $I_{i,n}^{(R)}$. We can define a local information as an average between the two,

$$I_{i,n} = \tfrac{1}{2}(I_{i,n}^{(L)} + I_{i,n}^{(R)}), \tag{4.22}$$

This means that the local information $I_{i,n}$, as well as the left- and right-handed versions, has a spatial average that equals the entropy of the system. In Figures 4.14 and 4.15, the local information is applied to the patterns generated by two cellular automaton rule, one being the irreversible class III rule R18, and the other almost reversible rule R60. The information pictures reveals a pattern not clearly seen in the space-time CA patterns.

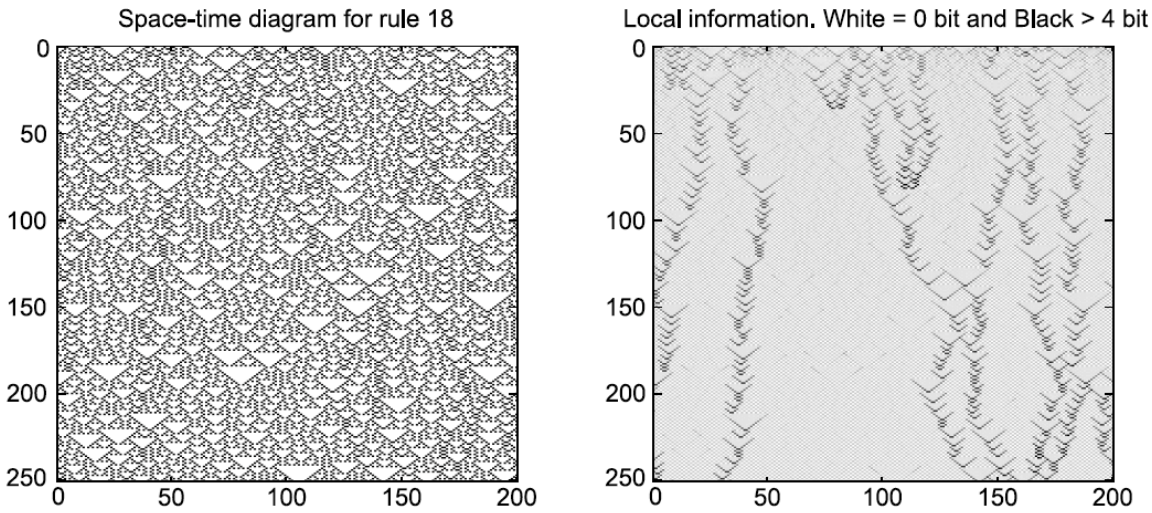


Figure 4.14. The space-time CA pattern for rule R18, and the corresponding information density $I_{i,n}$ picture.

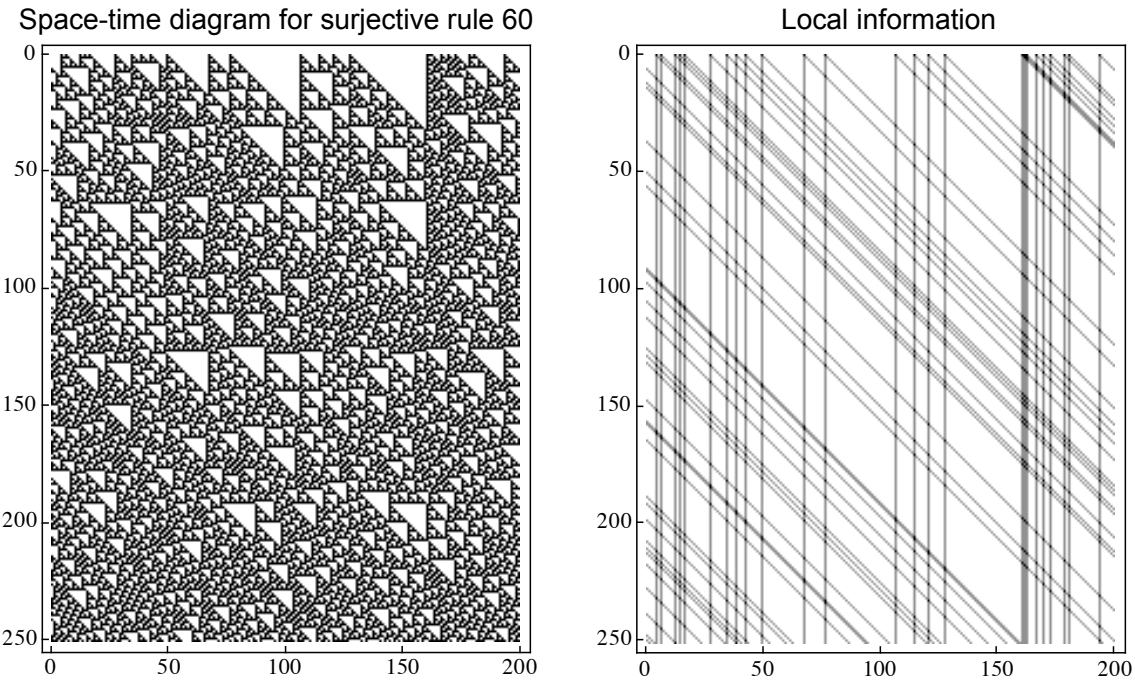
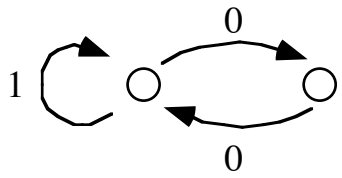


Figure 4.15. The space-time CA pattern for rule R60, and the corresponding information density picture. Here the informationdensity $I_{i,n}$ is derived analytically. A numerical estimate would not be computationally possible because of the linearly increasing correlation lengths.

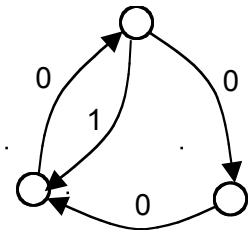
4.6 Exercises

4.1 Suppose that the initial state of a cellular automaton is generated by the following automaton, where the probability for choice of arc is 1/2.



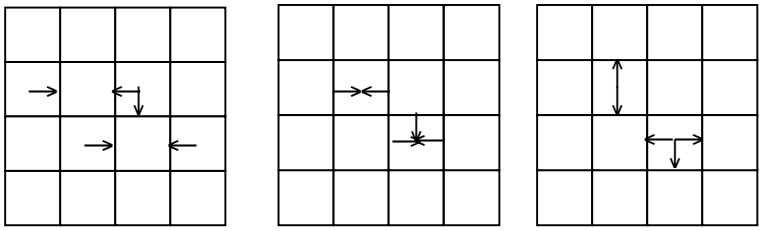
How large is the entropy initially? If the rule is R18, how does the automaton look like that describe the system after one time step, and what is the entropy s? What will the entropy be at this time step if the rule instead is R22?

4.2 **CA entropy.** Consider a one-dimensional cellular automaton given by elementary rule 71 (where configurations 110, 010, 001, and 000 result in a 1 and the rest give 0). Let the initial state be characterised by the following finite state automaton



where the probabilities for choosing an arc is always the same (1/2) if there is a choice. What is the initial entropy ($t = 0$), and what is the entropy at $t = 1$ and $t = 2$?

4.3 **Lattice gas entropy.** Consider an infinite 2-dimensional lattice gas constructed in the following way. The space is a square lattice and in each cell up to four particles may be present (one in each direction). The system evolves in discrete time, and in each time step there is movement and collision. Particles move from one cell to the next according to the direction of the particle. A collision occurs if and only if exactly two particles enter a cell with opposite directions, and then the direction of these particles are shifted so that they leave perpendicular to their initial directions. The two processes in a single time step is illustrated in the following figure:



Suppose we have a system where we initially (at $t = 0$) have equal densities of the four particle directions ($\rho/4$ each with ρ being the overall particle density), but where particles initially are present only in cells where *all* particle directions are present. Assume that these cells, each containing four particles, are randomly distributed over the whole lattice.

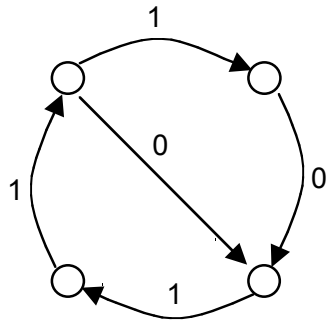
Consider the entropy s of the spatial configuration of particles, based on the 2-dimensional block entropy

$$s = \lim_{m\to\infty} \frac{1}{m^2} S_{m\times m}$$

What is the entropy s at $t = 0$? How does the entropy change in the time evolution?

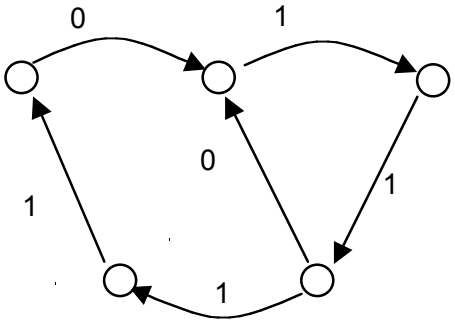
If one would estimate the entropy using a finite block size m after very long time T, with $T >> m$, what result should one expect? What is the explanation?

4.4 **CA entropy.** Consider a one-dimensional cellular automaton given by elementary rule 128 (where configurations 111 results in a 1 and the rest give 0). Let the initial state be characterized by the following finite state automaton



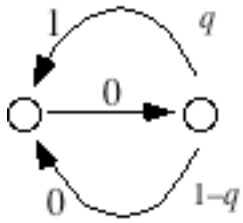
where the probabilities for choosing an arc is always the same (1/2) if there is a choice. Determine the finite state automaton that characterizes the state at time $t = 1$. What is the initial entropy ($t = 0$), and what is the entropy at $t = 1$ and $t = 2$?

4.5 **CA entropy.** Consider a one-dimensional cellular automaton given by elementary rule 192 (where configurations 111 and 110 result in a 1 and the rest give 0). Let the initial state be characterized by the following finite state automaton



where the probabilities for choosing an arc is always the same (1/2) if there is a choice. Determine the finite state automata that characterize the state at time $t = 1$ and at time $t = 2$, respectively. What is the initial entropy ($t = 0$), and what is the entropy at $t = 1$, $t = 2$, and $t = 3$?

4.6 Suppose that the initial state ($t = 0$) for the elementary CA rule 18 is described by the finite automaton:

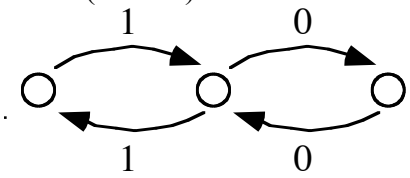


First, suppose that $q = 1/2$. What does the finite automaton look like that describes the state at time $t = 1$? How has the entropy changed between these time steps? What is the entropy at $t = 2$?

If $q > 1/2$, what is then the entropy s at $t = 2$?

4.7 Consider a binary periodic sequence ...01010101... . If noise is added so that symbols in this sequence are flipped (0 —> 1 and 1 —> 0) with a certain probability $q < ½$, how does the correlation complexity change?

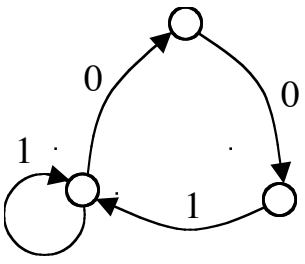
4.8 Suppose that elementary rule 129 governs the time evolution of a cellular automaton, and that the initial state (at $t = 0$) is characterised by the following finite automaton



where the probabilities for choosing 0 and 1 from the middle node are equal (1/2). What is the entropy s after one time step ($t = 1$) and after two steps ($t = 2$)?

Suppose instead that the probabilities for choosing 0 and 1 in the middle node differ, say $p(0) < ½ < p(1)$. Describe in a qualitative way how this affects entropy and correlations at the first time step ($t = 1$).

4.9 Suppose that elementary rule 68 governs the time evolution of a cellular automaton, and that the initial state (at $t = 0$) is characterised by the following finite automaton



where the probabilities for choosing 0 and 1 from the bottom left node are equal (1/2). What is the initial entropy, and how does it change over time?

5 Physics and Information Theory

When observing, or measuring on, a physical system we often gain knowledge on some macroscopic characteristics of the system, for example its energy or molecular composition. We seldom have full information on the exact microscopic configuration of the system, but we may, based on what we know regarding the macroscopic properties, assign a probabilistic description of the system. Such a desciption could be in the form of a probability distribution over the possible microstates of the system. Following the discussion in Chapter 2, such an assignment of probabilities should not include more knowledge than we have, and thus we should use a maximum entropy approach.

Before describing in detail how the maximum entropy approach connects information theory with statistical mechanics, we will study a simpler example in which the information is directly connected to the work that can be extracted out of an ordered, non-equilibrium, system, when bringing that system reversibly to an equilibrium, i.e., to a fully disordered state. That example will be based on some basic thermodynamic definitions and relations that will first be summarised below.

5.1 Basic thermodynamics

In the following section we will consider ideal gases (pure or mixed gases). In an ideal gas molecules are assumed to be point particles that do not interact. For such a system the ideal gas law holds,

$$pV = Nk_{\mathrm{B}}T. \tag{5.1}$$

Here, p is the pressure, V the volume, N the number of molecules in the gas, T the temperature, and k_{B} Boltzmann's constant (1.38×10^{-23} J/K).

We will use this relation in order to determine the work that is required (or that can be extracted) when a gas volume is changed. If one assumes that the system may interact with the environment (the world outside the system defined by the volume V) by receiving energy in the form of heat Q or deliver energy in the form of work W, then we can use Figure 5.1 to illustrate the first law of thermodynamics.

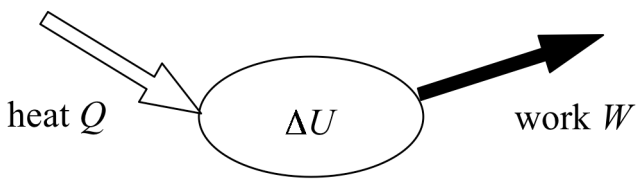


Figure 5.1. The first law of thermodynamics states that energy is a conserved quantity. The change in internal energy U is determined by the net energy difference resulting from the inflow of heat Q and work extracted from the system W.

The change ΔU of internal energy U must equal the difference between heat received and work delivered. (Heat and work may, of course, go in the other direction, when the quantity is negative.) This is the first law of thermodynamics stating that energy is a conserved quantity,

$$\Delta U = Q - W \,. \tag{5.2}$$

If the work is pressure-volume work, i.e., work dW extracted from an ideal gas from an infinitesimal volume change dV, then the work is

$$dW = pdV \,. \tag{5.3}$$

Thermodynamic entropy S_T is a quantity that follows a heat flow but that is not present in the energy flow associated with the work. (We will use the notation S_T for thermodynamic entropy to distinguish that from the information theoretic entropy. As will be seen, there is a simple relation between these.) With a heat flow dQ of temperature T follows an entropy increase of the system dS_T,

$$dS_\mathrm{T} = \frac{dQ}{T} \,. \tag{5.4}$$

Similarly, a heat flow leaving a system decreases the entropy of the system (but increases the entropy in the system that receives the heat). In general, the change of entropy of a system receiving heat is larger than or equal to the right-hand side of Eq. (5.4), where the inequality comes from irreversible processes taking place in the system which results in an increased entropy. For a closed system, the second law of thermodynamics states that, *the entropy always increases or stays constant*.

5.1.1 Intensive and extensive variables

We will consider a system of volume V that contains M different molecular species, each with a certain number of molecules N_i (i=1,…, M). Further, we assume that the system is characteised by a certain internal energy U. If we assume that the system is in internal equilibrium, the thermodynamic entropy S_T of the system is well defined by the state variables U, V, and N_i.

These state variables are extensive variables, i.e., they increase linearly with system size. One can characterise the same system using intensive variables instead, related to these three types of extensive variables. Sometimes the relations below, based on the how the thermodynamic entropy S_T depends on the extensive variables, are used as definitions of the intensive variables temperature T, pressure p, and chemical potential g_i. The relations are partial derivatives thermodynamic entropy S_T with respect to one of the extensive variables (U, V, or N_i), with the other variables kept constant (as indicated by the notation),

$$\left(\frac{\partial S_\mathrm{T}}{\partial U}\right)_{V,N_i} = \frac{1}{T} \,, \tag{5.5}$$

$$\left(\frac{\partial S_\mathrm{T}}{\partial V}\right)_{U,N_i} = \frac{p}{T} \,, \tag{5.6}$$

$$\left(\frac{\partial S_{\mathrm{T}}}{\partial N_j}\right)_{U,V,N_{i\neq j}} = -\frac{g_j}{T} \ . \tag{5.7}$$

We will use these relations when interpreting the maximum entropy approach from information theory within statistical mechanics in Section 5.3.

5.2 Work and information — an extended example

In this section we consider the following experiment dealing with mixing of ideal gases. First, the two gases are separeted within a container by a wall. When the wall is removed the gases expand into the full volume of the container, and finally the system consists of the mixture of the two gases occupying the full volume. In the mixed system we have less information about the positions of gas molecules compared to the case when the gases are seprated by the wall. It turns out that the additional information we have in the separated case corresponds to the amount of work that we can extract when bringing the system to the mixed state in a controlled way. Below we shall see how the information is related to the energy that can be extracted in the form of work.

Let us consider a gas container of volume V with a mixture of two different ideal gases 1 and 2, see Figure 5.2. The total number of molecules is N, of which N_1 are of type 1 and N_2 are of type 2. The normalised concentration can be written $x_1 = N_1/N$ and $x_2 = N_2/N$. If we pick a molecule from a certain part of the container, we do not know which molecule we will get, but we describe the chances by the probability distribution given by the concentrations $P = \{x_1, x_2\}$. The information-theoretic entropy S can then be used to quantify our lack of knowledge about the system (which molecule we will get),

$$S = x_1 \log\frac{1}{x_1} + x_2 \log\frac{1}{x_2}. \tag{5.8}$$

Consider now the case where the two gases are separated. The container is divided in two parts with volumes V_1 and V_2, so that the pressure is the same everywhere, i.e., $V_1 = x_1 V$ and $V_2 = x_2 V$, see Fig. 5.2. If we now pick a molecule from a certain place, we know which molecule it is and our uncertainty (entropy) is zero. By mixing the gases, we make an information loss of $\Delta S_{\mathrm{I}} = S_{\mathrm{I}}$ per molecule.

Figure 5.2. The to gases are mixed in the container to the left, but to the right they are separated by a wall so that the pressure is equal in the two parts.

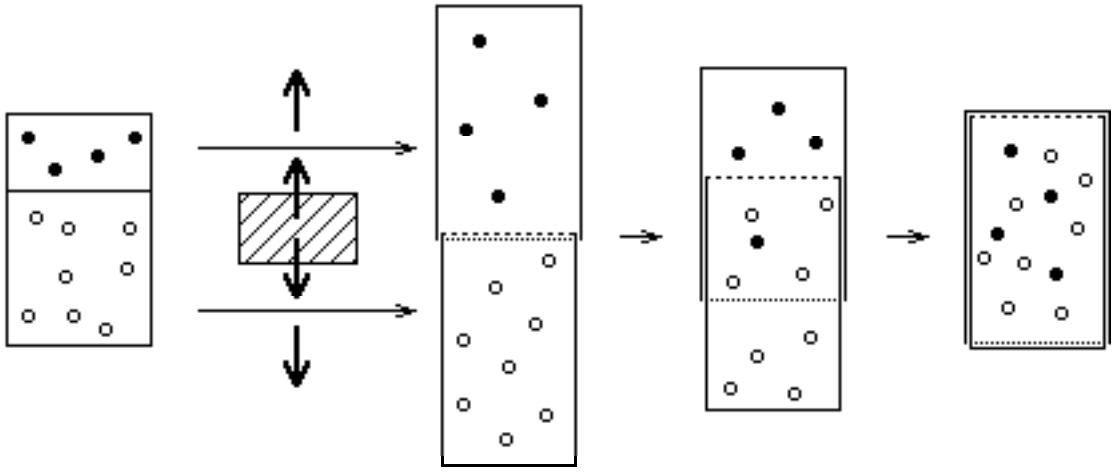


Figure 5.3. The two gases are mixed reversibly by an isothermal expansion in which work is done, followed by a fully reversible mixing without any energy transfer.

We will now calculate the maximum amount of work that can be derived from a process mixes the two gases. Then we may compare that with the information-theoretic loss we make by mixing the gases. We will also derive the thermodynamic increase of entropy, to be compared with the information-theoretic one.

Suppose that the gas container is in thermodynamic equilibrium with the environment at temperature T_0. Since the initial volume and the final volume of our system are equal, there will be no net pressure-volume work done on the environment and we may assume that the environmental pressure is zero in our calculations. Now we shall mix the gases in a reversible way as follows. First, let the separated gas volumes expand isothermally, each to the full volume V, by allowing a heat flow from the environment to keep the temperature constant, see Fig. 5.3. Next, we let the gases mix reversibly by letting them pass the semi-permeable walls, when the two volumes are pushed together. Note that there is no net force involved in this mixing, since the molecules in the first volume do not recognise the wall of the second, and *vice versa*. There is no heat flow involved in the mixing, so the entropy is unchanged in this part of the process. The only work (and entropy change) in this process comes from the expansion. The work W_i, for volumes $i = 1$ and 2, is given by, using the ideal gas law, Eq. (5.1), and Eq. (5.3),

$$W_i = \int_{V_i}^{V} p_i \, d\nu = \int_{V_i}^{V} N_i k_B T_0 \frac{d\nu}{\nu} = N_i k_B T_0 \ln \frac{V}{V_i} = N k_B T_0 \, x_i \ln \frac{1}{x_i}, \tag{5.9}$$

which results in

$$W = W_1 + W_2 = N k_B T_0 (x_1 \ln \frac{1}{x_1} + x_2 \ln \frac{1}{x_2}) = N k_B T_0 (\ln 2) \, S, \tag{5.10}$$

where the factor (ln 2) enters only if the information-theoretic entropy is measured in bits, as in Eq. (5.8). The relation between extractable work per molecule w and information loss when the system is mixed ΔS is then

$$w = k_B T_0 (\ln 2) \Delta S. \tag{5.11}$$

We have assumed an isothermal process (i.e., constant temperature). Since the internal energy U of an ideal gas is a function of temperature only (since there is no interaction between the molecules), energy conservation implies that the work W done by the system must equal the heat flow Q into the system. The change of thermodynamic entropy $\Delta S_T = Q/T_0$ for the whole system can then be written

$$\Delta S_T = \frac{Q}{T_0} = N k_B (\ln 2) \Delta S. \tag{5.12}$$

This illustrates the fact that the conversion factor between information-theoretic and thermodynamic entropy per molecule s_T is $k_B(\ln 2)$,

$$s_T = k_B (\ln 2) S. \tag{5.13}$$

Note that the information-theoretic quantity here was also expressed on the level of a single molecule, since it quantified information associated with observing *one* molecule from the system. One bit of information thus has a very low thermodynamic value. This example also illustrates that the information loss when mixing two gases is not really from the "mixing", but it comes from the fact that each gas, after the mixing, is distributed over a larger volume.

5.3 From information theory to statistical mechanics and thermodynamics

Information theory can be used as a tool to determine the macrostate (the probability distribution over microstates) in physical systems when averages of certain physical variables like energy and number of particles are known. According to the maximum entropy principle, we should choose the macrostate that has the largest entropy and that is consistent with the known variables. In this section, we shall demonstrate the connection between concepts in information theory and concepts in statistical mechanics and thermodynamics.

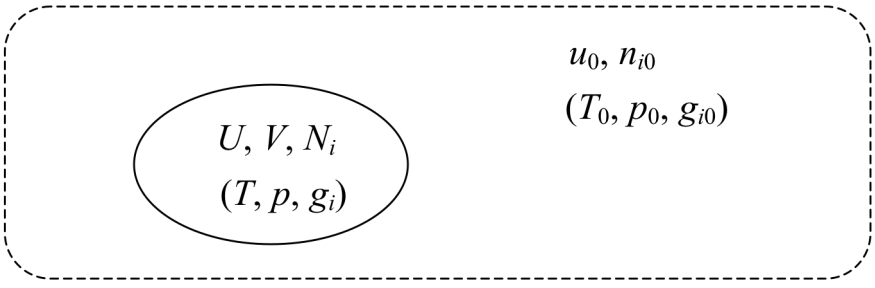


Figure 5.4. We consider a system om volume V, with an expected internal energy U and M different molecular species, each with an expected number of molecules N_i. The system is placed in a large environment characterised by a certain energy density u_0 and certain molecular densities (concentrations). Both the system and the enviornment can also be characterised by their corresponding intensive variables, temperature T, the pressure p, and the chemical potentials g_k.

We will consider a system characterised by a certain volume V, and M different molecular species. The system is part of a larger surrounding equilibrium system with possibly different energy per unit volume as well as different concentrations n_i (i=1,…, M), as indicated in Fig. 5.4. Our knowledge about the studied system of volume V is that its composition is given by expectation values (or averages) N_i (i=1,…, M) of the number of molecules of the different species. We also know that the the expectation value U of the internal energy is known.

This is the type of situation discussed in the Section 2.2 on the maximum entropy formalism. We do have some knowledge about the system in terms of averages, or expectation values, of certain properties of the system, but we do not know the microstate, neither the probability distribution over the microstates. The maximum entropy formalism can now be applied to derive a distribution over the microstates. Like in Section 2.2, the maximisation problem can be formulated as

Choose $P = \{p_i\}_{i\in\{\text{microstates}\}}$ so that
the entropy $S[P] = \sum_i p_i \ln \frac{1}{p_i}$ is maximised, subject to
the energy constraint, $\sum_i p_i h(i) = U$,
the molecular number constraints, $\sum_i p_i f_k(i) = N_k, \quad k = 1,...,M$, and
the normalisation constraint, $\sum_i p_i = 1$. (5.14)

Here the microscopic function $h(i)$ determines the energy of a certain microstate i, and in average that should equal internal energy U. Similarly, the function $f_k(i)$ determines the number of molecules of species k in microstate i, which in average should equal N_k. In order to solve the problem and to derive some important relations, this level of description is sufficient, and we do not need to specify these functions further.

Following the Lagrangian formalism in Section 2.2, we introduce Lagrangian variables: β for the energy constraint, λ_k (k=1,…, M) for the M different constraints on number of molecules, and μ–1 for the normalisation constraint. The solution is then a Gibbs distribution, cf., Eq. (2.15),

$$p_i = \exp\left(-\mu - \beta\, h(i) - \sum_k \lambda_k f_k(i)\right), \tag{5.15}$$

The entropy of the Gibbs distribution is easily determined,

$$S = \mu + \beta U + \sum_{k=1}^{M} \lambda_k N_k \ . \tag{5.16}$$

At this point, we can determine how the Langrangian variable μ depends on volume V. If we multiply the system by a certain factor, increasing all extensive variables (energy, number of molecules), for example by putting together two identical system (i.e., multiplying with a factor of 2), then all variables S, U, and N_k, will increase with the same factor. This means that also μ must have such a linear dependence on volume: $\mu = \mu(V) = \mu_1 V$, where μ_1 does not depend on volume. If we now also use the relation between thermodynamic entropy and the information-theoretic one (with information expressed in natural units), $S_\mathrm{T} = k_\mathrm{B} S$, we can write Eq. (5.16),

$$\frac{S_\mathrm{T}}{k_B} = \mu_1 V + \beta U + \sum_{k=1}^{M} \lambda_k N_k \ . \tag{5.17}$$

By using this equation together with the thermodynamic definitions on temperature T, pressure p, and chemical potential g_k, from Eqs. (5.5-5.7), we find that the Lagrangian variables can be expressed in these intensive variables,

$$\left(\frac{\partial S_\mathrm{T}}{\partial U}\right)_{V,N_i} = k_B \beta = \frac{1}{T} \quad \rightarrow \quad \beta = \frac{1}{k_B T} \ , \tag{5.18}$$

$$\left(\frac{\partial S_\mathrm{T}}{\partial V}\right)_{U,N_i} = k_B \mu_1 = \frac{p}{T} \quad \rightarrow \quad \mu_1 = \frac{p}{k_B T} \ , \tag{5.19}$$

$$\left(\frac{\partial S_\mathrm{T}}{\partial N_j}\right)_{U,V,N_{i\neq j}} = k_B \lambda_j = -\frac{g_j}{T} \quad \rightarrow \quad \lambda_j = -\frac{g_j}{k_B T} \ . \tag{5.20}$$

By inserting this in Eq. (5.17), we get the *Gibbs equation*,

$$U = T S_\mathrm{T} - pV + \sum_{k=1}^{M} g_k N_k . \tag{5.21}$$

By using the maximum entropy formalism to derive a probability distribution over microstates, the Gibbs distribution, we have derived one of the fundamental relations in thermodynamics.

5.3.1 Comparing two different Gibbs distributions

There are often reasons for comparing two different Gibbs distributions, for example, when a smaller physical system, characterised by a certain Gibbs distribution P, deviates from an environment (equilibrium) system characterised by the distribution P_0 that would describe the system if it would be in equilibrium with the environment. In that equilibrium situation, the system would have other values on its extensive variables (energy and number of molecules of different kinds). With the notation from Fig. 5.4, we would have for the internal energy $U_0 = u_0 V$, for the number of molecules $N_k = n_k V$. The environment (and the system in equilibrium with it) would also be characterised by intensive variables, the temperature T_0, the pressure p_0, and the chemical potentials g_{k0} (k=1,…, M). Again, using the maximum entropy

formalism to derive the distribution describing the equilibrium situation, the distrbution P_0, characterised by the probabilities

$$p_{i0} = \exp\left(-\mu_0 V - \beta_0\, h(i) - \sum_k \lambda_{k0} f_k(i)\right) . \tag{5.22}$$

Since the constraints for the equilibrium distribution is different, we have other Lagrnagian variables now, as indicated by the subscript 0. Like in Eqs. (5.18-20), we can replace the Lagrangian variables for the equilibrium situation with the corresponding intensive thermodynamic variables,

$$\beta_0 = \frac{1}{k_B T_0} , \tag{5.23}$$

$$\mu_0 = \frac{p_0}{k_B T_0} , \tag{5.24}$$

$$\lambda_{j0} = -\frac{g_{j0}}{k_B T_0} . \tag{5.25}$$

One can now ask the question: If we *a priori* assume that the system (in volume V) will be in equilibrium, characterised by the Gibbs distribution of Eq. (5.22), but after observation learn that the correct description is the non-equilibrium Gibbs distribution of Eq. (5.15), how much information have we gained? Below we will relate such an information gain with thermodynamic properties of the non-equilibrium situation.

The Kullback information $K[P_0; P]$ between the *a priori* equilibrium description and the observed one, both being Gibbs distributions, can be written

$$\begin{aligned} K[P_0;P] &= \sum_i p_i \left(-\mu_1 V - \beta\, h(i) - \sum_{k=1}^{M} \lambda_k f_k(i) + \mu_0 V + \beta_0\, h(i) + \sum_{k=1}^{M} \lambda_{k0} f_k(i)\right) = \\ &= (\mu_0 - \mu_1)V + (\beta_0 - \beta)U + \sum_{k=1}^{M} (\lambda_{k0} - \lambda_k) N_k = \\ &= \frac{1}{k_B T_0}\left(p_0 V + U - \sum_{k=1}^{M} g_{k0} N_k\right) - \frac{S_\mathrm{T}}{k_B} \end{aligned} . \tag{5.26}$$

where we in the last step have used Eq. (5.17). If we multiply by $k_B T_0$, we get

$$k_B T_0 K[P_0;P] = U + p_0 V - T_0 S_\mathrm{T} - \sum_{k=1}^{M} g_{k0} N_k . \tag{5.27}$$

The expression on the right-hand side quantifies the amount of work that can be extracted from a process that brings the system characterised by the distribution P (or the extensive variables U, V, and N_i) into equilibrium with the environment characterised by P_0 (or the

corresponding intensive variables T_0, p_0, and g_{i0}). Willard Gibbs introduced this general form of free energy already in 1875. This energy is often called the *exergy* of the system in the given environment. We will use the term *exergy* of a system as the most general form of free energy, which is then defined by the maximum amount of work that can be extracted when the system is reversibly brought to equilibrium with its environment. Thus, the exergy E equals the Kullback information between the equilibrium state and the actual one, multiplied with $k_B T_0$,

$$E = k_B T_0 K[P_0;P] \ . \tag{5.28}$$

By combining the expression for exergy, Eq. (5.27), with the Gibbs equation (5.21) for internal energy, the exergy can be written

$$E = S_{\mathrm{T}}(T - T_0) - V(p - p_0) + \sum_{i=1}^{M} N_i(g_i - g_{i0}). \tag{5.29}$$

5.3.2 Information and free energy in non-equilibrium concentrations

Let us now introduce chemical concentrations, $c_i = N_i/V$ and $c_{i0} = N_{i0}/V$, for the system and its environment, respectively. Here we will focus on informtion and thermodynamic characteristics of a system that deviates from equilibrium in its concentrations of different molecules. Suppose, therefore, that the system has the same temperature and pressure as the environment, $T = T_0$ and $p = p_0$. Further, we assume that the chemical potential can be written as for an ideal solution,

$$\frac{g_i}{k_{\mathrm{B}}T} = C + \ln c_i \ . \tag{5.30}$$

Here, C is a constant. Then the exergy can be written as a Kullback information between the concentration distribution in the equilibrium state (environment) and in the system, using the normalised concentrations $c_i V/N_i$ and $c_{i0} V/N_{i0}$, respectively,

$$E = k_{\mathrm{B}}T_0 V \sum_{i=1}^{M} c_i \ln \frac{c_i}{c_{i0}} = k_{\mathrm{B}}T_0 N K[c_0 V/N; cV/N] \ , \tag{5.31}$$

where N is the total number of molecules in the system, $N = \sum_k N_k$, and c and c_0 denotes the concentration distributions. We have now written the exergy as a Kullback information again, but now on a macroscopic level, since it is based on the concentrations observed in the system (and in the environment). This Kullback information can serve as a starting point for examining spatial structure in chemical systems. It has the advantage that it is in an information-theoretic form at the same time as it connects to statistical mechanics and thermodynamics through the basic concept of exergy. The second law of thermodynamics tells us that, for a closed system, the entropy increases (or is constant), or, equivalently, that the exergy decreases (or stays constant). Such physical restrictions may be important in the

description and analysis of chemical systems that exhibit self-organisation. We shall return to such an analysis in a later chapter.

5.4 Microscopic and macroscopic entropy

The second law of thermodynamics originates from Clausius (1850) and Kelvin (1852), who found a thermodynamic quantity which is non-decreasing in time, and Clausius (1865) introduced a term for it — *entropy*. At this time, statistical mechanics had not been established, but there was a need for an understanding of the relation between thermodynamics and microscopic properties of gases. A first step was taken by Krönig (1856) who discussed macroscopic phenomena in terms of microscopic properties. Soon Clausius, and later Maxwell and Boltzmann developed a kinetic theory for gases.

A problem which remained unsolved was how to understand the second law of thermodynamics from kinetic gas theory. In an attempt to solve this, Boltzmann (1872) introduced a function of the microstate in a gas, the H-function, which under certain assumptions was proven to be decreasing in time until the system reaches an equilibrium given by the Maxwell-Boltzmann distribution law for molecular velocities. This is called the H-theorem, and the idea was that this should correspond to the second law of thermodynamics, with the H-function being equal to the entropy with opposite sign. This approach met with difficulties, especially when it was applied to other systems than dilute gases, which led to the development of *ensemble theory* by Gibbs (1902) and Einstein (1902, 1903).

In ensemble theory one considers a large number of systems, in which certain macroscopic variables are known, either as an average or exactly. In the *microcanonical ensemble* all systems (elements in the ensemble) have exactly the same energy, while the *canonical ensemble* only prescribes an average energy for the whole ensemble. A probability p_i is associated with each microstate in the ensemble, which is then described by a distribution $P = \{ p_i \}$, and the entropy is defined as $k \Sigma p_i \ln(1/p_i)$, i.e. an ensemble average of $k \ln(1/p_i)$. Quantities like energy or number of particles can, of course, be interpreted both at the microscopic and at the macroscopic level. But for entropy it seems at first that it is a macroscopic property only, and that it is more difficult to associate it with a certain microstate. The physical state, the macrostate, is defined to be the probability distribution over the microstates, although a single physical system will always be found in a single microstate.

In our discussion of entropy in symbol sequences, based on internal statistics of a single but very long sequence (microstate), we have considered that entropy as a characterisation of the internal randomness or the disorder. It then seems plausible that there should exist a microscopic property that corresponds to the macroscopic entropy, and hence also yields the other thermodynamic properties of the system, at least of the system (or the microstate) is large enough to provide sufficient internal statistics. In fact, one can show under some, fairly general, conditions that a system in which each microstates can be described as a symbol sequence σ, the (internal) randomness in form of the entropy $s(\sigma)$ serves as a microscopic

entropy $s_{\text{micro}}(\sigma)$. The ensemble average of the entropy $s(\sigma)$ for an individual microstates σ is, in the thermodynamic limit, equal to the thermodynamic entropy s per symbol (times a constant),

$$s = k_{\text{B}} \sum_{\sigma} p(\sigma)\, s(\sigma), \tag{5.32}$$

where the summation is over all microstates in the ensemble. In fact, for almost all microstates in the ensemble, the microscopic entropy equals the thermodynamic entropy,

$$s = s(\sigma)\text{, for almost all } \sigma. \tag{5.33}$$

This can be understood if one realises that each microstate contains all correlations that are necessary for determining the whole ensemble. The internal correlations of a microstate give restrictions to its randomness, expressed by its (internal) entropy $s(\sigma)$, as well as to the uncertainty of the macrostate, expressed by the thermodynamic entropy S. (In the thermodynamic limit the exceptions to this is of measure zero.)

In conclusion, the thermodynamic entropy, which is usually interpreted as a measure of disorder or uncertainty on which microstate a certain physical occupies, also has an interpretation on the micro level. The microscopic entropy $s(\sigma)$ quantifies the internal disorder in the microstate σ on the basis of internal correlations.

5.4.1 Microscopic entropy in spin systems

In this section, we shall illustrate how the microscopic entropy can be used to determine the equilibrium state in simple one-dimensional spin systems.

In a previous section, we used the maximum entropy formalism to determine the macrostate, the probability distribution, which characterises a certain physical system. Here we shall apply the maximum entropy formalism on the microscopic entropy of a spin system, or a system that can be described as a symbol sequence. This should then result in a probabilistic description of blocks of symbols in the microstate. Again, the maximisation is done under certain constraints, usually in the form of an average energy in the system. There are also constraints regarding typical properties that hold for probabilities of symbol sequences, for example, regarding summation over last or first index, see Eqs. (3.5) and (3.6).

In a spin system, the local interactions between spins (symbols) give rise to certain energy contributions, depending on the specific local configuration. This is usually characterised by an energy function that associates a certain energy value $h(x_1,\ldots,x_n)$ for each sequence of spins $x_1,\ldots,x_n$ within the interaction distance n. A given value of the internal energy u (per spin) in the system, then results in the constraint

$$\sum_{x_1 \ldots x_n} p(x_1,\ldots,x_n) h(x_1,\ldots,x_n) = u. \tag{5.34}$$

The maximisation problem then takes the form: Find a set of probability distributions P_m over m-length sequences so that the entropy

$$s = \lim_{m\to\infty} \Delta S_m \tag{5.35}$$

is maximised, under the energy constraint (5.34).

This may look very complicated, but if there are no restrictions on longer sequences than interaction distance n, as stated by (5.34), then the entropy s equals ΔS_n, which means that we need only consider sequences of length up to n. In order to see this, let us suppose that we get convergence only at ΔS_m, with $m > n$. Then this means that there is correlation information over blocks of length m, i.e., $k_m > 0$. But this is an unnecessary correlation information that reduces the maximum entropy, since we may reduce k_m to zero by selecting the m-block probability as $p(x_1,\ldots, x_n) = p(x_1,\ldots, x_{n-1})\, p(x_2,\ldots, x_n) / p(x_2,\ldots, x_{n-1})$, see Eq. (3.17), without affecting the energy constraint. This procedure may then be repeated down to length n, showing that it is sufficient to examine

$$\max \Delta S_n\,, \tag{5.36}$$

under the constraints above.

As a more specific example, consider the one-dimensional Ising model (without an external field). Suppose that our system is composed by an infinite sequence of spins, pointing either up or down,

$$\ldots\uparrow\downarrow\downarrow\downarrow\uparrow\downarrow\downarrow\uparrow\uparrow\uparrow\downarrow\downarrow\uparrow\uparrow\uparrow\uparrow\downarrow\downarrow\downarrow\uparrow\downarrow\uparrow\uparrow\downarrow\downarrow\downarrow\downarrow\downarrow\uparrow\uparrow\uparrow\uparrow\downarrow\uparrow\uparrow\downarrow\downarrow\downarrow\ldots,$$

and that the energy in the system is determined by nearest neighbour interactions only: Configurations $\downarrow\downarrow$ and $\uparrow\uparrow$ contribute with energy $-J$, while $\uparrow\downarrow$ and $\downarrow\uparrow$ contribute with $+J$. Since h only depends on pairs of spins, it is sufficient to find maximum of ΔS_2, in order to determine the probability distributions. Then there are three probabilities to consider: $p_0 = P(\downarrow\downarrow)$, $p_1 = P(\downarrow\uparrow) = P(\uparrow\downarrow)$, and $p_2 = P(\uparrow\uparrow)$, where we have used the symmetry $P(\downarrow\uparrow) = P(\uparrow\downarrow)$ that must hold in an infinite binary symbol sequence. Probabilities for single spins can then be written $P(\downarrow) = p_0 + p_1$ and $P(\uparrow) = p_1 + p_2$.

To solve the maximisation problem, we introduce the Lagrange function

$$L(p_0, p_1, p_2, \lambda, \mu) = \Delta S_2 + \beta(u - J(2p_1 - p_0 - p_2)) + \mu(1 - p_0 - 2p_1 - p_2). \tag{5.37}$$

Here β is the Lagrange variable related to the energy constraint, while μ is related to the normalisation constraint. The block entropy difference can be written

$$\Delta S_2 = p_0 \ln\frac{p_0 + p_1}{p_0} + p_1 \ln\frac{p_0 + p_1}{p_1} + p_1 \ln\frac{p_1 + p_2}{p_1} + p_2 \ln\frac{p_1 + p_2}{p_2}\,. \tag{5.38}$$

The derivation of the following solution is left as an exercise

$$p_0 = p_2 = \frac{1}{2(1+e^{-2\beta J})}, \tag{5.39}$$

$$p_1 = \frac{1}{2(1+e^{2\beta J})}. \tag{5.40}$$

Here we have chosen to keep the Lagrange variable β, instead of the energy u. As in the general Gibbs distribution, see (5.18), it is related to temperature, $\beta = (k_B T)^{-1}$. It is clear that in the limit $T \to 0$, spins are arranged in parallel, $p_0 = p_2$, while in the limit of $T \to \infty$, all p are equal, and we get a completely disordered state.

5.5 Exercises

5.1 **A small spin system.** Consider a system of 2x2 spins, in which a microstate is a configuration like the one below.

Each spin has two neighbours, as the example indicates, one in the vertical direction and one in the horizontal direction. Each such neighbour interaction contributes with an energy +J for parallel spins and –J for anti-parallel spins (so that in the example above the energy is 0). If the average energy is u, what is the equilibrium distribution over microstates? (Use the maximum entropy formalism. You may give the answer as a function of temperature instead of energy.)

5.2 **Spin system.** Suppose that in a one-dimensional discrete spin system, described by a row of spins, up or down, the interaction is with third nearest neighbours only (position x interacts with position $x + 3$). Parallel spins at this distance contribute with an energy –J and anti-parallel spins with an energy +J. What is the equilibrium state of this system? Use the maximum entropy formalism, and express the probabilities that characterise the equilibrium as functions of temperature (or of β).

5.3 Consider a one-dimensional spin system with 4 states per position: ←, →, ↓ och ↑. Interaction is with nearest neighbours only, such that parallel spins contribute with energy $-J < 0$, anti-parallel spins with the energy $J > 0$, while perpendicular spins give no energy contribution. Characterise an equilibrium distribution as function of temperature T. What is the entropy in the limit $T \to 0$? (Make use of the fact that finite interaction distance in one dimension does not allow for any phase transitions, i.e., all symmetries are kept in the equilibrium description.)

5.4 Consider a one-dimensional spin system with 2 spin states: ↓ och ↑ . Interaction is with both nearest and next nearest neighbours, so that parallel spins at distance 1 (e.g., ↑↑) contribute with the energy $-J<0$ and anti-parallel spins with energy J, while parallel spins at distance 2 (e.g., ↑_↑) contribute with energy J and anti-parallel spins with energy $-J$. What is the equilibrium description? (The solution needs not be explicit, but it is sufficient to determine the equations from which the probabilities can be derived.) What is the entropy in the zero temperature limit ($T \to 0$), and how does a typical state look like in that situation?

5.5 **Happy agents.** Consider an infinite one-dimensional lattice system (of cells) in which each cell may be inhabited by a red (R) individual, a blue (B) individual, a pair (RB of agents with different colours), or the cell may be empty. Assume equal densities of individuals R and B of 1/4 each. Each individual has a happiness level being the sum of the happiness from the relation with the closest neighbours. If there is a **pair in a cell** the happiness of that cell is $4H$ (with H being a positive "happiness" constant), and then there is no contribution from the neighbouring cells. If there is a **single individual in a cell**, the happiness of that individual get a contribution of $H/2$ from each single living neighbour of opposite colour (to the left and to the right, but no contribution from neighbour pairs). Empty cells do not contribute to happiness.

If you know that the average happiness is w, how would you guess that the system looks like in equilibrium, using information-theoretic arguments. <u>You may answer in terms of a set of equations that you need *not* solve.</u> Discuss what happends if the "temperature" is low (limit of zero temperature), with the interpretation that a low temperature corresponds to a high "happiness". What is the entropy in this limit?

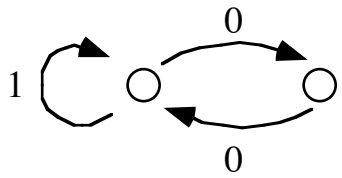
(If you prefer, all this can be thought of as molecules A and B that may aggregate to a larger molecules AB, with the interpretation of H as a negative interaction energy constant.)

5.6 **An infinite "spin" system.** Consider an infinite one-dimensional lattice system in which each position can be in one of three possible states: A, B, or C. Assume that neighbouring symbols contribute with energy +J if they are the same or if they are A and C (i.e., AA, BB, CC, AC, and CA) while all other neighbour combinations contribute with –J, where J is a positive constant.

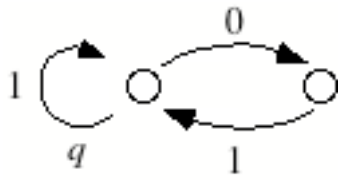
(a) If the average energy per position is u, what is the equilibrium distribution? Use the maximum entropy formalism to determine the equations that give the probabilistic description. **You need not solve the equations derived.**

(b) Calculate the entropy for zero temperature. (This can be done without solving the full problem in a.)

5.7 Could the automaton below possibly represent the equilibrium state in a one-dimensional spin system, in which the average internal energy u fulfils the constraint $u = \sum_{i_1 \ldots i_m} p(i_1 \ldots i_m)\, h(i_1 \ldots i_m)$?



Suppose instead that the automaton that represents the equilibrium state is as follows:



What spin system (or other one-dimensional system with energy constraints) may this automaton represent? (Must be shown.) How does the probability q connect to inverse temperature β and/or interaction energy constant J ?

5.8 Suppose that in a one-dimensional discrete system, described by a row of "land pieces" or cells, there live these "dot" agents. By inspecting what they are doing you quickly find the following characteristics. There is a density ρ of "dots", some of them living alone at a "piece of land", and some of them that have joined in a "marriage" so that two share the same "piece of land". You also find that whenever there is a married couple in one cell, the adjacent cells (to the left and to the right) are always free — the married couple really want to be on their own or others avoid them. You also find by studying several systems of this kind that the fraction of "dots" that are married is a certain constant α.

Now, before analysing more in detail the statistics of the "dot world", you want to design a probabilistic description of the system that is consistent with the observations above and that has a maximum in its (Shannon) entropy.

Use the maximum entropy formalism to find a probabilistic description that obeys the constraints described above, including the parameters ρ and α. Assume that the system is of infinite length.

5.9 **Dots are back.** Consider an infinite two-dimensional lattice system (of square cells) in which dots are distributed. The cells can be either empty or inhabited (by one or two dots). Free living dots have an "energy level" of 0, while pairs (two dots in one cell) have the energy –J (where J is a positive constant). Assume that there is a certain density of dots ρ, and that there is a given average energy u (per cell). Use the maximum entropy formalism to characterise an "equilibrium" state of this system as a function of an "inverse temperature" (β). What is the entropy? What happens in the limit of a "zero temperature"?

6 Geometric information theory

In the first chapters we have focused on an informaiton perspective in which the total information of the studied system can be decomposed into contributions from different correlation lengths as well as the remaining random information, the entropy. In the present Chapter, we are aiming for a different decomposition of the ordered information. We ask the questions: At what *length scales* do we find information? At what *positions* in the system is information located? We will consider systems in a continues space of arbitrary dimension. The application that we have in mind is to pattern formation in chemical systems that will be treated in Chapter 7. Another application is to the analysis of images, but that will not be part of this course.

6.1 Information decomposition with respect to position and resolution

A chemical self-organising system may build up spatial structure in the form of concentration variations. Information-theoretic quantities for analysing such structure formation may be based on probability distributions (densities) that correspond to the spatially distributed concentrations in the system. In order to be able to analyse how information is distributed over different length scales, we introduce a "resolution" operator that can be used to modify how sharp a distribution appears. The goal is to present a formalism that can be used to decompose the total information in a pattern into contributions from both different positions and different length scales.

6.1.1 Resolution dependent probability density

In our analysis we will use a normalised probability density $p(x)$, as in Eq. (3.27-28),

$$\int_{-\infty}^{\infty} dx\, p(x) = 1 \,. \tag{6.1}$$

In case of picture, $p(x)$ can be the light intensity of a certain color, or for a chemical system, the concentration profile for a certain component; in both cases scaled to a normalised probability density.

The formalism is here presented for a one-dimensional system, but the extension to higher dimensions is straightforward. When applied to systems of finite length L, the integral limits are 0 and L, respectively. We also have a reference, or *a priori*, distribution $p_0(x)$. In case of a finite system, we usually assume that the *a priori* description is a uniform distribution.

The amount of information that we get when we observe a spatial distribution $p(x)$ describing the spatial structure of the system is given by the Kullback information with the a priori $p_0(x)$ as the reference,

$$K[p_0;p] = \int dx\, p(x) \ln \frac{p(x)}{p_0(x)} \ . \tag{6.2}$$

We now introduce the resolution dependent distribution, where a resolution parameter r determines how well the original distribution p can be detected. The resolution dependent distribution is the result of a *Gaussian blur* operation applied to the original one — an operation available in various types of image analysis software. Mathematically, the reduced resolution is achieved by taking the convolution of the original distribution with a Gaussian of width r,

$$p(r;x) = \frac{1}{\sqrt{2\pi}\, r} \int_{-\infty}^{\infty} dw\, e^{-w^2/2r^2}\, p(x-w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dz\, e^{-z^2/2} p(x+rz). \tag{6.3}$$

Parameter $r = 0$ means perfect resolution while $r \to \infty$ leads to a Gaussian with increasing width approaching a completely uniform pattern. Any of the two expressions in Eq. (6.3) can be used to calculate the resolution dependent distribution. For some of the theory we develop, it will be useful to express the Gaussian blur operation as a differential operator. This can be achieved by making a series expansion of $p(x + rz)$ as follows,

$$\begin{aligned} p(r;x) &= \frac{1}{\sqrt{2\pi}} \int dz\, e^{-z^2/2} p(x+rz) = \\ &= \frac{1}{\sqrt{2\pi}} \int dz\, e^{-z^2/2} \left(p(x) + rz\, p'(x) + \frac{(rz)^2}{2} p''(x) + \ldots \right) = \\ &= \exp\left[\frac{r^2}{2} \frac{d^2}{dx^2} \right] p(x). \end{aligned} \tag{6.4}$$

The exponential function in the final expression should be viewed as a series expansion, involving terms of even powers of the differential operation with respect to x. By taking the derivative of the last expression with respect to r we find that the probability density fulfills

$$\left(-r \frac{\partial}{\partial r} + r^2 \frac{d^2}{dx^2} \right) p(r;x) = 0 \ . \tag{6.5}$$

It is also clear that the following properties hold,

$$\begin{aligned} &p(r;x) > 0, \\ &p(0;x) = p(x), \\ &\int dx\, p(r;x) = 1, \\ &\int dx\, x\, p(r;x) = \int dx\, x\, p(0;x), \end{aligned} \tag{6.6}$$

where the last one states that the Gaussian blur does not shift the “centre of mass” of the distribution.

As a first example, consider a Gaussian distribution of width b,

$$p(x) = \frac{1}{\sqrt{2\pi}\, b} \exp\left(-\frac{x^2}{2b^2} \right). \tag{6.7}$$

By using Eq. (6.3), we find that the resolution dependent distribution $p(r; x)$ is another Gaussian with a larger width $(b^2 + r^2)^{1/2}$,

$$p(r;x) = \frac{1}{\sqrt{2\pi}\, \sqrt{b^2 + r^2}} \exp\left(-\frac{x^2}{2(b^2 + r^2)} \right). \tag{6.8}$$

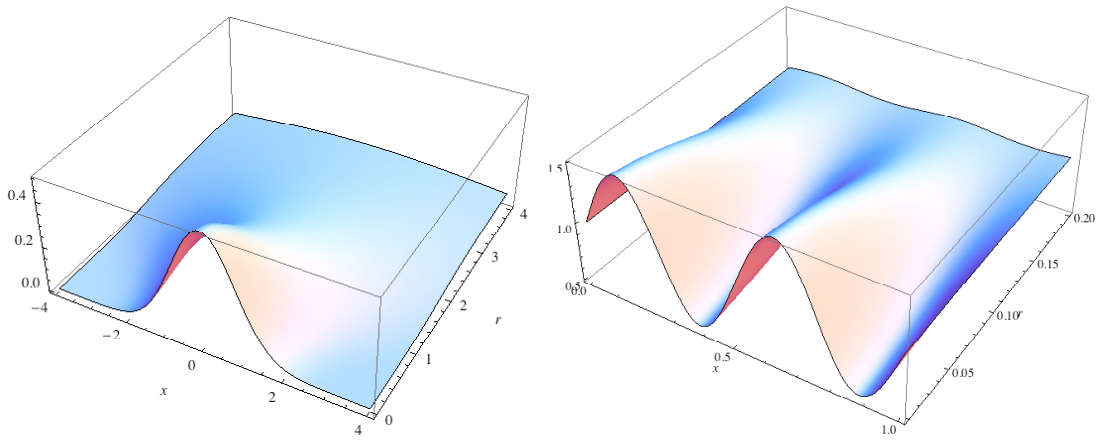


Figure 6.1 (a) The Gaussian probability density $p(r; x)$ as function of position x and resolution r. (b) Same plot for sinusoidal probability density. Periodic boundary conditions are assumed at x=0 and x=1.

For another example, consider a sinusoidal probability density,

$$p(x) = 1 + \tfrac{1}{2} \sin(4\pi x) \ , \tag{6.9}$$

on the interval $x \in [0, 1]$, which makes it normalised. One can show, using Eq. (6.3) and assuming periodic boundary conditions, that the resolution dependent density in this case takes the form

$$p(r;x) = 1 + \tfrac{1}{2} e^{-8\pi^2 r^2} \sin(4\pi x) \ . \tag{6.10}$$

The example distributions are illustrated in Fig. 6.1.

The worsening of the resolution can be seen as a diffusion process applied to the resolution dependent distribution $p(r; x)$. If we make a variable transformation in Eq. (6.5), replacing $r^2/2$ by t, we get a diffusion equation,

$$\frac{d}{dt} p(t;x) = \frac{d^2}{dx^2} p(t;x) \ . \tag{6.11}$$

This means that the distribution $p(r; x)$ can be derived by running a diffusion process the time $t = r^2/2$. In this perspective, Fig. (6.1) illustrates how diffusion would affect the original distributions as time goes on along the r axis.

6.1.2 Decomposition of information

We will now consider the total information in the pattern, expressed by the Kullback information in Eq. (6.2), and we will derive a decomposition of this quantity into contributions from different positions and different length scales, or resolution levels r.

First we assume that the resolution dependent distributions $p(r; x)$ and $p_0(r; x)$ are indistinguishable in the limit $r \to \infty$,

$$\lim_{r\to\infty} \frac{p(r;x)}{p_0(r;x)} = 1. \tag{6.12}$$

That this is a reasonable assumption can be seen from the fact that the ratio of any two Gaussian distributions (with centres of mass being finitely separated) always approches one when the resolution r goes to infinity, cf. Eq. (6.8).

We can use these relations to make a decomposition of the Kullback information, Eq. (6.2), with respect to both position and resolution length,

$$\begin{aligned} K[p_0;p] &= K[p_0(0;\bullet);p(0;\bullet)] - K[p_0(\infty;\bullet);p(\infty;\bullet)] = \\ &= -\int_0^\infty dr \frac{\partial}{\partial r} K[p_0(r;\bullet);p(r;\bullet)] = \\ &= -\int_0^\infty dr \int dx \; \frac{\partial}{\partial r}\left(p(r;x) \ln \frac{p(r;x)}{p_0(r;x)} \right) = \\ &= \int_0^\infty dr\, r \int dx \left[\frac{p(r;x)}{p_0(r;x)} \frac{d^2}{dx^2} p_0(r;x) - \left(1 + \ln \frac{p(r;x)}{p_0(r;x)} \right) \frac{d^2}{dx^2} p(r;x) \right]. \end{aligned} \tag{6.13}$$

In the last step we have used the the equivalence between $\partial/\partial r$ and $r\, d^2/dx^2$ from Eq. (6.5). By partial integration of the last expression this can be rewritten as

$$\begin{aligned} K[p_0;p] &= \int_0^\infty \frac{dr}{r} \int dx \; k(r,x) = \\ &= \int_0^\infty \frac{dr}{r} \int dx \; p(r;x) \nu(r,x)^2, \end{aligned} \tag{6.14}$$

where

$$\nu(r;x) = r \frac{d}{dx} \ln \frac{p(r;x)}{p_0(r;x)} \; . \tag{6.15}$$

It is then clear that Eq. (6.14) is a double decomposition of the Kullback information with respect to both resolution and position. The "local" weighted information $k(r, x)$ is always positive. In the next chapter on chemical self-organising systems, we will use this formalism to study the flow of information, both in space and in length scales. In the next section on fractals, we will relate resolution dependent information quantities to the concept of fractal dimension.

In most cases we will consider a reference distribution that is uniform, at least approximately. For a limited system, i.e., a finite interval of the real axis, this is reasonable to assume. This is also the case that will be studied for chemical pattern formation in the next Chapter. For an unbounded space, we typically assume p_0 to be a very broad Gaussian, which can be approximated with a uniform distribution over the region where p typically exhibits spatial pattern. In these situations, assuming a constant p_0, the local information density $k(r, x)$ can be expressed as

$$k(r,x) = r^2 p(r;x)\left(\frac{d}{dx}\ln p(r;x)\right)^2 . \tag{6.16}$$

From this expression, it is clear that the locla information is high at the slopes of the probability density. For an image this means that information is primarily located at the edges of the objects, with edges being defined as slopes in the intensity (or probability density) p.

For the examples discussed above, we find, using Eq. (6.16), the for the Gaussian, the information density takes the form,

$$k_{\text{Gaussian}}(r,x) = \frac{1}{(2\pi)^{3/2}(b^2+r^2)^{7/2}} r^2 x^2 \exp\left(-\frac{3x^2}{2(b^2+r^2)}\right), \tag{6.17}$$

and for the sinusodial density, we get

$$k_{\text{sinusoidal}}(r,x) = 4e^{-16\pi^2 r^2}\pi^2 r^2 \cos(4\pi x)^2\left(1+\tfrac{1}{2}e^{-8\pi^2 r^2}\sin(4\pi x)\right) . \tag{6.18}$$

For both these examples, it is clear, see Fig. 6.2, that there are certain length scales and positions for which we have a higher information density. These length scales are then related to the width of the structure.

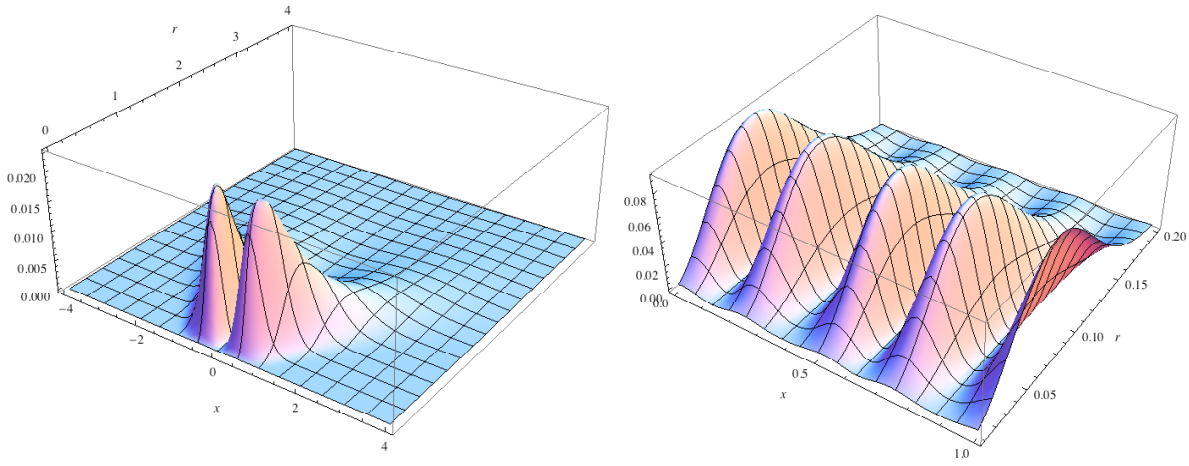


Figure 6.2 The information density $k(r, x)$ for the two examples of Fig. 6.1: (a) the Gaussian probability density, and (b) the sinusoidal probability density.

6.2 Fractals patterns, dimension, and information

6.2.1 Dimensions

The word "dimension" has several definitions in mathematics. Behind the length of the coastal line of Sweden, there may be several dimension numbers hiding. A line or a curve can be considered a one-dimensional object — characterised by the fact that it can be divided into two objects by removal of a point. At the same time the object may be placed in a space of higher dimension. Furthermore, even if the coastal line is one-dimensional and constrained by a two- or three-dimensional space, it may be so irregular that its length is infinite. If a curve is sufficiently dwindling it may be sufficiently close to any point in a two-dimensional object, and one may consider the curve two-dimensional.

The first dimension number we mentioned is related to the topology of an object — a curve may be divided by removing a point, a surface may be divided by removing a curve, etc — and therefore we call this the *topological dimension*, D_T, of the object. The object is placed in a space with a certain *Euclidean dimension*, D_E. The third dimension number is related to how dwindling, for example, a curve is, and as we have indicated this number may be larger than the topological dimension of the object. This number is actually defined so that it does not need to take only integer numbers, and therefore it is called *the fractal dimension*, D_T, of the object. An object is said to be *fractal* if its fractal dimension is larger than the topological dimension, $D_F > D_T$. In general, for these dimension numbers, we have

$$D_T \le D_F \le D_E \tag{6.19}$$

There is also another perspective on the dimension of an object, related to the resolution with which the object is observed. The following example, by Benoit Mandelbrot (1983), illustrates this:

> "… a ball of 10 cm diameter made of thick thread of 1 mm diameter possesses (in latent fashion) several effective dimensions. To an observer placed far away, the ball appears as a zero-dimensional figure: a point. (…) As seen from a distance of 10 cm resolution, the

ball of thread is a three-dimensional figure. At 10 mm it is a mess of one-dimensional threads. At 0.1 mm, each thread becomes a column and the whole becomes a three-dimensional figure again. At 0.01 mm, each column dissolves into fibers, and the ball again becomes one-dimensional, and so on, with the dimensionality crossing over repeatedly from one value to another. When the ball is represented by a finite number of atomlike pinpoints, it becomes zero-dimensional again."

6.2.2 Fractal dimension

In the following, we sketch on a definition of the fractal dimension, based on how characteristic lengths of objects scale when resolution is changed. Suppose that we have an object in a D_E-dimensional space, and that we want to cover this with spheres of a certain radius r, corresponding to the resolution. Let us assume that we find that $N(r)$ spheres are needed for this, where, of course, a smaller r requires a larger number of spheres, see Figure 7.1.

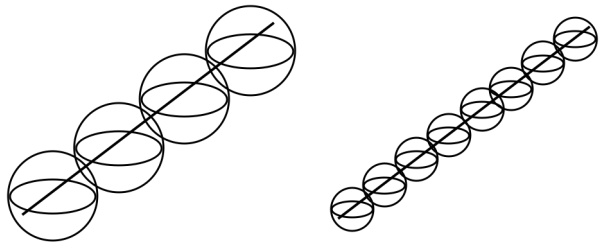


Figure 6.2. The number of spheres with radius r that are needed to cover a curve is proportional to r^{-D}, where D is the fractal dimension of the curve. In the figure, the number of spheres doubles as the radius is reduced to ahlf the original, implying a curve dimension $D = 1$.

If the object is a straight line the number of spheres $N(r)$ will be proportional to $1/r$, see Figure 6.2, while in the case of a flat two-dimensional surface $N(r)$ will be proportional to $1/r^2$. We introduce the resolution dependent *fractal dimension*, $D(r)$, characterised by

$$N(r) \sim r^{-D(r)} \tag{6.20}$$

We use this relation to define the fractal dimension by

$$D(r) = -\frac{\partial \ln N(r)}{\partial \ln r} \tag{6.21}$$

In general, the fractal dimension depends on the resolution r, but in most constructed objects found in the literature on fractal patterns there is a scale invariance that make the fractal dimension independent of the resolution, see, for example, Figure 6.3.

The fractal dimension may also be defined by scaling of the measured length (in case of $D_T = 1$) of an object when resolution is varied. One may assume that the length scales as

$$L(r) \sim r^{1-D(r)} \tag{6.22}$$

For the coast line of England one has found, within certain limits of r, that the scaling leads to a fractal dimension of $D = 1.2$, while for a straight line, the dimension is 1 since the length does not depend on the resolution (at least if resolution is finer than the length of the object). A curve for which new dwindling structures appear when resolution is increased will have a resolution dependent length and a fractal dimension larger than 1.

The fractal dimension of the Koch curve in Figure 6.3, can be derived as follows. At each step, resolution is increased from r to $r/3$, i.e., an improved resolution by a factor of 3, and the observed length is increased from L to $4L/3$, or by a factor of 4/3. The dimension D can then be written as

$$D(r) = 1 - \frac{\partial \ln L(r)}{\partial \ln r} = 1 - \frac{\ln L - \ln(4L/3)}{\ln r - \ln(r/3)} = \frac{\ln 4}{\ln 3} \approx 1.26 \ . \tag{6.23}$$

In this example the fractal dimension does not depend on the resolution. The object is constructed so that new structure (with the same characteristics) appears whenever resolution is increased.

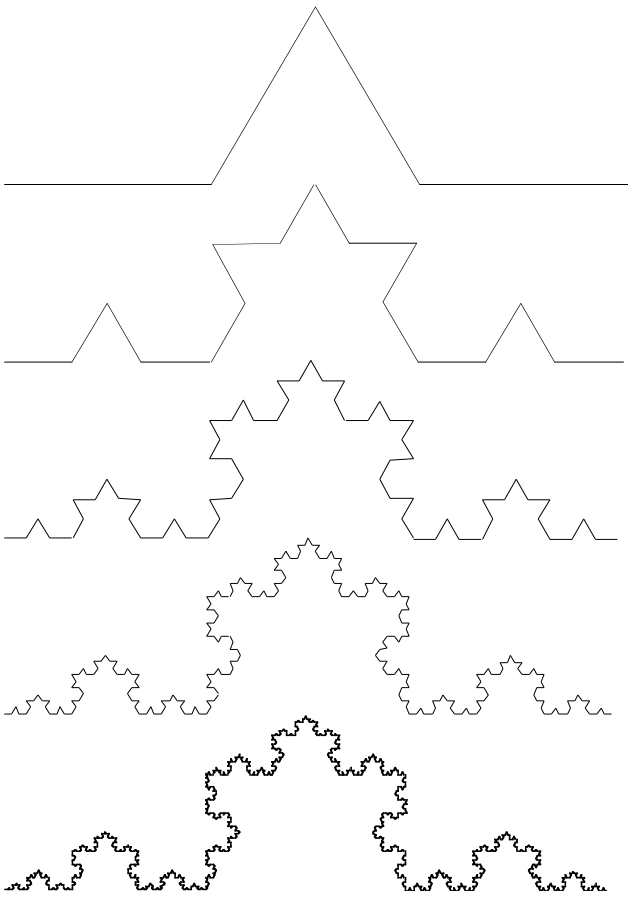


Figure 6.3. The Koch curve is a fractal object with dimension $\ln 4/\ln 3 = 1.26$, in which new structures appear as soon as resolution is increased.

6.2.3 Dimension and information

If resolution is improved when we observe an object, we may gain information, or the other way around, if the resolution becomes worse (r increases) we may loose information. In a three-dimensional space, a point looses information in all three directions, since we get less accurate description on three of its coordinates. For a line, on the other hand, a blurred picture only reduces information in two directions, perpendicular to the extension of the line. This indicates that there may be some connection between the dimensionality of an object and how information quantities change when resolution is varied.

Suppose that our object can be characterised by a probability density $p(\mathbf{x})$ in a D_E-dimensional Euclidean space, and further that we have an *a priori* probability density p_0 that is uniform. The decomposition of the Kullback information with respect to both position and resolution, Eq. (6.13), can then be written

$$K[p_0;p] = \int_0^\infty \frac{dr}{r} \int d\mathbf{x}\; p(r;\mathbf{x}) \left(r \frac{d}{d\mathbf{x}} \ln p(r;\mathbf{x}) \right)^2 . \tag{6.24}$$

We now define the information-theoretic dimension

$$\begin{aligned} d(r) &= \int d\mathbf{x}\; p(r;\mathbf{x}) \left(D_E - \left(r \frac{d}{d\mathbf{x}} \ln p(r;\mathbf{x}) \right)^2 \right) = \\ &= D_E - \int d\mathbf{x}\; p(r;\mathbf{x}) \left(r \frac{d}{d\mathbf{x}} \ln p(r;\mathbf{x}) \right)^2 . \end{aligned} \tag{6.25}$$

as the dimensionality of the probability density at resolution r (Eriksson and Lindgren, 1987). The following argument illustrate that this is the information-theoretic counterpart to the fractal dimension defined in Eq (6.21). By partial integration, assuming $p'(r;\mathbf{x}) \ln p(r;\mathbf{x}) \to 0$ when $\mathbf{x} \to \infty$, we can rewrite Eq. (6.25) as

$$\begin{aligned} d(r) &= D_E + \int d\mathbf{x}\, r^2 \frac{d^2 p(r;\mathbf{x})}{d\mathbf{x}^2} \ln p(r;\mathbf{x}) = \\ &= D_E - r \frac{\partial}{\partial r} \int d\mathbf{x}\; p(r;\mathbf{x}) \ln \frac{1}{p(r;\mathbf{x})} . \end{aligned} \tag{6.26}$$

In order to illustrate how this is related to the fractal dimension defined in Eq. (6.21), we approximate the probability density $p(r;\mathbf{x})$ with a uniform distribution over the $N(r)$ spheres covering the object as it was illustrated in Figure 6.2, i.e.,

$$p(r;\mathbf{x}) \sim \frac{1}{N(r)\, r^{D_E}} . \tag{6.27}$$

Then the dimension $d(r)$, Eq (6.26), can be simplified,

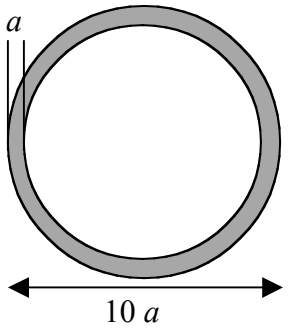
$$d(r) = D_E - r \frac{\partial}{\partial r} \ln(N(r)\, r^{D_E}) = -\frac{\partial \ln(N(r))}{\partial \ln r} = D(r) , \tag{6.28}$$

illustrating the connection between information-theoretic dimension $d(r)$ and fractal dimension $D(r)$, see Eq. (6.21).

6.3 Exercises

6.1 What is the dimension $d(r)$ of a Gaussian distribution of width b (variance b^2)? How can this be interpreted? Discuss the relation between r and b.

6.2 **Dimension of a ring.** Consider a pattern in a two-dimensional space that is described as a ring with a certain thickness a and a certain diameter $10a$, schematically depicted in the figure below.



The ring may be described by a uniform probability distribution $p(x, y)$ that is constant within the grey area and zero elsewhere. By introducing the resolution dependent probability density $p(r; x, y)$, one may study how the entropy $S(r)$ changes when the resolution is made worse (r increases).

Discuss how the quantity

$$r\frac{\partial}{\partial r}\int dxdy\; p(r;x,y)\ln\frac{1}{p(r;x,y)}$$

depends on r. It should not be necessary with calculations (which are quite complicated), but you should be able to discuss this in a more schematic way, relating resolution level r to characteristic lengths in the system, for example a.

7 Pattern formation in chemical systems

In this chapter we apply information theory to chemical pattern formation, often called chemical self-organising systems. Here self-organisation refers to a system that spontaneously builds up or sustain spatio-temporal structure in the form of concentration variations. The patterns formed and the dynamics is not controlled by anything outside the system. The first theoretical investigation of chemical pattern formation was done by Alan Turing in his classic paper (1952), where he demonstrated that simple mechanisms in reaction-diffusion dynamics can account for symmetry breaking necessary for morhogenesis. Self-organising systems, or *dissipative structures*, that also go beyond chemical pattern formation have been extensively studied by, for example, Prigogine and Nicolis (1977) and Haken (1984).

There are several types of information quantities involved in pattern formation processes, and they are typically of different orders of magnitude. A first piece of information when one is presented with a specific chemical system is the selection of molecules involved, both those that are present in the system and those that are allowed to flow over the system border. All this can loosely be referred to as genetic information, i.e., information on which components and thus also which processes that will be part of the self-organising chemical system. The amount of this information is not vary large, as is reflected for example by the size of genomes in living organisms with the order of 10^4 genes.

In biological systems all necessary information is not genetically encoded, but there is also compositional information in the transfer of chemicals and structures from the mother to the daughter cells. There are proposals expressing the idea that this type of information may have played an important role in the origin of life (Segré et al, 2000).

Another type of information enters when a specific self-organising system starts to develop a pattern. The typical form of the pattern may be determined by the reaction scheme involved, but in many cases fluctuations in concentrations or other disturbances may affect the exact pattern that is formed. An example of that is the difference in finger prints between identical twins. One may view this as an information flow from fluctuations to the actual pattern that is observed. This flow is of the same character as the flow from micro to macro that we find in chaotic systems (Shaw, 1981). This information flow can characterised by the Lyapunov exponent, an important quantity for the analysis of chaotic dynamical systems. This perspective will be brought up in the next Chapter presenting an information-theoretic perspective on low-dimensional chaotic systems.

The focus of the approach presented here is a third type of information quantity, the information capacity in free energy or exergy, based on the information-theoretic formalism presented in Chapter 6. This is then combined with the geometric information theory of Chapter 7. The starting point is the free energy of a chemical system, involving both the deviation from homogeneity when a spatial pattern is present and a deviation from equilibrium (when the system is stirred). This free energy is expressed as the total information of the system, and in our approach we decompose this into information contributions from both different positions and different length scales. The connection with thermodynamics then

allows us to view the inflow of free energy, due to the fact that the system is open to an inflow of a fuel and outflow of waste products, as an inflow of information capacity. This inflow allows for an accumulation of information in the system when a pattern is formed. Entropy production due to chemical reactions and diffusion leads to desctruction of information – an information loss that can be balanced by the information capacity inflow to maintain the chemical pattern. The following presentation is based on work previously published in (Eriksson and Lindgren, 1987; Eriksson et al, 1987) which applied to closed chemcial systems, which was later extended to open reaction-diffusion systems (Lindgren et al, 2004).

7.1 Information analysis of chemical pattern formation

A closed self-organising system is driven by the exergy that is initially present in the form of chemical energy. We assume that the system has the same pressure and the same temperature as the environment. According to Eq. (5.20), the exergy can be written as a Kullback information

$$E = V k_{\mathrm{B}} T_0 \sum_{i=1}^{M} c_i \ln \frac{c_i}{c_{i0}} = k_{\mathrm{B}} T_0 N K[P_0;P] \ , \tag{7.1}$$

where the distributions P_0 and P denote the normalised distributions $p_{i0} = Vc_{i0}/N$ and $p_i = Vc_i/N$, respectively. The reference system P_0 represent a system in reaction equilibrium, and the non-equilibrium of the present system, characterised by P, can be interpreted as the system being prepared with an abundancy of a fuel that may be used in the process of structure formation. In the case of an open system where chemical substances are allowed to pass the system border, the concentrations in the system may be kept off equilibrium so that exergy (free energy) is available both for the built-up and the maintenance of the spatial patterns.

7.1.1 Chemical and spatial information

We shall use Eq. (7.1) as a starting point for our combined information-theoretic and thermodynamic analysis of the system. Since homogeneity is assumed to be broken, we introduce spatially varying concentrations $c_i(\mathbf{x})$, expressed in the probabilistic form $p_i(\mathbf{x}) = Vc_i(\mathbf{x})/N$. (We assume that the number of molecules per unit volume, N/V, does not depend on position.) This results in probability distributions $p_i(\mathbf{x})$ over the different molecules that are normalised at each position. Then, Eq. (7.1) is replaced by

$$E = k_{\mathrm{B}} T_0 \frac{N}{V} K \ , \tag{7.2}$$

where the information K is now an integral over Kullback information quantities for each position in the system,

$$K = \int_V d\mathbf{x}\, K[P_0;P(\mathbf{x})] = \int_V d\mathbf{x} \sum_{i=1}^{M} p_i(\mathbf{x}) \ln \frac{p_i(\mathbf{x})}{p_{i0}} \ . \tag{7.3}$$

We shall use the average concentration within the system, defined by

$$\overline{p}_i = \frac{1}{V}\int_V d\mathbf{x}\, p_i(\mathbf{x}), \tag{7.4}$$

in order to decompose the total information K in Eq. (7.2) into two terms, one that quantifies the deviation of the average concentrations from equilibrium, K_{chem}, and one that quantifies the deviation from homogeneity, i.e., the presence of spatial structure, K_{spatial},

$$\begin{aligned} K &= \int_V d\mathbf{x} \sum_{i=1}^{M} p_i(\mathbf{x}) \ln \frac{p_i(\mathbf{x})}{\overline{p}_i} \frac{\overline{p}_i}{p_{i0}} = \\ &= \int_V d\mathbf{x} \sum_{i=1}^{M} p_i(\mathbf{x}) \ln \frac{p_i(\mathbf{x})}{\overline{p}_i} + V \sum_{i=1}^{M} \overline{p}_i \ln \frac{\overline{p}_i}{p_{i0}} . \end{aligned} \tag{7.5}$$

Therefore we define the *spatial information*, K_{spatial},

$$K_{\text{spatial}} = \int_V d\mathbf{x} \sum_{i=1}^{M} p_i(\mathbf{x}) \ln \frac{p_i(\mathbf{x})}{\overline{p}_i} \geq 0, \tag{7.6}$$

and the *chemical information*, K_{chem},

$$K_{\text{chem}} = V \sum_{i=1}^{M} \overline{p}_i \ln \frac{\overline{p}_i}{p_{i0}} \geq 0, \tag{7.7}$$

and we can write the total information K as

$$K = K_{\text{spatial}} + K_{\text{chem}} . \tag{7.8}$$

Thermodynamically, the chemical information is related to the presence of a chemical non-equilibrium even when spatial variations are not taken into account. This means that there is an abundance of "fuel" and a low level of "waste" products in the system. The spatial information reflects the presence of a non-equilibrium for diffusion processes, i.e., that there is some spatial pattern in the system.

7.1.2 Decomposition of spatial information in a chemical pattern

We shall continue our analysis by further decompose the structural information into contributions from position *and* length scales, similar to what we did in Chapter 6. For simplicity, let us assume that we have a chemical system, characterised by concentrations $c_i(\mathbf{x}, t)$ for the different molecules that are normalised at each position $\mathbf{x}$,

$$\sum_{i=1}^{M} c_i(\mathbf{x}, t) = 1 . \tag{7.9}$$

In order to be able to analyse contributions from different length scales, we introduce a resolution dependent concentration $\tilde{c}_i$, cf. Eq. (6.3), by convolution of the original one c_i with a Gaussian,

$$\begin{aligned}\tilde{c}_i(r,\mathbf{x}) &= \frac{1}{(2\pi)^{n/2}}\int d\mathbf{z}\, e^{-\mathbf{z}^2/2} c_i(\mathbf{x}+r\mathbf{z}) = \\ &= \exp(\tfrac{r^2}{2}\nabla^2)c_i(\mathbf{x})\,.\end{aligned} \tag{7.10}$$

We will use the last expression as the resolution operator, but when calculating $\tilde{c}_i$, the first expression will be used. We assume that this operation handles the boundary conditions so that in the limit of $r \to \infty$, i.e., complete loss of position information, the concentrations equal the average concentrations in the system[10],

$$\tilde{c}_i(\infty,x) = \bar{c}_i. \tag{7.11}$$

The derivation of a decomposition of the total information K into different contributions will now be slightly different from the one in Chapter 6, since the distributions are now normalised in each position (instead of over the system volume). Our starting point, in Chapter 6, for decomposing the total information K,

$$\begin{aligned}\int_0^\infty dr \frac{\partial}{\partial r}\int dx K[c_0;\tilde{c}(r,x)] &= \int dx K[c_0;\bar{c}] - \int dx K[c_0;\tilde{c}(0,x)] = \\ &= K_{\text{chem}} - K\end{aligned} \tag{7.12}$$

illustrates that the chemical information K_{chem} can be detected regardless of how bad the resolution is. This is obvious, since complete loss of resolution results in average concentrations in the system which determines the chemical information. This means that the spatial information K_{chem} can be decomposed in a similar way as in Chapter 6,

$$\begin{aligned}K_{\text{spatial}} &= -\int_0^\infty dr \frac{\partial}{\partial r}\int d\mathbf{x} K[c_0;\tilde{c}(r,\mathbf{x})] = -\int_0^\infty dr \int d\mathbf{x} \frac{\partial}{\partial r}\sum_i \tilde{c}_i(r,\mathbf{x}) \ln\frac{\tilde{c}_i(r,\mathbf{x})}{c_{i0}} = \\ &= -\int_0^\infty dr \int d\mathbf{x} \sum_i \frac{\partial \tilde{c}_i(r,\mathbf{x})}{\partial r}(\ln\frac{\tilde{c}_i(r,\mathbf{x})}{c_{i0}} + 1) = -\int_0^\infty dr \int d\mathbf{x} \sum_i r\nabla^2 \tilde{c}_i(r,\mathbf{x}) \ln\frac{\tilde{c}_i(r,\mathbf{x})}{c_{i0}} = \\ &= \int_0^\infty \frac{dr}{r}\int dx \sum_i \tilde{c}_i(r,\mathbf{x})\left(r\nabla \ln\frac{\tilde{c}_i(r,\mathbf{x})}{c_{i0}}\right)^2\end{aligned} \tag{7.13}$$

In the last step a partial integration is invoved, including an assumption of periodic (or non-flow) boundary conditions. The equilibrium concentration c_{i0} in the last expressions could be removed (as it disappears under the differential operation), but we keep it as it will serve a

[10] In the simple examples we study, this follows immediately from the assumption on preiodic boundary conditions.

purpose in later derivations. In the last expression we recognize a local information density, $k(r, x)$, which can be written in three different forms, of which the last one will be used later,

$$\begin{aligned} k(r,\mathbf{x}) &= r^2 \sum_i \tilde{c}_i(r,\mathbf{x}) \left(\nabla \ln \frac{\tilde{c}_i(r,\mathbf{x})}{c_{i0}} \right)^2 = \\ &= r^2 \sum_i \frac{(\nabla \tilde{c}_i(r,\mathbf{x}))^2}{\tilde{c}_i(r,\mathbf{x})} = \\ &= \left(-r \frac{\partial}{\partial r} + r^2 \nabla^2 \right) \sum_i \tilde{c}_i(r,\mathbf{x}) \ln \frac{\tilde{c}_i(r,\mathbf{x})}{c_{i0}} . \end{aligned} \tag{7.14}$$

The information density $k(r, \mathbf{x})$ can be integrated over space so that we achieve the spatial information at a certain length scale,

$$k_{\text{spatial}}(r) = \int d\mathbf{x}\, k(r,\mathbf{x}). \tag{7.15}$$

A schematic illustration of this decmposition is shown in Fig. 7.1, illustrating a chemical patterns in a two-dimensional space with its variations along the axis of resolution r.

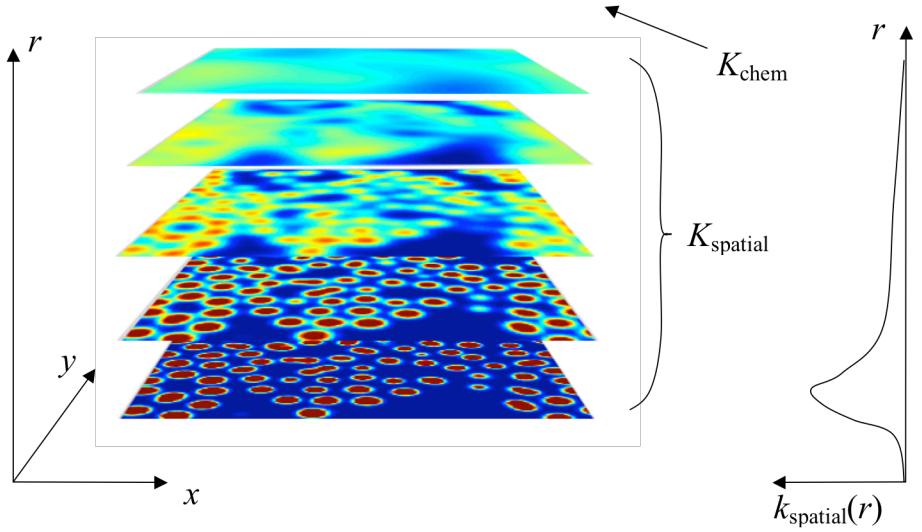


Figure 7.1. An illustration of a two-dimensional chemical pattern at different levels of resolution r, along with a schematic illustration on how the spatial information may be decomposed between different length scales. The chemical information can be detected regardsless of how bad the resolution is and it is therefore located at infinite r.

7.1.3 Reaction-diffusion dynamics

The concentrations are time-dependent since they vary in time due to diffusion, characterised by diffusion constants D_i, and chemical reactions, characterised by reaction functions $F_i(\mathbf{c}(\mathbf{x}, t)$, where $\mathbf{c} = (c_1, \ldots, c_M)$. The equations of motion governing the dynamics is then the

ordinary reaction-diffusion equations plus a term $B_i(c_i(x, t))$ capturing flows across the system border in the case of an open system,

$$\dot{c}_i(\mathbf{x},t) = \frac{d}{dt}c_i(\mathbf{x},t) = D_i\nabla^2 c_i(\mathbf{x},t) + F_i(\mathbf{c}(\mathbf{x},t)) + B_i(c_i(\mathbf{x},t)) \ . \tag{7.16}$$

The term $B_i(c_i(\mathbf{x}, t))$ typically is in the form of diffusion-controlled flows. We assume that the in- and out-flow is directly connected to the whole system. For example, in a two-dimensional system, this means that there is a flow across the "surface" of the system in the direction of a third dimension. We can view this as if the reaction volume everywhere is in contact with a reservoir having a constant concentration $c_{i,\,\mathrm{res}}$ which results in an inflow

$$B_i(c_i(\mathbf{x},t)) = b_i(c_{i,\,\mathrm{res}} - c_i(\mathbf{x},t)) . \tag{7.17}$$

with b_i being a diffusion constant. (A negative value of that expression reflects an outflow.)

The equations of motion for the resolution dependent concentrations are derived from Eq. (7.16) by applying the resolution operator, see Eq. (7.10), to both sides of the equation,

$$\begin{aligned}\dot{\tilde{c}}_i(r,\mathbf{x},t) &= D_i\nabla^2\tilde{c}_i(r,\mathbf{x},t) + \exp(\tfrac{r^2}{2}\nabla^2)F_i(\mathbf{c}(\mathbf{x},t)) + \exp(\tfrac{r^2}{2}\nabla^2)B_i(c_i(x,t)) = \\ &= D_i\nabla^2\tilde{c}_i(r,\mathbf{x},t) + \exp(\tfrac{r^2}{2}\nabla^2)F_i(\mathbf{c}(\mathbf{x},t)) + b_i(c_{i,\,\mathrm{res}} - \tilde{c}_i(\mathbf{x},t))\end{aligned} \tag{7.18}$$

Since the reaction terms typically are non-linear the resolution operator need to remain in front of the reaction function F_i.

7.1.4 Flows of information in a closed chemical systems

In a closed homogenous chemical system, prepared in an out-of-equilibrium state, all information is initially in the form of chemical information, K_{chem}, corresponding to an initial amount of exergy. Chemical reactions consume the exergy, according to the 2nd law of thermodynamics, and the chemical information thus decays. If spatial structure is built up, we find that some of the information, at least temporarily, is transformed to spatial information, K_{spatial}. If the system remains closed, though, such spatial structures cannot remain and the system approaches a homogenous equiliobrium state. In an open system, a steady inflow of chemical energy may keep the chemical information at a high level, and spatial structures may be supported. First we will discuss the closed system and derive a set of information flow quantities, and in the next section the system will be open for molecular flows across the boundary which will result in additional effects. So to start with we neglect the flow term B in Eqs. (7.16) and (7.18).

Let us discuss briefly some thermodynamic characteristics of the system. The chemical system, described by the dynamics Eq. (7.16) but excluding the result of the in- and out-flow (which will be treated separately), results in a thermodynamic entropy production σ (in units of Boltzmann's constant) according to

$$\sigma(\mathbf{x},t) = \sum_i \left(D_i \frac{(\nabla c_i(\mathbf{x},t))^2}{c_i(\mathbf{x},t)} - \left(\ln \frac{c_i(\mathbf{x},t)}{c_{i0}} \right) F_i(\mathbf{c}(\mathbf{x},t)) \right) . \tag{7.19}$$

The entropy production is determined by one term corresponding to the entropy produced due to diffusion within the system and one term given by the reactions that tend to even out the chemical non-equilibrium in the system. The entropy production certainly leads to a decay of the information in the system — decay of structural information as well as of chemical information.

We shall make an information-theoretic description of how information is flowing in the system that connects to the thermodynamic loss of information due to entropy production. This will be formulated in a continuity equation for information density $k(r, \mathbf{x}, t)$, taking into account flows both in scale (r) and in space ($\mathbf{x}$), see Fig. 7.2.

In the limit of $r \to \infty$ we cannot distinguish any spatial structure, but the chemical information K_{chem} is still present, unaffected by the resolution parameter. In a decomposition of the total information this part can therefore be considered as present at the $r \to \infty$ limit. The chemical information will be consumed by the chemical reactions and we should therefore expect that a proper definition of information flow shows how information will flow in the direction towards smaller length scales r. If the system gives rise to spatial structure, that should be captured in the continuity equation, resulting in a temporal accumulation of structural information.

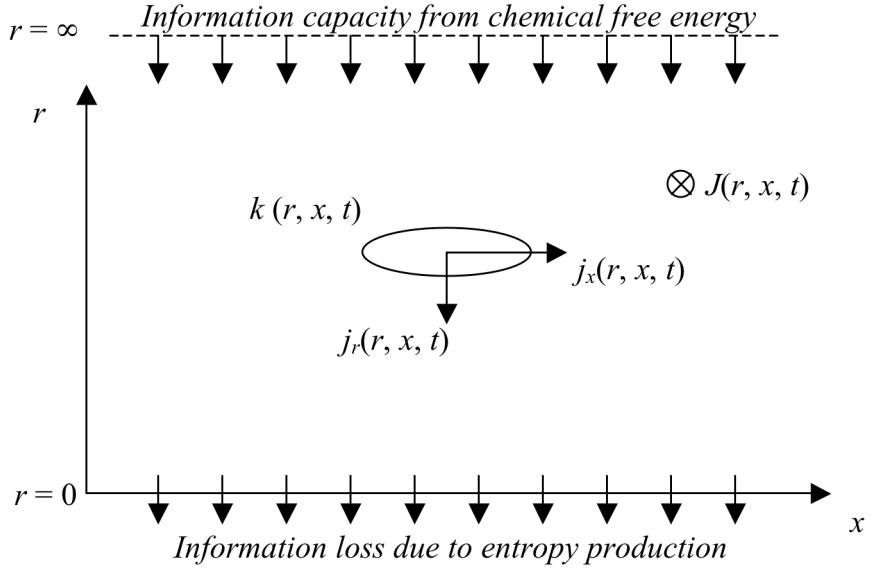


Figure 7.2. A schematic picture of information (capacity) flows in a chemical pattern formation system. The pattern is characterised by an information density $k(r, x, t)$ distributed over spatial dimensions as well as over different length scales r. Information flows both in space and in scale, where the flow is destroyed when it gets down to the microscopic level. Here information disappears into microscopic degrees of freedom due to entropy production. Information capacity enters the system at the very large scale due to a diffusion-controlled inflow of chemical information or Gibbs free energy. A pattern is formed when information that flows downwards in scale is aggregated at certain positions as described by the continuity equation.

It is reasonable to think that information is leaving the system, through the thermodynamic entropy production, at the finest length scales of the system, i.e., at $r = 0$. At this point information leaves the macroscopic description that we have of our system, and the information is spread out on microscopic degrees of freedom. Therefore, we define the information flow $j_r(r, \mathbf{x}, t)$ in the direction of smaller r, at the border $r = 0$, to be equal to the chemical entropy production,

$$j_r(0,\mathbf{x},t) = \sigma(\mathbf{x},t) \ . \tag{7.20}$$

To define this flow for general resolution values r, we generalise by introducing the resolution operator into the expression for entropy production,

$$j_r(r,\mathbf{x},t) = \sum_i \left(D_i \frac{(\nabla \tilde{c}_i(r,\mathbf{x},t))^2}{\tilde{c}_i(r,\mathbf{x},t)} - \left(\ln \frac{\tilde{c}_i(r,\mathbf{x},t)}{c_{i0}} \right) \exp(\tfrac{r^2}{2}\nabla^2) F_i(\mathbf{c}(\mathbf{x},t)) \right) . \tag{7.21}$$

In the limit of $r \to \infty$, this information flow can be written

$$\begin{aligned} j_r(\infty,\mathbf{x},t) &= \lim_{r\to\infty} \sum_i \left(-\left(\ln \frac{\tilde{c}_i(r,\mathbf{x},t)}{c_{i0}} \right) \exp(\tfrac{r^2}{2}\nabla^2) F_i(\mathbf{c}(\mathbf{x},t)) \right) = \\ &= -\sum_i \dot{\tilde{c}}_i(\infty,\mathbf{x},t) \ln \frac{\tilde{c}_i(\infty,\mathbf{x},t)}{c_{i0}} = -\sum_i \frac{d\bar{c}_i(t)}{dt} \ln \frac{\bar{c}_i(t)}{c_{i0}}, \end{aligned} \tag{7.22}$$

where we have used the dynamics, Eq. (7.18), and Eq. (7.11). Note that this expression is equal to the decay of chemical information, $-dk_{\text{chem}}/dt$, since

$$\dot{k}_{\text{chem}}(t) = \frac{d}{dt} k_{\text{chem}} = \sum_i \dot{\tilde{c}}_i(\infty,\mathbf{x},t) \left(\ln \frac{\tilde{c}_i(\infty,\mathbf{x},t)}{c_{i0}} + 1 \right) = -j_r(\infty,\mathbf{x},t) , \tag{7.23}$$

since $\Sigma_i \, dc_i/dt = 0$ due to the normalisation. In general it holds that $dk_{\text{chem}}/dt \leq 0$ (except for extreme cases), which means that there is a positive flow of information $j_r(\infty, \mathbf{x}, t)$ at the infinite length scale limit in the direction of smaller length scales, similar to what we have at the other limit, $r = 0$. The interpretation is then that information flows from its origin as chemical information down through the length scales, temporarily halting if spatial structure is built up, but sooner or later continuing to the finest length scale where it disappears as entropy production.

7.1.5 A continuity equation for information in the case of a closed system

We can view the information density k as a generalised form of exergy (or free energy). By the flow across length scales j_r we have accounted for the destruction of information from entropy production in chemical reactions and diffusion within the system. We will now derive a continuity equation for information which will connect the previously defined flow across length scales, $j_r(r, \mathbf{x}, t)$, with the change of local information $k(r, \mathbf{x}, t)$ and a flow of

information in space $\mathbf{j}(r, \mathbf{x}, t)$. For a closed chemical system such a continuity equation takes the form

$$\dot{k}(r,\mathbf{x},t) = r\frac{\partial}{\partial r} j_r(r,\mathbf{x},t) - \nabla \cdot \mathbf{j}(r,\mathbf{x},t). \tag{7.24}$$

The equation states that, for a closed system, the local information density k changes due to accumulation of of the information flows across length scales (in the downward direction) and across space. This continuity equation then implies a definition of the $\nabla \cdot \mathbf{j}$ term,

$$\begin{aligned} &\nabla \cdot \mathbf{j}(r,\mathbf{x},t) = r\frac{\partial}{\partial r} j_r(r,\mathbf{x},t) - \dot{k}(r,\mathbf{x},t) = \ldots = \\ &= -r^2\nabla^2 \sum_i \left[\left(\ln \frac{\tilde{c}_i(r,\mathbf{x},t)}{c_{i0}} \right) \exp(\tfrac{r^2}{2}\nabla^2) F_i(\mathbf{c}(\mathbf{x},t)) \right]. \end{aligned} \tag{7.25}$$

If we require that $\mathbf{j}(r, \mathbf{x}, t) = 0$ when $\tilde{c}_i$ is uniform and that $\mathbf{j}$ is rotation-free (i.e., does not contain any term $\nabla \times \mathbf{A}$), the spatial flow is defined

$$\mathbf{j}(r,\mathbf{x},t) = -r^2\nabla \sum_i \left[\left(\ln \frac{\tilde{c}_i(r,\mathbf{x},t)}{c_{i0}} \right) \exp(\tfrac{r^2}{2}\nabla^2) F_i(\mathbf{c}(\mathbf{x},t)) \right]. \tag{7.26}$$

Note that the spatial flow depends on the presence of reactions. This means that when reactions are not present, the only flow is the one across length scales. This is a direct consequence of the fact that the resolution operator, the Gaussian blur, is equivalent to a diffusion process. The flow across length scales, though, depends on both reactions and diffusion. This should be expected since it is a generalisation of the entropy production, and entropy is produced in both these processes.

7.1.6 A continuity equation for information in the case of an open system

Finally, we open the system for inflow and outflow of molecules, given by the term B in the reaction-diffusion dynamics, Eq. (7.16). In addition to the terms in the continuity equation for the closed system, we need to introduce a local source/sink term $J(r, x, t)$ which will capture the effects from the system being open. The continuity equation for an open system then takes the following form,

$$\dot{k}(r,\mathbf{x},t) = r\frac{\partial}{\partial r} j_r(r,\mathbf{x},t) - \nabla \cdot \mathbf{j}(r,\mathbf{x},t) + J(r,x,t)\ . \tag{7.27}$$

For an open system, there are two ways in which the information characteristics of the system is directly affected from the flow of molecules across the system border. First, an inflow and an outflow of components changes the average concentrations of the transported components and in that way the chemical information is changed. For a driven chemical system, with an inflow of a fuel component and an outflow of a product, the flow across the system boundary typically keeps the chemical information at a level sufficient for driving the information flows in the system. But, there is also a direct influence that the inflow may have on spatial patterns

in the system. For example, diffusion into or out of the system of a component that has a spatial variation, directly leads to that spatial pattern being less accentuated, and in that way the flow destroys the local information density $k(r,\mathbf{x},t)$.

The negative effect on informaiton density *k* from the diffusion over the system boundary is captured by the sink term *J* in the continuity equation. For a diffusion controlled flow, as in Eq. (7.17), we get (after some calculations) the following expression for *J*,

$$J(r,x,t) = -\sum_i b_i(\tilde{c}_i + c_{i,\,\mathrm{res}})[r\nabla \ln \tilde{c}_i]^2 \le 0\,, \tag{7.28}$$

which shows that *J* is a sink for information. In conclusion we have

$$\begin{aligned}
k(r,\mathbf{x},t) &= r^2 \sum_i \frac{(\nabla \tilde{c}_i(r,\mathbf{x},t))^2}{\tilde{c}_i(r,\mathbf{x},t)}\,,\\
j_r(r,\mathbf{x},t) &= \sum_i \left(D_i \frac{(\nabla \tilde{c}_i(r,\mathbf{x},t))^2}{\tilde{c}_i(r,\mathbf{x},t)} - \left(\ln \frac{\tilde{c}_i(r,\mathbf{x},t)}{c_{i0}} \right) \exp(\tfrac{r^2}{2}\nabla^2) F_i(\mathbf{c}(\mathbf{x},t)) \right),\\
\mathbf{j}(r,\mathbf{x},t) &= -r^2 \nabla \sum_i \left[\left(\ln \frac{\tilde{c}_i(r,\mathbf{x},t)}{c_{i0}} \right) \exp(\tfrac{r^2}{2}\nabla^2) F_i(\mathbf{c}(\mathbf{x},t)) \right],\\
J(r,x,t) &= -\sum_i b_i(\tilde{c}_i + c_{i,\,\mathrm{res}})[r\nabla \ln \tilde{c}_i]^2 \le 0\,.
\end{aligned} \tag{7.29}$$

These quantities can be integrated over either position space **x** or resolution length scale *r* in order to derive aggregated information quantities like *k*(*r*, *t*), i.e., the resolution-dependent information, or *k*(**x**, *t*) and **j**(**x**, *t*) which are the normal information density and the corresponding information flow. If we integrate the continuity equation (7.27) over resolution lengths *r*, we get the following balance equation for the information density *k*(**x**, *t*),

$$k(\mathbf{x},t) + \nabla \cdot \mathbf{j}(\mathbf{x},t) + J(x,t) + \sigma(\mathbf{x},t) + \dot{k}_{\mathrm{chem}}(\mathbf{x},t) = 0\ . \tag{7.30}$$

Here the entropy production and the change in chemical information come from the upper and lower limits of the flow j_r.

In the next section these information-theoretic concepts will be applied to a simple model of a chemical self-organising system.

7.2 Application to the self-replicating spots dynamics

We apply the formalism to the pattern formation of the "self-replicating spots" system (Gray and Scott, 1984; Pearson, 1993; Lee *et al*, 1993),

$$\mathrm{U} + 2\mathrm{V} \rightarrow 3\mathrm{V}, \tag{7.31}$$

with the reaction-diffusion dynamics

$$\begin{aligned}\dot{c}_{\mathrm{U}} &= D_{\mathrm{U}}\nabla^2 c_{\mathrm{U}} - (c_{\mathrm{U}} - k_{back}c_{\mathrm{V}})c_{\mathrm{V}}{}^2 + \hat{D}(1-c_{\mathrm{U}}) \\ \dot{c}_{\mathrm{V}} &= D_{\mathrm{V}}\nabla^2 c_{\mathrm{V}} + (c_{\mathrm{V}} - k_{back}c_{\mathrm{V}})c_{\mathrm{V}}{}^2 - (\hat{D}+k)c_{\mathrm{V}}\end{aligned} \tag{7.32}$$

We have introduced a very slow back reaction ($k_{back} = 10^{-5}$) in order to get the relationship between equilibrium concentrations of U and V defined by the reactions.

In Fig. 7.3, the dynamics is illustrated starting from an initial state (left) with a square of high concentration of V. As the system evolves four concentration peaks (spots) emerges from the square, and these spots reproduce by growing and splitting until the system is filled with spots (middle and right). In the process, spots may disappear, which leaves space for other spots to reproduce. In the lower part of the figure, the decomposition of the information in the pattern with respect to scale is plotted for the three snapshots above (at time 0, 1000, and 7000, respectively).

It is clear that the initial state has a longer characteristic length as detected by the information density. When the system produces the spots at the significantly shorter length scale, information is found both at the old length, now due to the size of the cluster (middle), and at the length scale of the spots. When the square distribution has been completely decomposed into spots, no information is left at the initial length scale.

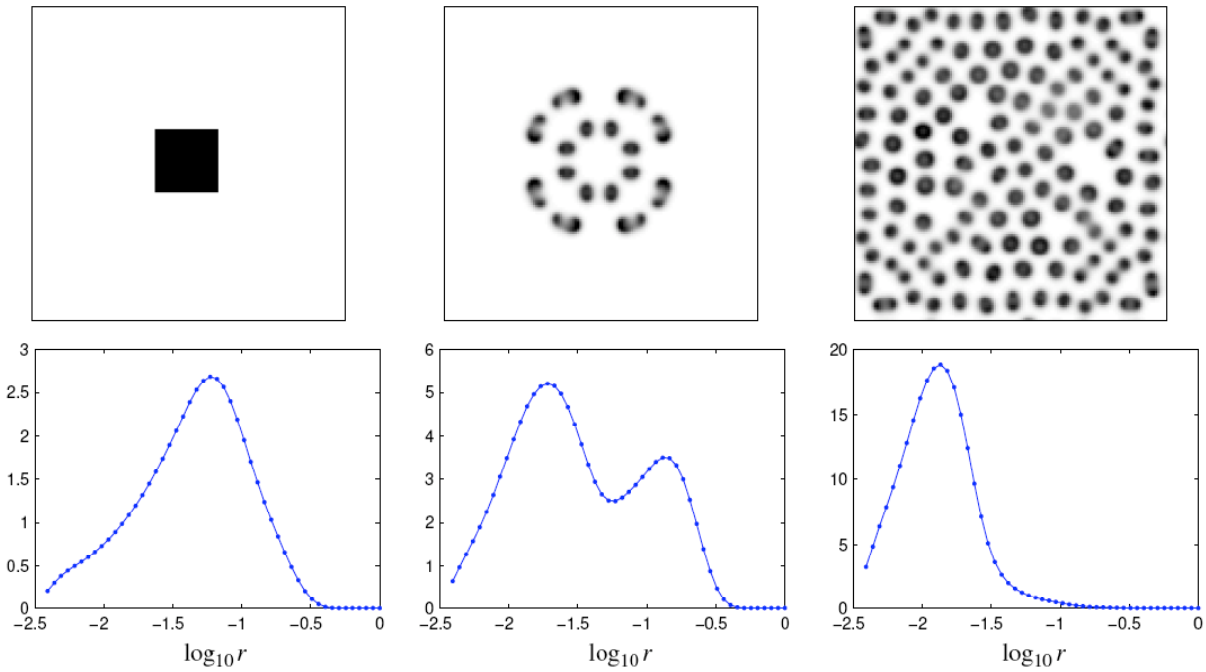


Figure 7.3. Top row: The concentration of the chemical V in the system at three times; t = 0, t = 1000, and t = 7000 steps. White corresponds to zero concentration, and black corresponds to a concentration of one half. **Bottom row:** The structural information $k(r, x)$, integrated over the system, as a function of the resolution r for the same time steps as above. The length of the system is 1, $D_{\mathrm{U}} = 2$ $D_{\mathrm{V}} = 0.05$, $\hat{D}_{\mathrm{U}} = 0.02$, $k = 0.058$. [Figure from (Lindgren *et al* 2004).]

In Fig. 7.4, we show the information density over the system for three different length scales after long time (upper part), and the information flow in scale, j_r, for the same state (lower part). At low resolution, or large r (right), the information density is low and captures structures of longer lengths, while at finer resolution, small r (left), the information density is large and reflects the pattern of spots. Note that each spot is seen as a circle in the information density picture, since the information is sensitive to gradients in the pattern. The information flow in scale, j_r, is close to homogenous for large r (right), but when information moves on to finer scales of resolution, the spatial flow **j** redistributes the flow so that a higher flow j_r is obtained at the concentration peaks. At finest resolution, $r = 0$, the information leaves the system as entropy production, which is mainly located to the concentration peaks where the chemical activity is high.

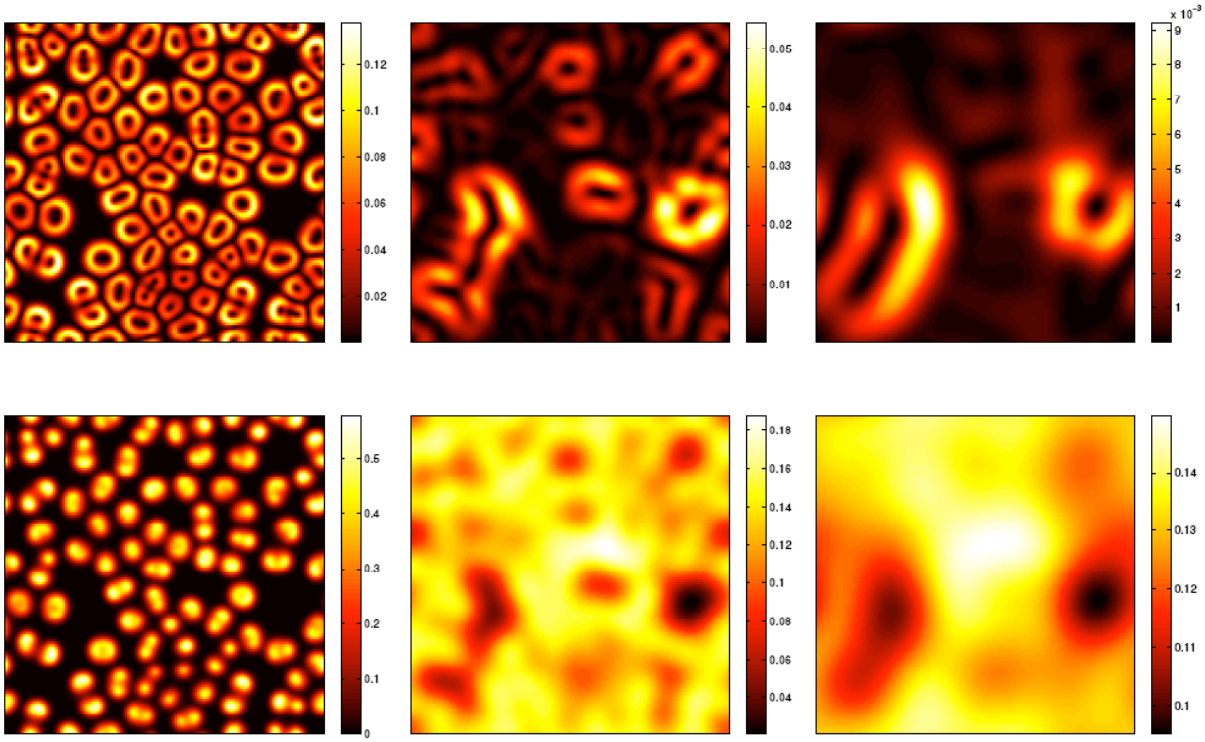


Figure 7.4. Top row: The structural information in the system at $t = 10\,000$ steps, for three values of the resolution r; $r = 0.01$, $r = 0.05$, and $r = 0.1$. **Bottom row:** The information flow $j_r(r, x, t)$. The length of the system is 1, $D_U = 2$, $D_V = 0.05$, $\hat{D}_U = 0.02$, $k = 0.058$. [Figure from (Lindgren *et al* 2004).]

7.3 Exercises

7.1 Prove the equality between the two last expressions in Eq. (7.14).

7.2 Show how the entropy production, Eq. (7.19), is related to the decay of total information $\int d\mathbf{x}\, K[c_o; c(\mathbf{x}, t)]$, using the reaction-diffusion dynamics for a closed system.

8 Chaos and information

In chemical self-organising systems one can identify a tendency of information to flow from larger to smaller length scales, as we have discussed in the last Chapter. Still, it is clear that noise or fluctuations are important to break symmetries and to initiate the formation of spatial structure. Quantitatively this information flow from the noise, or from the microscales of the system, is very small and is hidden by the much larger thermodynamically related flow *towards* the microscales.

In chaotic systems, noise is also of crucial importance. Chaos may even be characterised by the extent to which a system is sensitive to noise. *The Lyapunov exponents* of a dynamical system quantifies how noise is amplified in the dynamics. In this Chapter, we shall make an information-theoretic interpretation of these exponents and relate them to an entropy concept, *the measure entropy*.

8.1 Basic concepts

8.1.1 Iterated maps, fixed points, and periodic orbits

In this presentation, we only consider time-discrete one-variable systems, with state $x(t) \in \mathrm{R}$, with a dynamics given by the simple equation

$$x(t+1) = f(x(t)) \ , \tag{8.1}$$

where f is a differentiable real-valued function on R, f: R → R. The formalism can be extended to higher dimensions and to continuous time, see for example the classic review by Eckman and Ruelle (1985).

A dynamical system of this type has a number of possible types of behaviour. The system can be in a *fixed point*, $x = f(x)$, and nothing changes. The trajectory of the system may also be on a *cycle* with a certain *period T*, characterised by $x(T) = f^T(x(0)) = x(0)$ and $x(k) \neq x(0)$ for $0 < k < T$. Fixed points and periodic orbits may be stable or unstable. Instability means that an arbitrarily small disturbance is sufficient for taking the system away from the point (or the orbit). A fixed point is unstable if $|f'(x^*)|>1$, since any disturbance will result in a trajectory that moves away from x^*. Similarly, a periodic orbit of period T is unstable if the corresponding fixed point of f^T is unstable.

The dynamics can be said to be *chaotic* if the trajectory does not approach a fixed point or a periodic orbit and if it does not diverge. In this case, the trajectory of the system is approaching a (usually complex) set of points in state space, called a *strange attractor*.

8.1.2 Probability densities and measures on the state space

Dynamical systems are often characterised by quantities that are averages over either time or state space (here R). Such a spatial average is based on a *probability measure* μ on the state space, such that $\mu(E)$ can be interpreted as the probability for finding the system in a certain

subset $E \subseteq \mathrm{R}$. We are usually interested in an *invariant measure*[11], i.e., a probability measure that does not change under the dynamics (9.1),

$$\mu(f^{-1}(E)) = \mu(E) \ , \tag{8.2}$$

where E is a set of points in R, and $f^{-1}(E)$ is the set of points that under the map f is transformed to E, i.e., $f(f^{-1}(E)) = E$. There may be several invariant measures for a dynamical system. If there is a fixed point x^*, then a point distribution $\delta(x - x^*)$ in that point is an invariant measure, even if the fixed point is unstable. Any combination of invariant measures is also an invariant measure. In order to eliminate invariant measures that correspond to unstable modes of the dynamics one introduce the so-called *physical measure*, by adding noise to the dynamics and using the resulting measure when reducing the noise to zero. Thus, the physical measure does not include contributions from any of the *single* unstable modes (fixed points or periodic orbits), but if the system is chaotic the resulting measure is reflecting how the states are spread over the attractor. The physical measure can be composed of several invariant measures (for example from several stable fixed points or periodic orbits). If an invariant measure cannot be decomposed into parts that are invariant, the measure is called *ergodic*. The implications of this is that an average (of any function φ) calculated using the ergodic measure is identical (almost always) to a temporal average following the trajectory of the system. This is the *ergodicity theorem*,

$$\int dx \mu(x) \varphi(x) = \lim_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} \varphi(f^k(x(0))) \ , \tag{8.3}$$

for almost all initial states $x(0)$.

8.2 Lyapunov exponent

A chaotic system is sensitive to small changes in the initial state. This tendency to amplify small perturbations is quantified by the Lyapunov exponent of the system (or several exponents in the case of systems in higher dimensions).

Suppose that there is a small change $\delta x(0)$ in the initial state $x(0)$. At time t this has changed to $\delta x(t)$ given by

$$\delta x(t) \approx \delta x(0) \left| \frac{df^t}{dx}(x(0)) \right| = \delta x(0) \left| f'(x(t-1)) \cdot f'(x(t-2)) \cdot \ldots \cdot f'(x(0)) \right|, \tag{8.4}$$

where we have used the chain rule to expand the derivative of f^t. In the limit of infinitesimal perturbations $\delta x(0)$ and infinite time, we get an average exponential amplification, the Lyapunov exponent λ,

[11] This corresponds to a stationary probability distribution when one has a system with a finite state space, for example a finite automaton.

$$\lambda = \lim_{t\to\infty}\frac{1}{t}\ln\left|\frac{\delta x(t)}{\delta x(0)}\right| = \lim_{t\to\infty}\frac{1}{t}\ln\left|\frac{df^t}{dx}(x(0))\right| = \lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1}\ln\left|f'(x(k))\right|. \tag{8.5}$$

If the system is characterised by an ergodic measure μ we can use the ergodicity theorem to express the Lyapunov exponent as

$$\lambda = \lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1}\ln\left|f'(x(k))\right| = \int dx\mu(x)\ln\left|f'(x)\right| \ . \tag{8.6}$$

If the Lyapunov exponent is larger than zero, the system amplifies small perturbations and we have chaos, while if the exponent is negative the system is stable in a fixed point or periodic orbit.

8.2.1 The Lyapunov exponent as an information flow from "micro" to "macro"

An illustration for the Lyapunov exponent as an information flow from smaller scales is given by the following example. Suppose that we determine the position of the system at a certain time t_0, with some resolution δ, implying that we can describe the position as a uniform probability density over a certain interval $[x_0-\delta, x_0+\delta]$. In n time steps, the dynamics will transform this uncertainty interval to

$$\left[f^n(x_0) - \delta\left|\frac{df^n(x_0)}{dx}\right|, f^n(x_0) + \delta\left|\frac{df^n(x_0)}{dx}\right|\right] , \tag{8.7}$$

if δ is small enough. Assuming a chaotic system, the new interval will be larger, which means that we have a less good descriptions of the future position of the system. Let us use this as an *a priori* description of the system position at this future point in time, using a uniform probability distribution that (by normalisation) equals

$$p_0 = \frac{1}{2\delta\left|\frac{df^n(x_0)}{dx}\right|} . \tag{8.8}$$

In these n time steps the dynamics will bring the system to some point within this interval. If we observe the system at the new time, using the same level of resolution, we get a new (better) estimate of the position, characterised by a smaller interval (length 2δ) and a probability $p = 1/(2\delta)$. The Kullback information associated with this observation is then

$$K[p_0;p] = \int dx\, p\ln\frac{p}{p_0} = \ln\left|\frac{df^n(x_0)}{dx}\right| .$$

In the limit of arbitrarily fine resolution and infinite time, the information I gained per time step is

$$I = \lim_{n\to\infty}\frac{1}{n}K[p_0;p] = \lim_{n\to\infty}\frac{1}{n}\ln\left|\frac{df^n(x_0)}{dx}\right| = \lambda\,.$$

This illustrates the interpretation of the Lyapunov exponent as a measure of information flow from finer to larger length scales.

8.3 Dynamical systems entropy and information flow

In earlier chapters we have characterised disorder in symbol sequences by the use of entropies. We shall make a similar approach here, to characterise the noise amplification in information-theoretic terms.

Consider again a dynamical system characterised by the iterative map $f(x)$, and suppose that it generates an ergodic measure μ. Let $\mathbf{A} = (A_1, A_2, \ldots, A_r)$ be a decomposition of the state space into non-overlapping subsets. For each part A_j we let $f^{-k}A_j$ denote the set of points that results in A_j if f is applied k times, i.e., $f^k(f^{-k}A_j) = A_j$. Based on this we construct a new decomposition $B^{(n)}$ of the state space, using the components $B_{i_1\ldots i_n}$, defined by

$$B_{i_1\ldots i_n} = A_{i_1}\bigcap f^{-1}A_{i_2}\bigcap\ldots\bigcap f^{-n+2}A_{i_{n-1}}\bigcap f^{-n+1}A_{i_n}\,, \tag{8.9}$$

where the indices $i_k \in \{1,\ldots, r\}$. A component B_σ in $B^{(n)}$ is then characterised by a sequence of indices $\sigma = i_1,\ldots, i_n$. The interpretation is as follows. If the system starts at time $t = 1$ with a state in B_σ, the system will in n consecutive time steps be found in the subsets: A_{i_1} at time 1, A_{i_2} at time 2, …, and finally A_{i_n} at time n, i.e., $x(t) \in A_{i_t}$. This means that the sequence $\sigma = i_1,\ldots, i_n$ can be viewed as a symbol sequence generated by the dynamics, given a certain basic partition of the state space **A**. In other words, the components in the partition $B^{(n)}$ correspond to subsets B_σ of state space that result in different symbol sequences σ generated by the dynamics. Each such subset has a certain probability measure, $\mu(B_\sigma)$, and we use this to define an entropy for the partition $B^{(n)}$,

$$H(B^{(n)}) = \sum_\sigma \mu(B_\sigma)\ln\frac{1}{\mu(B_\sigma)}\,, \tag{8.10}$$

corresponding to the block entropy in Chapter 4. We now define an entropy, similar to the Shannon entropy, by taking the limit of infinite time (or symbol block length),

$$h(\mu,\mathbf{A}) = \lim_{n\to\infty}\frac{1}{n}H(B^{(n)}) = \lim_{n\to\infty}\left(H(B^{(n+1)}) - H(B^{(n)})\right)\,. \tag{8.11}$$

This entropy depends on how we have chosen to make the partition **A** and which invariant measure that is used. Next, we take the limit of this entropy when the partition is made arbitrarily fine, i.e., the size of the subsets in **A** tends to zero, which defines the *measure entropy*

$$s_\mu = \lim_{diam(\mathbf{A})\to 0} h(\mu,\mathbf{A}), \tag{8.12}$$

where $diam(\mathbf{A})$ denotes the largest distance between two points in a subset A_j.

The entropy $h(\mu, \mathbf{A})$ is the average entropy per time step when observing the symbols (defined by the partition **A**) that are generated by the dynamics. If the system stabilises at a fixed point or a periodic orbit the symbol sequence will certainly have zero entropy. If the system is chaotic, the result may depend on how the partition is made, and therefore the limit used in the definition of measure entropy is needed. If $s_\mu > 0$, we cannot (always) tell which part of **A** the system will visit in the next time step, regardless of how many previous steps we have observed, and regardless of how fine we make the partition. Thus, there is a "creation" of information through the dynamics, and s_μ can be said to quantify the *average rate of creation of information*.

In some cases the limit $diam(\mathbf{A}) \to 0$ needs not be taken. If the partition **A** is done so that $diam(B^{(n)}) \to 0$ when $n \to \infty$, then the measure entropy is given already by the entropy based on **A**, Eq. (9.9). If this holds, **A** is a *generating partition*. In this case we have a finite decomposition of state space, each part represented by a certain symbol. The dynamics $x(t)$ is then represented by a sequence of symbols $s(t)$. The longer sequence we observe, the more accurate will our knowledge be about the real position of the system at $t = 0$. Each time step brings us new information, in average s_μ. A chaotic system has $s_\mu > 0$, and this can be viewed as an information flow from smaller length scales. It is reasonable to think that if the Lyapunov exponent corresponds to a divergence of a factor of two ($\lambda = \ln 2$) then there is one bit of information creation per time step ($s_\mu = \ln 2$). In fact, it often holds (in all cases that we will encounter) that the measure entropy equals the Lyapunov exponent if it is larger than 0,

$$s_\mu = \lambda, \tag{8.13}$$

but, of course, not if $\lambda < 0$.

8.3.1 Extended example of a generated partition for a skew roof map

Consider an iterated map defined by $f(0)=\alpha$, $f(\alpha)=1$, $f(1)=0$ (where $0<\alpha<1$), with $f(x)$ being linear on the intervals $[0, \alpha]$ and $[\alpha, 1]$, as illustrated in Fig. 8.1. By defining a partition so that points x in the first interval generate symbol A and in the second interval symbol B, the iterated map is transformed into a symbolic dynamics with the alphabet {A, B}. We let the symbols denote the intervals they correspond to: A = $[0, \alpha]$ and B = $]\alpha, 1]$. Following the procedure summarized in Eq. (8.9), we contruct the subsets of A and B corresponding to increasing lengths of sequences generated in the symbolic dynamics. For example, the set BBA is the subset that starts in B, in one iteration maps to B again, but in the next iteration is mapped to A,

$$\mathrm{BBA} = \mathrm{B} \cap f^{-1}(\mathrm{B}) \cap f^{-2}(\mathrm{A}). \tag{8.14}$$

By taking such sequences $x_1x_2\ldots x_m$ of increasing length m, where $x_j \in \{A, B\}$, we see that the corresponding sets decreases in size ($diam(x_1x_2\ldots x_m) \to 0$, as $m \to \infty$. This means that we have a generating partition.

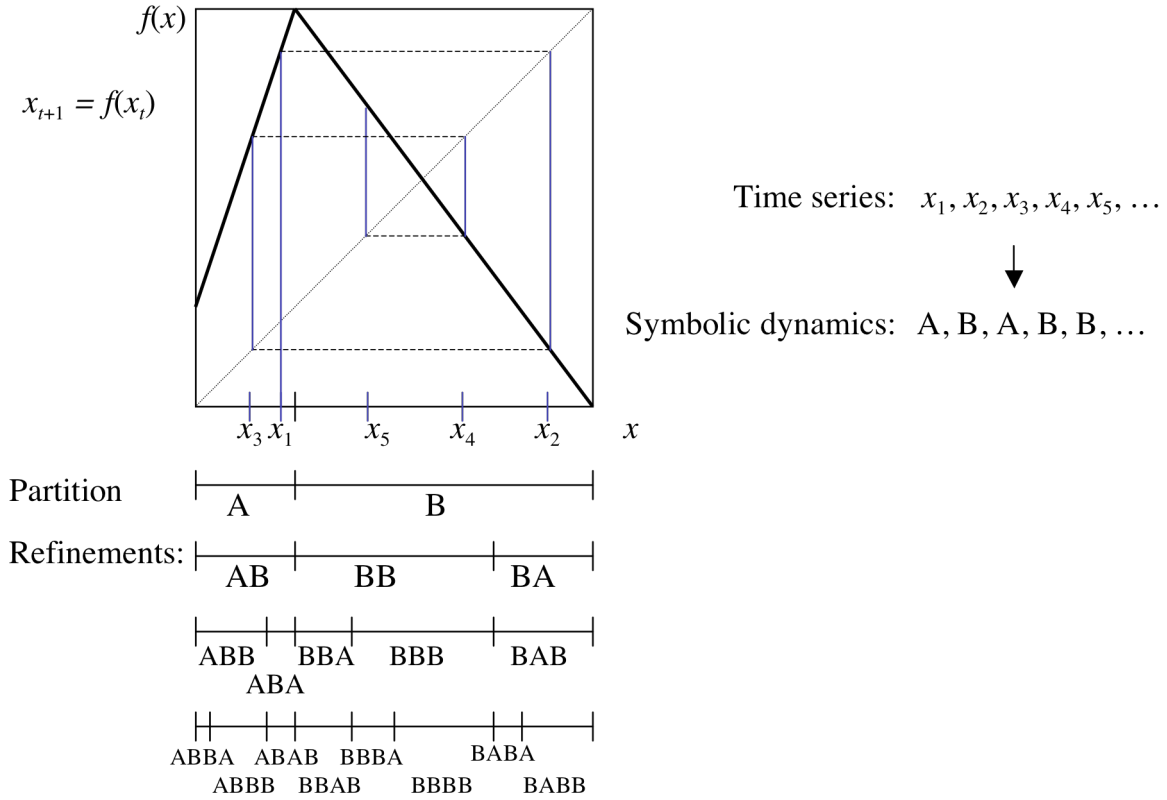


Figure 8.1 The graph shows an iterated map with $f(0)=\alpha$, $f(\alpha)=1$, $f(1)=0$, along with a partion associating points to the left and to the right of the peak with A and B, respectively. Two the right the corresponding symbolic dynamics is illustrated. The sets corresponding to A and B are then refined by finding subsets that in one iteration are mapped to the A and B, respectively. In this case it turns out that no points in A are mapped to A so there is no such subset. A continued refinement leads to decreasing subsets sizes, and we have a generating partition.

In this system, with the present choice of partition, we have a have a simple way in which the sets A and B map onto these sets or unions of them. We note that the set A maps onto the set B, so that if we have a uniform probability density (uniform measure) over A, that is evenly distributed in one iteration over the set B. Similarly, a uniform measure over B is evenly distributed over A∪B, the whole unit interval, in one iteration. The fraction of the probability density from B that ends up in A is then the |A|, i.e., the length of the A interval, and the corresponding fraction for B is |B|. Because of this, we can easily find an invariant distribution, since the requirement on invariant measure from Eq. (8.2),

$$\begin{aligned} \mu(f^{-1}(\mathrm{B})) &= \mu(\mathrm{B}) \\ \mu(f^{-1}(\mathrm{A})) &= \mu(\mathrm{A}) \end{aligned} \tag{8.15}$$

together with the result for this specific map f

$$\begin{aligned}\mu(f^{-1}(\mathrm{B})) &= \mu(\mathrm{A}) + |\mathrm{B}|\mu(\mathrm{B}) \\ \mu(f^{-1}(\mathrm{A})) &= |\mathrm{A}|\mu(\mathrm{B})\end{aligned} \tag{8.16}$$

results in the equations

$$\begin{aligned}\mu(\mathrm{B}) &= \mu(\mathrm{A}) + |\mathrm{B}|\mu(\mathrm{B}) \\ \mu(\mathrm{A}) &= |\mathrm{A}|\mu(\mathrm{B})\end{aligned} \tag{8.17}$$

The constraint that the total measure is 1, $\mu(\mathrm{A}) + \mu(\mathrm{B}) = 1$, replaces the last equation and we find that $\mu(\mathrm{B}) = 1/(2 - |\mathrm{B}|)$ and $\mu(\mathrm{A}) = (1 - |\mathrm{B}|)/(2 - |\mathrm{B}|)$. This procedure, with its solution, is equivalent to our analysis of finite state automata and the stationary distribution over the nodes, discussed in Chapter 3. In this example, the symbolic dynamics is in the form of a Markov process, defined by Eq. (8.17), and illustrated in Fig. 8.2. This type of partition, resulting in a Markov process for the symbolic dynamics is also called a *Markov partition*. The invariant measure for the sets A and B correspond to the stationary probability distribution over the nodes A and B, respectively.

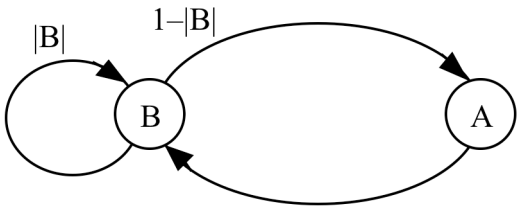


Figure 8.2 The finite state automaton representing the symbolic dynamics of Fig. 8.1 with an invariant measure that is uniform within A and B.

This means that the measure entropy s_μ, for this invariant measure, equals the entropy s of the stochastic process of Fig. 8.2. Based on our previous discussion on these processes, we conclude that

$$s_\mu = s = \mu(\mathrm{B})\left(|\mathrm{B}|\ln\frac{1}{|\mathrm{B}|} + (1 - |\mathrm{B}|)\ln\frac{1}{1 - |\mathrm{B}|}\right). \tag{8.18}$$

From Eq. (8.13), we can determine the Lyapunov exponent, $\lambda = s_\mu$.

When we know the invariant measure, we can also calculate the Lyapunov exponent directly from Eq. (8.6), using the fact that the slope $|f'(x)|$ equals $(1–|\mathrm{A}|)/|\mathrm{A}| = |\mathrm{B}|/(1–|\mathrm{B}|)$ in A and $1/|\mathrm{B}|$ in B. Then λ can be written

$$\begin{aligned}\lambda &= \int dx\mu(x)\ln|f'(x)| = \int_{\mathrm{A}} dx\mu(x)\ln\frac{|\mathrm{B}|}{1 - |\mathrm{B}|} + \int_{\mathrm{B}} dx\mu(x)\ln\frac{1}{|\mathrm{B}|} = \\ &= \frac{1 - |\mathrm{B}|}{2 - |\mathrm{B}|}\ln\frac{|\mathrm{B}|}{1 - |\mathrm{B}|} + \frac{1}{2 - |\mathrm{B}|}\ln\frac{1}{|\mathrm{B}|} = s\end{aligned} \tag{8.19}$$

Next, we discuss in what way a badly chosen partition can go wrong.

8.3.2 A partition that is *not* generating

Consider the tent map, $f(x)=1-|2x-1|$, see Fig. 8.3. Since this is a map with slope $|f'| = 2$ everywhere we know that all fixed points and periodic orbits are unstable. Furthermore, Eq. (8.6) says that any ergodic invariant measure results in a Lyapunov exponent $\lambda = \ln 2$. Then, from Eq. (8.13) we also know that the measure entropy = ln 2.

All this can be verified with the approach above, for example, by making a partition dividing the unit interval in two equal halves. Here, though, we will discuss what happens when one chooses a different partion, that are not necessarily generating. Therefore, we make a partition by dividing the unit interval at $x = 3/4$, so that A is the interval [0, 3/4] and B is]3/4, 1]. This partition with two steps of refinements are shown in Fig. 8.3. We note already at symbol sequences of length 2 that there is a set AA (corresponding to points for which the map generates a pair of A´s in two iterations) that is not connected. A further refinement, going to symbol sequences of length 3 shows that this set is further split up in disconnected part and same holds now for the set AAB. This means that there will exist sequences $x_1x_2\ldots x_m$ for which we do *not* have $diam(x_1x_2\ldots x_m) \to 0$, as $m \to \infty$. (The largest distance between two points in the set will typically be finitely separated since many sequences will have disconnected parts spread out over the unit interval.) Thus, this partition is *not* generating.

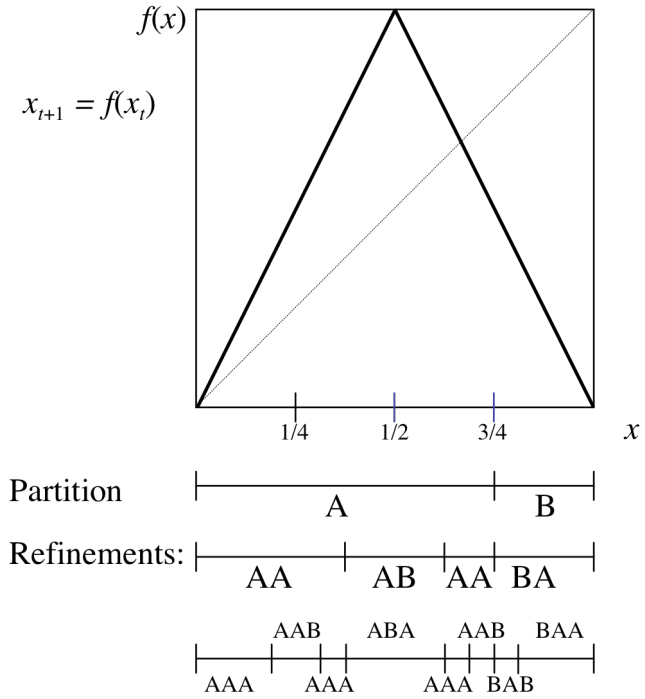


Figure 8.3 The graph shows a partition (A, B), for the tent map, that is not generating.

The symbolic dynamics resulting from this partion can still be analysed using the approach of the previous section. If the interval A is divided in three equal parts (at the points x=1/4 and x=1/2), A_1, A_2, and A_3, then the partition (A_1, A_2, A_3, B) is in fact generating. From the tent map we see that these intervals are mapped in one iteration as follows,

$$
\begin{aligned}
A_1 &\rightarrow A_1 \cup A_2 \\
A_2 &\rightarrow A_3 \cup B \\
A_3 &\rightarrow A_3 \cup B \\
B &\rightarrow A_1 \cup A_2
\end{aligned}
\tag{8.20}
$$

This means, using similar arguments as for the skew roof map, that the partition results in the symbolic dynamics being a Markov process with an automaton respresentation as in Fig. 8.4a. Since all intervals are of equal length, the transition probabilities are all ½. This Markov process has an entropy of ln 2, and gives the correct values for s_μ and thus also for λ.

The non-generating partition (A, B) has a symbolic dynamics that results from changing any symbol A_k to A. This means that the symbolic dynamics is a Hidden Markov model, shown in Fig. 8.4b, with an equivalent, more compact, representation in Fig. 8.4c. Even though the entropy involved in the choices in this automaton is still ln 2 the entropy of the symbol seqence (the Hidden Markov model) is smaller. This follows from the fact that there are different combinations of choices in the two A nodes (in Fig. 8.4c) that result in the same symbol sequence, i.e., some of the choices made does not show up in the symbol sequence, and the entropy of the symbolic dynamics is then smaller.

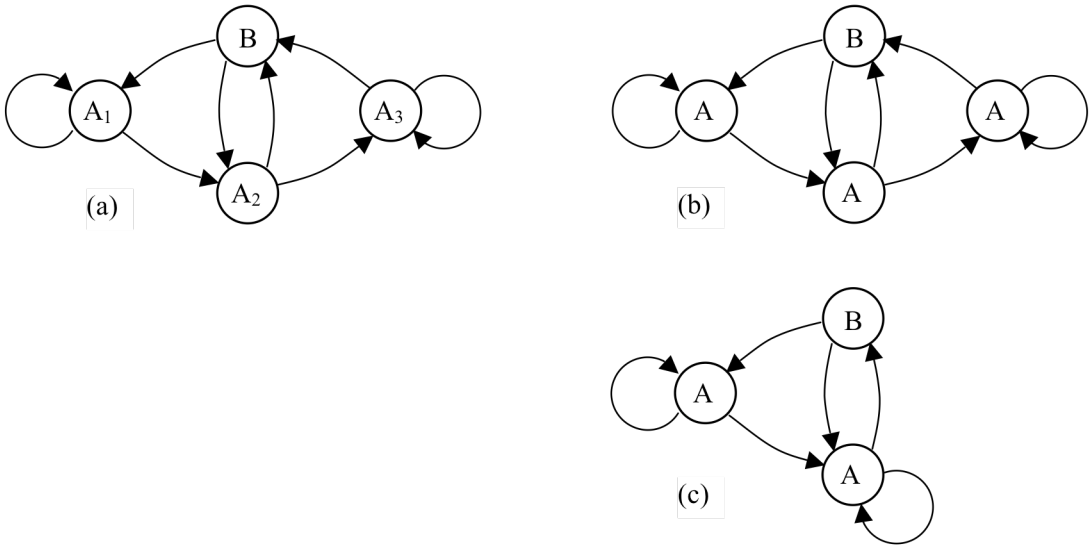
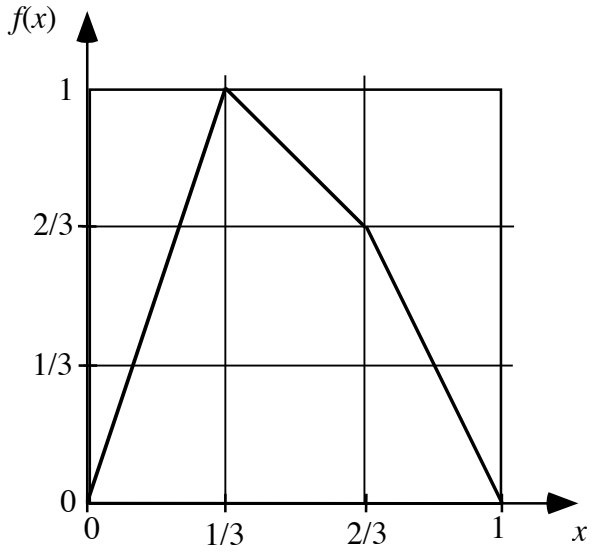


Figure 8.4 (a) The finite state automaton corresponding to the symbolic dynamics of the generating partition (A_1, A_2, A_3, B). (b) The hidden Markov model corresponding to the symbolic dynamics of the non-generating partition (A, B) of Fig. 8.3, and (c) an equivalent representation of the hidden Markov model.

8.4 Exercises

8.1 Let a mapping $f(x)$ be defined by the figure below, where $f(1/3) = 1$, $f(2/3) = 2/3$, and $f(0) = f(1) = 0$.



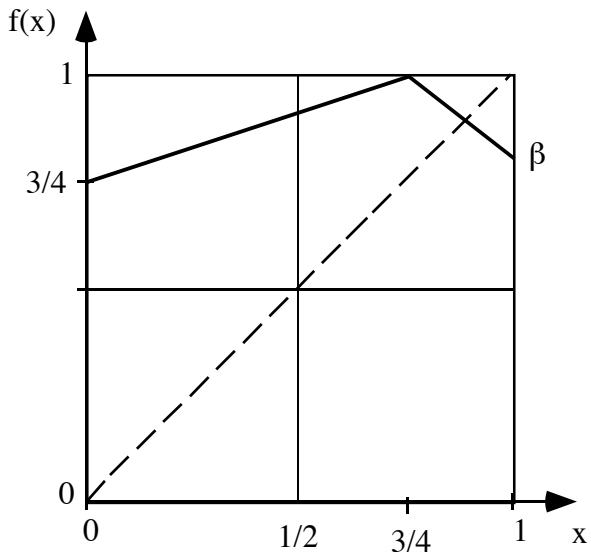
Consider the dynamical system

$$x_{t+1} = f(x_t) .$$

Find the invariant measure μ that characterises the chaotic behaviour, and determine the corresponding Lyapunov exponent λ by using that measure. Show that there is a partition that has a symbolic dynamics with a measure entropy s_μ that equals the Lyapunov exponent (as one should expect).

Suppose that we at a certain time t observe the system in the region given by $x > 2/3$. If we find the system in this region again two time steps later, how much information do we gain by this observation?

8.2 Let a piecewise linear mapping $f(x)$ be defined by the figure below, where $0 < \beta < 1$, and where the mapping is determined by $f(0) = 3/4$, $f(3/4) = 1$, and $f(1) = \beta$.



Consider the dynamical system

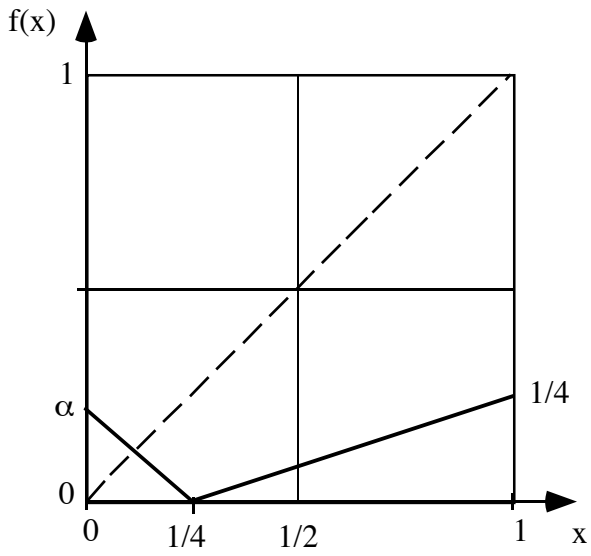
$$x_{t+1} = f(x_t) .$$

(a) At what value on β does the system become chaotic (starting with β being close to 1)? Characterize the dynamics for β above the critical value for chaos (stable/unstable fixed point and/or periodic). Is there any other value on β that corresponds to a change in the dynamics in the non-chaotic regime?

(b) Suppose that $\beta = 0$. Determine the invariant measure that characterizes the chaotic behaviour, and calculate the Lyapunov exponent λ. Calculate also the measure entropy from the finite state automaton describing the symbolic dynamics (for a generating partition).

(c) If you know that the system is in the region $x > 3/4$ at time t, how much information do you get if you observe the system in the region $x < 3/4$ at time $t+3$ (again assuming $\beta = 0$)?

8.3 Let a piecewise linear mapping $f(x)$ be defined by the figure below, where $0 < \alpha < 1$, and where the mapping is determined by $f(0) = \alpha$, $f(1/4) = 0$, and $f(1) = 1/4$.



Consider the dynamical system

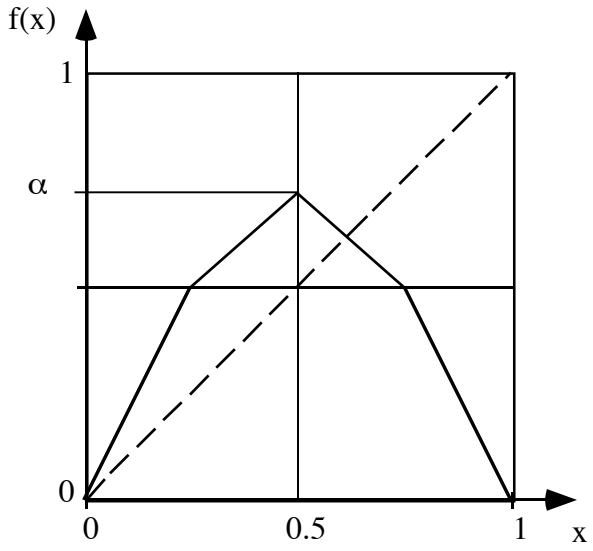
$$x_{t+1} = f(x_t) .$$

(a) At what value on α does the system become chaotic (starting with α being small)? Characterize the dynamics for α below the critical value for chaos (stable/unstable fixed point and/or periodic). Is there any other value on α that correspond to a drastic change in the dynamics in the non-chaotic regime?

(b) Suppose that $\alpha = 1$. Determine the invariant measure that characterizes the chaotic behaviour, and calculate the Lyapunov exponent λ. Calculate also the measure entropy from the finite state automaton describing the symbolic dynamics (for a generating partition).

(c) If you know that the system is in the region $x < 1/4$ at time t, how much information do you get if you observe the system in the same region again at time $t+3$ (again assuming $\alpha = 1$)?

8.4 Let a mapping $f(x)$ be defined by the figure below, where $1/2 < \alpha < 1$.



Consider the dynamical system

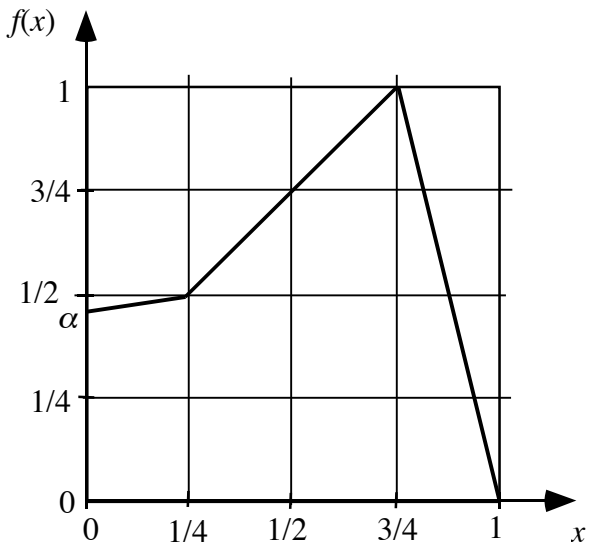
$x_{t+1} = f(x_t)$.

At what value on α does the system become chaotic?

Suppose that $\alpha = 7/8$. Determine the invariant measure that characterizes the chaotic behaviour, and calculate the Lyapunov exponent λ. Calculate also the measure entropy from the finite state automaton describing the symbolic dynamics (for a generating partition).

If you know that the system is in the region $x < 1/4$ at time t, how much information do you get if you observe the system in the same region again at time $t+3$?

8.5 Let a mapping $f(x)$ be defined by the figure below (so that $f(0) = \alpha$, $f(1/4) = 1/2$, $f(3/4) = 1$, $f(1) = 0$, with $0 \le \alpha \le 1/2$.



Consider the dynamical system

$$x_{t+1} = f(x_t) .$$

a) At which value of $\alpha < 1/2$ does the system become chaotic when decreasing from 1/2?

b) What is the behaviour for α close to 1/2 (above the critical value derived above). Describe qualitatively only.

Assume from now on that $\alpha = 0$.

c) Find the invariant measure μ that characterises the chaotic behaviour, and determine the corresponding Lyapunov exponent λ by using that measure. Find a partition that has a symbolic dynamics with a measure entropy s_μ that equals the Lyapunov exponent (as one should expect).

d) Suppose that we at a certain time t observe the system in the region given by $x > 3/4$. If we find the system in this region again three time steps later (at $t + 3$), how much information do we gain by this observation?

8.6 Suppose that for a continuous mapping $f(x)$, $0 \le x \le 1$, holds $f^2(0) = 1$, $f^3(0) = 0$, and that the function is linearly increasing in the interval $0 \le x \le f(0)$, and linearly decreasing in the interval $f(0) \le x \le 1$.

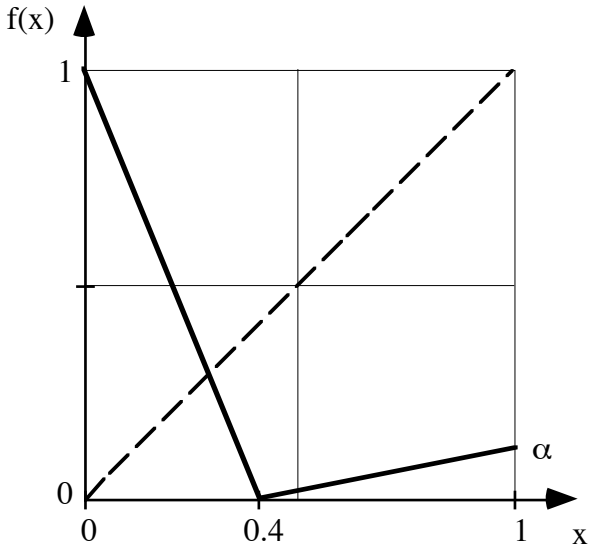
Consider the dynamical system

$$x_{t+1} = f(x_t) .$$

Determine at what value $f(0)$ the Lyapunov exponent λ has its maximum. What is the value of λ?

If an observer knows that the system at time t is found in the interval $0 \le x \le f(0)$, how much information does the observer gain when she also learns that the same holds at time $t + 4$?

8.7 Let a mapping $f(x)$ be defined by the figure below, where $f(1) = \alpha$ and $f(0.4) = 0$.



Consider the dynamical system

$$x_{t+1} = f(x_t) .$$

Describe the behaviour when α is small? At what value on α becomes the system chaotic?

Suppose that $\alpha = 2/7$. Calculate the Lyapunov exponent λ. What does a generating partition of the interval [0, 1] look like, and what is the measure entropy s_μ ?

What is the Lyapunov exponent if $\alpha = 2/5$?

9 Algorithmic information theory

The information-theoretic properties discussed so far requires a probabilistic description of the system studied. Such probabilities may come from a statistical analysis, either from a collection of data from a study of several systems, or, as in the case for symbol sequences, from some internal statistics of the system. The internal statistics may detect correlations between symbols which can be exploited to encode the system in a more compact way, as was illustrated in Chapter 3. But there are also other internal structures in a symbol sequence that could be used for making a more compact representation of the system. One example is if the symbol sequence represent the digits in a comupatable irrational number, like the decimal digits in π. In such a situation there is a very short algorithm generating the symbols, if the sequence is long, compared to what one could get by a statistically based compression.

In algorithmic information theory the basic idea is to define information as the length of the most compact description of the system. Formally this is done by using an abstract general computing device, a Turing machine, that transforms a description, in the form of a computer programme, into the desired system as an output. The algorithmic information of a system is then defined as the length of the smallest computer program that produces the system as an output. The formal definition of these concepts and some extensions is presented in the following sections. Sometimes the shortest description, the algorithmic information, is also referred to as the *algorithmic complexity*.

9.1 The Turing machine

The type of Turing machine that we need for the definition of algorithmic information is a universal one. This means that it is capable of performing any computation. Any *computational structure* that is present in the system can then be exploited to find a more compact description. A computationally universal Turing machine can also simulate any other Turing machine.

We will use the following, fairly general definition, of a computationally universal Turing machine. There is a *control unit* that is communicating with a *program tape* and a another tape, the *output tape*, which is used for both memory and output, as indicated in Fig. 9.1. If one need an input, separate from the program, that is initially placed on the output tape. The output tape is unbounded (infinite). We assume that a binary alphabet is used.

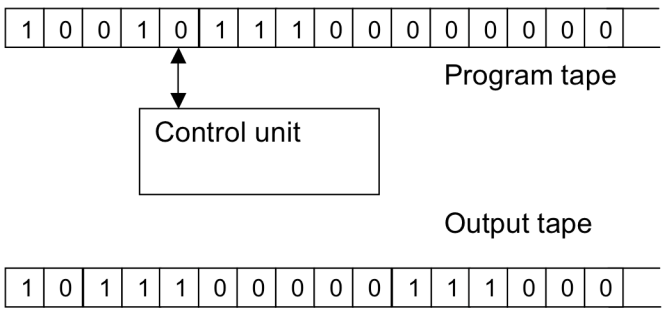


Figure 9.1 The Turing machine

At each computational step, the control unit is located at certain position on the output tape. The program determines the action to be performed, which depends on the internal state q of the control unit and the symbol at the current position on the output tape (the input for the computational step). The actions are: move left (L) or right (R) on the output tape, write a symbol (0) or (1) on the current position, or halt (H).

The set of states Q of the control unit contains one start state q_0, one final state s (the halt state), and a number of intermediary states q_i. The set I of inputs is {0, 1}, and the set A of actions is {L, R, 0, 1}. A program step is a mapping from Q x I to Q x A and can be written as

$$q_i \; i \; a \; q_j \,, \tag{9.1}$$

where q_i is the internal state, I the symbol of the current output tape position, a the action to be chosen, and q_j the new internal state. If $q_j = s$, the program halts. For example, if the tape head is positioned within a block of 1's on the output tape, the following programme shifts that sequence one step to the right,

q_0 0 0 s
q_0 1 L q_1
q_1 1 L q_1
q_1 0 R q_2
q_2 1 0 q_3
q_3 1 R q_3
q_3 0 1 s

The first three lines move the control unit to the leftmost 1 of the sequence, then that is changed to a 0. The remaining part of the program finds the first 0 to the right and replaces that with a 1 before the program halts.

The *halting problem* from computation theory applies to the universal computer, stating that there is no general procudere that can determine whether a computer programme will ever halt.

9.2 Algorithmic information

Based on the computationally universal Turing machine, denoted U, described above, we will now make a formal definition of the algorithmic information quantity. The algorithmic information $H_U(\alpha_m)$ of a symbol sequence α_m of lengt m is defined by the size of the smallest sum of the lengths of program P and input X that generates the sequence $\alpha_m = U(P, X)$ as an output,

$$H_U(\alpha_m) = \min_{U(P,X)=\alpha_m} l(P) + l(X) \,. \tag{9.2}$$

The abstract computer U described in the previous section is only one example of a computationally universal machine. But a computationally universal machine can simulate

any computational process, which includes the simulation of any other machine. The simulation program of another machine has a certain, but finite, length. Therefore, one can say that a general algorithmic information quantity equals any quantity based on a specific universal computer C up to a constant of order $O(1)$,

$$H(\alpha_m) = H_C(\alpha_m) + O(1). \tag{9.3}$$

A sequence α_m of length m that cannot be described more compactly than itself, e.g., by a program saying "print the input sequence" which can be done by directly entering the halt state with the desired sequence α_m placed as an input on the output tape. Then $l(X)=l(\alpha_m)=m$ and $H(\alpha_m) = m + O(1)$.

9.3 Relations between algorithmic and Shannon-based information

Because of the halting problem, the algorithmic information is not a generally computable quantity. There is no general procedure to tell you whether a given program with its input is the smallest. One cannot always test all combinations of program and input since some of them may lead to compuational processes that never halts, but for which we cannot know that.

However, for seqences α (in the limit of inifinite length) generated by some stationary stochastic process, we will hardly ever find one wich happen to contain some computational structure (e.g., from the digits in a computable irrational number). We define the average algorithmic information per symbol,

$$h(\alpha) = \lim_{m\to\infty} \frac{1}{m} H(\alpha_m). \tag{9.4}$$

In this situation, one can only hope for a statistical structure, coming from the stochastic process, that makes it possible to find a more compact representation than the seqence itself. By letting the program P serve as a decoder working on the most compressed coded sequence, we will get a finite program of length l_P taking as input a coded sequence x_m of length $s{\cdot}m$ (for large m). Here s is the ordinary information theory entropy per symbol of the sequence α. This means that (Zhvonkin and Levin 1970, Brudno 1977, Lindgren 1987), for seqences generated by a stochastic process, almost allways the algorithmic information equals the the information-theoretic entropy per symbol,

$$h(\alpha) = \lim_{m\to\infty} \frac{1}{m}\left(l_P + m\cdot s + O(1)\right) = s. \tag{9.5}$$

Following Chaitin (1979), we can define a measure of computational structure at different lengths. Let $\omega_d(\alpha_m)$ be a decomposition of α_m into non-overlapping sub-sequences of lengths $\leq d$. Further, let $H_U(\gamma)$, be the algorithmic information of a sub-sequence γ in $\omega_d(\alpha_m)$. Define the composed algorithmic information $L_U(\omega_d)$ as the sum of the description lengths from the sub-sequences γ, where the length for each γ is its length $l(\gamma)$ or its algorithmic information, whichever is the smallest,

$$L_U(\omega_d) = \sum_{\gamma \in \omega_d} \min\bigl(H_U(\gamma), l(\gamma)\bigr). \tag{9.6}$$

Now, we define the smallest description, defined in this way, that only exploits structures at lengths up to d by taking the minimum over all possible decompositions $\omega_d(\alpha_m)$,

$$L_U^{(d)}(\alpha_m) = \min_{\omega_d}\bigl(L_U(\omega_d)\bigr). \tag{9.7}$$

We note that

$$m \geq L_U^{(m)}(\alpha_m) = H_U(\alpha_m) + O(1), \tag{9.8}$$

and that $L_U^{(1)}(\alpha_m) = m$. It is also clear from Eqs. (9.6) and (9.7) that

$$L_U^{(d)}(\alpha_m) \leq L_U^{(d-1)}(\alpha_m). \tag{9.9}$$

The difference between two such description lengths, comparing d and d–1, quantifies the algorthmic information $C_U^{(d)}(\alpha_m)$ in "computational structures" of length d,

$$C_U^{(d)}(\alpha_m) = L_U^{(d-1)}(\alpha_m) - L_U^{(d)}(\alpha_m), \qquad d = 2, 3, \ldots, m. \tag{9.10}$$

The sum of all these quantities results in $L_U^{(1)}(\alpha_m) - L_U^{(m)}(\alpha_m) = m - L_U^{(m)}(\alpha_m)$. Therefore we can view the following expression as a decomposition of the total information m of α_m into contributions from computational structures of different lengths and the remaining algoritmic information, $L_U^{(m)}(\alpha_m)$,

$$m = \sum_{k=2}^{m} C_U^{(d)}(\alpha_m) + L_U^{(m)}(\alpha_m). \tag{9.11}$$

From this it is clear that one may interpret the algorithmic information $L_U^{(m)}(\alpha_m)$ as "algorithmic randomness". This decomposition is the computational equivalent to the information decomposition of Eq. (3.23). But since computational structures are not necessarily statistically detectable, there is no direct relation between the concepts in Chapter 3 and those presented here.

9.4 Exercises

9.1 What can be said bout the change in the algorithmic information $h(\alpha)$ in one time-step of a deterministic CA rule when α is the (inifinte) state of the CA?

10 Hints and answers to selected problems

Answers to some problems in Chapter 2

2.1 log 51

2.2 $\frac{1}{2}\left(-1+\frac{b_1^{\,2}}{b_2^{\,2}}\right)+\log\left[\frac{b_2}{b_1}\right]$

2.3 - log (2(1 – F(2))), where F(2) = P(x<2) for the standardised normal distribution (Gaussian with mean 0 and variance 1).

2.4 Homework problem 1.1

2.5 $p(k) = 1/3$, $k = 0$, 1, and 2.

2.6 $-\log[e^{-1}]$
$-\log[1 - e^{-1}]$

2.7 $e^{-\lambda\,\Delta t} = ½$, gives $\Delta t = (\ln 2)/\lambda$

2.8 $p(k_1 \ldots k_n) = a^K(1\text{-}a)^{n-K}$ with $K = \Sigma_j\, k_j$

2.9 $\log b \;+\; (1 + \log(2\pi))/2$

Answers to some problems in Chapter 3

3.1 We only aim at minimising the length of the coded message in the limit of an infinite sequence, and we do not attempt in minimising the start and end of the coding. Therefore, (i) initially use 0's as 1's uncoded until this first 1 appears; (ii) from then on use 1 and 0 only when preceeding symbol is one; if preceeding symbol is 0 no code word is needed; (iii) if the message ends with a sequence of 0's all of them are expressed in the coded sequence (as in the initial part). In this way the average length is compressed down to a fraction $s = 2/3$ of the originial sequence.

3.2 $s = 1$ (bit)

3.3 $s = 1/2$ (bit), with arbitrarily long correlations.

3.4 Homework problem 2.1

3.6 $s = -(p \log p + (1 - p) \log(1 - p))/(2 + p)$, vary p to maximise s.

3.9 Homework problem 2.2

Answers to some problems in Chapter 4

4.1 $s(0) = s(1) = 2/3$ (bit); same holds for R22.

4.2 $s(0) = s(1) = s(2) = 2/5$ (bit); from exam 2005, solution on web page.

4.3 $s(0) = S[\{\rho/4, 1-\rho/4\}]$, s is conserved in the time evolution.
For entropy estimate based on finite blocks (after long time), see solution on web page (from 2005 exam).

4.4 $s(0) = s(1) = 2/7$ (bit); $s(2) = 0$

4.5 Homework problem 3.1

4.6 $s(0) = s(1) = s(2) = 1/2$ (bit); in general $s(t) = \frac{1}{2} \cdot S[\{q, 1-q\}]$

4.7 $\eta = S[\{q, 1-q\}] + \log 2$

4.8 $s(0) = s(1) = s(2) = 1/2$ (bit); for transition probability $\neq \frac{1}{2}$ one gets correlation information appearing in the same way as for R60 (in the example in the lecture notes).

4.9 $s(t) = 1/2$ (bit); the CA reaches a fixed point after one time step.

Answers to some problems in Chapter 5

5.1 Homework problem 4.1

5.2 This results in three decoupled systems, each with a solution as in Eqs. (5.39-40).

5.3

$$p_0 = P(\uparrow\uparrow), p_1 = P(\uparrow\downarrow), p_2 = P(\uparrow\rightarrow);$$

$$p_2 = e^{-\mu} = \frac{1}{4(e^{\beta J} + e^{-\beta J} + 2)}, p_0 = e^{-\mu+\beta J}, p_2 = e^{-\mu-\beta J}$$

5.4 Discussed in the lecture. Summary:

$$p_0 = P(\uparrow\uparrow\downarrow), p_1 = P(\uparrow\uparrow\uparrow), p_2 = P(\uparrow\downarrow\uparrow);$$

$$p_0^{\,2} = p_1 p_2 e^{4\beta J}, \frac{p_1}{p_0 + p_1} = e^{-\mu}, \frac{p_2}{p_0 + p_2} = e^{-\mu-2\beta J}$$

5.5 Solution on course web page, see exam March 2005.

5.6 Formulate problem by setting up the following table:

config.	energy	notation	multipl.
AA	J	p_0	1
BB	J	p_1	1
CC	J	p_0	1
AC	J	p_2	2
AB	–J	p_3	2
BC	–J	p_3	2

p(A) = $p_0 + p_2 + p_3$, p(B) = $p_1 + 2p_3$, p(C) = p(A).
$s = S_2 - S_1$,
$L = s + \beta(u - J(2p_0 + p_1 + 2p_2 - 4p_3)) + \mu(1 - 2p_0 - p_1 - 2p_2 - 4p_3)$.
Derivation w resp. to p_k gives four eq. + normalisation, which solves the problem…

5.7 (a) No, this automaton contains arbitrarily long correlations, while the equilibrium state is limited to lengths up to m.
(b) See course web site (exam problem 2002-2b).

5.8 Homework problem 4.2

5.9 Single cell probabilities: P(empty) = p_0, P(single) = p_1, P(pair) = p_2.
$p_0 = 1 - \rho + p_2$, $p_1 = \rho - 2p_2$,
$(\rho - 2p_2)^2 = (1 - \rho + p_2)\, p_2 \exp(-\beta J)$.
which gives p_2 and the rest. At T=0 (infinite β), $p_2 = \rho/2$, $p_0 = 1 - \rho/2$, $p_2 = 0$.

11 Literature

Brillouin, L. (1956). *Science and information theory*, Academic Press, New York.

Brudno, A. A. (1977). "On the complexity of trajectories in dynamical systems", (in Russian) *Usp. Mat. Nauk.* **33**, 207.

Boltzmann, L. (1872). "Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen", *Sitzungsberichte der Mathematisch-Naturwissenschaftlichen Classe der Kaiserlichen Akademie der Wissenschaften* **66**, II Abth., 275-370.

Capurro, R., and B. Hjørland (2003). Chapter 8, pp. 343-411 in *Annual Review of Information Science and Technology* Ed. B. Cronin, Vol. 37.

Chaitin, G. J. (1966). "On the length of programs for computing finite binary sequences", *Journal of the Association for Computing Machinery* **13**, 547-569.

Chomsky, N. (1956). "Three models for the description of language", *IRE Transaction on Information Theory* **2**, 113.

Clausius, R. (1850). "Über die bewegende Kraft der Wärmeund die Gesetze welche sich daraus für die Wärmelehre selbst ableiten lassen", *Annalen der Physik und Chemie* **79**, 368-397 and 500-524.

Clausius, R. (1865). "Über verschiedene für die Anwendung bequeme Formen der Hauptgleichung der mechanischen Wärmetheorie", *Annalen der Physik und Chemie* **125**, 353-400.

Cook, M. (2004). Universality in elementary cellular automata. *Complex Systems* **15**, 1-40.

Cvitanovic, P. (red.) (1984). *Universality in Chaos*, Adam Hilger Ltd., Bristol.

Eckman, J.-P. and D. Ruelle (1985). "Ergodic theory of chaos and strange attractors", *Reviews of Modern Physics* **57**, 617.

Einstein, A. (1902). Kinetische Theorie des Wärmegleichgewichtes und des zweiten Hauptsatzes der Thermodynamik, *Annalen der Physik* **9**, 417-433.

Einstein, A. (1903). "Eine Theorie der Grundlagen der Thermodynamik", *Annalen der Physik* **11**, 170-187.

Eriksson, K.-E. and K. Lindgren (1987). Structural information in self-organizing systems, *Physica Scripta* **35**, 388-397.

Eriksson, K.-E., K. Lindgren and B. Å. Månsson (1987) *Structure, Context, Complexity, Organization*, World Scientific, Singapore.

Gibbs, J. W. (1873). Page 53 in *Collected Works*, Vol. 1, Yale University Press, New Haven, 1948. First published in *Transactions of the Connecticut Academy*, Vol. 2, 382.

Gibbs, J. W. (1902). In *Collected Works*, Vol. 2, Yale University Press, New Haven, 1948.

Gödel, K. (1931). Über formal unentschiedbare Sätze der Principia Mathematica und verwandter Systeme I, *Monatshefte für Mathematic und Physik*. English translation in M. Davies, *The Undecidable*, Raven Press, 1965.

Grassberger, P. (1986). Towards a quantitative theory of self-generated complexity, *International Journal of Theoretical Physics* **25**, 907-938.

Gray, P. and Scott, S. K. (1984). "Autocatalytic reactions in the isothermal, continuous stirred tank reactor – oscillations and instabilities in the system A+2B→3B, B→C." *Chem. Eng. Sci*., **39**(6):1087–1097.

Haken, H. (1984). *Advanced Synergetics*, Springer Verlag, Berlin.

Hartley, R.V.L. (1928). Transmission of Information, *Bell System Technical Journal*, Vol. 7, July 1928, pp. 535-563.

Helvik, T., K. Lindgren, and M.G. Nordahl (2004). "Local information in one-dimensional cellular automata," in proceedings for ACRI-2004: *From individual to collective behaviour*, Amsterdam, October, 2004. Published in *Springer Lecture Notes in Computer Science*, Vol. 3305, pp. 121-130 (Springer).

Helvik, T., K. Lindgren, and M. G. Nordahl (2007). Continuity of information transport in surjective cellular automata, *Communications in Mathematical Physics* **272**, 53-74.

Hopcroft, J. E. and J. D. Ullman (1979). *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Reading, Massachusets.

Jaynes, E. T. (1957). Information theory and statistical mechanics", *Physical Review* **106**, 620.

Kelvin, Lord [Thomson, W.] (1852). On a universal tendency in Nature to the dissipation of mechanical energy, *Proceedings of the Royal Society of Edinburgh* **3**, 139-142; *Philosophical Magazine*, ser. 4, **4**, 304-306.

Kolmogorov, A. N. (1965) Three approaches to the quantitative definition of information, *Problemy Peredachi Informatsii* **1**, 3-11 (In Russian). English translation in *Problems of Information Transmission* **1**, 1-7.

Krönig, A. K. (1856). Grundzüge einer Theorie der Gase, *Annalen der Physik und Chemie* **99**, 315-322.

Kullback, S. (1959). *Information Theory and Statistics*, Wiley, New York.

Lee, K. J., McCormick, W. D., Ouyang, Q., and Swinney, H. L. (1993). "Pattern formation by interacting chemical fronts." *Science*, **261**(5118):192–194.

Lindgren, K. (1987). Correlations and random information of cellular automata, *Complex Systems* **1**, 529-543.

Lindgren, K. (1988). Microscopic and macroscopic entropy, *Physical Review* **A38**, 4794-4798.

Lindgren, K. and M. G. Nordahl (1988). Complexity measures and cellular automata, *Complex Systems* **2**, 409-440.

Lindgren, K., Eriksson, A., and Eriksson, K.-E. (2004). "Flows of information in spatially extended chemical dynamics." *Artificial Life IX Proceedings*, Edited by J. Pollack, M. Bedau, P. Husbands, T. Ikegami and R. A. Watson (MIT Press).

Mandelbrot, B. B. (1983). *The Fractal Geometry of Nature*, W H Freeman, New York.

Nicolis, G. and I. Prigogine (1977). *Self-Organization in Non-Equilibrium Systems*, John Wiley and Sons, New York.

Nyquist, H. (1924). Certain Factors Affecting Telegraph Speed, *Bell System Technical Journal*, Vol. 3, April 1924, pp. 324-346.

Nyquist, H. (1928). Certain Topics in Telegraph Transmission Theory, *A.I.E.E. Transactions*, Vol. 47, April 1928, pp. 617-644.

Oseledec, V. I. (1968). A multiplicative ergodic theorem. Ljapunov characteristic numbers for dynamical systems, *Trans. Moscow Math. Soc.* **19**, 197.

Pearson, J. E. (1993). "Complex patterns in a simple system." *Science*, **261**(5118):189–192.

Shannon, C. E. (1948). A Mathematical Theory of Communication, *Bell System Technical Journal*, Vol. 27, July 1948, pp. 379-423 and October 1948, pp. 623-656.

Tribus, M. and E.C. McIrvine (1971). Energy and information, *Scientific American*, 224 (September 1971), 178–184.

Wehrl, A. (1978). General properties of entropy, *Rev. Mod. Phys.* **50**, 221–260.

Segré, D., D. Ben-Eli, and D. Lancet (2000). "Compositional genomes: Prebiotic information transfer in mutually catalytic noncovalent assemblies." *PNAS* **97**, 4112-4117.

Shaw, R. (1981). Strange attractors, chaotic behavior, and information flow", *Zeitschrift für Naurforschung* **36a**, 80.

Solomonoff, R. J. (1964). A formal theory of inductive inference, *Information and Control* **7**, 1-22.

Steinbeck, J. (1937). *Of Mice and Men*, Viking Press.

Turing, A. (1936-37). On computable numbers, with application to the Entscheidungsproblem, *Proceedings of London Mathematical Society* **42**, 230 and **43**, 544.

Von Neumann, J. (1966). *Theory of Self-Reproducing Automata*, edited by A. W. Burks, University of Illinois, Urbana.

Wolfram, S. (1983). Statistical Mechanics of Cellular Automata, *Reviews of Modern Physics* **55**, 601-644.

Wolfram, S. (1984). Computtaion theory of cellular automata, *Communications in Mathematical Physics* **96**, 15.

Wolfram, S. (Ed.) (1986) *Theory and Applications of Cellular Automata*, World Scientific, Singapore.

von Neumann, J., and A. W. Burks (1966). *Theory of Self-Reproducing Automata*, Univ. of Illionois Press, Urbana.

Zhvonkin, A. K. and L. A. Levin (1970). The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms, *Russian Mathematical Surveys* **25**, 83-124.