1

INTRODUCTION TO INFORMATION THEORY

{ch:intro_info}

This chapter introduces some of the basic concepts of information theory, as well as the definitions and notations of probabilities that will be used throughout the book. The notion of entropy, which is fundamental to the whole topic of this book, is introduced here. We also present the main questions of information theory, data compression and error correction, and state Shannon's theorems.

1.1 Random variables

The main object of this book will be the behavior of large sets of **discrete random variables**. A discrete random variable X is completely defined[1] by the set of values it can take, $\mathcal{X}$, which we assume to be a finite set, and its **probability distribution** $\{p_X(x)\}_{x\in\mathcal{X}}$. The value $p_X(x)$ is the probability that the random variable X takes the value x. The probability distribution $p_X : \mathcal{X} \to [0,1]$ must satisfy the normalization condition

$$\sum_{x\in\mathcal{X}} p_X(x) = 1 \ . \tag{1.1}$$ {proba_norm}

We shall denote by $\mathbb{P}(A)$ the probability of an **event** $A \subseteq \mathcal{X}$, so that $p_X(x) = \mathbb{P}(X = x)$. To lighten notations, when there is no ambiguity, we use $p(x)$ to denote $p_X(x)$.

If $f(X)$ is a real valued function of the random variable X, the **expectation value** of $f(X)$, which we shall also call the average of f, is denoted by:

$$\mathbb{E}\, f = \sum_{x\in\mathcal{X}} p_X(x) f(x) \ . \tag{1.2}$$

While our main focus will be on random variables taking values in finite spaces, we shall sometimes make use of **continuous random variables** taking values in $\mathbb{R}^d$ or in some smooth finite-dimensional manifold. The probability measure for an 'infinitesimal element' dx will be denoted by $dp_X(x)$. Each time p_X admits a density (with respect to the Lebesgue measure), we shall use the notation $p_X(x)$ for the value of this density at the point x. The total probability $\mathbb{P}(X \in \mathcal{A})$ that the variable X takes value in some (Borel) set $\mathcal{A} \subseteq \mathcal{X}$ is given by the integral:

[1] In probabilistic jargon (which we shall avoid hereafter), we take the probability space $(\mathcal{X}, \mathsf{P}(\mathcal{X}), p_X)$ where $\mathsf{P}(\mathcal{X})$ is the σ-field of the parts of $\mathcal{X}$ and $p_X = \sum_{x\in X} p_X(x)\, \delta_x$.

Info-Phys Comp, Draft 9

$$\mathbb{P}(X \in \mathcal{A}) = \int_{x\in\mathcal{A}} dp_X(x) = \int \mathbb{I}(x \in \mathcal{A})\, dp_X(x)\ , \tag{1.3}$$

where the second form uses the **indicator function** $\mathbb{I}(s)$ of a logical statement s,which is defined to be equal to 1 if the statement s is true, and equal to 0 if the statement is false.

The expectation value of a real valued function $f(x)$ is given by the integral on $\mathcal{X}$:

$$\mathbb{E}\, f(X) = \int f(x)\ dp_X(x)\ . \tag{1.4}$$

Sometimes we may write $\mathbb{E}_X f(X)$ for specifying the variable to be integrated over. We shall often use the shorthand **pdf** for the **probability density function** $p_X(x)$.

Example 1.1 A fair dice with M faces has $\mathcal{X} = \{1, 2, ..., M\}$ and $p(i) = 1/M$ for all $i \in \{1, ..., M\}$. The average of x is $\mathbb{E}\, X = (1 + ... + M)/M = (M + 1)/2$.

Example 1.2 Gaussian variable: a continuous variable $X \in \mathbb{R}$ has a Gaussian distribution of mean m and variance σ^2 if its probability density is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{[x-m]^2}{2\sigma^2}\right)\ . \tag{1.5}$$

One has $\mathbb{E}X = m$ and $\mathbb{E}(X - m)^2 = \sigma^2$.

The notations of this chapter mainly deal with discrete variables. Most of the expressions can be transposed to the case of continuous variables by replacing sums $\sum_x$ by integrals and interpreting $p(x)$ as a probability density.

Exercise 1.1 Jensen's inequality. Let X be a random variable taking value in a set $\mathcal{X} \subseteq \mathbb{R}$ and f a convex function (i.e. a function such that $\forall x, y$ and $\forall \alpha \in [0, 1]$: $f(\alpha x + (1 - \alpha y)) \leq \alpha f(x) + (1 - \alpha) f(y)$). Then

{eq:Jensen}

$$\mathbb{E} f(X) \geq f(\mathbb{E}X)\ . \tag{1.6}$$

Supposing for simplicity that $\mathcal{X}$ is a finite set with $|\mathcal{X}| = n$, prove this equality by recursion on n.

{se:entropy}

1.2 Entropy

The **entropy** H_X of a discrete random variable X with probability distribution $p(x)$ is defined as

{S_def}

$$H_X \equiv -\sum_{x\in\mathcal{X}} p(x) \log_2 p(x) = \mathbb{E} \log_2 \left[\frac{1}{p(X)}\right]\ , \tag{1.7}$$

Info 's Draft

where we define by continuity $0 \log_2 0 = 0$. We shall also use the notation $H(p)$ whenever we want to stress the dependence of the entropy upon the probability distribution of X.

In this Chapter we use the logarithm to the base 2, which is well adapted to digital communication, and the entropy is then expressed in **bits**. In other contexts one rather uses the natural logarithm (to base $e \approx 2.7182818$). It is sometimes said that, in this case, entropy is measured in **nats**. In fact, the two definitions differ by a global multiplicative constant, which amounts to a change of units. When there is no ambiguity we use H instead of H_X.

Intuitively, the entropy gives a measure of the uncertainty of the random variable. It is sometimes called the missing information: the larger the entropy, the less a priori information one has on the value of the random variable. This measure is roughly speaking the logarithm of the number of typical values that the variable can take, as the following examples show.

Example 1.3 A fair coin has two values with equal probability. Its entropy is 1 bit.

Example 1.4 Imagine throwing M fair coins: the number of all possible outcomes is 2^M. The entropy equals M bits.

Example 1.5 A fair dice with M faces has entropy $\log_2 M$.

Example 1.6 Bernouilli process. A random variable X can take values $0, 1$ with probabilities $p(0) = q$, $p(1) = 1 - q$. Its entropy is

$$H_X = -q \log_2 q - (1-q) \log_2 (1-q) \ , \tag{1.8}$$ {S_bern}

it is plotted as a function of q in fig.1.1. This entropy vanishes when $q = 0$ or $q = 1$ because the outcome is certain, it is maximal at $q = 1/2$ when the uncertainty on the outcome is maximal.

Since Bernoulli variables are ubiquitous, it is convenient to introduce the function $\mathcal{H}(q) \equiv -q \log q - (1-q) \log(1-q)$, for their entropy.

Exercise 1.2 An unfair dice with four faces and $p(1) = 1/2$, $p(2) = 1/4$, $p(3) = p(4) = 1/8$ has entropy $H = 7/4$, smaller than the one of the corresponding fair dice.

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

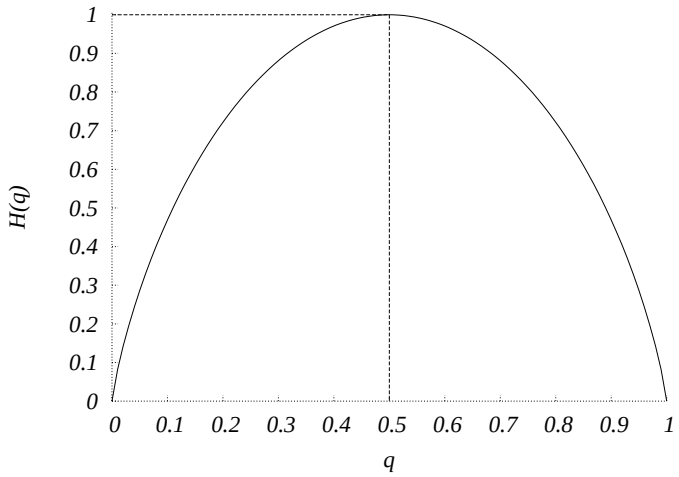

FIG. 1.1. The entropy $\mathcal{H}(q)$ of a binary variable with $p(X = 0) = q$, $p(X = 1) = 1 - q$, plotted versus q {fig_bernouilli}

Exercise 1.3 DNA is built from a sequence of bases which are of four types, A,T,G,C. In natural DNA of primates, the four bases have nearly the same frequency, and the entropy per base, if one makes the simplifying assumptions of independence of the various bases, is $H = -\log_2(1/4) = 2$. In some genus of bacteria, one can have big differences in concentrations: $p(G) = p(C) = 0.38$, $p(A) = p(T) = 0.12$, giving a smaller entropy $H \approx 1.79$.

Exercise 1.4 In some intuitive way, the entropy of a random variable is related to the 'risk' or 'surprise' which are associated to it. In this example we discuss a simple possibility for making these notions more precise.

Consider a gambler who bets on a sequence of bernouilli random variables $X_t \in \{0, 1\}$, $t \in \{0, 1, 2, \dots\}$ with mean $\mathbb{E}X_t = p$. Imagine he knows the distribution of the X_t's and, at time t he bets a fraction $w(1) = p$ of his money on 1 and a fraction $w(0) = (1-p)$ on 0. He looses whatever is put on the wrong number, while he doubles whatever has been put on the right one. Define the average doubling rate of his wealth at time t as

$$W_t = \frac{1}{t}\,\mathbb{E}\,\log_2\left\{\prod_{t'=1}^{t} 2w(X_{t'})\right\}\,. \tag{1.9}$$

It is easy to prove that the expected doubling rate $\mathbb{E}W_t$ is related to the entropy of X_t: $\mathbb{E}W_t = 1 - \mathcal{H}(p)$. In other words, it is easier to make money out of predictable events.

Another notion that is directly related to entropy is the **Kullback-Leibler**

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

(KL) divergence between two probability distributions $p(x)$ and $q(x)$ over the same finite space $\mathcal{X}$. This is defined as:

$$D(q||p) \equiv \sum_{x\in\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} \tag{1.10}$$

where we adopt the conventions $0 \log 0 = 0$, $0 \log(0/0) = 0$. It is easy to show that: (i) $D(q||p)$ is convex in $q(x)$; (ii) $D(q||p) \geq 0$; (iii) $D(q||p) > 0$ unless $q(x) \equiv p(x)$. The last two properties derive from the concavity of the logarithm (i.e. the fact that the function $-\log x$ is convex) and Jensen's inequality (1.6): if $\mathbb{E}$ denotes expectation with respect to the distribution $q(x)$, then $-D(q||p) = \mathbb{E} \log[p(x)/q(x)] \leq \log \mathbb{E}[p(x)/q(x)] = 0$. The KL divergence $D(q||p)$ thus looks like a distance between the probability distributions q and p, although it is not symmetric.

The importance of the entropy, and its use as a measure of information, derives from the following properties:

1. $H_X \geq 0$.
2. $H_X = 0$ if and only if the random variable X is certain, which means that X takes one value with probability one.
3. Among all probability distributions on a set $\mathcal{X}$ with M elements, H is maximum when all events x are equiprobable, with $p(x) = 1/M$. The entropy is then $H_X = \log_2 M$.
   Notice in fact that, if $\mathcal{X}$ has M elements, then the KL divergence $D(p||\overline{p})$ between $p(x)$ and the uniform distribution $\overline{p}(x) = 1/M$ is $D(p||\overline{p}) = \log_2 M - \mathcal{H}(p)$. The thesis follows from the properties of the KL divergence mentioned above.
4. If X and Y are two **independent** random variables, meaning that $p_{X,Y}(x,y) = p_X(x)p_Y(y)$, the total entropy of the pair X, Y is equal to $H_X + H_Y$:

$$\begin{aligned} H_{X,Y} &= -\sum_{x,y} p(x,y) \log_2 p_{X,Y}(x,y) = \\ &= -\sum_{x,y} p_X(x)p_Y(y) \left(\log_2 p_X(x) + \log_2 p_Y(y)\right) = H_X + H_Y \end{aligned} \tag{1.11}$$

5. For any pair of random variables, one has in general $H_{X,Y} \leq H_X + H_Y$, and this result is immediately generalizable to n variables. (The proof can be obtained by using the positivity of the KL divergence $D(p_1||p_2)$, where $p_1 = p_{X,Y}$ and $p_2 = p_X p_Y$). ⋆
6. Additivity for composite events. Take a finite set of events $\mathcal{X}$, and decompose it into $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$, where $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$. Call $q_1 = \sum_{x\in\mathcal{X}_1} p(x)$ the probability of $\mathcal{X}_1$, and q_2 the probability of $\mathcal{X}_2$. For each $x \in \mathcal{X}_1$, define as usual the conditional probability of x, given that $x \in \mathcal{X}_1$, by $r_1(x) = p(x)/q_1$ and define similarly $r_2(x)$ as the conditional probability

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

of x, given that $x \in \mathcal{X}_2$. Then the total entropy can be written as the sum of two contributions $H_X = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) = H(q) + H(r)$, where:

$$H(q) = -q_1 \log_2 q_1 - q_2 \log_2 q_2 \tag{1.12}$$

$$H(r) = -q_1 \sum_{x\in\mathcal{X}_1} r_1(x)\log_2 r_1(x) - q_2 \sum_{x\in\mathcal{X}_1} r_2(x)\log_2 r_2(x) \tag{1.13}$$

⋆ The proof is obvious by just substituting the laws r_1 and r_2 by their expanded definitions. This property is interpreted as the fact that the average information associated to the choice of an event x is additive, being the sum of the relative information $H(q)$ associated to a choice of subset, and the information $H(r)$ associated to the choice of the event inside the subsets (weighted by the probability of the subsetss). It is the main property of the entropy, which justifies its use as a measure of information. In fact, this is a simple example of the so called chain rule for conditional entropy, which will be further illustrated in Sec. 1.4.

Conversely, these properties together with some hypotheses of continuity and monotonicity can be used to define axiomatically the entropy.

{sec:RandomVarSequences}

1.3 Sequences of random variables and entropy rate

In many situations of interest one deals with a random process which generates **sequences of random variables** $\{X_t\}_{t\in\mathbb{N}}$, each of them taking values in the same finite space $\mathcal{X}$. We denote by $P_N(x_1,\dots,x_N)$ the joint probability distribution of the first N variables. If $A \subset \{1,\dots,N\}$ is a subset of indices, we shall denote by $\overline{A}$ its complement $\overline{A} = \{1,\dots,N\} \setminus A$ and use the notations $\underline{x}_A = \{x_i, i \in A\}$ and $\underline{x}_{\overline{A}} = \{x_i, i \in \overline{A}\}$. The **marginal distribution** of the variables in A is obtained by summing P_N on the variables in $\overline{A}$:

$$P_A(\underline{x}_A) = \sum_{\underline{x}_{\overline{A}}} P_N(x_1,\dots,x_N)\ . \tag{1.14}$$

Example 1.7 The simplest case is when the X_t's are **independent**. This means that $P_N(x_1,\dots,x_N) = p_1(x_1)p_2(x_2)\dots p_N(x_N)$. If all the distributions p_i are identical, equal to p, the variables are **independent identically distributed**, which will be abbreviated as **iid**. The joint distribution is

$$P_N(x_1,\dots,x_N) = \prod_{t=1}^{N} p(x_i)\,. \tag{1.15}$$

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

Example 1.8 The sequence $\{X_t\}_{t\in\mathbb{N}}$ is said to be a **Markov chain** if

$$P_N(x_1,\dots,x_N) = p_1(x_1)\prod_{t=1}^{N-1} w(x_t\to x_{t+1})\,. \tag{1.16}$$

Here $\{p_1(x)\}_{x\in\mathcal{X}}$ is called the **initial state**, and $\{w(x\to y)\}_{x,y\in\mathcal{X}}$ are the **transition probabilities** of the chain. The transition probabilities must be non-negative and normalized:

$$\sum_{y\in\mathcal{X}} w(x\to y) = 1\,, \qquad \text{for any } y\in\mathcal{X}. \tag{1.17}$$

When we have a sequence of random variables generated by a certain process, it is intuitively clear that the entropy grows with the number N of variables. This intuition suggests to define the **entropy rate** of a sequence $\{X_t\}_{t\in\mathbb{N}}$ as

$$h_X = \lim_{N\to\infty} H_{\underline{X}_N}/N\,, \tag{1.18}$$

if the limit exists. The following examples should convince the reader that the above definition is meaningful.

Example 1.9 If the X_t's are i.i.d. random variables with distribution $\{p(x)\}_{x\in\mathcal{X}}$, the additivity of entropy implies

$$h_X = H(p) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)\,. \tag{1.19}$$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Example 1.10 Let $\{X_t\}_{t\in\mathbb{N}}$ be a Markov chain with initial state $\{p_1(x)\}_{x\in\mathcal{X}}$ and transition probabilities $\{w(x\to y)\}_{x,y\in\mathcal{X}}$. Call $\{p_t(x)\}_{x\in\mathcal{X}}$ the marginal distribution of X_t and assume the following limit to exist independently of the initial condition:

$$p^*(x) = \lim_{t\to\infty} p_t(x)\,. \tag{1.20}$$

As we shall see in chapter 4, this turns indeed to be true under quite mild hypotheses on the transition probabilities $\{w(x\to y)\}_{x,y\in\mathcal{X}}$. Then it is easy to show that

$$h_X = -\sum_{x,y\in\mathcal{X}} p^*(x)\,w(x\to y)\log w(x\to y)\,. \tag{1.21}$$

If you imagine for instance that a text in English is generated by picking letters randomly in the alphabet $\mathcal{X}$, with empirically determined transition probabilities $w(x\to y)$, then Eq. (1.21) gives a first estimate of the entropy of English. But if you want to generate a text which looks like English, you need a more general process, for instance one which will generate a new letter x_{t+1} given the value of the k previous letters $x_t, x_{t-1}, ..., x_{t-k+1}$, through transition probabilities $w(x_t, x_{t-1}, ..., x_{t-k+1} \to x_{t+1})$. Computing the corresponding entropy rate is easy. For $k = 4$ one gets an entropy of 2.8 bits per letter, much smaller than the trivial upper bound $\log_2 27$ (there are 26 letters, plus the space symbols), but many words so generated are still not correct English words. Some better estimates of the entropy of English, through guessing experiments, give a number around 1.3.

{se:CorrelatedVariables}

1.4 Correlated variables and mutual entropy

Given two random variables X and Y, taking values in $\mathcal{X}$ and $\mathcal{Y}$, we denote their joint probability distribution as $p_{X,Y}(x,y)$, which is abbreviated as $p(x,y)$, and the conditional probability distribution for the variable y given x as $p_{Y|X}(y|x)$, abbreviated as $p(y|x)$. The reader should be familiar with Bayes' classical theorem:

$$p(y|x) = p(x,y)/p(x)\ . \tag{1.22}$$

When the random variables X and Y are independent, $p(y|x)$ is x-independent. When the variables are dependent, it is interesting to have a measure on their degree of dependence: how much information does one obtain on the value of y if one knows x? The notions of conditional entropy and mutual entropy will be useful in this respect.

Let us define the **conditional entropy** $H_{Y|X}$ as the entropy of the law $p(y|x)$, averaged over x:

{Scond_def}

$$H_{Y|X} \equiv -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x)\log_2 p(y|x)\,. \tag{1.23}$$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

The total entropy $H_{X,Y} \equiv -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log_2 p(x,y)$ of the pair of variables x, y can be written as the entropy of x plus the conditional entropy of y given x:

$$H_{X,Y} = H_X + H_{Y|X}\,. \tag{1.24}$$

In the simple case where the two variables are independent, $H_{Y|X} = H_Y$, and $H_{X,Y} = H_X + H_Y$. One way to measure the correlation of the two variables is the **mutual entropy** $I_{X,Y}$ which is defined as:

$$I_{X,Y} \equiv \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)}\ . \tag{1.25}$$ {Smut_def}

It is related to the conditional entropies by:

$$I_{X,Y} = H_Y - H_{Y|X} = H_X - H_{X|Y}\ , \tag{1.26}$$

which shows that $I_{X,Y}$ measures the reduction in the uncertainty of x due to the knowledge of y, and is symmetric in x, y.

Proposition 1.11 *$I_{X,Y} \geq 0$. Moreover $I_{X,Y} = 0$ if and only if X and Y are independent variables.*

Proof: Write $-I_{X,Y} = \mathbb{E}_{x,y}\log_2 \frac{p(x)p(y)}{p(x,y)}$. Consider the random variable $u = (x,y)$ with probability distribution $p(x,y)$. As the logarithm is a concave function (i.e. -log is a convex function), one and applies Jensen's inequality (1.6). This gives the result $I_{X,Y} \geq 0$ □

Exercise 1.5 A large group of friends plays the following game (telephone without cables). The guy number zero chooses a number $X_0 \in \{0,1\}$ with equal probability and communicates it to the first one without letting the others hear, and so on. The first guy communicates the number to the second one, without letting anyone else hear. Call X_n the number communicated from the n-th to the $(n+1)$-th guy. Assume that, at each step a guy gets confused and communicates the wrong number with probability p. How much information does the n-th person have about the choice of the first one?

We can quantify this information through $I_{X_0,X_n} \equiv I_n$. A simple calculation shows that $I_n = 1 - \mathcal{H}(p_n)$ with p_n given by $1 - 2p_n = (1-2p)^n$. In particular, as $n \to \infty$

$$I_n = \frac{(1-2p)^{2n}}{2\log 2}\left[1 + O((1-2p)^{2n})\right]\ . \tag{1.27}$$

The 'knowledge' about the original choice decreases exponentially along the chain.

The mutual entropy gets degraded when data is transmitted or processed. This is quantified by:

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Proposition 1.12 ***Data processing inequality***.

Consider a Markov chain $X \to Y \to Z$ (so that the joint probability of the three varaibles can be written as $p_1(x)w_2(x \to y)w_3(y \to z)$). Then: $I_{X,Z} \leq I_{X,Y}$. In particular, if we apply this result to the case where Z is a function of Y, $Z = f(Y)$, we find that applying f degrades the information: $I_{X,f(Y)} \leq I_{X,Y}$.

Proof: Let us introduce, in general, the mutual entropy of two varaibles conditioned to a third one: $I_{X,Y|Z} = H_{X|Z} - H_{X,(YZ)}$. The mutual information between a variable X and a pair of varaibles (YZ) can be decomposed in a sort of **chain rule**: $I_{X,(YZ)} = I_{X,Z} + I_{X,Y|Z} = I_{X,Y} + I_{X,Z|Y}$. If we have a Markov chain $X \to Y \to Z$, X and Z are independent when one conditions on the value of Y, therefore $I_{X,Z|Y} = 0$. The result follows from the fact that $I_{X,Y|Z} \geq 0$. □

1.5 Data compression

Imagine an information source which generates a sequence of symbols $\underline{X} = \{X_1, \ldots, X_N\}$ taking values in a finite alphabet $\mathcal{X}$. Let us assume a probabilistic model for the source: this means that the X_i's are taken to be random variables. We want to store the information contained in a given realization $\underline{x} = \{x_1 \ldots x_N\}$ of the source in the most compact way.

This is the basic problem of **source coding**. Apart from being an issue of utmost practical interest, it is a very instructive subject. It allows in fact to formalize in a concrete fashion the intuitions of 'information' and 'uncertainty' which are associated to the definition of entropy. Since entropy will play a crucial role throughout the book, we present here a little *detour* into source coding.

1.5.1 *Codewords*

We first need to formalize what is meant by "storing the information". We define[2] therefore a **source code** for the random variable $\underline{X}$ to be a mapping w which associates to any possible information sequence in $\mathcal{X}^N$ a string in a reference alphabet which we shall assume to be $\{0, 1\}$:

$$\begin{aligned} w : \mathcal{X}^N &\to \{0,1\}^* \\ \underline{x} &\mapsto w(\underline{x})\,. \end{aligned} \tag{1.28}$$

Here we used the convention of denoting by $\{0, 1\}^*$ the set of binary strings of arbitrary length. Any binary string which is in the image of w is called a **codeword**.

Often the sequence of symbols $X_1 \ldots X_N$ is a part of a longer stream. The compression of this stream is realized in three steps. First the stream is broken into blocks of length N. Then each block is encoded separately using w. Finally the codewords are glued to form a new (hopefully more compact) stream. If the original stream consisted in the blocks $\underline{x}^{(1)}, \underline{x}^{(2)}, \ldots, \underline{x}^{(r)}$, the output of the

[2]The expert will notice that here we are restricting our attention to "fixed-to-variable" codes.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

encoding process will be the concatenation of $w(\underline{x}^{(1)}), \dots, w(\underline{x}^{(r)})$. In general there is more than one way of parsing this concatenation into codewords, which may cause troubles to any one willing to recover the compressed data. We shall therefore require the code w to be such that any concatenation of codewords can be parsed unambiguously. The mappings w satisfying this property are called **uniquely decodable codes**.

Unique decodability is surely satisfied if, for any pair $\underline{x}, \underline{x}' \in \mathcal{X}^N$, $w(\underline{x})$ is not a prefix of $w(\underline{x}')$. If this stronger condition is verified, the code is said to be **instantaneous** (see Fig. 1.2). Hereafter we shall focus on instantaneous codes, since they are both practical and (slightly) simpler to analyze.

Now that we precised how to store information, namely using a source code, it is useful to introduce some figure of merit for source codes. If $l_w(x)$ is the length of the string $w(x)$, the **average length** of the code is:

$$L(w) = \sum_{\underline{x} \in \mathcal{X}^N} p(\underline{x})\, l_w(\underline{x}) \ . \tag{1.29}$$ {avlength}

Example 1.13 Take $N = 1$ and consider a random variable X which takes values in $\mathcal{X} = \{1, 2, \dots, 8\}$ with probabilities $p(1) = 1/2$, $p(2) = 1/4$, $p(3) = 1/8$, $p(4) = 1/16$, $p(5) = 1/32$, $p(6) = 1/64$, $p(7) = 1/128$, $p(8) = 1/128$. Consider the two codes w_1 and w_2 defined by the table below

x	$p(x)$	$w_1(x)$	$w_2(x)$
1	1/2	000	0
2	1/4	001	10
3	1/8	010	110
4	1/16	011	1110
5	1/32	100	11110
6	1/64	101	111110
7	1/128	110	1111110
8	1/128	111	11111110

(1.30)

These two codes are instantaneous. For instance looking at the code w_2, the encoded string 10001101110010 can be parsed in only one way since each symbol 0 ends a codeword. It thus corresponds to the sequence $x_1 = 2, x_2 = 1, x_3 = 1, x_4 = 3, x_5 = 4, x_6 = 1, x_7 = 2$. The average length of code w_1 is $L(w_1) = 3$, the average length of code w_2 is $L(w_2) = 247/128$. Notice that w_2 achieves a shorter average length because it assigns the shortest codeword (namely 0) to the most probable symbol (i.e. 1).

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

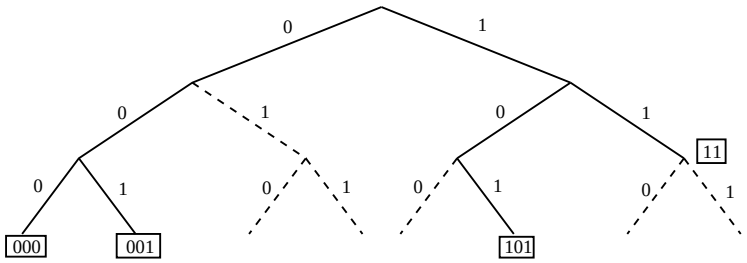

FIG. 1.2. An instantaneous source code: each codeword is assigned to a node in a binary tree in such a way that no one among them is the ancestor of another one. Here the four codewords are framed. {fig_kraft}

Example 1.14 A useful graphical representation of source code is obtained by drawing a binary tree and associating each codeword to the corresponding node in the tree. In Fig. 1.2 we represent in this way a source code with $|\mathcal{X}^N| = 4$. It is quite easy to recognize that the code is indeed instantaneous. The codewords, which are framed, are such that no codeword is the ancestor of any other codeword in the tree. Given a sequence of codewords, parsing is immediate. For instance the sequence `00111000101001` can be parsed only in `001, 11, 000, 101, 001`

1.5.2 *Optimal compression and entropy*

Suppose to have a 'complete probabilistic characterization' of the source you want to compress. What is the 'best code' w for this source? What is the shortest achievable average length?

This problem was solved (up to minor refinements) by Shannon in his celebrated 1948 paper, by connecting the best achievable average length to the entropy of the source. Following Shannon we assume to know the probability distribution of the source $p(\underline{x})$ (this is what 'complete probabilistic characterization' means). Moreover we interpret 'best' as 'having the shortest average length'.

{theorem:ShannonSource}

Theorem 1.15 *Let L_N^* the shortest average length achievable by an instantaneous code for $\underline{X} = \{X_1, \dots, X_N\}$, and $H_{\underline{X}}$ the entropy of the same variable. Then*

1. *For any $N \geq 1$:*

{Shcomp1}
$$H_{\underline{X}} \leq L_N^* \leq H_{\underline{X}} + 1\,. \tag{1.31}$$

2. *If the source has a finite entropy rate $h = \lim_{N\to\infty} H_{\underline{X}}/N$, then*

{Shcomp2}
$$\lim_{N\to\infty} \frac{1}{N} L_N^* = h\,. \tag{1.32}$$

Proof: The basic idea in the proof of Eq. (1.31) is that, if the codewords were too short, the code wouldn't be instantaneous. 'Kraft's inequality' makes

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

this simple remark more precise. For any instantaneous code w, the lengths $l_w(\underline{x})$ satisfy:

$$\sum_{\underline{x}\in\mathcal{X}^N} 2^{-l_w(\underline{x})} \leq 1\,. \tag{1.33}$$ {kraft}

This fact is easily proved by representing the set of codewords as a set of leaves on a binary tree (see fig.1.2). Let L_M be the length of the longest codeword. Consider the set of all the 2^{L_M} possible vertices in the binary tree which are at the generation L_M, let us call them the 'descendants'. If the information $\underline{x}$ is associated with a codeword at generation l (i.e. $l_w(\underline{x}) = l$), there can be no other codewords in the branch of the tree rooted on this codeword, because the code is instantaneous. We 'erase' the corresponding 2^{L_M-l} descendants which cannot be codewords. The subsets of erased descendants associated with each codeword are not overlapping. Therefore the total number of erased descendants, $\sum_{\underline{x}} 2^{L_M-l_w(\underline{x})}$, must be smaller or equal to the total number of descendants, 2^{L_M}. This establishes Kraft's inequality.

Conversely, for any set of lengths $\{l(\underline{x})\}_{\underline{x}\in\mathcal{X}^N}$ which satisfies the inequality (1.33) there exist at least a code, whose codewords have the lengths $\{l(\underline{x})\}_{\underline{x}\in\mathcal{X}^N}$. A possible construction is obtained as follows. Consider the smallest length $l(\underline{x})$ and take the first allowed binary sequence of length $l(\underline{x})$ to be the codeword for $\underline{x}$. Repeat this operation with the next shortest length, and so on until you have exhausted all the codewords. It is easy to show that this procedure is successful if Eq. (1.33) is satisfied.

The problem is therefore reduced to finding the set of codeword lengths $l(\underline{x}) = l^*(\underline{x})$ which minimize the average length $L = \sum_{\underline{x}} p(\underline{x})l(\underline{x})$ subject to Kraft's inequality (1.33). Supposing first that $l(\underline{x})$ are real numbers, this is easily done with Lagrange multipliers, and leads to $l(\underline{x}) = -\log_2 p(\underline{x})$. This set of optimal lengths, which in general cannot be realized because some of the $l(\underline{x})$ are not integers, gives an average length equal to the entropy $H_{\underline{X}}$. This gives the lower bound in (1.31). In order to build a real code with integer lengths, we use

$$l^*(\underline{x}) = \lceil -\log_2 p(\underline{x}) \rceil\,. \tag{1.34}$$

Such a code satisfies Kraft's inequality, and its average length is less or equal than $H_{\underline{X}} + 1$, proving the upper bound in (1.31).

The second part of the theorem is a straightforward consequence of the first one. □

The code we have constructed in the proof is often called a **Shannon code**. For long strings ($N \gg 1$), it gets close to optimal. However it has no reason to be optimal in general. For instance if only one $p(x)$ is very small, it will code it on a very long codeword, while shorter codewords are available. It is interesting to know that, for a given source $\{X_1, \ldots, X_N\}$, there exists an explicit construction of the optimal code, called Huffman's code.

At first sight, it may appear that Theorem 1.15, together with the construction of Shannon codes, completely solves the source coding problem. But this is far from true, as the following arguments show.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

From a computational point of view, the encoding procedure described above is unpractical. One can build the code once for all, and store it somewhere, but this requires $O(|\mathcal{X}|^N)$ memory. On the other hand, one could reconstruct the code each time a string requires to be encoded, but this takes $O(|\mathcal{X}|^N)$ time. One can use the same code and be a bit smarter in the encoding procedure, but this does not improve things dramatically.

From a practical point of view, the construction of a Shannon code requires an accurate knowledge of the probabilistic law of the source. Suppose now you want to compress the complete works of Shakespeare. It is exceedingly difficult to construct a good model for the source 'Shakespeare'. Even worse: when you will finally have such a model, it will be of little use to compress Dante or Racine.

Happily, source coding has made tremendous progresses in both directions in the last half century.

{sec:DataTransmission}

1.6 Data transmission

In the previous pages we considered the problem of encoding some information in a string of symbols (we used bits, but any finite alphabet is equally good). Suppose now we want to communicate this string. When the string is transmitted, it may be corrupted by some noise, which depends on the physical device used in the transmission. One can reduce this problem by adding redundancy to the string. The redundancy is to be used to correct (some) transmission errors, in the same way as redundancy in the English language can be used to correct some of the typos in this book. This is the field of channel coding. A central result in information theory, again due to Shannon's pioneering work in 1948, relates the level of redundancy to the maximal level of noise that can be tolerated for error-free transmission. The entropy again plays a key role in this result. This is not surprising in view of the symmetry between the two problems. In data compression, one wants to reduce the redundancy of the data, and the entropy gives a measure of the ultimate possible reduction. In data transmission, one wants to add some well tailored redundancy to the data.

1.6.1 *Communication channels*

The typical flowchart of a communication system is shown in Fig. 1.3. It applies to situations as diverse as communication between the earth and a satellite, the cellular phones, or storage within the hard disk of your computer. Alice wants to send a message m to Bob. Let us assume that m is a M bit sequence. This message is first encoded into a longer one, a N bit message denoted by $\underline{x}$ with $N > M$, where the added bits will provide the redundancy used to correct for transmission errors. The encoder is a map from $\{0,1\}^M$ to $\{0,1\}^N$. The encoded message is sent through the communication channel. The output of the channel is a message $\underline{y}$. In a noiseless channel, one would simply have $\underline{y} = \underline{x}$. In a realistic channel, $\underline{y}$ is in general a string of symbols different from $\underline{x}$. Notice that $\underline{y}$ is not even necessarily a string of bits. The **channel** will be described by the transition probability $Q(\underline{y}|\underline{x})$. This is the probability that the received signal is

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

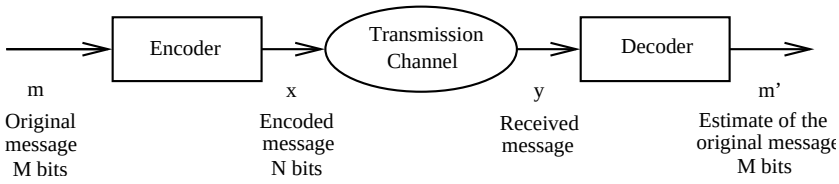

FIG. 1.3. Typical flowchart of a communication device. {fig_channel}

$\underline{y}$, conditional to the transmitted signal being $\underline{x}$. Different physical channels will be described by different $Q(\underline{y}|\underline{x})$ functions. The decoder takes the message $\underline{y}$ and deduces from it an estimate m' of the sent message.

Exercise 1.6 Consider the following example of a channel with **insertions**. When a bit `x` is fed into the channel, either `x` or `x0` are received with equal probability 1/2. Suppose that you send the string `111110`. The string `1111100` will be received with probability $2 \cdot 1/64$ (the same output can be produced by an error either on the 5^{th} or on the 6^{th} digit). Notice that the output of this channel is a bit string which is always longer or equal to the transmitted one.

A simple code for this channel is easily constructed: use the string `100` for each `0` in the original message and `1100` for each `1`. Then for instance you have the encoding

$$\texttt{01101} \mapsto \texttt{100110011001001100}\,. \tag{1.35}$$

The reader is invited to define a decoding algorithm and verify its effectiveness.

Hereafter we shall consider **memoryless** channels. In this case, for any input $\underline{x} = (x_1, ..., x_N)$, the output message is a string of N letters, $\underline{y} = (y_1, ..., y_N)$, from an alphabet $\mathcal{Y} \ni y_i$ (not necessarily binary). In memoryless channels, the noise acts independently on each bit of the input. This means that the conditional probability $Q(\underline{y}|\underline{x})$ factorizes:

$$Q(\underline{y}|\underline{x}) = \prod_{i=1}^{N} Q(y_i|x_i)\ , \tag{1.36}$$

and the transition probability $Q(y_i|x_i)$ is i independent.

Example 1.16 Binary symmetric channel (BSC). The input x_i and the output y_i are both in $\{0, 1\}$. The channel is characterized by one number, the probability p that an input bit is transmitted as the opposite bit. It is customary to represent it by the diagram of Fig. 1.4.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

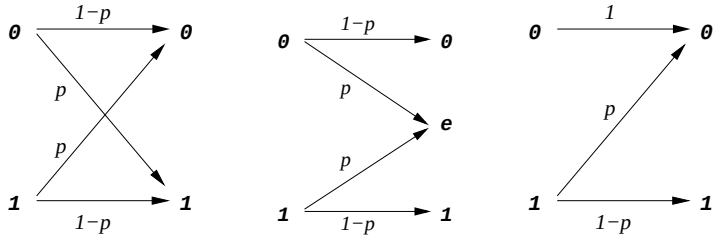

FIG. 1.4. Three communication channels. *Left*: the binary symmetric channel. An error in the transmission, in which the output bit is the opposite of the input one, occurs with probability p. *Middle*: the binary erasure channel. An error in the transmission, signaled by the output `e`, occurs with probability p. *Right*: the Z channel. An error occurs with probability p whenever a `1` is transmitted. {fig_bsc}

Example 1.17 Binary erasure channel (BEC). In this case some of the input bits are erased instead of being corrupted: x_i is still in $\{0, 1\}$, but y_i now belongs to $\{0, 1, e\}$, where `e` means erased. In the symmetric case, this channel is described by a single number, the probability p that a bit is erased, see Fig. 1.4.

Example 1.18 Z channel. In this case the output alphabet is again $\{0, 1\}$. Moreover, a `0` is always transmitted correctly, while a `1` becomes a `0` with probability p. The name of this channel come from its graphical representation, see Fig. 1.4.

A very important characteristics of a channel is the **channel capacity** C. It is defined in terms of the mutual entropy I_{XY} of the variables X (the bit which was sent) and Y (the signal which was received), through:

{capadef}
$$C = \max_{p(x)} I_{XY} = \max_{p(x)} \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{1.37}$$

We recall that I measures the reduction on the uncertainty of x due to the knowledge of y. The capacity C gives a measure of how faithful a channel can be: If the output of the channel is pure noise, x and y are uncorrelated and $C = 0$. At the other extreme if $y = f(x)$ is known for sure, given x, then $C = \max_{\{p(x)\}} H(p) = 1$ bit. The interest of the capacity will become clear in section 1.6.3 with Shannon's coding theorem which shows that C characterizes the amount of information which can be transmitted faithfully in a channel.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Example 1.19 Consider a binary symmetric channel with flip probability p. Let us call q the probability that the source sends $x = \mathtt{0}$, and $1 - q$ the probability of $x = \mathtt{1}$. It is easy to show that the mutual information in Eq. (1.37) is maximized when zeros and ones are transmitted with equal probability (i.e. when $q = 1/2$).

Using the expression (1.37), we get, $C = 1 - \mathcal{H}(p)$ bits, where $\mathcal{H}(p)$ is the entropy of Bernouilli's process with parameter p (plotted in Fig. 1.1).

Example 1.20 Consider now the binary erasure channel with error probability p. The same argument as above applies. It is therefore easy to get $C = 1-p$.

Exercise 1.7 Compute the capacity of the Z channel.

1.6.2 *Error correcting codes*

{sec:ECC}

The only ingredient which we still need to specify in order to have a complete definition of the channel coding problem, is the behavior of the information source. We shall assume it to produce a sequence of uncorrelated unbiased bits. This may seem at first a very crude model for any real information source. Surprisingly, Shannon's source-channel separation theorem assures that there is indeed no loss of generality in treating this case.

The sequence of bits produced by the source is divided in blocks $m_1, m_2, m_3, \ldots$ of length M. The **encoding** is a mapping from $\{0,1\}^M \ni m$ to $\{0,1\}^N$, with $N \geq M$. Each possible M-bit message m is mapped to a **codeword** $\underline{x}(m)$ which is a point in the N-dimensional unit hypercube. The codeword length N is also called the **blocklength**. There are 2^M codewords, and the set of all possible codewords is called the **codebook**. When the message is transmitted, the codeword $\underline{x}$ is corrupted to $\underline{y} \in \mathcal{Y}^N$ with probability $Q(\underline{y}|\underline{x}) = \prod_{i=1}^N Q(y_i|x_i)$. The output alphabet $\mathcal{Y}$ depends on the channel. The **decoding** is a mapping from $\mathcal{Y}^N$ to $\{0,1\}^M$ which takes the received message $\underline{y} \in \mathcal{Y}^N$ and maps it to one of the possible original messages $m' = d(\underline{y}) \in \{0,1\}^M$.

An **error correcting code** is defined by the set of two functions, the encoding $\underline{x}(m)$ and the decoding $d(\underline{y})$. The ratio

$$R = \frac{M}{N} \tag{1.38}$$

of the original number of bits to the transmitted number of bits is called the **rate** of the code. The rate is a measure of the redundancy of the code. The smaller the rate, the more redundancy is added to the code, and the more errors one should be able to correct.

The **block error probability** of a code on the input message m, denoted by $\mathrm{P_B}(m)$, is given by the probability that the decoded messages differs from the one which was sent:

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

$$\mathrm{P_B}(m) = \sum_{\underline{y}} Q(\underline{y}|\underline{x}(m)) \;\; \mathbb{I}(d(\underline{y}) \neq m) \; . \tag{1.39}$$

Knowing thee probability for each possible transmitted message is an exceedingly detailed characterization of the code performances. One can therefore introduce a **maximal block error probability** as

$$\mathrm{P_B^{max}} \equiv \max_{m \in \{0,1\}^M} \mathrm{P_B}(m) \, . \tag{1.40}$$

This corresponds to characterizing the code by its 'worst case' performances. A more optimistic point of view consists in averaging over the input messages. Since we assumed all of them to be equiprobable, we introduce the **average block error probability** as

$$\mathrm{P_B^{av}} \equiv \frac{1}{2^M} \sum_{m \in \{0,1\}^M} \mathrm{P_B}(m) \, . \tag{1.41}$$

Since this is a very common figure of merit for error correcting codes, we shall call it block error probability and use the symbol $\mathrm{P_B}$ without further specification hereafter.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Example 1.21 Repetition code. Consider a BSC which transmits a wrong bit with probability p. A simple code consists in repeating k times each bit, with k odd. Formally we have $M = 1$, $N = k$ and

$$\underline{x}(\mathtt{0}) = \underbrace{\mathtt{000\ldots00}}_{k}, \tag{1.42}$$

$$\underline{x}(\mathtt{1}) = \underbrace{\mathtt{111\ldots11}}_{k} \tag{1.43}$$

For instance with $k = 3$, the original stream `0110001` is encoded as `00011111100000 0000111`. A possible decoder consists in parsing the received sequence in groups of k bits, and finding the message m' from a majority rule among the k bits. In our example with $k = 3$, if the received group of three bits is `111` or `110` or any permutation, the corresponding bit is assigned to `1`, otherwise it is assigned to `0`. For instance if the channel output is `000101111011000010111`, the decoding gives `0111001`.

This $k = 3$ repetition code has rate $R = M/N = 1/3$. It is a simple exercise to see that the block error probability is $\mathrm{P_B} = p^3 + 3p^2(1-p)$ independently of the information bit.

Clearly the $k = 3$ repetition code is able to correct mistakes induced from the transmission only when there is at most one mistake per group of three bits. Therefore the block error probability stays finite at any nonzero value of the noise. In order to improve the performances of these codes, k must increase. The error probability for a general k is

$$\mathrm{P_B} = \sum_{r=\lceil k/2 \rceil}^{k} \binom{k}{r} (1-p)^{k-r} p^r \ . \tag{1.44}$$

Notice that for any finite k, $p > 0$ it stays finite. In order to have $\mathrm{P_B} \to 0$ we must consider $k \to \infty$. Since the rate is $R = 1/k$, the price to pay for a vanishing block error probability is a vanishing communication rate!

Happily enough much better codes exist as we will see below.

1.6.3 *The channel coding theorem*

{sec:channeltheorem}

Consider a communication device in which the channel capacity (1.37) is C. In his seminal 1948 paper, Shannon proved the following theorem.

{theorem:Shannon_channel}

Theorem 1.22 *For every rate $R < C$, there exists a sequence of codes $\{\mathcal{C}_N\}$, of blocklength N, rate R_N, and block error probability $\mathrm{P_{B,}}_N$, such that $R_N \to R$ and $\mathrm{P_{B,}}_N \to 0$ as $N \to \infty$. Conversely, if for a sequence of codes $\{\mathcal{C}_N\}$, one has $R_N \to R$ and $\mathrm{P_{B,}}_N \to 0$ as $N \to \infty$, then $R < C$.*

In practice, for long messages (i.e. large N), reliable communication is possible if and only if the communication rate stays below capacity. We shall not give the

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

proof here but differ it to Chapters 6 and **???**. Here we keep to some qualitative comments and provide the intuitive idea underlying this result.

First of all, the result is rather surprising when one meets it for the first time. As we saw on the example of repetition codes above, simple minded codes typically have a finite error probability, for any non-vanishing noise strength. Shannon's theorem establishes that it is possible to achieve zero error probability, while keeping the communication rate finite.

One can get an intuitive understanding of the role of the capacity through a qualitative reasoning, which uses the fact that a random variable with entropy H 'typically' takes 2^H values. For a given codeword $\underline{x}(m) \in \{0,1\}^N$, the channel output $\underline{y}$ is a random variable with an entropy $H_{\underline{y}|\underline{x}} = NH_{y|x}$. There exist of order $2^{NH_{y|x}}$ such outputs. For a perfect decoding, one needs a decoding function $d(\underline{y})$ that maps each of them to the original message m. Globally, the typical number of possible outputs is 2^{NH_y}, therefore one can send at most $2^{N(H_y - H_{y|x})}$ codewords. In order to have zero maximal error probability, one needs to be able to send all the $2^M = 2^{NR}$ codewords. This is possible only if $R < H_y - H_{y|x} < C$.

Notes

There are many textbooks introducing to probability and to information theory. A standard probability textbook is the one of Feller (Feller, 1968). The original Shannon paper (Shannon, 1948) is universally recognized as the foundation of information theory. A very nice modern introduction to the subject is the book by Cover and Thomas (Cover and Thomas, 1991). The reader may find there a description of Huffman codes which did not treat in the present Chapter, as well as more advanced topics in source coding.

We did not show that the six properties listed in Sec. 1.2 provide in fact an alternative (axiomatic) definition of entropy. The interested reader is referred to (Csiszár and Körner, 1981). An advanced information theory book with much space devoted to coding theory is (Gallager, 1968). The recent (and very rich) book by MacKay (MacKay, 2002) discusses the relations with statistical inference and machine learning.

The information-theoretic definition of entropy has been used in many contexts. It can be taken as a founding concept in statistical mechanics. Such an approach is discussed in (Balian, 1992).

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

2

STATISTICAL PHYSICS AND PROBABILITY THEORY

{chap:StatisticalPhysicsIntro

One of the greatest achievement of science has been to realize that matter is made out of a small number of simple elementary components. This result seems to be in striking contrast with our experience. Both at a simply perceptual level and with more refined scientific experience, we come in touch with an ever-growing variety of states of the matter with disparate properties. The ambitious purpose of statistical physics (and, more generally, of a large branch of condensed matter physics) is to understand this variety. It aims at explaining how complex behaviors can emerge when large numbers of identical elementary components are allowed to interact.

We have, for instance, experience of water in three different states (solid, liquid and gaseous). Water molecules and their interactions do not change when passing from one state to the other. Understanding how the same interactions can result in qualitatively different macroscopic states, and what rules the change of state, is a central topic of statistical physics.

The foundations of statistical physics rely on two important steps. The first one consists in passing form the deterministic laws of physics, like Newton's law, to a probabilistic description. The idea is that a precise knowledge of the motion of each molecule in a macroscopic system is inessential to the understanding of the system as a whole: instead, one can postulate that the microscopic dynamics, because of its chaoticity, allows for a purely probabilistic description. The detailed justification of this basic step has been achieved only in a small number of concrete cases. Here we shall bypass any attempt at such a justification: we directly adopt a purely probabilistic point of view, as a basic postulate of statistical physics.

The second step starts from the probabilistic description and recovers determinism at a macroscopic level by some sort of law of large numbers. We all know that water boils at 100° Celsius (at atmospheric pressure) or that its density (at 25° Celsius and atmospheric pressures) is $1\,\mathrm{gr/cm^3}$. The regularity of these phenomena is not related to the deterministic laws which rule the motions of water molecule. It is instead the consequence of the fact that, because of the large number of particles involved in any macroscopic system, the fluctuations are "averaged out". We shall discuss this kind of phenomena in Sec. 2.4 and, more mathematically, in Ch. 4.

The purpose of this Chapter is to introduce the most basic concepts of this discipline, for an audience of non-physicists with a mathematical background. We adopt a somewhat restrictive point of view, which keeps to classical (as opposed to quantum) statistical physics, and basically describes it as a branch

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

of probability theory (Secs. 2.1 to 2.3). In Section 2.4 we focus on large systems, and stress that the statistical physics approach becomes particularly meaningful in this regime. Theoretical statistical physics often deal with highly idealized mathematical models of real materials. The most interesting (and challenging) task is in fact to understand the *qualitative* behavior of such systems. With this aim, one can discard any "irrelevant" microscopic detail from the mathematical description of the model. This modelization procedure is exemplified on the case study of ferromagnetism through the introduction of the Ising model in Sec. 2.5. It is fair to say that the theoretical understanding of Ising ferromagnets is quite advanced. The situation is by far more challenging when Ising spin glasses are considered. Section 2.6 presents a rapid preview of this fascinating subject.

{se:Boltzmann}

2.1 The Boltzmann distribution

The basic ingredients for a probabilistic description of a physical system are:

- A **space of configurations** $\mathcal{X}$. One should think of $x \in \mathcal{X}$ as giving a complete microscopic determination of the state of the system under consideration. We are not interested in defining the most general mathematical structure for $\mathcal{X}$ such that a statistical physics formalism can be constructed. Throughout this book we will in fact consider only two very simple types of configuration spaces: (i) finite sets, and (ii) smooth, compact, finite-dimensional manifolds. If the system contains N 'particles', the configuration space is a product space:

$$\mathcal{X}_N = \underbrace{\mathcal{X} \times \cdots \times \mathcal{X}}_{N} . \tag{2.1}$$

  The configuration of the system has the form $x = (x_1, \ldots, x_N)$. Each coordinate $x_i \in \mathcal{X}$ is meant to represent the state (position, orientation, etc) of one of the particles.
  But for a few examples, we shall focus on configuration spaces of type (i). We will therefore adopt a discrete-space notation for $\mathcal{X}$. The generalization to continuous configuration spaces is in most cases intuitively clear (although it may present some technical difficulties).
- A set of **observables**, which are real-valued functions on the configuration space $\mathcal{O} : x \mapsto \mathcal{O}(x)$. If $\mathcal{X}$ is a manifold, we shall limit ourselves to observables which are smooth functions of the configuration x. Observables are physical quantities which can be measured through an experiment (at least in principle).
- Among all the observables, a special role is played by the **energy function** $E(x)$. When the system is a N particle system, the energy function generally takes the form of sums of terms involving few particles. An energy function of the form:

$$E(x) = \sum_{i=1}^{N} E_i(x_i) \tag{2.2}$$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

corresponds to a **non-interacting** system. An energy of the form

$$E(x) = \sum_{i_1,..,i_k} E_{i_1,..,i_k}(x_{i_1}, ..., x_{i_k}) \tag{2.3}$$

is called a k**-body** interaction. In general, the energy will contain some pieces involving k-body interactions, with $k \in \{1, 2, ..., K\}$. An important feature of real physical systems is that K is never a large number (usually $K = 2$ or 3), even when the number of particles N is very large. The same property holds for all measurable observables. However, for the general mathematical formulation which we will use here, the energy can be any real valued function on $\mathcal{X}$.

Once the configuration space $\mathcal{X}$ and the energy function are fixed, the probability $p_\beta(x)$ for the system to be found in the configuration x is given by the **Boltzmann distribution**:

$$p_\beta(x) = \frac{1}{Z(\beta)} e^{-\beta E(x)} \quad ; \quad Z(\beta) = \sum_{x \in \mathcal{X}} e^{-\beta E(x)} . \tag{2.4}$$

The real parameter $T = 1/\beta$ is the **temperature** (and one refers to β as the inverse temperature)[3]. The normalization constant $Z(\beta)$ is called the **partition function**. Notice that Eq. (2.4) defines indeed the density of the Boltzmann distribution with respect to some reference measure. The reference measure is usually the counting measure if $\mathcal{X}$ is discrete or the Lebesgue measure if $\mathcal{X}$ is continuous. It is customary to denote the expectation value with respect to Boltzmann's measure by brackets: the expectation value $\langle \mathcal{O}(x) \rangle$ of an observable $\mathcal{O}(x)$, also called its **Boltzmann average** is given by:

$$\langle \mathcal{O} \rangle = \sum_{x \in \mathcal{X}} p_\beta(x) \mathcal{O}(x) = \frac{1}{Z(\beta)} \sum_{x \in \mathcal{X}} e^{-\beta E(x)} \mathcal{O}(x) . \tag{2.5}$$

[3]In most books of statistical physics, the temperature is defined as $T = 1/(k_B \beta)$ where k_B is a constant called Boltzmann's constant, whose value is determined by historical reasons. Here we adopt the simple choice $k_B = 1$ which amounts to a special choice of the temperature scale

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Example 2.1 One intrinsic property of elementary particles is their spin. For 'spin 1/2' particles, the spin σ takes only two values: $\sigma = \pm 1$. A localized spin 1/2 particle, in which the only degree of freedom is the spin, is described by $\mathcal{X} = \{+1, -1\}$, and is called an **Ising spin**. The energy of the spin in the state $\sigma \in \mathcal{X}$ in a magnetic field B is

$$E(\sigma) = -B\,\sigma \tag{2.6}$$ {eq:Ising_energy_1spin}

Boltzmann's probability of finding the spin in the state σ is

$$p_\beta(\sigma) = \frac{1}{Z(\beta)}\, e^{-\beta E(\sigma)} \qquad Z(\beta) = e^{-\beta B} + e^{\beta B} = 2\cosh(\beta B)\,. \tag{2.7}$$ {eq:boltz_spin}

The average value of the spin, called **the magnetization** is

$$\langle \sigma \rangle = \sum_{\sigma \in \{1,-1\}} p_\beta(\sigma)\,\sigma = \tanh(\beta B)\,. \tag{2.8}$$ {eq:mag_tanh_beta_B}

At high temperatures, $T \gg |B|$, the magnetization is small. At low temperatures, the magnetization its close to its maximal value, $\langle \sigma \rangle = 1$ if $B > 0$. Section 2.5 will discuss the behaviors of many Ising spins, with some more complicated energy functions.

Example 2.2 Some spin variables can have a larger space of possible values. For instance a **Potts spin** with q states takes values in $\mathcal{X} = \{1, 2, ..., q\}$. In presence of a magnetic field of intensity h pointing in direction $r \in \{1, .., q\}$, the energy of the Potts spin is

$$E(\sigma) = -B\,\delta_{\sigma,r}\ . \tag{2.9}$$

In this case, the average value of the spin in the direction of the field is

$$\langle \delta_{\sigma,r} \rangle = \frac{\exp(\beta B)}{\exp(\beta B) + (q-1)}\ . \tag{2.10}$$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Example 2.3 Let us consider a single water molecule inside a closed container, for instance, inside a bottle. A water molecule H_2O is already a complicated object. In a first approximation, we can neglect its structure and model the molecule as a point inside the bottle. The space of configurations reduces then to:

$$\mathcal{X} = \texttt{BOTTLE} \subset \mathbb{R}^3 \,, \tag{2.11}$$

where we denoted by BOTTLE the region of $\mathbb{R}^3$ delimited by the container. Notice that this description is not very accurate at a microscopic level.

The description of the precise form of the bottle can be quite complex. On the other hand, it is a good approximation to assume that all positions of the molecule are equiprobable: the energy is independent of the particle's position $x \in \texttt{BOTTLE}$. One has then:

$$p(x) = \frac{1}{Z} \,, \qquad Z = |\mathcal{X}| \,, \tag{2.12}$$

and the Boltzmann average of the particle's position, $\langle x \rangle$, is the barycentre of the bottle.

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

Example 2.4 In assuming that all the configurations of the previous example are equiprobable, we neglected the effect of gravity on the water molecule. In the presence of gravity our water molecule at position x has an energy:

$$E(x) = w\,\mathrm{he}(x)\,, \tag{2.13}$$

where $\mathrm{he}(x)$ is the height corresponding to the position x and w is a positive constant, determined by terrestrial attraction, which is proportional to the mass of the molecule. Given two positions x and y in the bottle, the ratio of the probabilities to find the particle at these positions is

$$\frac{p_\beta(x)}{p_\beta(y)} = \exp\{-\beta w[\mathrm{he}(x) - \mathrm{he}(y)]\} \tag{2.14}$$

For a water molecule at a room temperature of 20 degrees Celsius ($T = 293$ degrees Kelvin), one has $\beta w \approx 7 \times 10^{-5}\,\mathrm{m}^{-1}$. Given a point x at the bottom of the bottle and y at a height of 20 cm, the probability to find a water molecule 'near' x is approximatively 1.000014 times larger than the probability to find it 'near' y. For a tobacco-mosaic virus, which is about 2×10^6 times heavier than a water molecule, the ratio is $p_\beta(x)/p_\beta(y) \approx 1.4 \times 10^{12}$ which is very large. For a grain of sand the ratio is so large that one never observes it floating around y. Notice that, while these ratios of probability densities are easy to compute, the partition function and therefore the absolute values of the probability densities can be much more complicated to estimate, depending on the shape of the bottle.

Example 2.5 In many important cases, we are given the space of configurations $\mathcal{X}$ and a stochastic dynamics defined on it. The most interesting probability distribution for such a system is the stationary state $p_{\mathrm{st}}(x)$ (we assume that it is unique). For sake of simplicity, we can consider a finite space $\mathcal{X}$ and a discrete time Markov chain with transition probabilities $\{w(x \to y)\}$ (in Chapter 4 we shall recall some basic definitions concerning Markov chains). It happens sometimes that the transition rates satisfy, for any couple of configurations $x, y \in \mathcal{X}$, the relation

$$f(x)w(x \to y) = f(y)w(y \to x)\,, \tag{2.15}$$

for some positive function $f(x)$. As we shall see in Chapter 4, when this condition, called the **detailed balance**, is satisfied (together with a couple of other technical conditions), the stationary state has the Boltzmann form (2.4) with $e^{-\beta E(x)} = f(x)$.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Exercise 2.1 As a particular realization of the above example, consider an 8×8 chessboard and a special piece sitting on it. At any time step the piece will stay still (with probability $1/2$) or move randomly to one of the neighboring positions (with probability $1/2$). Does this process satisfy the condition (2.15)? Which positions on the chessboard have lower (higher) "energy"? Compute the partition function.

From a purely probabilistic point of view, one can wonder why one bothers to decompose the distribution $p_\beta(x)$ into the two factors $e^{-\beta E(x)}$ and $1/Z(\beta)$. Of course the motivations for writing the Boltzmann factor $e^{-\beta E(x)}$ in exponential form come essentially from physics, where one knows (either exactly or within some level of approximation) the form of the energy. This also justifies the use of the inverse temperature β (after all, one could always redefine the energy function in such a way to set $\beta = 1$).

However, it is important to stress that, even if we adopt a mathematical viewpoint, and if we are interested in a particular distribution $p(x)$ which corresponds to a particular value of the temperature, it is often illuminating to embed it into a one-parameter family as is done in the Boltzmann expression (2.4). Indeed, (2.4) interpolates smoothly between several interesting situations. As $\beta \to 0$ (**high-temperature limit**), one recovers the flat probability distribution

$$\lim_{\beta \to 0} p_\beta(x) = \frac{1}{|\mathcal{X}|} \,. \tag{2.16}$$

Both the probabilities $p_\beta(x)$ and the observables expectation values $\langle \mathcal{O}(x) \rangle$ can be expressed as convergent Taylor expansions around $\beta = 0$. For small β the Boltzmann distribution can be thought as a "softening" of the original one.

In the limit $\beta \to \infty$ (**low-temperature limit**), the Boltzmann distribution concentrates over the global maxima of the original one. More precisely, one says $x_0 \in \mathcal{X}$ to be a **ground state** if $E(x) \geq E(x_0)$ for any $x \in \mathcal{X}$. The minimum value of the energy $E_0 = E(x_0)$ is called the **ground state energy**. We will denote the set of ground states as $\mathcal{X}_0$. It is elementary to show that

$$\lim_{\beta \to \infty} p_\beta(x) = \frac{1}{|\mathcal{X}_0|} \, \mathbb{I}(x \in \mathcal{X}_0) \,, \tag{2.17}$$

where $\mathbb{I}(x \in \mathcal{X}_0) = 1$ if $x \in \mathcal{X}_0$ and $\mathbb{I}(x \in \mathcal{X}_0) = 0$ otherwise. The above behavior is summarized in physicists jargon by saying that, at low temperature, "low energy configurations dominate" the behavior of the system.

2.2 Thermodynamic potentials

{se:Potentials}

Several properties of the Boltzmann distribution (2.4) are conveniently summarized through the thermodynamic potentials. These are functions of the temperature $1/\beta$ and of the various parameters defining the energy $E(x)$. The most important thermodynamic potential is the **free energy**:

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

$$F(\beta) = -\frac{1}{\beta} \log Z(\beta) \,, \tag{2.18}$$

where $Z(\beta)$ is the partition function already defined in Eq. (2.4). The factor $-1/\beta$ in Eq. (2.18) is due essentially to historical reasons. In calculations it is sometimes more convenient to use the **free entropy**[4] $\Phi(\beta) = -\beta F(\beta) = \log Z(\beta)$.

Two more thermodynamic potentials are derived from the free energy: the **internal energy** $U(\beta)$ and the **canonical entropy** $S(\beta)$:

$$U(\beta) = \frac{\partial}{\partial\beta}(\beta F(\beta)) \,, \qquad S(\beta) = \beta^2 \frac{\partial F(\beta)}{\partial\beta} \,. \tag{2.19}$$

By direct computation one obtains the following identities concerning the potentials defined so far:

$$F(\beta) = U(\beta) - \frac{1}{\beta} S(\beta) = -\frac{1}{\beta}\Phi(\beta) \,, \tag{2.20}$$

$$U(\beta) = \langle E(x) \rangle \,, \tag{2.21}$$

$$S(\beta) = -\sum_x p_\beta(x) \log p_\beta(x) \,, \tag{2.22}$$

$$-\frac{\partial^2}{\partial\beta^2}(\beta F(\beta)) = \langle E(x)^2 \rangle - \langle E(x) \rangle^2 \,. \tag{2.23}$$

Equation (2.22) can be rephrased by saying that the canonical entropy is the Shannon entropy of the Boltzmann distribution, as we defined it in Ch. 1. It implies that $S(\beta) \geq 0$. Equation (2.23) implies that the free entropy is a convex function of the temperature. Finally, Eq. (2.21) justifies the name "internal energy" for $U(\beta)$.

In order to have some intuition of the content of these definitions, let us reconsider the high- and low-temperature limits already treated in the previous Section. In the high-temperature limit, $\beta \to 0$, one finds

$$F(\beta) = -\frac{1}{\beta} \log |\mathcal{X}| + \langle E(x) \rangle_0 + \Theta(\beta) \,, \tag{2.24}$$

$$U(\beta) = \langle E(x) \rangle_0 + \Theta(\beta) \,, \tag{2.25}$$

$$S(\beta) = \log |\mathcal{X}| + \Theta(\beta) \,. \tag{2.26}$$

{ch:Notation}

(The symbol Θ means 'of the order of'; the precise definition is given in Appendix). The interpretation of these formulae is straightforward. At high temperature the system can be found in any possible configuration with similar probabilities (the probabilities being exactly equal when $\beta = 0$). The entropy counts the number of possible configurations. The internal energy is just the average value of the energy over the configurations with flat probability distribution.

[4] Unlike the other potentials, there is no universally accepted name for $\Phi(\beta)$; because this potential is very useful, we adopt for it the name 'free entropy'

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

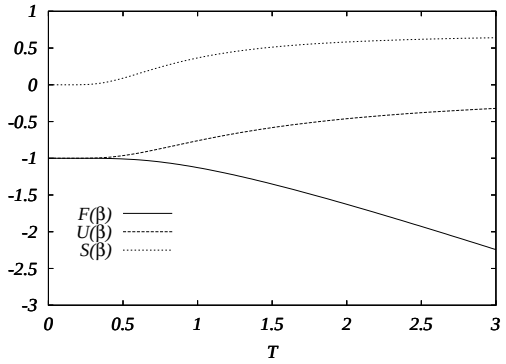

FIG. 2.1. Thermodynamic potentials for a two-level system with $\epsilon_1 = -1$, $\epsilon_2 = +1$ as a function of the temperature $T = 1/\beta$. {fig:twolevel}

While the high temperature expansions (2.24)–(2.26) have the same form both for a discrete and a continuous configuration space $\mathcal{X}$, in the low temperature case, we must be more careful. If $\mathcal{X}$ is finite we can meaningfully define the **energy gap** $\Delta E > 0$ as follows (recall that we denoted by E_0 the ground-state energy)

$$\Delta E = \min\{E(y) - E_0 \,:\, y \in \mathcal{X}\backslash\mathcal{X}_0\}\,. \tag{2.27}$$

With this definition we get

$$F(\beta) = E_0 - \frac{1}{\beta}\,\log|\mathcal{X}_0| + \Theta(e^{-\beta\Delta E})\,, \tag{2.28}$$

$$E(\beta) = E_0 + \Theta(e^{-\beta\Delta E})\,, \tag{2.29}$$

$$S(\beta) = \log|\mathcal{X}_0| + \Theta(e^{-\beta\Delta E})\,. \tag{2.30}$$

The interpretation is that, at low temperature, the system is found with equal probability in any of the ground states, and nowhere else. Once again the entropy counts the number of available configurations and the internal energy is the average of their energies (which coincide with the ground state).

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Exercise 2.2 A two level system. This is the simplest non-trivial example: $\mathcal{X} = \{1, 2\}$, $E(1) = \epsilon_1$, $E(2) = \epsilon_2$. Without loss of generality we assume $\epsilon_1 < \epsilon_2$. It can be used as a mathematical model for many physical systems, like the spin 1/2 particle discussed above.

Derive the following results for the thermodynamic potentials ($\Delta = \epsilon_2 - \epsilon_1$ is the energy gap):

$$F(\beta) = \epsilon_1 - \frac{1}{\beta} \log(1 + e^{-\beta\Delta}) \,, \tag{2.31}$$

$$U(\beta) = \epsilon_1 + \frac{e^{-\beta\Delta}}{1 + e^{-\beta\Delta}} \Delta \,, \tag{2.32}$$

$$S(\beta) = \frac{e^{-\beta\Delta}}{1 + e^{-\beta\Delta}} \beta\Delta + \log(1 + e^{-\beta\Delta}) \,. \tag{2.33}$$

The behavior of these functions is presented in Fig. 2.1. The reader can work out the asymptotics, and check the general high and low temperature behaviors given above.

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

Exercise 2.3 We come back to the example of the previous section: one water molecule, modeled as a point, in a bottle. Moreover, we consider the case of a cylindric bottle of base $B \subset \mathbb{R}^2$ (surface $|B|$) and height d.

Using the energy function (2.13), derive the following explicit expressions for the thermodynamic potentials:

$$F(\beta) = -\frac{1}{\beta} \log |B| - \frac{1}{\beta} \log \frac{1 - e^{-\beta w d}}{\beta w} \,, \tag{2.34}$$

$$U(\beta) = \frac{1}{\beta} - \frac{wd}{e^{\beta w d} - 1} \,, \tag{2.35}$$

$$S(\beta) = \log |Bd| + 1 - \frac{\beta w d}{e^{\beta w d} - 1} - \log \left(\frac{\beta w d}{1 - e^{-\beta w d}} \right) . \tag{2.36}$$

Notice that the internal energy formula can be used to compute the average height of the molecule $\langle \mathrm{he}(x) \rangle = U(\beta)/w$. This is a consequence of the definition of the energy, cf. Eq. (2.13) and of Eq. (2.21). Plugging in the correct w constant, one may find that the average height descends below 49.99% of the bottle height $d = 20\,\mathrm{cm}$ only when the temperature is below $3.2^{\mathrm{o}}\,K$.

Using the expressions (2.34)–(2.36) one obtains the low-temperature expansions for the same quantities:

$$F(\beta) = -\frac{1}{\beta} \log \left(\frac{|B|}{\beta w} \right) + \Theta(e^{-\beta w d}) \,, \tag{2.37}$$

$$U(\beta) = \frac{1}{\beta} + \Theta(e^{-\beta w d}) \,, \tag{2.38}$$

$$S(\beta) = \log \left(\frac{|B| e}{\beta w} \right) + \Theta(e^{-\beta w d}) \,. \tag{2.39}$$

In this case $\mathcal{X}$ is continuous, and the energy has no gap. But these results can be understood as follows: at low temperature the molecule is confined to a layer of height of order $1/(\beta w)$ above the bottom of the bottle. It occupies therefore a volume of size $|B|/(\beta w)$. Its entropy is approximatively given by the logarithm of such a volume.

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

Exercise 2.4 Let us reconsider the above example and assume the bottle to have a different shape, for instance a sphere of radius R. In this case it is difficult to compute explicit expressions for the thermodynamic potentials but one can easily compute the low-temperature expansions. For the entropy one gets at large β:

$$S(\beta) = \log\left(\frac{2\pi e^2 R}{\beta^2 w^2}\right) + \Theta(1/\beta)\,. \tag{2.40}$$

The reader should try understand the difference between this result and Eq. (2.39) and provide an intuitive explanation as in the previous example. Physicists say that the low-temperature thermodynamic potentials reveal the "low-energy structure" of the system.

{se:free_energy}

2.3 The fluctuation dissipation relations

It often happens that the energy function depends smoothly upon some real parameters. They can be related to the experimental conditions under which a physical system is studied, or to some fundamental physical quantity. For instance, the energy of a water molecule in the gravitational field, cf. Eq. (2.13), depends upon the weight w of the molecule itself. Although this is a constant number in the physical world, it is useful, in the theoretical treatment, to consider it as an adjustable parameter.

It is therefore interesting to consider an energy function $E_\lambda(x)$ which depends smoothly upon some parameter λ and admit the following Taylor expansion in the neighborhood of $\lambda = \lambda_0$:

$$E_\lambda(x) = E_{\lambda_0}(x) + (\lambda - \lambda_0)\left.\frac{\partial E}{\partial \lambda}\right|_{\lambda_0}(x) + O((\lambda-\lambda_0)^2)\,. \tag{2.41}$$

The dependence of the free energy and of other thermodynamic potentials upon λ in the neighborhood of λ_0 is easily related to the explicit dependence of the energy function itself. Let us consider the partition function, and expand it to first order in $\lambda - \lambda_0$:

$$\begin{aligned} Z(\lambda) &= \sum_x \exp\left(-\beta\left[E_{\lambda_0}(x) + (\lambda-\lambda_0)\left.\frac{\partial E}{\partial\lambda}\right|_{\lambda_0}(x) + O((\lambda-\lambda_0)^2)\right]\right) \\ &= Z(\lambda_0)\left[1 - \beta(\lambda-\lambda_0)\langle\left.\frac{\partial E}{\partial\lambda}\right|_{\lambda_0}\rangle_0 + O((\lambda-\lambda_0)^2)\right] \end{aligned} \tag{2.42}$$

where we denoted by $\langle\cdot\rangle_0$ the expectation with respect to the Boltzmann distribution at $\lambda = \lambda_0$.

This shows that the free entropy behaves as:

$$\left.\frac{\partial\Phi}{\partial\lambda}\right|_{\lambda_0} = -\beta\,\langle\left.\frac{\partial E}{\partial\lambda}\right|_{\lambda_0}\rangle_0\,, \tag{2.43}$$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

One can also consider the λ dependence of the expectation value of a generic observable $A(x)$. Using again the Taylor expansion one finds that

$$\left.\frac{\partial \langle A\rangle_\lambda}{\partial\lambda}\right|_{\lambda_0} = -\beta \,\langle\, A \,;\, \left.\frac{\partial E}{\partial\lambda}\right|_{\lambda_0} \rangle_0 \,. \tag{2.44}$$

where we denoted by $\langle A;B\rangle$ the **connected correlation function**: $\langle A;B\rangle = \langle AB\rangle - \langle A\rangle\langle B\rangle$. A particular example of this relation was given in Eq. (2.23).

The result (2.44) has important practical consequences and many generalizations. Imagine you have an experimental apparatus that allows you to tune some parameter λ (for instance the pressure of a gas, or the magnetic or electric field acting on some material) and to monitor the value of the observable $A(x)$ (the volume of the gas, the polarization or magnetization of the material). The quantity on the left-hand side of Eq. (2.44) is the response of the system to an infinitesimal variation of the tunable parameter. On the right-hand side, we find some correlation function within the "unperturbed" system. One possible application is to measure correlations within a system by monitoring its response to an external perturbation. Such a relation between a correlation and a response is called a **fluctuation dissipation relation**.

2.4 The thermodynamic limit

{se:Thermodynamic}

The main purpose of statistical physics is to understand the macroscopic behavior of a large number, $N \gg 1$, of simple components (atoms, molecules, etc) when they are brought together.

To be concrete, let us consider a few drops of water in a bottle. A configuration of the system is given by the positions and orientations of all the H_2O molecules inside the bottle. In this case $\mathcal{X}$ is the set of positions and orientations of a single molecule, and N is typically of order 10^{23} (more precisely 18 gr of water contain approximatively $6\cdot 10^{23}$ molecules). The sheer magnitude of such a number leads physicists to focus on the $N\to\infty$ limit, also called the **thermodynamic limit**.

As shown by the examples below, for large N the thermodynamic potentials are often proportional to N. One is thus lead to introduce the **intensive thermodynamic potentials** as follows. Let us denote by $F_N(\beta)$, $U_N(\beta)$, $S_N(\beta)$ the free energy, internal energy and canonical entropy for a system with N 'particles'. The **free energy density** is defined by

$$f(\beta) = \lim_{N\to\infty} F_N(\beta)/N \,, \tag{2.45}$$

if the limit exists [5]. One defines analogously the **energy density** $u(\beta)$ and the **entropy density** $s(\beta)$.

The free energy $F_N(\beta)$, is, quite generally, an analytic function of β in a neighborhood of the real β axis. This is a consequence of the fact that $Z(\beta)$

[5] The limit usually exist, at least if the forces between particles decrease fast enough at large inter-particle distances

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

is analytic throughout the entire β plane, and strictly positive for real β's. A question of great interest is whether analyticity is preserved in the thermodynamic limit (2.45), under the assumption that the limit exists. Whenever the free energy density $f(\beta)$ is non-analytic, one says that a **phase transition** occurs. Since the free entropy density $\phi(\beta) = -\beta f(\beta)$ is convex, the free energy density is necessarily continuous whenever it exists.

In the simplest cases the non-analyticities occur at isolated points. Let β_c be such a point. Two particular type of singularities occur frequently:

- The free energy density is continuous, but its derivative with respect to β is discontinuous at β_c. This singularity is named a **first order phase transition**.
- The free energy and its first derivative are continuous, but the second derivative is discontinuous at β_c. This is called a **second order phase transition**.

Higher order phase transitions can be defined as well on the same line.

Apart from being interesting mathematical phenomena, phase transitions correspond to *qualitative* changes in the underlying physical system. For instance the transition from water to vapor at 100°C at normal atmospheric pressure is modeled mathematically as a first order phase transition in the above sense. A great part of this book will be devoted to the study of phase transitions in many different systems, where the interacting 'particles' can be very diverse objects like information bits or occupation numbers on the vertices of a graph.

When N grows, the volume of the configuration space increases exponentially: $|\mathcal{X}_N| = |\mathcal{X}|^N$. Of course, not all the configurations are equally important under the Boltzmann distribution: lowest energy configurations have greater probability. What is important is therefore the number of configurations at a given energy. This information is encoded in the **energy spectrum** of the system:

$$\mathcal{N}_\Delta(E) = |\Omega_\Delta(E)|\,; \qquad \Omega_\Delta(E) \equiv \{x \in \mathcal{X}_N \,:\, E \leq E(x) < E + \Delta\}\,. \quad (2.46)$$

In many systems of interest, the energy spectrum diverges exponentially as $N \to \infty$, if the energy is scaled linearly with N. More precisely, there exist a function $s(e)$ such that, given two fixed numbers e and $\delta > 0$,

$$\lim_{N\to\infty} \frac{1}{N} \log \mathcal{N}_{N\delta}(Ne) = \sup_{e' \in [e, e+\delta]} s(e')\,. \quad (2.47)$$

The function $s(e)$ is called **microcanonical entropy density**. The statement (2.47) is often rewritten in the more compact form:

$$\mathcal{N}_\Delta(E) \doteq_N \exp\left[Ns\left(\frac{E}{N}\right)\right]\,. \quad (2.48)$$

The notation $A_N \doteq_N B_N$ is used throughout the book to denote that two quantities A_N and B_N, which normally behave exponentially in N, are equal **to leading exponential order** when N is large, meaning: $\lim_{N\to\infty}(1/N)\log(A_N/B_N) =$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

0. We often use $\doteq$ without index when there is no ambiguity on what the large variable N is.

The microcanonical entropy density $s(e)$ conveys a great amount of information about the system. Furthermore it is directly related to the intensive thermodynamic potentials through a fundamental relation:

{prop:micro_cano}

Proposition 2.6 *If the microcanonical entropy density (2.47) exists for any e and if the limit in (2.47) uniform in e, then the free entropy density (2.45) exists and is given by:*

$$\phi(\beta) = \max_e [s(e) - \beta e] \,. \tag{2.49}$$

If the maximum of the $s(e)-\beta e$ is unique, then the internal energy density equals $\arg\max[s(e) - \beta e]$.

Proof: For a rigorous proof of this statement, we refer the reader to (Galavotti, 1999; Ruelle, 1999). The basic idea is to write the partition function as follows

$$Z_N(\beta) \doteq \sum_{k=-\infty}^{\infty} \mathcal{N}_\Delta(k\Delta)\, e^{-\beta\Delta} \doteq \int de\, \exp\{N s(e) - N\beta e\} \,, \tag{2.50}$$

and to evaluate the last integral by saddle point. □.

Example 2.7 Let us consider N identical two-level systems: $\mathcal{X}_N = \mathcal{X}\times\cdots\times\mathcal{X}$, with $\mathcal{X} = \{1, 2\}$. We take the energy to be the sum of single-systems energies: $E(x) = E_{\rm single}(x_1) + \cdots + E_{\rm single}(x_N)$, with $x_i \in \mathcal{X}$. As in the previous Section we set $E_{\rm single}(1) = \epsilon_1$, and $E_{\rm single}(2) = \epsilon_2 > \epsilon_1$ and $\Delta = \epsilon_2 - \epsilon_1$.

The energy spectrum of this model is quite simple. For any energy $E = N\epsilon_1 + n\Delta$, there are $\binom{N}{n}$ configurations x with $E(x) = E$. Therefore, using the definition (2.47), we get

$$s(e) = \mathcal{H}\left(\frac{e-\epsilon_1}{\Delta}\right) \,. \tag{2.51}$$

Equation (2.49) can now be used to get

$$f(\beta) = \epsilon_1 - \frac{1}{\beta}\log(1 + e^{-\beta\Delta}) \,, \tag{2.52}$$

which agrees with the result obtained directly from the definition (2.18).

The great attention paid by physicists to the thermodynamic limit is extremely well justified by the huge number of degrees of freedom involved in a macroscopic piece of matter. Let us stress that the interest of the thermodynamic limit is more general than these huge numbers might suggest. First of all, it often happens that fairly small systems are well approximated by the thermodynamic limit. This is extremely important for numerical simulations of physical systems:

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

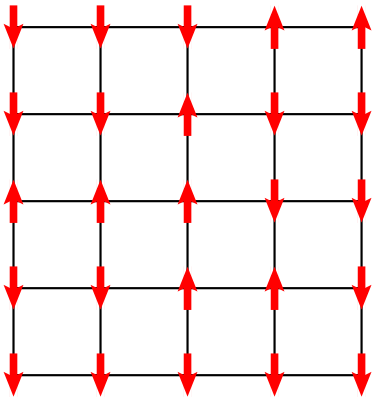

FIG. 2.2. A configuration of a two dimensional Ising model with $L = 5$. There is an Ising spin σ_i on each vertex i, shown by an arrow pointing up if $\sigma_i = +1$, pointing down if $\sigma_i = -1$. The energy (2.53) is given by the sum of two types of contributions: (i) A term $-\sigma_i\sigma_j$ for each edge (ij) of the graph, such that the energy is minimized when the two neighboring spins σ_i and σ_j point in the same direction; (ii) A term $-B\sigma_i$ for each site i, due to the coupling to an external magnetic field. The configuration depicted here has energy $-8 + 9B$

{fig:ising_def}

one cannot of course simulate 10^{23} molecules on a computer! Even the cases in which the thermodynamic limit *is not* a good approximation are often fruitfully analyzed as *violations* of this limit. Finally, the insight gained in analyzing the $N \to \infty$ limit is always crucial in understanding moderate-size systems.

{se:ising}

2.5 Ferromagnets and Ising models

Magnetic materials contain molecules with a magnetic moment, a three-dimensional vector which tends to align with the magnetic field felt by the molecule. Moreover, the magnetic moments of two distinct molecules interact with each other. Quantum mechanics plays an important role in magnetism. Because of its effects, the space of possible configurations of a magnetic moment becomes discrete. It is also at the origin of the so-called exchange interaction between magnetic moments. In many materials, the effect of the exchange interactions are such that the energy is lower when two moments align. While the behavior of a single magnetic moment in an external field is qualitatively simple, when we consider a bunch of interacting moments, the problem is much richer, and exhibits remarkable collective phenomena.

A simple mathematical model for such materials is the Ising model. It describes the magnetic moments by Ising spins localized at the vertices of a certain region of the d-dimensional cubic lattice. To keep things simple, let us consider a region $\mathbb{L}$ which is a cube of side L: $\mathbb{L} = \{1, \dots, L\}^d$. On each site $i \in \mathbb{L}$ there is an Ising spin $\sigma_i \in \{+1, -1\}$.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

A configuration $\underline{\sigma} = (\sigma_1 \dots \sigma_N)$ of the system is given by assigning the values of all the spins in the system. Therefore the space of configurations $\mathcal{X}_N = \{+1, -1\}^{\mathbb{L}}$ has the form (2.1) with $\mathcal{X} = \{+1, -1\}$ and $N = L^d$.

The definition of ferromagnetic Ising models is completed by the definition of the energy function. A configuration $\underline{\sigma}$ has an energy:

$$E(\underline{\sigma}) = -\sum_{(ij)} \sigma_i \sigma_j - B \sum_{i \in \mathbb{L}} \sigma_i \,, \tag{2.53}$$

where the sum over (ij) runs over all the (unordered) couples of sites $i, j \in \mathbb{L}$ which are nearest neighbors. The real number B measures the applied external magnetic field.

Determining the free energy density $f(\beta)$ in the thermodynamic limit for this model is a non-trivial task. The model was invented by Wilhem Lenz in the early twenties, who assigned the task of analyzing it to his student Ernst Ising. In his dissertation thesis (1924) Ising solved the $d = 1$ case and showed the absence of phase transitions. In 1948, Lars Onsager brilliantly solved the $d = 2$ case, exhibiting the first soluble "finite-dimensional" model with a second order phase transition. In higher dimensions the problem is unsolved although many important features of the solution are well understood.

Before embarking in any calculation, let us discuss what we expect to be the qualitative properties of this model. Two limiting cases are easily understood. At infinite temperature, $\beta = 0$, the energy (2.53) no longer matters and the Boltzmann distribution weights all the configurations with the same factor 2^{-N}. We have therefore an assembly of completely independent spins. At zero temperature, $\beta \to \infty$, the Boltzmann distribution concentrates onto the ground state(s). If there is no magnetic field, $h = 0$, there are two degenerate ground states: the configurations $\underline{\sigma}^{(+)}$ with all the spins pointing up, $\sigma_i = +1$, and the configuration $\underline{\sigma}^{(-)}$ with all the spins pointing down, $\sigma_i = -1$. If the magnetic field is set to some non-zero value, one of the two configuration dominates: $\underline{\sigma}^{(+)}$ for $h > 0$ and $\underline{\sigma}^{(-)}$ for $h < 0$.

Notice that the reaction of the system to the external magnetic field h is quite different in the two cases. To see this fact, define a "rescaled" magnetic field $x = \beta h$ and take the limits $\beta \to 0$ or $\beta \to \infty$ keeping x fixed. The expected value of any spin in $\mathbb{L}$, in the two limits, is:

$$\langle \sigma_i \rangle = \begin{cases} \tanh(x) & \text{for } \beta \to 0 \\ \tanh(Nx) & \text{for } \beta \to \infty \end{cases} . \tag{2.54}$$

Each spin reacts independently for $\beta \to 0$. On the contrary, they react as a whole as $\beta \to \infty$: one says that the response is cooperative.

A useful quantity for describing the response of the system to the external field is the **average magnetization**:

$$M_N(\beta, B) = \frac{1}{N} \sum_{i \in \mathbb{L}} \langle \sigma_i \rangle \,. \tag{2.55}$$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Because of the symmetry between the up and down directions, $M_N(\beta, B)$ is an odd function of B. In particular $M_N(\beta, 0) = 0$. A cooperative response can be evidenced by considering the **spontaneous magnetization**

$$M_+(\beta) = \lim_{B\to 0+} \lim_{N\to\infty} M_N(\beta, B)\,. \tag{2.56}$$

It is important to understand that a non-zero spontaneous magnetization can appear only in an infinite system: the order of the limits in Eq. (2.56) is crucial. Our analysis so far has shown that the spontaneous magnetization exists at $\beta = \infty$: $M_+(\infty) = 1$. On the other hand $M_+(0) = 0$. It can be shown that actually the spontaneous magnetization $M(\beta)$ is always zero in a high temperature phase $\beta < \beta_c(d)$ (such a phase is called **paramagnetic**). In one dimension ($d = 1$), we will show below that $\beta_c(1) = \infty$. The spontaneous magnetization is always zero, except at zero temperature ($\beta = \infty$): one speaks of a zero temperature phase transition. In dimensions $d \geq 2$, $\beta_c(d)$ is finite, and $M(\beta)$ becomes non zero in the so called **ferromagnetic phase** $\beta > \beta_c$: a phase transition takes place at $\beta = \beta_c$. The temperature $T_c = 1/\beta_c$ is called the **critical temperature**. In the following we shall discuss the $d = 1$ case, and a variant of the model, called the Curie Weiss model, where each spin interacts with all the other spins: this is a solvable model which exhibits a finite temperature phase transition.

ec:OneDimensionalIsing}

2.5.1 *The one-dimensional case*

The $d = 1$ case has the advantage of being simple to solve. We want to compute the partition function (2.4) for a system of N spins with energy $E(\underline{\sigma}) = -\sum_{i=1}^{N-1} \sigma_i\sigma_{i+1} - B\sum_{i=1}^{N} \sigma_i$. We will use a method called the transfer matrix method, which belongs to the general 'dynamic programming' strategy familiar to computer scientists.

We introduce the partial partition function where the configurations of all spins $\sigma_1, \ldots, \sigma_p$ have been summed over, at fixed σ_{p+1}:

$$z_p(\beta, B, \sigma_{p+1}) \equiv \sum_{\sigma_1,\ldots,\sigma_p} \exp\left[\beta\sum_{i=1}^{p} \sigma_i\sigma_{i+1} + \beta B\sum_{i=1}^{p} \sigma_i\right]\,. \tag{2.57}$$

The partition function (2.4) is given by $Z_N(\beta, B) = \sum_{\sigma_N} z_{N-1}(\beta, B, \sigma_N)\exp(\beta B\sigma_N)$. Obviously z_p satisfies the recursion relation

$$z_p(\beta, B, \sigma_{p+1}) = \sum_{\sigma_p=\pm 1} T(\sigma_{p+1}, \sigma_p) z_{p-1}(\beta, B, \sigma_p) \tag{2.58}$$

where we define the so-called **transfer matrix** $T(\sigma, \sigma') = \exp[\beta\sigma\sigma' + \beta B\sigma']$, which is a 2×2 matrix:

$$T = \begin{pmatrix} e^{\beta+\beta B} & e^{-\beta-\beta B} \\ e^{-\beta+\beta B} & e^{\beta-\beta B} \end{pmatrix} \tag{2.59}$$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Introducing the two component vectors $\psi_{\rm L} = \begin{pmatrix} \exp(\beta B) \\ \exp(-\beta B) \end{pmatrix}$ and $\psi_{\rm R} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, and the standard scalar product between such vectors $(a, b) = a_1 b_1 + a_2 b_2$, the partition function can be written in matrix form:

$$Z_N(\beta, B) = (\psi_{\rm L}, T^{N-1} \psi_{\rm R}) \, . \tag{2.60}$$

Let us call λ_1, λ_2 the eigenvalues of T, and ψ_1, ψ_2 the corresponding eigenvectors. Since ψ_1, ψ_2 can be chosen to be linearly independent, $\psi_{\rm R}$ can be decomposed as $\psi_{\rm R} = u_1 \psi_1 + u_2 \psi_2$. The partition function is then expressed as:

$$Z_N(\beta, B) = u_1 \, (\psi_{\rm L}, \psi_1) \, \lambda_1^{N-1} + u_2 \, (\psi_{\rm L}, \psi_2) \, \lambda_2^{N-1} \, . \tag{2.61}$$

The diagonalization of the matrix T gives:

$$\lambda_{1,2} = e^{\beta} \cosh(\beta B) \pm \sqrt{e^{2\beta} \sinh^2 \beta B + e^{-2\beta}} \, . \tag{2.62}$$

For β finite, in the large N limit, the partition function is dominated by the largest eigenvalue λ_1, and the free entropy density is given by $\phi = \log \lambda_1$.

$$\phi(\beta, B) = \log \left[e^{\beta} \cosh(\beta B) + \sqrt{e^{2\beta} \sinh^2 \beta B + e^{-2\beta}} \right] \, . \tag{2.63}$$

Using the same transfer matrix technique we can compute expectation values of observables. For instance the expected value of a given spin is

$$\langle \sigma_i \rangle = \frac{1}{Z_N(\beta, B)} \, (\psi_{\rm L}, T^{i-1} \hat{\sigma} T^{N-i} \psi_{\rm R}) \, , \tag{2.64}$$

where $\hat{\sigma}$ is the following matrix:

$$\hat{\sigma} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \, . \tag{2.65}$$

Averaging over the position i, one can compute the average magnetization $M_N(\beta, B)$. In the thermodynamic limit we get

$$\lim_{N \to \infty} M_N(\beta, B) = \frac{\sinh \beta B}{\sqrt{\sinh^2 \beta h + e^{-4\beta}}} = \frac{1}{\beta} \frac{\partial \phi}{\partial B} (\beta, B) \, . \tag{2.66}$$

Both the free energy and the average magnetization turn out to be analytic functions of β and h for $\beta < \infty$. In particular the spontaneous magnetization vanishes at any non-zero temperature:

$$M_+(\beta) = 0 \, , \qquad \forall \beta < \infty \, . \tag{2.67}$$

In Fig. 2.3 we plot the average magnetization $M(\beta, B) \equiv \lim_{N \to \infty} M_N(\beta, B)$ as a function of the applied magnetic field h for various values of the temperature

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

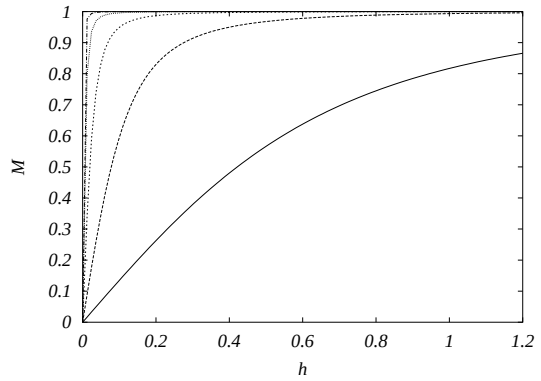

FIG. 2.3. The average magnetization of the one dimensional Ising model, as a function of the magnetic field B, at inverse temperatures $\beta = 0.5, 1, 1.5, 2$ (from bottom to top) {fig:ising1d_mag}

β. The curves become steeper and steeper as β increases. This statement can be made more quantitative by computing the **susceptibility** associated to the average magnetization:

$$\chi_M(\beta) = \frac{\partial M}{\partial h}(\beta, 0) = \beta\, e^{2\beta}\,. \tag{2.68}$$

This result can be interpreted as follows. A single spin in a field has susceptibility $\chi(\beta) = \beta$. If we consider N spins constrained to take the the same value, the corresponding susceptibility will be $N\beta$, as in Eq (2.54). In the present case the system behaves as if the spins were blocked into groups of $\chi(\beta)/\beta$ spins each. The spins in each group are constrained to take the same value, while spins belonging to different blocks are independent.

This qualitative interpretation receives further support by computing a **correlation function**. For $h = 0$ and $\delta N < i < j < (1-\delta)N$, one finds, at large
⋆ N:

$$\langle \sigma_i \sigma_j \rangle = e^{-|i-j|/\xi(\beta)} + \Theta(e^{-\alpha N})\,, \tag{2.69}$$

with $\xi(\beta) = -1/\log\tanh\beta$. Notice that $\xi(\beta)$ gives the typical distance below which two spins in the system are well correlated. For this reason it is usually called the **correlation length** of the model. This correlation length increases when the temperature decreases: spins become correlated at larger and larger distances. The result (2.69) is clearly consistent with our interpretation of the susceptibility. In particular, as $\beta \to \infty$, $\xi(\beta) \approx e^{2\beta}/2$ and $\chi(\beta) \approx 2\beta\xi(\beta)$.

The connection between correlation length and susceptibility is very general and can be understood as a consequence of the fluctuation-dissipation theorem (2.44):

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

$$\chi_M(\beta) = \beta N \left\langle \left(\frac{1}{N}\sum_{i=1}^N \sigma_i\right) ; \left(\frac{1}{N}\sum_{i=1}^N \sigma_i\right) \right\rangle$$
$$= \frac{\beta}{N}\sum_{i,j=1}^N \langle \sigma_i \, ; \, \sigma_j \rangle = \frac{\beta}{N}\sum_{i,j=1}^N \langle \sigma_i \sigma_j \rangle \, , \tag{2.70}$$

where the last equality comes from the fact that $\langle \sigma_i \rangle = 0$ when $B = 0$. Using (2.69), we get

$$\chi_M(\beta) = \beta \sum_{i=-\infty}^{+\infty} e^{-|i|/\xi(\beta)} + \Theta(e^{-\alpha N}) \, . \tag{2.71}$$

It is therefore evident that a large susceptibility must correspond to a large correlation length.

2.5.2 *The Curie-Weiss model*

{se:CurieWeiss}

The exact solution of the one-dimensional model, lead Ising to think that there couldn't be a phase transition in any dimension. Some thirty years earlier a qualitative theory of ferromagnetism had been put forward by Pierre Curie. Such a theory assumed the existence of a phase transition at non-zero temperature T_c (the so-called the "Curie point") and a non-vanishing spontaneous magnetization for $T < T_c$. The dilemma was eventually solved by Onsager solution of the two-dimensional model.

Curie theory is realized exactly within a rather abstract model: the so-called **Curie-Weiss model**. We shall present it here as one of the simplest solvable models with a finite temperature phase transition. Once again we have N Ising spins $\sigma_i \in \{\pm 1\}$ and a configuration is given by $\underline{\sigma} = (\sigma_1, \ldots, \sigma_N)$. However the spins no longer sits on a d-dimensional lattice: they all interact in pairs. The energy function, in presence of a magnetic field B, is given by:

$$E(\underline{\sigma}) = -\frac{1}{N}\sum_{(ij)} \sigma_i \sigma_j - B\sum_{i=1}^N \sigma_i \, , \tag{2.72}$$

where the sum on (ij) runs over all the couples of spins. Notice the peculiar $1/N$ scaling in front of the exchange term. The exact solution presented below shows that this is the only choice which yields a non-trivial free-energy density in the thermodynamic limit. This can be easily understood intuitively as follows. The sum over (ij) involves $O(N^2)$ terms of order $O(1)$. In order to get an energy function scaling as N, we need to put a $1/N$ coefficient in front.

In adopting the energy function (2.72), we gave up the description of any finite-dimensional geometrical structure. This is a severe simplification, but has the advantage of making the model exactly soluble. The Curie-Weiss model is the first example of a large family: the so-called **mean-field models**. We will explore many instances of this family throughout the book.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

A possible approach to the computation of the partition function consists in observing that the energy function can be written in terms of a simple observable, the **instantaneous magnetization**:

$$m(\underline{\sigma}) = \frac{1}{N}\sum_{i=1}^{N}\sigma_i\,. \tag{2.73}$$

Notice that this is a function of the configuration $\underline{\sigma}$, and shouldn't be confused with its expected value, the average magnetization, cf. Eq. (2.55). It is a "simple" observable because it is equal to the sum of observables depending upon a single spin.

We can write the energy of a configuration in terms of its instantaneous magnetization:

$$E(\underline{\sigma}) = \frac{1}{2}N - \frac{1}{2}N\,m(\underline{\sigma})^2 - NB\,m(\underline{\sigma})\,. \tag{2.74}$$

This implies the following formula for the partition function

$$Z_N(\beta,B) = e^{-N\beta/2}\sum_m \mathcal{N}_N(m)\,\exp\left\{\frac{N\beta}{2}m^2 + N\beta Bm\right\}\,, \tag{2.75}$$

where the sum over m runs over all the possible instantaneous magnetizations of N Ising spins: $m = -1+2k/N$ with $0 \le k \le N$ an integer number, and $\mathcal{N}_N(m)$ is the number of configurations having a given instantaneous magnetization. This is given by a binomial coefficient whose large N behavior is given in terms of the entropy function of a Bernoulli process:

$$\mathcal{N}_N(m) = \binom{N}{N\frac{1+m}{2}} \doteq \exp\left[N\,\mathcal{H}\left(\frac{1+m}{2}\right)\right]\,. \tag{2.76}$$

To leading exponential order in N, the partition function can thus be written as:

$$Z_N(\beta,B) \doteq \int_{-1}^{+1} dm\;\; e^{N\phi_{\rm mf}(m;\beta,B)} \tag{2.77}$$

where we have defined

$$\phi_{\rm mf}(m;\beta,B) = -\frac{\beta}{2}(1-m^2) + \beta Bm + \mathcal{H}\left(\frac{1+m}{2}\right)\,. \tag{2.78}$$

The integral in (2.77) is easily evaluated by Laplace method, to get the final result for the free-energy density

$$\phi(\beta,B) = \max_{m\in[-1,+1]}\phi_{\rm mf}(m;\beta,B)\,. \tag{2.79}$$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

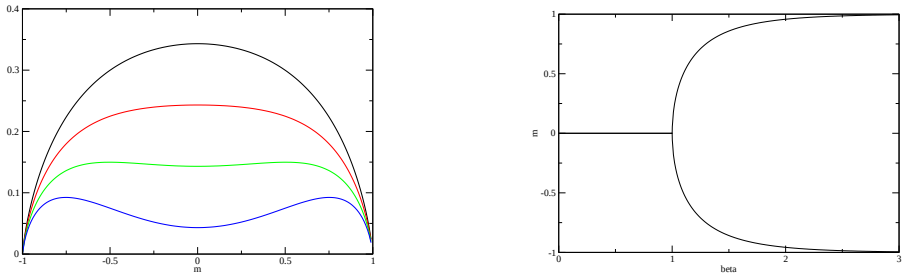

FIG. 2.4. Left: the function $\phi_{\rm mf}(m;\beta,B=0)$ is plotted versus m, for $\beta = .7, .9, 1.1, 1.3$ (from top to bottom). For $\beta < \beta_{\rm c} = 1$ there is a unique maximum at $m = 0$, for $\beta < \beta_{\rm c} = 1$ there are two degenerate maxima at two symmetric values $\pm m_+(\beta)$.Right: values of m which maximize $\phi_{\rm mf}(m;\beta,B=0)$ are plotted versus β. The phase transition at $\beta_{\rm c} = 1$ is signaled by the bifurcation. {fig:phiCW}

One can see that the maximum is obtained away from the boundary points, so that the corresponding m must be a stationary point of $\phi_{\rm mf}(m;\beta,B)$, which satisfies the **saddle-point equation** $\partial\phi_{\rm mf}(m;\beta,B)/\partial m = 0$:

$$m_* = \tanh(\beta m_* + \beta B)\,. \tag{2.80}$$

In the above derivation we were slightly sloppy at two steps: substituting the binomial coefficient with its asymptotic form and changing the sum over m into an integral. The mathematically minded reader is invited to show that these passages are indeed correct. ⋆

With a bit more work the above method can be extended to expectation values of observables. Let us consider for instance the average magnetization $M(\beta,B)$. It can be easily shown that, whenever the maximum of $\phi_{\rm mf}(m;\beta,B)$ ⋆ over m is non-degenerate,

$$M(\beta,B) \equiv \lim_{N\to\infty} \langle m(\underline{\sigma})\rangle = m_*(\beta,B) \equiv \arg\max_m \phi_{\rm mf}(m;\beta,B)\,, \tag{2.81}$$

We can now examine the implications that can be drawn from Eqs. (2.79) and (2.80). Let us first consider the $B = 0$ case (see Fig.2.4). The function $\phi_{\rm mf}(m;\beta,0)$ is symmetric in m. For $0 \le \beta \le 1 \equiv \beta_{\rm c}$, it is also concave and achieves its unique maximum in $m_*(\beta) = 0$. For $\beta > 1$, $m = 0$ remains a stationary point but becomes a local minimum, and the function develops two degenerate global maxima at $m_\pm(\beta)$ with $m_+(\beta) = -m_-(\beta) > 0$. These two maxima bifurcate continuously from $m = 0$ at $\beta = \beta_{\rm c}$.

A phase transition takes place at $\beta_{\rm c}$. Its meaning can be understood by computing the expectation value of the spins. Notice that the energy function (2.72) is symmetric a spin-flip transformation which maps $\sigma_i \to -\sigma_i$ for all i's. Therefore $\langle\sigma_i\rangle = \langle(-\sigma_i)\rangle = 0$ and the average magnetization vanishes $M(\beta,0) = 0$. On the other hand, the spontaneous magnetization, defined in (2.56), is zero

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

in the paramagnetic phase $\beta < \beta_c$, and equal to $m_+(\beta)$ in the ferromagnetic phase $\beta > \beta_c$. The physical interpretation of this phase is the following: for any finite N the pdf of the instantaneous magnetization $m(\underline{\sigma})$ has two symmetric peaks, at $m_\pm(\beta)$, which become sharper and sharper as N increases. Any external perturbation which breaks the symmetry between the peaks, for instance a small positive magnetic field B, favors one peak with respect to the other one, and therefore the system develops a spontaneous magnetization. Notice that, in mathematical terms, the phase transition is a property of systems in the thermodynamic limit $N \to \infty$.

In physical magnets the symmetry breaking can come for instance from impurities, subtle effects of dipolar interactions together with the shape of the magnet, or an external magnetic field. The result is that at low enough temperatures some systems, the ferromagnets develop a spontaneous magnetization. If you heat a magnet made of iron, its magnetization disappears at a critical temperature $T_c = 1/\beta_c = 770$ degrees Celsius. The Curie Weiss model is a simple solvable case exhibiting the phase transition.

Exercise 2.5 Compute the expansion of $m_+(\beta)$ and of $\phi(\beta, B = 0)$ near $\beta = \beta_c$, and show that the transition is of second order. Compute the low temperature behavior of the spontaneous magnetization.

{ex:Ising_inhom}

Exercise 2.6 Inhomogeneous Ising chain. The one dimensional Ising problem does not have a finite temperature phase transition, as long as the interactions are short range and translational invariant. But when the couplings in the Ising chain grow fast enough at large distance, one can have a phase transition. This is not a very realistic model from the point of view of physics, but it is useful as a solvable example of phase transition.

Consider a chain of Ising spins $\sigma_0, \sigma_1, \dots, \sigma_N$ with energy $E(\underline{\sigma}) = -\sum_{n=0}^{N-1} J_n \sigma_n \sigma_{n+1}$. Suppose that the coupling constants J_n form a positive, monotonously increasing sequence, growing logarithmically. More precisely, we assume that $lim_{n\to\infty} J_n / \log n = 1$. Denote by $\langle . \rangle_+$ (resp. $\langle . \rangle_-$) the expectation value with respect to Boltzmann's probability distribution when the spin σ_N is fixed to $\sigma_N = +1$ (resp. fixed to $\sigma_N = -1$).

(*i*) Show that , for any $n \in \{0. \dots, N-1\}$, the magnetization is $\langle \sigma_n \rangle_\pm = \prod_{p=n}^{N-1} \tanh(\beta J_p)$

(*ii*) Show that the critical inverse temperature $\beta_c = 1/2$ separates two regimes, such that: for $\beta < \beta_c$, one has $\lim_{N\to\infty} \langle \sigma_n \rangle_+ = \lim_{N\to\infty} \langle \sigma_n \rangle_- = 0$; for $\beta > \beta_c$, one has $\lim_{N\to\infty} \langle \sigma_n \rangle_\pm = \pm M(\beta)$, and $M(\beta) > 0$.

Notice that in this case, the role of the symmetry breaking field is played by the choice of boundary condition.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

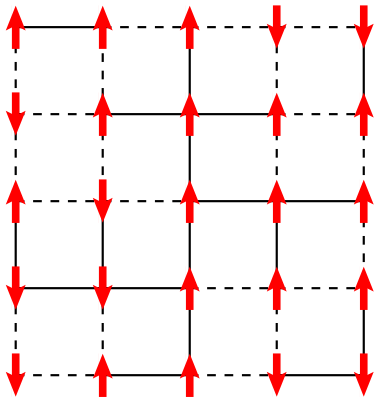

FIG. 2.5. A configuration of a two dimensional Edwards-Anderson model with $L = 5$. Spins are coupled by two types of interactions: ferromagnetic ($J_{ij} = +1$), indicated by a continuous line, and antiferromagnetic ($J_{ij} = -1$), indicated by a dashed line. The energy of the configuration shown here is $-14 - 7h$. {fig:ea_def}

{sec:SpinGlass}

2.6 The Ising spin glass

In real magnetic materials, localized magnetic moments are subject to several sources of interactions. Apart from the exchange interaction mentioned in the previous Section, they may interact through intermediate conduction electrons, etc... As a result, depending on the material which one considers, their interaction can be either ferromagnetic (their energy is minimized when they are parallel) or **antiferromagnetic** (their energy is minimized when they point *opposite* to each other). **Spin glasses** are a family of materials whose magnetic properties are particularly complex. They can be produced by diluting a small fraction of a 'transition magnetic metal' like manganese into a 'noble metal' like copper in a ratio 1 : 100. In such an alloy, magnetic moments are localized at manganese atoms, which are placed at random positions in a copper background. Depending on the distance of two manganese atoms, the net interaction between their magnetic moments can be either ferromagnetic or antiferromagnetic.

The **Edwards-Anderson model** is a widely accepted mathematical abstraction of these physical systems. Once again, the basic degrees of freedom are Ising spins $\sigma_i \in \{-1, +1\}$ sitting at the corners of a d-dimensional cubic lattice $\mathbb{L} = \{1, \dots, L\}^d$, $i \in \mathbb{L}$. The configuration space is therefore $\{-1, +1\}^{\mathbb{L}}$. As in the Ising model, the energy function reads

$$E(\underline{\sigma}) = -\sum_{(ij)} J_{ij}\sigma_i\sigma_j - B\sum_{i\in\mathbb{L}} \sigma_i \,, \tag{2.82}$$

where $\sum_{(ij)}$ runs over each edge of the lattice. Unlike in the Ising ferromagnet, a different coupling constant J_{ij} is now associated to each edge (ij), and its

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

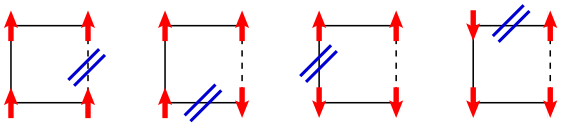

FIG. 2.6. Four configurations of a small Edwards-Anderson model: continuous lines indicate ferromagnetic interactions ($J_{ij} = +1$), while dashed lines are for antiferromagnetic interactions ($J_{ij} = -1$). In zero magnetic field ($h = 0$), the four configurations are degenerate and have energy $E = -2$. The bars indicate the unsatisfied interaction. Notice that there is no configuration with lower energy. This system is frustrated since it is impossible to satisfy simultaneously all constraints.

{fig:frustr}

sign can be positive or negative. The interaction between spins σ_i and σ_j is ferromagnetic if $J_{ij} > 0$ and antiferromagnetic if $J_{ij} < 0$.

A pictorial representation of this energy function is given in Fig. 2.5. The Boltzmann distribution is given by

$$p_\beta(\underline{\sigma}) = \frac{1}{Z(\beta)} \exp\left\{ \beta \sum_{(ij)} J_{ij}\sigma_i\sigma_j + \beta B \sum_{i\in\mathbb{L}} \sigma_i \right\}, \tag{2.83}$$

$$Z(\beta) = \sum_{\underline{\sigma}} \exp\left\{ \beta \sum_{(ij)} J_{ij}\sigma_i\sigma_j + \beta B \sum_{i\in\mathbb{L}} \sigma_i \right\}. \tag{2.84}$$

It is important to notice that the couplings $\{J_{ij}\}$ play a completely different role from the spins $\{\sigma_i\}$. The couplings are just parameters involved in the definition of the energy function, as the magnetic field B, and they are not summed over when computing the partition function. In principle, for any particular sample of a magnetic material, one should estimate experimentally the values of the J_{ij}'s, and then compute the partition function. We could have made explicit the dependence of the partition function and of the Boltzmann distribution on the couplings by using notations such as $Z(\beta, B; \{J_{ij}\})$, $p_{\beta,B;\{J_{ij}\}}(\underline{\sigma})$. However, when it is not necessary, we prefer to keep to lighter notations.

The present understanding of the Edwards-Anderson model is much poorer than for the ferromagnetic models introduced in the previous Section. The basic reason of this difference is **frustration** and is illustrated in Fig. 2.6 on an $L = 2$, $d = 2$ model (a model consisting of just 4 spins). A spin glass is frustrated whenever there exist local constraints that are in conflict, meaning that it is not possible to all of them satisfy simultaneously. In the Edwards Anderson model, a plaquette is a group of four neighbouring spins building a square. A plaquette is frustrated if and only if the product of the J_{ij} along all four edges of the plaquette is negative. As shown in Fig. 2.6, it is then impossible to minimize simultaneously all the four local energy terms associated with each edge. In a spin glass, the presence of a finite density of frustrated plaquettes generates a

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

very complicated energy landscape. The resulting effect of all the interactions is not obtained by 'summing' the effects of each of them separately, but is is the outcome of a complex interplay. The ground state spin configuration (the one satisfying the largest possible number of interactions) is difficult to find: it cannot be guessed on symmetry grounds. It is also frequent to find in a spin glass a configuration which is very different form the ground state but has an energy very close to the ground state energy. We shall explore these and related issues throughout the book.

Notes

There are many good introductory textbooks on statistical physics and thermodynamics, for instance the books by Reif (Reif, 1965) or Huang (Huang, 1987). Going towards more advanced texts, one can suggest the books by Ma (Ma, 1985) and Parisi (Parisi, 1998*b*). A more mathematically minded presentation can be found in the books by Gallavotti (Galavotti, 1999) and Ruelle (Ruelle, 1999).

The two dimensional Ising model at vanishing external field can also be solved by a transfer matrix technique, see for instance (Baxter, 1982). The transfer matrix, which passes from a column of the lattice to the next, is a $2^L \times 2^L$ matrix, and its dimension diverges exponentially with the lattice size L. Finding its largest eigenvalue is therefore a complicated task. Nobody has found the solution so far for $B \neq 0$.

Spin glasses will be a recurring theme in this book, and more will be said about them in the next Chapters. An introduction to this subject from a physicist point of view is provided by the book of Fischer and Hertz (Fischer and Hetz, 1993) or the review by Binder and Young (Binder and Young, 1986). The concept of frustration was introduced in a beautiful paper by Gerard Toulouse (Toulouse, 1977).

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

3

INTRODUCTION TO COMBINATORIAL OPTIMIZATION

{ch:intro_optim}

This Chapter provides an elementary introduction to some basic concepts in theoretical computer science. Which computational tasks can/cannot be accomplished efficiently by a computer? How much resources (time, memory, etc.) are needed for solving a specific problem? What are the performances of a specific solution method (an algorithm), and, whenever more than one method is available, which one is preferable? Are some problems intrinsically harder than others? This are some of the questions one would like to answer.

One large family of computational problems is formed by combinatorial optimization problems. These consist in finding a member of a finite set which maximizes (or minimizes) an easy-to-evaluate objective function. Several features make such problems particularly interesting. First of all, most of the time they can be converted into decision problems (questions which require a YES/NO answer), which are the simplest problems allowing for a rich theory of computational complexity. Second, optimization problems are ubiquitous both in applications and in pure sciences. In particular, there exist some evident connections both with statistical mechanics and with coding theory. Finally, they form a very large and well studied family, and therefore an ideal context for understanding some advanced issues. One should however keep in mind that computation is more than just combinatorial optimization. A distinct (and in some sense larger) family consists of counting problems. In this case one is asked to count how many elements of a finite set have some easy-to-check property. We shall say something about such problems in later Chapters. Another large family on which we will say basically nothing consists of continuous optimization problems.

This Chapter is organized as follows. The study of combinatorial optimization is introduced in Sec. 3.1 through a simple example. This section also contains the basic definition of graph theory that we use throughout the book. General definitions and terminology are given in Sec. 3.2. These definitions are further illustrated in Sec. 3.3 through several additional examples. Section 3.4 provides an informal introduction to some basic concepts in computational complexity. As mentioned above, combinatorial optimization problems often appear in pure sciences and applications. The examples of statistical physics and coding are briefly discussed in Secs. 3.5 and 3.6.

{sec:MST}

3.1 A first example: minimum spanning tree

The minimum spanning tree problem is easily stated and may appear in many practical applications. Suppose for instance you have a bunch of computers in a

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

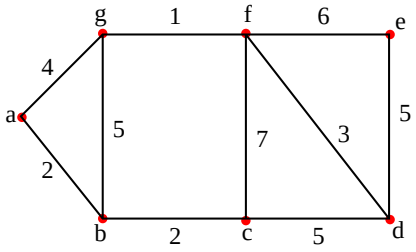

FIG. 3.1. This graph has 7 vertices (labeled a to g) and 10 edges. The 'cost' of each edge is indicated next to it. In the Minimum Spanning Tree problem, one seeks a subgraph connecting all vertices, without any loop, of minimum cost. {fig:MSTree}

building. You may want to connect them pairwise in such a way that the resulting network is completely connected and the amount of cable used is minimum.

3.1.1 *Definition of the problem and basics of graph theory*

A mathematical abstraction of the above practical problem requires us to first define basic graph theoretic definitions. A **graph** is a set $\mathcal{V}$ of vertices, labeled by $\{1, 2, \ldots, |\mathcal{V}|\}$ and a set $\mathcal{E}$ of edges connecting them: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The vertex set can be any finite set but one often takes the set of the first $|\mathcal{V}|$ integers: $\mathcal{V} = \{1, 2, \ldots, |\mathcal{V}|\}$. The edges are simply unordered couples of distinct vertices $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. For instance an edge joining vertices i and j is identified as $e = (i, j)$. A **weighted** graph is a graph where a cost (a real number) is associated with every edge. The **degree** of a vertex is the number of edges connected to it. A **path** between two vertices i and j is a set of edges $\{(j, i_2), (i_2, i_3), (i_3, i_4), \ldots, (i_{r-1}, i_r), (i_r, j)$. A graph is **connected** if, for every pair of vertices, there is a path which connects them. A **completely connected** graph, or **complete** graph, also called a **clique**, is a graph where all the $|\mathcal{V}|(|\mathcal{V}| - 1)/2$ edges are present. A **cycle** is a path starting and ending on the same vertex. A **tree** is a connected graph without a cycle.

Consider the graph in Fig. 3.1. You are asked to find a tree (a subset of the edges buiding a cycle-free subgraph) such that any two vertices are connected by exactly one path (in this case the tree is said to be spanning). To find such a subgraph is an easy task. The edges $\{(a, b); (b, c); (c, d); (b, g); (d, e)\}$, for instance, do the job. However in our problem a cost is associated with each edge. The cost of a subgraph is assumed to be equal to the sum of the costs of its edges. Your problem is to find the spanning tree with minimum cost. This is a non-trivial problem.

In general, an instance of the **minimum spanning tree** (MST) problem is given by a connected weighted graph (each edge e has a cost $w(e) \in \mathbb{R}$). The optimization problem consists in finding a spanning tree with minimum cost. What one seeks is an algorithm which, given an instance of the MST problem, outputs the spanning tree with lowest cost.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

3.1.2 *An efficient algorithm for the minimum spanning tree problem*

{sec:efficient}

The simple minded approach would consist in enumerating all the spanning trees for the given graph, and comparing their weights. However the number of spanning trees grows very rapidly with the size of the graph. Consider, as an example, the complete graph on N vertices. The number of spanning trees of such a graph is, according to the Cayley formula, N^{N-2}. Even if the cost of any such tree were evaluated in 10^{-3} sec, it would take 2 years to find the MST of a $N = 12$ graph, and half a century for $N = 13$. At the other extreme, if the graph is very simple, it may contain a small number of spanning trees, a single one in the extreme case where the graph is itself a tree. Nevertheless, in most interesting examples the situation is nearly as dramatic as in the complete graph case.

{thm:MSTtheorem}

A much better algorithm can be obtained from the following theorem:

Theorem 3.1 *Let $\mathcal{U} \subset \mathcal{V}$ be a proper subset of the vertex set $\mathcal{V}$ (such that neither $\mathcal{U}$ nor $\mathcal{V}\backslash\mathcal{U}$ are empty). Let us consider the subset $\mathcal{F}$ of edges which connect a vertex in $\mathcal{U}$ to a vertex in $\mathcal{V}\backslash\mathcal{U}$, and let $e \in \mathcal{F}$ be an edge of lowest cost in this subset: $w(e) \leq w(e')$ for any $e' \in \mathcal{F}$. If there are several such edges, e can be any of them. Then there exists a minimum spanning tree which contains e.*

Proof: Consider a MST $\mathcal{T}$, and suppose that it does not contain the edge e. This edge is such that $e = (i, j)$ with $i \in \mathcal{U}$ and $j \in \mathcal{V}\backslash\mathcal{U}$. The spanning tree $\mathcal{T}$ must contain a path between i and j. This path contains at least one edge f connecting a vertex in $\mathcal{U}$ to a vertex in $\mathcal{V}\backslash\mathcal{U}$, and f is distinct from e. Now consider the subgraph $\mathcal{T}'$ built from $\mathcal{T}$ by removing the edge f and adding the edge e. We leave to the reader the exercise of showing that $\mathcal{T}'$ is a spanning tree. Moreover $E(\mathcal{T}') = E(\mathcal{T}) + w(e) - w(f)$. Since $\mathcal{T}$ is a MST, $E(\mathcal{T}') \geq E(\mathcal{T})$. On the other hand e has minimum cost within $\mathcal{F}$, hence $w(e) \leq w(f)$. Therefore $w(e) = w(f)$ and $\mathcal{T}'$ is a MST containing e. □

This result allows to construct a minimum spanning tree of a graph incrementally. One starts from a single vertex. At each step a new edge can be added to the tree, whose cost is minimum among all the ones connecting the already existing tree with the remaining vertices. After $N-1$ iterations, the tree will be spanning.

MST algorithm ((Prim, 1957))

Input: A non-empty connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and a weight function $w : \mathcal{E} \to \mathbb{R}_+$.

Output: A minimum spanning tree $\mathcal{T}$ and its cost $E(\mathcal{T})$.

1. Set $\mathcal{U} := \{1\}$, $\mathcal{T} := \emptyset$ and $E = 0$.
2. While $\mathcal{V}\backslash\mathcal{U}$ is not empty
   - 2.1 Let $\mathcal{F} := \{e = (ij) \in \mathcal{E}$ such that $i \in \mathcal{U},\ j \in \mathcal{V}\backslash\mathcal{U}\}$.
   - 2.2 Find $e_* := \arg\min_{e\in\mathcal{F}}\{w(e)\}$. Let $e_* := (i_*, j_*)$ with $i_* \in \mathcal{U}$, $j_* \in \mathcal{V}\backslash\mathcal{U}$.
   - 2.3 Set $\mathcal{U} := \mathcal{U} \cup i_*$, $\mathcal{T} := \mathcal{T} \cup e_*$, and $E := E + w(e_*)$.
3. Output the spanning tree $\mathcal{T}$ and its cost E.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

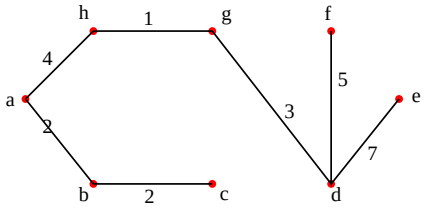

FIG. 3.2. A minimum spanning tree for the graph defined in Fig. 3.1. The cost of this tree is $E = 17$. {fig:MSTree_sol}

Figure 3.2 gives the MST for the problem described in Fig. 3.1. It is easy to obtain it by applying the above algorithm. ⋆

Exercise 3.1 Show explicitly that the algorithm `MST` always outputs a minimum spanning tree.

Theorem 3.1 establishes that, for any $\mathcal{U} \subset \mathcal{V}$, and any lowest cost edge e among the ones connecting $\mathcal{U}$ to $\mathcal{V}\backslash\mathcal{U}$, there exists a MST containing e. This does not guarantee that, when two different sets $\mathcal{U}_1$ and $\mathcal{U}_2$, and the corresponding lowest cost edges e_1 and e_2 are considered, there exists a MST containing *both* e_1 and e_2. The above algorithm works by constructing a sequence of such $\mathcal{U}$'s and adding to the tree the corresponding lowest weight edges. It is therefore not obvious a priori that it will output a MST (unless this is unique).

Let us analyze the number of elementary operations required by the algorithm to construct a spanning tree on an N nodes graph. By 'elementary operation' we mean comparisons, sums, multiplications, etc, all of them counting as one. Of course, the number of such operations depends on the graph, but we can find a simple upper bound by considering the completely connected graph. Most of the operations in the above algorithm are comparisons among edge weights for finding e_* in step `2.2`. In order to identify e_*, one has to scan at most $|\mathcal{U}| \times |\mathcal{V}\backslash\mathcal{U}| = |\mathcal{U}| \times (N - |\mathcal{U}|)$ edges connecting $\mathcal{U}$ to $\mathcal{V}\backslash\mathcal{U}$. Since $|\mathcal{U}| = 1$ at the beginning and is augmented of one element at each iteration of the cycle `2.1-2.3`, the number of comparisons is upper bounded by $\sum_{U=0}^{N} U(N-U) \leq N^3/6$[6]. This is an example of a polynomial algorithm, whose computing time grows like a power law of the number of vertices. The insight gained from the theorem provides an algorithm which is much better than the naive one, at least when N gets large.

3.2 General definitions {sec:gendef}

MST is an example of a **combinatorial optimization problem**. This is defined by a set of possible instances. An instance of MST is defined by a connected

[6]The algorithm can be easily improved by keeping an ordered list of the edges already encountered

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

weighted graph. In general, an **instance** of a combinatorial optimization problem is described by a finite set $\mathcal{X}$ of allowed **configurations** and a **cost function** E defined on this set and taking values in $\mathbb{R}$. The optimization problem consists in finding the **optimal** configuration $C \in \mathcal{X}$, namely the one with the smallest cost $E(C)$. Any set of such instances defines a combinatorial optimization problem. For a particular instance of MST, the space of configurations is simply the set of spanning trees on the given graph, while the cost function associated with each spanning tree is the sum of the costs of its edges.

We shall say that an algorithm solves an optimization problem if, for every instance of the optimization problem, it gives the optimal configuration, or if it computes its cost. In all the problems which we shall discuss, there is a 'natural' measure of the size of the problem N (typically a number of variables used to define a configuration, like the number of edges of the graph in MST), and the number of configurations scales, at large N like c^N, or in some cases even faster, e. g. like $N!$ or N^N. Notice that, quite generally, evaluating the cost function on a particular configuration is an easy task. The difficulty of solving the combinatorial optimization problem comes therefore essentially from the size of the configuration space.

It is a generally accepted practice to estimate the **complexity** of an algorithm as the number of 'elementary operations' required to solve the problem. Usually one focuses onto the asymptotic behavior of this quantity as $N \to \infty$. It is obviously of great practical interest to construct algorithms whose complexity is as small as possible.

One can solve a combinatorial optimization problem at several levels of refinement. Usually one distinguishes three types of problems:

- The **optimization** problem: Find an optimal configuration C^*.
- The **evaluation** problem: Determine the cost $E(C^*)$ of an optimal configuration.
- The **decision** problem: Answer to the question: "Is there a configuration of cost less than a given value E_0?"

{sec:Examples}

3.3 More examples

The general setting described in the previous Section includes a large variety of problems having both practical and theoretical interest. In the following we shall provide a few selected examples.

3.3.1 *Eulerian circuit*

One of the oldest documented examples goes back to the 18th century. The old city of Königsberg had seven bridges (see Fig. 3.3), and its habitants were wondering whether it was possible to cross once each of this bridges and get back home. This can be generalized and translated in graph-theoretic language as the following decision problem. Define a **multigraph** exactly as a graph but for the fact that two given vertices can be connected by several edges. The problem consists in finding whether there is there a circuit which goes through all edges

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

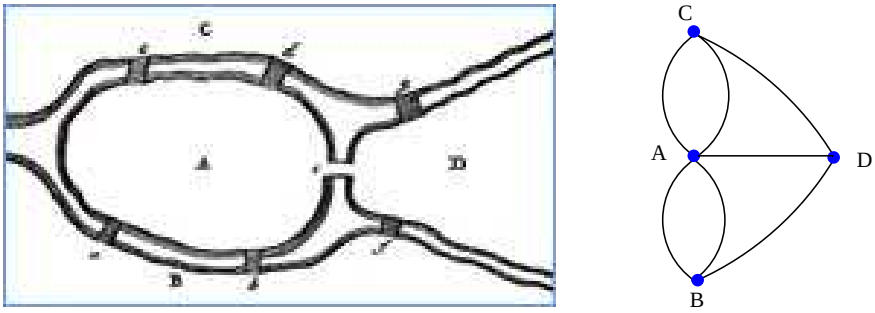

FIG. 3.3. Left: a map of the old city of Königsberg, with its seven bridges, as drawn in Euler's paper of 1736. The problem is whether one can walk along the city, crossing each bridge exactly once and getting back home. Right: a graph summarizing the problem. The vertices A, B, C, D are the various parts of lands separated by a river, an edge exists between two vertices whenever there is a bridge. The problem is to make a closed circuit on this graph, going exactly once through every edge. {fig:seven-bridges}

of the graph only once, and returns to its starting point. Such a circuit is now called a **Eulerian circuit**, because this problem was solved by Euler in 1736, when he proved the following nice theorem. As for ordinary graphs, we define the **degree** of a vertex as the number of edges which have the vertex as an end-point.

Theorem 3.2 *Given a connected multigraph, there exists an Eulerian circuit if and only if every vertex has an even degree.* {th:euler}

This theorem automatically provides an algorithm for the decision problem whose complexity grows linearly with the number of vertices of the graph: just go through all the vertices of the graph and check their degree.

Exercise 3.2 Show that, if an Eulerian circuit exists the degrees are necessarily even.

Proving the inverse implication is slightly more difficult. A possible approach consists in showing the following slightly stronger result. If all the vertices of a connected graph $\mathcal{G}$ have even degree but i and j, then there exists a path from i to j that visits once each edge in $\mathcal{G}$. This can be proved by induction on the number of vertices. [Hint: Stat from i and make a step along the edge (i, i'). Show that it is possible to choose i' in such a way that the residual graph $\mathcal{G}\backslash(i, i')$ is connected.]

3.3.2 *Hamiltonian cycle*

More than a century after Euler's theorem, the great scientist sir William Hamilton introduced in 1859 a game called the icosian game. In its generalized form,

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

it basically asks whether there exists, in a graph, a **Hamiltonian cycle**, which is a path going once through every vertex of the graph, and getting back to its starting point. This is another decision problem, and, at a first look, it seems very similar to the Eulerian circuit. However it turns out to be much more complicated. The best existing algorithms for determining the existence of an Hamiltonian cycle on a given graph run in a time which grows exponentially with the number of vertices N. Moreover, the theory of computational complexity, which we shall describe later in this Chapter, strongly suggests that this problem is in fact intrinsically difficult.

3.3.3 *Traveling salesman*

Given a complete graph with N points, and the distances d_{ij} between all pairs of points $1 \leq i < j \leq N$, the famous **traveling salesman problem** (TSP) is an optimization problem: find a Hamiltonian cycle of minimum total length. One can consider the case where the points are in a portion of the plane, and the distances are Euclidean distances (we then speak of a Euclidean TSP), but of course the problem can be stated more generally, with d_{ij} representing general costs, which are not necessarily distances. As for the Hamiltonian cycle problem, the best algorithms known so far for the TSP have a running time which grows exponentially with N at large N. Nevertheless Euclidean problems with thousands of points can be solved.

3.3.4 *Assignment*

Given N persons and N jobs, and a matrix C_{ij} giving the affinity of person i for job j, the **assignment** problem consists in finding the assignment of the jobs to the persons (an exact one-to-one correspondence between jobs and persons) which maximizes the total affinity. A configuration is characterized by a permutation of the N indices (there are thus $N!$ configurations), and the cost of the permutation π is $\sum_i C_{i\pi(i)}$. This is an example of a polynomial problem: there exists an algorithm solving it in a time growing like N^3.

3.3.5 *Satisfiability*

In the **satisfiability** problem one has to find the values of N Boolean variables $x_i \in \{T, F\}$ which satisfy a set of logical constraints. Since each variable can be either true or false, the space of configurations has size $|\mathcal{X}| = 2^N$. Each logical constraint, called in this context a **clause**, takes a special form: it is the logical `OR` (for which we use the symbol $\vee$) of some variables or their negations. For instance $x_1 \vee \overline{x}_2$ is a 2-clause (2-clause means a clause of length 2, i.e. which involves exactly 2 variables), which is satisfied if either $x_1 = T$, or $x_2 = F$, or both. $\overline{x}_1 \vee \overline{x}_2 \vee x_3$ is a 3-clause, which is satisfied by all configurations of the three variables except $x_1 = x_2 = T$, $x_3 = F$. The problem is to determine whether there exists a configuration which satisfies all constraints (decision problem), or to find the configuration which minimizes the number of violated constraints (optimization problem). The decision problem is easy when all the clauses have length smaller or equal to 2: there exists an algorithm running in a time growing

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

linearly with N. In other cases, all known algorithms solving the satisfiability decision problem run in a time which grows exponentially with N.

3.3.6 *Coloring and vertex covering*

Given a graph and an integer q, the famous **q-coloring** problem asks if it is possible to color the vertices of the graph using q colors, in such a way that two vertices connected by an edge have different colors. In the same spirit, the **vertex-cover** problem asks to cover the vertices with 'pebbles', using the smallest possible number of pebbles, in such a way that every edge of the graph has at least one of its two endpoints covered by a pebble.

3.3.7 *Number partitioning*

Number partitioning is an example which does not come from graph theory. An instance is a set $\mathcal{S}$ of N integers $\mathcal{S} = \{x_1, .., x_N\}$. A configuration is a partition of these numbers into two groups $\mathcal{A}$ and $\mathcal{S} \setminus \mathcal{A}$. Is there a partition such that $\sum_{i\in\mathcal{A}} x_i = \sum_{i\in\mathcal{S}\setminus\mathcal{A}} x_i$?

3.4 Elements of the theory of computational complexity

{sec:Complexity}

One main branch of theoretical computer science aims at constructing an intrinsic theory of computational complexity. One would like, for instance, to establish which problems are harder than others. By 'harder problem', we mean a problem that takes a longer running time to be solved. In order to discuss rigorously the computational complexity of a problem, we would need to define a precise *model of computation* (introducing, for instance, Turing machines). This would take us too far. We will instead evaluate the running time of an algorithm in terms of 'elementary operations': comparisons, sums, multiplications, etc. This informal approach is essentially correct as long as the size of the operands remains uniformly bounded.

3.4.1 *The worst case scenario*

As we already mentioned in Sec. 3.2, a combinatorial optimization problem, is defined by the set of its possible instances. Given an algorithm solving the problem, its running time will vary from instance to instance, even if the instance 'size' is fixed. How should we quantify the overall hardness of the problem? A crucial choice of computational complexity theory consists in considering the 'worst' (i.e. the one which takes longer time to be solved) instance among all the ones having the same size.

This choice has two advantages: (*i*) It allows to construct a 'universal' theory. (*ii*) Once the worst case running time of a given algorithm is estimated, this provides a performance guarantee on any instance of the problem.

3.4.2 *Polynomial or not?*

A second crucial choice consists in classifying algorithms in two classes: (*i*) **Polynomial**, if the running time is upper bounded by a fixed polynomial in the size

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

of the instance. In mathematical terms, let T_N the number of operations required for solving an instance of size N in the worst case. The algorithm is polynomial when there exist a constant k such that $T_N = O(N^k)$. (*ii*) **Super-polynomial**, if no such upper bound exists. This is for instance the case if the time grows exponentially with the size of the instance (we shall call algorithms of this type **exponential**), i.e. $T_N = \Theta(k^N)$ for some constant k.

Example 3.3 In 3.1.2, we were able to show that the running time of the `MST` algorithm is upper bounded by N^3, with N the number of vertices tin the graph. This implies that such an algorithm is polynomial.

Notice that we did not give a precise definition of the 'size' of a problem. One may wonder whether, changing the definition, a particular problem can be classified both as polynomial an as super-polynomial. Consider, for instance, the assignment problem with $2N$ points. One can define the size as being N, or $2N$, or even N^2 which is the number of possible person-job pairs. The last definition would be relevant if one would work for instance with occupation numbers $n_{ij} \in \{0, 1\}$, the number n_{ij} being one if and only if the job i is assigned to person j. However, any two of these 'natural' definitions of size are a polynomial function one of the other. Therefore they do not affect the classification of an algorithm as polynomial or super-polynomial. We will discard other definitions (such as e^N or $N!$) as 'unnatural', without any further ado. The reader can convince himself on each of the examples of the previous Section.

3.4.3 *Optimization, evaluation, decision*

In order to get a feeling of their relative levels of difficulty, let us come back for a while to the three types of optimization problems defined in Sec. 3.2, and study which one is the hardest.

Clearly, if the cost of any configuration can be computed in polynomial time, the evaluation problem is not harder than the optimization problem: if one can find the optimal configuration in polynomial time, one can compute its cost also in polynomial time. The decision problem (deciding whether there exists a configuration of cost smaller than a given E_0) is not harder than the evaluation problem. So the order of increasing difficulty is: decision, evaluation, optimization.

But actually, in many cases where the costs take discrete values, the evaluation problem is not harder than the decision problem, in the following sense. Suppose that we have a polynomial algorithm solving the decision problem, and that the costs of all configurations can be scaled to be integers in an interval $[0, E_{\max}]$ of length $E_{\max} = \exp\{O(N^k)\}$ for some $k > 0$. An algorithm solving the decision problem can be used to solve the evaluation problem by dichotomy: one first takes $E_0 = E_{\max}/2$. If there exists a configuration of energy smaller than E_0, one iterates with E_0 the center of the interval $[0, E_{\max}/2]$. In the opposite case, one iterates with E_0 the center of the interval $[E_{\max}/2, E_{\max}]$. Clearly

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

this procedure finds the cost of the optimal configuration(s) in a time which is also polynomial.

3.4.4 *Polynomial reduction*

{sub:polred}

One would like to compare the levels of difficulty of various *decision problems*. The notion of polynomial reduction formalizes the sentence "not harder than" which we used so far, and helps to get a classification of decision problems.

Roughly speaking, we say that a problem $\mathcal{B}$ is not harder than $\mathcal{A}$ if any efficient algorithm for $\mathcal{A}$ (if such an algorithm existed) could be used as a subroutine of an algorithm solving efficiently $\mathcal{B}$. More precisely, given two decision problems $\mathcal{A}$ and $\mathcal{B}$, one says that $\mathcal{B}$ is **polynomially reducible** to $\mathcal{A}$ if the following conditions hold:

1. There exists a mapping R which transforms any instance $\mathtt{I}$ of problem $\mathcal{B}$ into an instance $R(\mathtt{I})$ of problem $\mathcal{A}$, such that the solution (yes/no) of the instance $R(\mathtt{I})$ of $\mathcal{A}$ gives the solution (yes/no) of the instance $\mathtt{I}$ of $\mathcal{B}$.

2. The mapping $\mathtt{I} \mapsto R(\mathtt{I})$ can be computed in a time which is polynomial in the size of $\mathtt{I}$.

3. The size of $R(\mathtt{I})$ is polynomial in the size of $\mathtt{I}$. This is in fact a consequence of the previous assumptions but there is no harm in stating it explicitly.

A mapping R satisfying the above requirements is called a polynomial reduction. Constructing a polynomial reduction among two problems is an important achievement since it effectively reduces their study to the study of just one of them. Suppose for instance to have a polynomial algorithm $\mathrm{Alg}_{\mathcal{A}}$ for solving $\mathcal{A}$. Then a polynomial reduction of $\mathcal{B}$ to $\mathcal{A}$ can be used for constructing a polynomial algorithm for solving $\mathcal{B}$. Given an instance $\mathtt{I}$ of $\mathcal{B}$, the algorithm just compute $R(\mathtt{I})$, feeds it into the $\mathrm{Alg}_{\mathcal{A}}$, and outputs the output of $\mathrm{Alg}_{\mathcal{A}}$. Since the size of $R(\mathtt{I})$ is polynomial in the size of $\mathtt{I}$, the resulting algorithm for $\mathcal{B}$ is still polynomial.

For concreteness, we will work out an explicit example. We will show that the problem of existence of a Hamiltonian cycle in a graph is polynomially reducible to the satisfiability problem.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Example 3.4 An instance of the Hamiltonian cycle problem is a graph with N vertices, labeled by $i \in \{1, ..., N\}$. If there exists a Hamiltonian cycle in the graph, it can be characterized by N^2 Boolean variables $x_{ri} \in \{0, 1\}$, where $x_{ri} = 1$ if vertex number i is the r'th vertex in the cycle, and $x_{ri} = 0$ otherwise (one can take for instance $x_{11} = 1$). We shall now write a number of constraints that the variables x_{ri} must satisfy in order for a Hamiltonian cycle to exist, and we shall ensure that these constraints take the forms of the clauses used in the satisfiability problem (identifying $x = 1$ as true, $x = 0$ as false):

- Each vertex $i \in \{1, ..., N\}$ must belong to the cycle: this can be written as the clause $x_{1i} \vee x_{2i} \vee .... \vee x_{Ni}$, which is satisfied only if at least one of the numbers $x_{1i}, x_{2i}, ..., x_{Ni}$ equals one.
- For every $r \in \{1, ..., N\}$, one vertex must be the r'th visited vertex in the cycle: $x_{r1} \vee x_{r2} \vee ... \vee x_{rN}$
- Each vertex $i \in \{1, ..., N\}$ must be visited only once. This can be implemented through the $N(N-1)/2$ clauses $\bar{x}_{rj} \vee \bar{x}_{sj}$, for $1 \leq r < s \leq N$.
- For every $r \in \{1, ..., N\}$, there must be only one r'th visited vertex in the cycle; This can be implemented through the $N(N-1)/2$ clauses $\overline{x}_{ri} \vee \overline{x}_{rj}$, for $1 \leq i < j \leq N$.
- For every pair of vertices $i < j$ which are not connected by an edge of the graph, these vertices should not appear consecutively in the list of vertices of the cycle. Therefore we add, for every such pair and for every $r \in \{1, ..., N\}$,the clauses $\overline{x}_{ri} \vee \overline{x}_{(r+1)j}$ and $\overline{x}_{rj} \vee \overline{x}_{(r+1)i}$ (with the 'cyclic' convention $N + 1 = 1$).

It is straightforward to show that the size of the satisfiability problem constructed in this way is polynomial in the size of the Hamiltonian cycle problem. We leave as an exercise to show that the set of all above clauses is a sufficient set: if the N^2 variables satisfy all the above constraints, they describe a Hamiltonian cycle.

3.4.5 *Complexity classes*

Let us continue to focus onto decision problems. The classification of these problems with respect to polynomiality is as follows:

- **Class P:** These are the **polynomial** problems, for which there exists an algorithm running in polynomial time. An example, cf. Sec. 3.1, is the decision version of the minimum spanning tree (which asks for a yes/no answer to the question: given a graph with costs on the edges, and a number E_0, is there a spanning tree with total cost less than E_0?).
- **Class NP:** This is the class of **non-deterministic polynomial** problems, which can be solved in polynomial time by a 'non deterministic' algorithm. Roughly speaking, such an algorithm can run in parallel on an arbitrarily large number of processors. We shall not explain this notion in detail here, but rather use an alternative and equivalent characterization. We say that a

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

problem is in the class NP if there exists a 'short' certificate which allows to check a 'yes' answer to the problem. A short certificate means a certificate that can be checked in polynomial time.

A polynomial problem like the minimum spanning tree describes above is automatically in NP so P $\subseteq$ NP. The decision version of the TSP is also in NP: if there is a TSP tour with cost smaller than E_0, the short certificate is simple: just give the tour, and its cost will be computed in linear time, allowing to check that it is smaller than E_0. Satisfiability also belongs to NP: a certificate is obtained from the assignment of variables satisfying all clauses. Checking that all clauses are satisfied is linear in the number of clauses, taken here as the size of the system. In fact there are many important problems in the class NP, with a broad spectrum of applications ranging from routing to scheduling, to chip verification, or to protein folding...

- **Class NP-complete:** These are the hardest problem in the NP class. A problem is **NP-complete** if: (i) it is in NP, (ii) any other problem in NP can be polynomially reduced to it, using the notion of polynomial reduction defined in Sec. 3.4.4. If $\mathcal{A}$ is NP-complete, then: for any other problem $\mathcal{B}$ in NP, there is a polynomial reduction mapping $\mathcal{B}$ to $\mathcal{A}$. So if we had a polynomial algorithm to solve $\mathcal{A}$, then all the problems in the broad class NP would be solved in polynomial time.

It is not *a priori* obvious whether there exist any NP-complete problem. A major achievement of the theory of computational complexity is the following theorem, obtained by Cook in 1971.

Theorem 3.5 *The satisfiability problem is NP-complete*

We shall not give here the proof of the theorem. Let us just mention that the satisfiability problem has a very universal structure (an example of which was shown above, in the polynomial reduction of the Hamiltonian cycle problem to satisfiability). A clause is built as the logical `OR` (denoted by $\vee$) of some variables, or their negations. A set of several clauses, to be satisfied simultaneously, is the logical `AND` (denoted by $\wedge$) of the clauses. Therefore a satisfiability problem is written in general in the form $(a_1 \vee a_2 \vee ...) \wedge (b_1 \vee b_2 \vee ...) \wedge ....$, where the a_i, b_i are 'literals', i.e. any of the original variables or their negations. This form is called a **conjunctive normal form** (CNF), and it is easy to see that any logical statement between Boolean variables can be written as a CNF. This universal decomposition gives some idea of why the satisfiability problem can play a central role.

3.4.6 *P=NP ?*

When a NP-complete problem $\mathcal{A}$ is known, one can relatively easily find other NP-complete problems: if there exists a polynomial reduction from $\mathcal{A}$ to another problem $\mathcal{B} \in$ NP, then $\mathcal{B}$ is also NP-complete. In fact, whenever $R_{\mathcal{A}\leftarrow\mathcal{P}}$ is a polynomial reduction from a problem $\mathcal{P}$ to $\mathcal{A}$ and $R_{\mathcal{B}\leftarrow\mathcal{A}}$ is a polynomial reduc-

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

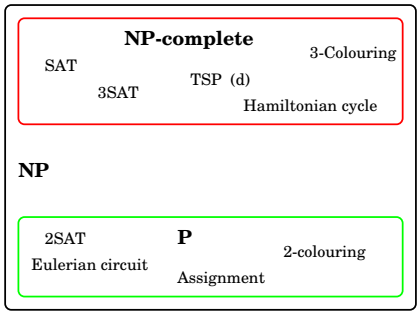

FIG. 3.4. Classification of some famous decision problems. If P $\neq$ NP, the classes P and NP-complete are disjoint. If it happened that $P = NP$, all the problems in NP, and in particular all those mentioned here, would be solvable in polynomial time.

tion from $\mathcal{A}$ to $\mathcal{B}$, then $R_{\mathcal{B}\leftarrow\mathcal{A}} \circ R_{\mathcal{A}\leftarrow\mathcal{P}}$ is a polynomial reduction from $\mathcal{P}$ to $\mathcal{B}$. Starting from satisfiability, it has been possible to find, with this method, thousands of NP-complete problems. To quote a few of them, among the problems we have encountered so far, Hamiltonian circuit, TSP, and 3-satisfiability (i.e. satisfiability with clauses of length 3 only) are NP-complete. Actually most of NP problems can be classified either as being in P, or being NP-complete. The precise status of some NP problems, like graph isomorphism, is still unknown.

Finally, those problems which, not being in NP are at least as hard as NP-complete problems, are usually called **NP-hard**. These includes both decision problems for which a short certificate does not exist, and non-decision problems. For instance the optimization and evaluation versions of TSP are NP-hard. However, in such cases, we shall chose among the expressions 'TSP is NP-complete' or 'TSP is NP-hard' rather freely.

One major open problem in the theory of computational complexity is whether the classes P and NP are distinct or not. It might be that P=NP=NP-complete: this would be the case if someone found a polynomial algorithm for one NP-complete problem. This would imply that no problem in the broad NP-class could be solved in polynomial time.

It is a widespread conjecture that there exist no polynomial algorithm for NP-complete problems. Then the classes P and NP-complete would be disjoint. In fact it is known that, if P $\neq$ NP, then there are NP problems which are neither in P nor in NP-complete.

3.4.7 *Other complexity classes*

Notice the fundamental asymmetry in the definition of the NP class: the existence of a short certificate is requested only for the yes answers. To understand the meaning of this asymmetry, consider the problem of unsatisfiability (which is the complement of the satisfiability problem) formulated as: "given a set of

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

clauses, is the problem unsatisfiable?". It is not clear if there exists a short certificate allowing to check a yes answer: it is very difficult to prove that a problem cannot be satisfied without checking an exponentially large number of possible configurations. So it is not at all obvious that unsatisfiability is in NP. Problems which are complements of those in NP define the class of co-NP problems, ans it is not known whether NP=co-NP or not, although it is widely believed that co-NP is different from NP. This consideration opens a Pandora box with many other classes of complexities, but we shall immediately close it since it would carry us too far.

3.5 Optimization and statistical physics

{sec:OptimizationPhysics}

3.5.1 *General relation*

There exists a natural mapping from optimization to statistical physics. Consider an optimization problem defined by a finite set $\mathcal{X}$ of allowed configurations, and a cost function E defined on this set with values in $\mathbb{R}$. While optimization consists in finding the configuration $C \in \mathcal{X}$ with the smallest cost, one can introduce a probability measure of the Boltzmann type on the space of configurations: For any β, each C is assigned a probability [7]

$$p_\beta(C) = \frac{1}{Z(\beta)} e^{-\beta E(C)} \quad ; \quad Z(\beta) = \sum_{C \in \mathcal{X}} e^{-\beta E(C)} \ . \tag{3.1}$$

{eq:boltzmann_optim}

The positive parameter β plays the role of an inverse temperature. In the limit $\beta \to \infty$, the probability distribution p_β concentrates on the configurations of minimum energy (ground states in the statistical physics jargon). This is the relevant limit for optimization problems. In the statistical physics approach one generalizes the problem to study properties of the distribution p_β at finite β. In many cases it is useful to follow p_β when β increases (for instance by monitoring the thermodynamic properties: internal energy, the entropy, and the specific heat). This may be particularly useful, both for analytical and for algorithmic purpose, when the thermodynamic properties evolve smoothly. An example of practical application is the simulated annealing method, which actually samples the configuration space at larger and larger values of β until it finds a ground state. It will be described in Chap. 4. Of course the existence of phase transitions pose major challenges to this kind of strategies, as we will see.

3.5.2 *Spin glasses and maximum cuts*

To give a concrete example, let us go back to the spin glass problem of Sec. 2.6. This involves N Ising spins $\sigma_1, \ldots, \sigma_N$ in $\{\pm 1\}$, located on the vertices of a graph, and the energy function is:

[7] Notice that there exist alternatives to the straightforward generalization (3.1). In some problems the configuration space involves hard constraints, which can also be relaxed in a finite temperature version.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

$$E(\underline{\sigma}) = -\sum_{(ij)} J_{ij}\sigma_i\sigma_j, \tag{3.2}$$

where the sum $\sum_{(ij)}$ runs over all edges of the graph and the J_{ij} variables are exchange couplings which can be either positive or negative. Given the graph and the exchange couplings, what is the ground state of the corresponding spin glass? This is a typical optimization problem. In fact, it very well known in computer science in a slightly different form.

Each spin configuration partitions the set of vertices into two complementary subsets: $V_\pm = \{i \,|\, \sigma_i = \pm 1\}$. Let us call $\gamma(V_+)$ the set of edges with one endpoint in V_+, the other in V_-. The energy of the configuration can be written as:

$$E(\underline{\sigma}) = -C + 2\sum_{(ij)\in\gamma(V_+)} J_{ij}, \tag{3.3}$$

where $C = \sum_{(ij)} J_{ij}$. Finding the ground state of the spin glass is thus equivalent to finding a partition of the vertices, $V = V_+ \cup V_-$, such that $\sum_{(ij)\in\gamma(V_+)} c_{ij}$ is maximum, where $c_{ij} \equiv -J_{ij}$. This problem is known as the **maximum cut** problem (MAX-CUT): the set of edges $\gamma(V_+)$ is a cut, each cut is assigned a weight $\sum_{(ij)\in\gamma(V_+)} c_{ij}$, and one seeks the cut with maximal weight.

Standard results on max-cut immediately apply: In general this is an NP-hard problem, but there are some categories of graphs for which it is polynomially solvable. In particular the max-cut of a planar graph can be found in polynomial time, providing an efficient method to obtain the ground state of a spin glass on a square lattice in two dimensions. The three dimensional spin glass problem falls into the general NP-hard class, but nice 'branch and bound' methods, based on its max-cut formulation, have been developed for it in recent years.

Another well known application of optimization to physics is the random field Ising model, which is a system of Ising spins with ferromagnetic couplings (all J_{ij} are positive), but with a magnetic field h_i which varies from site to site taking positive and negative values. Its ground state can be found in polynomial time thanks to its equivalence with the problem of finding a maximal flow in a graph.

{sec:OptimizationCoding}

3.6 Optimization and coding

Computational complexity issues are also crucial in all problems of information theory. We will see it recurrently in this book, but let us just give here some small examples in order to fix ideas.

Consider the error correcting code problem of Chapter 1. We have a code, which maps an original message to a codeword $\underline{x}$, which is a point in the N-dimensional hypercube $\{0, 1\}^N$. There are 2^M codewords (with $M < N$), which we assume to be *a priori* equiprobable. When the message is transmitted, the codeword $\underline{x}$ is corrupted to -say- a vector $\underline{y}$ with probability $Q(\underline{y}|\underline{x})$. The decoding

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

maps the received message $\underline{y}$ to one of the possible original codewords $\underline{x}' = d(\underline{y})$. As we saw, a measure of performance is the average block error probability:

$$\mathrm{P}_{\mathrm{B}}^{\mathrm{av}} \equiv \frac{1}{2^M} \sum_{\underline{x}} \sum_{\underline{y}} Q(\underline{y}|\underline{x}) \;\; \mathbb{I}(d(\underline{y}) \neq \underline{x}) \tag{3.4}$$

A simple decoding algorithm would be the following: for each received message $\underline{y}$, consider all the 2^N codewords, and determine the most likely one: $d(\underline{y}) = \arg\max_{\underline{x}} Q(\underline{y}|\underline{x})$. It is clear that this algorithm minimizes the average block error probability.

For a general code, there is no better way for maximizing $Q(\underline{y}|\underline{x})$ than going through all codewords and computing their likelihood one by one. This takes a time of order 2^M, which is definitely too large. Recall in fact that, to achieve reliable communication, M and N have to be large (in data transmission application one may use N as large as 10^5). One may object that 'decoding a general code' is too a general optimization problem. Just for specifying a single instance we would need to specify all the codewords, which takes $N\,2^M$ bits. Therefore, the complexity of decoding could be a trivial consequence of the fact that even reading the input takes a huge time. However, it can be proved that also decoding codes possessing a concise (polynomial in the blocklength) specification is NP-hard. Examples of such codes will be given in the following chapters.

Notes

We have left aside most algorithmic issues in this chapter. In particular many optimization algorithms are based on linear programming. There exist nice theoretical frameworks, and very efficient algorithms, for solving continuous optimization problems in which the cost function, and the constraints, are linear functions of the variables. These tools can be successfully exploited for addressing optimization problems with discrete variables. The idea is to relax the integer constraints. For instance, in the MAX-CUT problem, one should assign a value $x_e \in \{0, 1\}$ to an edge e, saying whether e is in the cut. If c_e is the cost of the edge, one needs to maximize $\sum_e x_e c_e$ over all feasible cuts. A first step consists in relaxing the integer constraints $x_e \in \{0, 1\}$ to $x_e \in [0, 1]$, enlarging the space search. One then solves the continuous problem using linear programming. If the maximum is achieved over integer x_e's, this yields the solution of the original discrete problem. In the opposite case one can add extra constraints in order to reduce again the space search until the a real MAX-CUT will be found. A general introduction to combinatorial optimization, including all these aspects, is provided by (Papadimitriou and Steiglitz, 1998).

A complete treatment of computational complexity theory can be found in (Garey and Johnson, 1979), or in the more recent (Papadimitriou, 1994). The seminal theorem by Cook was independently rediscovered by Levin in 1973. The reader can find its proof in one of the above books.

Euler discussed the Könisberg's 7 bridges problem in (Euler, 1736).

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

The TSP, which is simple to state, difficult to solve, and lends itself to nice pictorial representations, has attracted lots of works. The interested reader can find many references, pictures of TSP's optimal tours with thousands of vertices, including tours among the main cities in various countries, applets, etc.. on the web, starting from instance from (Applegate, Bixby, Chvátal and Cook,).

The book (Hartmann and Rieger, 2002) focuses on the use of optimization algorithms for solving some problems in statistical physics. In particular it explains the determination of the ground state of a random field Ising model with a maximum flow algorithm. A recent volume edited by these same authors (Hartmann and Rieger, 2004) addresses several algorithmic issues connecting optimization and physics; in particular chapter 4 by Liers, Jünger, Reinelt and Rinaldi describes the branch-and-cut approach to the maximum cut problem used for spin glass studies.

An overview classical computational problems from coding theory is the review by Barg (Barg, 1998). Some more recent issues are addressed by Spielman (Spielman, 1997). Finally, the first proof of NP-completeness for a decoding problem was obtained by Berlekamp, McEliecee and van Tilborg (Berlekamp, McEliecee and van Tilborg, 1978).

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

4

PROBABILISTIC TOOLBOX

{ch:Bridges}

The three fields that form the subject of this book, all deal with large sets of random variables. Not surprisingly, they possess common underlying structures and techniques. This Chapter describes some of them, insisting on the mathematical structures, large deviations on one hand, and Markov chains for Monte Carlo computations on the other hand. These tools will reappear several times in the following Chapters.

Since this Chapter is more technical than the previous ones, we devote the entire Section 4.1 to a qualitative introduction to the subject. In Sec. 4.2 we consider the large deviation properties of simple functions of many independent random variables. In this case many explicit results can be easily obtained. We present a few general tools for correlated random variables in Sec. 4.3 and the idea of Gibbs free energy in Sec. 4.4. Section 4.5 provide a simple introduction to the Monte Carlo Markov chain method for sampling configurations from a given probability distribution. Finally, in Sec. 4.6 we show how sinulated annealing exploits Monte Carlo techniques for solving optimization problems.

4.1 Many random variables: a qualitative preview

{sec:Preview}

Consider a set of random variables $\underline{x} = (x_1, x_2, \dots, x_N)$, with $x_i \in \mathcal{X}$ and an N dependent probability distribution

$$P_N(\underline{x}) = P_N(x_1, \dots, x_N)\,. \tag{4.1}$$

This could be for instance the Boltzmann distribution for a physical system with N degrees of freedom. The entropy of this law is $H_N = -\mathbb{E}\log P_N(\underline{x})$. It often happens that this entropy grows linearly with N at large N. This means that the entropy per variable $h_N = H_N/N$ has a finite limit $\lim_{N\to\infty} h_N = h$. It is then natural to characterize any particular realization of the random variables $(x_1, \dots, x_N)$ by computing the quantity

$$f(\underline{x}) = \frac{1}{N}\log\left[\frac{1}{P_N(\underline{x})}\right]\,, \tag{4.2}$$

{eq:Deff}

which measures how probable the event $(x_1, \dots, x_N)$ is.. The expectation of f is $\mathbb{E}f(\underline{x}) = h_N$. One may wonder if $f(\underline{x})$ fluctuates a lot, or if its distribution is strongly peaked around $f = h_N$. The latter hypothesis turns out to be the correct one in many cases: When $N \gg 1$, it often happens that the probability distribution of f, $Q_N(f)$ behaves exponentially:

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

$$Q_N(f) \doteq e^{-NI(f)} \ . \tag{4.3}$$ {eq:larged_ex}

where $I(f)$ has a non-degenerate minimum at $f = h$, and $I(h) = 0$. This means that, with large probability, a randomly chosen configuration $\underline{x}$ has $f(\underline{x})$ 'close to' h, and, because of the definition (4.2) its probability is approximatively $\exp(-Nh)$. Since the total probability of realizations $\underline{x}$ such that $f(\underline{x}) \approx h$ is close to one, their number must behave as $\mathcal{N} \doteq \exp(Nh)$. In other words, the whole probability is carried by a small fraction of all configurations (since their number, $\exp(Nh)$, is in general exponentially smaller than $|\mathcal{X}|^N$), and these configurations all have the same probability. When such a property (often called 'asymptotic equipartition') holds, it has important consequences.

Suppose for instance one is interested in compressing the information contained in the variables $(x_1, \dots, x_N)$, which is a sequence of symbols produced by an information source. Clearly, one should focus on those 'typical' sequences $\underline{x}$ such that $f(\underline{x})$ is close to h, because all the other sequences have vanishing small probability. Since there are $\exp(Nh)$ such typical sequences, one must be able to encode them in $Nh/\log 2$ bits by simply numbering them.

Another very general problem consists in sampling from the probability distribution $P_N(\underline{x})$. With r realizations $\underline{x}^1, \dots, \underline{x}^r$ drawn independently from $P_N(\underline{x})$, one can estimate an expectation values $\mathbb{E}\,\mathcal{O}(\underline{x}) = \sum_{\underline{x}} P_N(\underline{x})\mathcal{O}(\underline{x})$ as $\mathbb{E}\,\mathcal{O}(\underline{x}) \approx \frac{1}{r}\sum_{k=1}^{r} \mathcal{O}(\underline{x}^k)$ without summing over $|\mathcal{X}|^N$ terms, and the precision usually improves like $1/\sqrt{r}$ at large r. A naive sampling algorithm could be the following. First 'propose' a configuration $\underline{x}$ from the uniform probability distribution $P_N^{\rm unif}(\underline{x}) = 1/|\mathcal{X}|^N$: this is simple to be sampled[8]. Then 'accept' the configuration with probability $P_N(\underline{x})$. Such an algorithm is totally unefficient: It is clear that, for the expectation values of 'well behaved' observables, we seek configurations $\underline{x}$ such that $f(\underline{x})$ is close to h. However, such configurations are exponentially rare, and the above algorithm will require a time of order $\exp[N(\log|\mathcal{X}| - h)]$ to find just one of them. The Monte Carlo method will provide a better alternative.

{sec:LargedevIID}

4.2 Large deviations for independent variables

A behavior of the type (4.3) is an example of a large deviation principle. One often encounters systems with this property, and it can also hold with more general functions $f(\underline{x})$. The simplest case where such behaviors are found, and the case where all properties can be controlled in great details, is that of independent random variables. We study this case in the present section.

4.2.1 *How typical is a series of observations?*

Suppose that you are given given the values $s_1, \dots, s_N$ of N i.i.d. random variables drawn from a finite space $\mathcal{X}$ according to a known probability distribution

[8] Here we are assuming that we have access to a source of randomness: $\lceil N \log_2 |\mathcal{X}| \rceil$ unbiased random bits are sufficient to sample from $P_N^{\rm unif}(\underline{x})$. In practice one replaces the source of randomness by a pseudorandom generator.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

$\{p(s)\}_{s\in\mathcal{X}}$. The s_i's could be produced for instance by an information source, or by some repeated measurements on a physical system. You would like to know if the sequence $\underline{s} = (s_1, \ldots, s_N)$ is a typical one, or if you found a rare event. If N is large, one can expect that the number of appearances of a given $x \in \mathcal{X}$ in a typical sequence should be close to $Np(x)$. The method of types allows to quantify this statement.

The **type** $q_{\underline{s}}(x)$ of the sequence $\underline{s}$ is the frequency of appearance of symbol x in the sequence:

$$q_{\underline{s}}(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x,s_i}\,, \tag{4.4}$$

where δ is the **Kronecker** symbol, such that $\delta_{x,y} = 1$ if $x = y$ and 0 otherwise. For any observation $\underline{s}$, the type $q_{\underline{s}}(x)$, considered as a function of x, has the properties of a probability distribution over $\mathcal{X}$: $q(x) \geq 0$ for any $x \in \mathcal{X}$ and $\sum_x q(x) = 1$. In the following we shall denote by $\mathfrak{M}(\mathcal{X})$ the space of probability distributions over $\mathcal{X}$: $\mathfrak{M}(\mathcal{X}) \equiv \{q \in \mathbb{R}^{\mathcal{X}} \text{ s.t. } q(x) \geq 0\,,\ \sum_x q(x) = 1\}$. Therefore $q_{\underline{s}} \in \mathfrak{M}(\mathcal{X})$.

The expectation of the type $q_{\underline{s}}(x)$ coincides with the original probability distribution:

$$\mathbb{E}\, q_{\underline{s}}(x) = p(x)\,. \tag{4.5}$$

Sanov's theorem estimates the probability that the type of the sequence differs from $p(x)$.

{thm:Sanov}

Theorem 4.1. (Sanov) *Let $x_1, \ldots, x_N \in \mathcal{X}$ be N i.i.d.'s random variables drawn from the probability distribution $p(x)$, and $K \subset \mathfrak{M}(\mathcal{X})$ a compact set of probability distributions over $\mathcal{X}$. If q is the type of $(x_1, \ldots, x_N)$, then*

$$\text{Prob}\,[q \in K] \doteq \exp[-ND(q^*||p)]\,, \tag{4.6}$$

where $q_* = \arg\min_{q\in K} D(q||p)$, and $D(q||p)$ is the KL divergence defined in Eq. (1.10) .

Basically this theorem means that the probability of finding a sequence with type q behaves at large N like $\exp[-ND(q||p)]$. Therefore, for large N, typical sequences have a type $q(x) = p(x)$, and those with a different type are exponentially rare. The proof of the theorem is a straightforward application of Stirling's formula and is left as an exercise for the reader. In Appendix 4.7 we give a ⋆ derivation using a 'field theoretical' method as used in physics. It may be an instructive simple example for the reader who wants to get used to these kinds of techniques, frequently used by physicists.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Example 4.2 Let the x_i's be the outcome of a biased coin: $\mathcal{X} = \{\texttt{head}, \texttt{tail}\}$, with $p(\texttt{head}) = 1 - p(\texttt{tail}) = 0.8$. What is the probability of getting 50 heads and 50 tails in 100 throws of the coin? Using the expression (4.6) and (1.10) with $N = 100$ and $q(\texttt{head}) = q(\texttt{tail}) = 0.5$, we get Prob[50 `tails`] $\approx 2.04 \cdot 10^{-10}$.

Example 4.3 Let us consider the reverse case: we take a fair coin ($p(\texttt{head}) = p(\texttt{tail}) = 0.5$) and ask what is the probability of getting 80 heads and 20 tails. Sanov theorem provides the estimate Prob[80 `heads`] $\approx 4.27 \cdot 10^{-9}$, which is much higher than the one computed in the previous example.

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

Example 4.4 A simple model of a column of the atmosphere consists in studying N particles in the earth gravitational field. The state of particle $i \in \{1, \dots, N\}$ is given by a single coordinate $z_i \geq 0$ which measures its height with respect to earth level. For the sake of simplicity, we assume z_i's to be integer numbers. We can, for instance, imagine to discretize the heights in terms of some small unit length (e.g. millimeters). The N-particles energy function reads, in properly chosen units:

$$E = \sum_{i=1}^{N} z_i \,. \tag{4.7}$$

The type of a configuration $\{x_1, \dots, x_N\}$ can be interpreted as the density profile $\rho(z)$of the configuration:

$$\rho(z) = \frac{1}{N} \sum_{i=1}^{N} \delta_{z,z_i} \,. \tag{4.8}$$

Using the Boltzmann probability distribution (2.4), it is simple to compute the expected density profile, which is usually called the 'equilibrium' profile:

$$\rho_{\rm eq}(z) \equiv \langle \rho(z) \rangle = (1 - e^{-\beta})\, e^{-\beta z} \,. \tag{4.9}$$

If we take a snapshot of the N particles at a given instant, their density will present some deviations with respect to $\rho_{\rm eq}(z)$. The probability of seeing a density profile $\rho(z)$ is given by Eq. (4.6) with $p(z) = \rho_{\rm eq}(z)$ and $q(z) = \rho(z)$. For instance, we can compute the probability of observing an exponential density profile, like (4.9) with a different parameter λ: $\rho_\lambda(x) = (1 - e^{-\lambda})\, e^{-\lambda x}$. Using Eq. (1.10) we get:

$$D(\rho_\lambda || \rho_{\rm eq}) = \log \left(\frac{1 - e^{-\lambda}}{1 - e^{-\beta}} \right) + \frac{\beta - \lambda}{e^\lambda - 1} \,. \tag{4.10}$$

The function $I_\beta(\lambda) \equiv D(\rho_\lambda || \rho_{\rm eq})$ is depicted in Fig. 4.1.

Exercise 4.1 The previous example is easily generalized to the density profile of N particles in an arbitrary potential $V(x)$. Show that the Kullback-Leibler divergence takes the form

$$D(\rho || \rho_{\rm eq}) = \beta \sum_x V(x) \rho(x) - \sum_x \rho(x) \log \rho(x) + \log z(\beta) \,. \tag{4.11}$$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

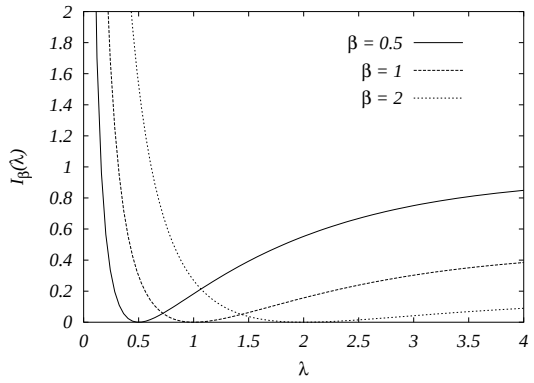

FIG. 4.1. Example 3: In an atmosphere where the equilibrium density profile is $\rho_{\rm eq}(z) \propto e^{-\beta z}$, the probability of observing an atypical profile $\rho(z) \propto e^{-\lambda z}$ is, for a large number of particles N, $\exp[-NI_\beta(\lambda)]$. The curves $I_\beta(\lambda)$, plotted here, show that small values of λ are very rare. {fig:profilefluc}

4.2.2 *How typical is an empirical average?*

The result (4.6) contains a detailed information concerning the large fluctuations of the random variables $\{x_i\}$. Often one is interested in monitoring the fluctuations of the empirical average of a measurement, which is a real number $f(x)$:

$$\overline{f} \equiv \frac{1}{N}\sum_{i=1}^{N} f(x_i)\,. \tag{4.12}$$

Of course $\overline{f}$, will be "close" to $\mathbb{E}\, f(x)$ with high probability. The following result quantifies the probability of rare fluctuations. {thm:EmpiricalAverage}

Corollary 4.5 *Let $x_1, \ldots, x_N$ be N i.i.d.'s random variables drawn from the probability distribution $p(x)$. Let $f : \mathcal{X} \to \mathbb{R}$ be a real valued function and $\overline{f}$ be its empirical average. If $A \subset \mathbb{R}$ is a closed interval of the real axis*

$$\text{Prob}\,[\overline{f} \in A] \doteq \exp[-NI(A)]\,, \tag{4.13}$$

where

$$I(A) = \min_q \left[D(q||p) \;\Big|\; \sum_{x\in\mathcal{X}} q(x) f(x) \in A \right]\,. \tag{4.14}$$

Proof: We apply Theorem 4.1 with the compact set

$$K = \{q \in \mathfrak{M}(\mathcal{X}) \,|\, \sum_{x\in\mathcal{X}} q(x) f(x) \in A\}\,. \tag{4.15}$$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

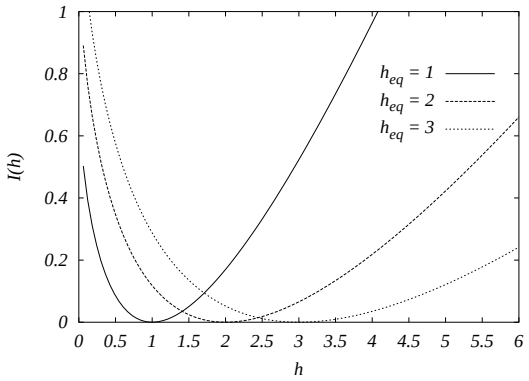

FIG. 4.2. Probability of an atypical average height for N particles with energy function (4.7). {fig:heightfluc}

This implies straightforwardly Eq. (4.13) with

$$I(\varphi) = \min \left[D(q||p) \left| \sum_{x \in \mathcal{X}} q(x) f(x) = \varphi \right. \right] . \tag{4.16}$$

The minimum in the above equation can be found by Lagrange multipliers method, yielding Eq. (4.14). □

Example 4.6 We look again at N particles in a gravitational field, as in Example 3, and consider the average height of the particles:

$$\overline{z} = \frac{1}{N} \sum_{i=1}^{N} z_i \, . \tag{4.17}$$

The expected value of this quantity is $\mathbb{E}(\overline{z}) = z_{\rm eq} = (e^{\beta} - 1)^{-1}$. The probability of a fluctuation of $\overline{z}$ is easily computed using the above Corollary. For $z > z_{\rm eq}$, one gets $P[\overline{z} > z] \doteq \exp[-N\, I(z)]$, with

$$I(z) = (1 + z) \, \log \left(\frac{1 + z_{\rm eq}}{1 + z} \right) + z \, \log \left(\frac{z}{z_{\rm eq}} \right) . \tag{4.18}$$

Analogously, for $z < z_{\rm eq}$, $P[\overline{z} < z] \doteq \exp[-N\, I(z)]$, with the same rate function $I(z)$. The function $I(z)$ is depicted in Fig. 4.2.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Exercise 4.2 One can construct a thermometer using the system of N particles with the energy function (4.7). Whenever the temperature is required, you take a snapshot of the N particles, compute $\overline{x}$ and estimate the inverse temperature βest using the formula $(e^{\beta_{\rm est}}-1)^{-1}=\overline{x}$. What is (for $N\gg 1$) the probability of getting a result $\beta_{\rm est}\neq\beta$?

{subsec:AEQ}

4.2.3 *Asymptotic equipartition*

The above tools can also be used for *counting* the number of configurations $\underline{s}=(s_1,\dots,s_N)$ with either a given type $q(x)$ or a given empirical average of some observable $\overline{f}$. One finds for instance:

{prop:counting}

Proposition 4.7 *The number $\mathcal{N}_{K,N}$ of sequences $\underline{s}$ which have a type belonging to the compact $K\subset\mathfrak{M}(\mathcal{X})$ behaves as $\mathcal{N}_{K,N}\doteq\exp\{NH(q_*)\}$, where $q_*=\arg\max\{H(q)\,|\,q\in K\}$.*

This result can be stated informally by saying that "there are approximately $e^{NH(q)}$ sequences with type q".

Proof: The idea is to apply Sanov's theorem, taking the "reference" distribution $p(x)$ to be the flat probability distribution $p_{\rm flat}(x)=1/|\mathcal{X}|$. Using Eq. (4.6), we get

$$\mathcal{N}_{K,N}=|\mathcal{X}|^N{\rm Prob}_{\rm flat}[q\in K]\doteq\exp\{N\log|\mathcal{X}|-ND(q_*||p_{\rm flat})\}=\exp\{NH(q_*)\}\;. \tag{4.19}$$

□

We now get back to a generic sequence $\underline{s}=(s_1,\dots,s_N)$ of$N$ iid variables with a probability distribution $p(x)$. As a consequence of Sanov's theorem, we know that the most probable type is $p(x)$ itself, and that deviations are exponentially rare in N. We expect that almost all the probability is concentrated on sequences having a type in some sense close to $p(x)$. On the other hand, because of the above proposition, the number of such sequences is exponentially smaller than the total number of possible sequences $|\mathcal{X}|^N$.

These remarks can be made more precise by defining what is meant by a sequence having a type 'close to $p(x)$'. Given the sequence $\underline{s}$, we introduce the quantity

$$r(\underline{s})\equiv-\frac{1}{N}\log P_N(\underline{s})=-\frac{1}{N}\sum_{i=1}^N\log p(x_i)\;. \tag{4.20}$$

Clearly, $\mathbb{E}\,r(\underline{s})=H(p)$. The sequence $\underline{s}$ is said to be **ε-typical** if and only if $|r(\underline{s})-H(p)|\leq\varepsilon$. Let $T_{N,\varepsilon}$ be the set of ε-typical sequences. It has the following properties:

Theorem 4.8 *(i)* $\lim_{N\to\infty}{\rm Prob}[\underline{s}\in T_{N,\varepsilon}]=1$.
(ii) For N large enough, $e^{N[H(p)-\varepsilon]}\leq|T_{N,\varepsilon}|\leq e^{N[H(p)+\varepsilon]}$.
(iii) For any $\underline{s}\in T_{N,\varepsilon}$, $e^{-N[H(p)+\varepsilon]}\leq P_N(\underline{s})\leq e^{-N[H(p)-\varepsilon]}$.

Proof: Since $r(\ \mathbf{s})$ is an empirical average, we can apply Corollary 4.5. This allows to estimate the probability of *not* being typical as ${\rm Prob}[\,\underline{s}\notin T_{N,\varepsilon}]\doteq\exp(-NI)$.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

The exponent is given by $I = \min_q D(q||p)$, the minimum being taken over all probability distributions $q(x)$ such that $\left| \sum_{x\in\mathcal{X}} q(x) \log[1/q(x)] - H(p) \right| \geq \varepsilon$. But $D(q||p) > 0$ unless $q = p$, and p does not belong to the of minimization. Therefore $I > 0$ and $\lim_{N\to\infty} \text{Prob}[\,\underline{s} \notin T_{N,\varepsilon}] = 0$, which proves (i).

The condition for $q(x)$ to be the type of a ε-typical sequence can be rewritten as $|D(q||p) + H(q) - H(p)| \leq \varepsilon$. Therefore, for any ε-typical sequence, $|H(q) - H(p)| \leq \varepsilon$ and Proposition 4.7 leads to (ii). Finally, (iii) is a direct consequence of the definition of ε-typical sequences. □

The behavior described in this proposition is usually denoted as **asymptotic equipartition property**. Although we proved it for i.i.d. random variables, this is not the only context in which it is expected to hold. In fact it will be found in many interesting systems throughout the book.

4.3 Correlated variables

{sec:CorrelatedVariables}

In the case of independent random variables on finite spaces, the probability of a large fluctuation is easily computed by combinatorics. It would be nice to have some general result for large deviations of non-independent random variables. In this Section we want to describe the use of Legendre transforms and saddle point methods to study the general case. As it often happens, this method corresponds to a precise mathematical statement: the Gärtner-Ellis theorem. We first describe the approach informally and apply it to a few of examples. Then we will state the theorem and discuss it.

4.3.1 *Legendre transformation*

To be concrete, we consider a set of random variables $\underline{x} = (x_1, \ldots, x_N)$, with $x_i \in \mathcal{X}$ and an N dependent probability distribution

$$P_N(\underline{x}) = P_N(x_1, \ldots, x_N)\,. \tag{4.21}$$

Let $f : \mathcal{X} \to \mathbb{R}$ be a real valued function. We are interested in estimating, at large N, the probability distribution of its empirical average

$$\overline{f}(\underline{x}) = \frac{1}{N} \sum_{i=1}^{N} f(x_i)\,. \tag{4.22}$$

In the previous Section, we studied the particular case in which the x_i's are i.i.d. random variables. We proved that, quite generally, a finite fluctuation of $\overline{f}(\underline{x})$ is exponentially unlikely. It is natural to expect that the same statement holds true if the x_i's are "weakly correlated". Whenever $P_N(\underline{x})$ is the Gibbs-Boltzmann distribution for some physical system, this expectation is supported by physical intuition. We can think of the x_i's as the microscopic degrees of freedom composing the system and of $\overline{f}(\underline{x})$ as a macroscopic observable (pressure, magnetization, etc.). It is a common observation that the relative fluctuations of macroscopic observables are very small.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Let us thus assume that the distribution of $\overline{f}$ follows a **large deviation principle**, meaning that the asymptotic behavior of the distribution at large N is:

$$P_N(\overline{f}) \doteq \exp[-NI(\overline{f})]\,, \tag{4.23}$$

with a **rate function** $I(\overline{f}) \geq 0$.

In order to determine $I(\overline{f})$, a useful method consists in "tilting" the measure $P_N(\cdot)$ in such a way that the rare events responsible for $O(1)$ fluctuations of $\overline{f}$ become likely. In practice we define the **(logarithmic) moment generating function** of $\overline{f}$ as follows

$$\psi_N(t) = \frac{1}{N} \log\left(\mathbb{E}\, e^{Nt\overline{f}(\underline{x})}\right) \quad , \quad t \in \mathbb{R}\,. \tag{4.24}$$

When the property (4.23) holds, we can evaluate the large N limit of $\psi_N(t)$ using the saddle point method:

$$\lim_{N\to\infty} \psi_N(t) = \lim_{N\to\infty} \frac{1}{N} \log\left\{\int e^{Nt\overline{f}-NI(\overline{f})} d\overline{f}\right\} = \psi(t)\,, \tag{4.25}$$

with

$$\psi(t) = \sup_{\overline{f}\in\mathbb{R}} \left[t\overline{f} - I(\overline{f})\right]\,. \tag{4.26}$$

$\psi(t)$ is the Legendre transform of $I(\overline{f})$, and it is a convex function of t by construction (this is proved by differentiating twice Eq. (4.24)). It is therefore natural to invert the Legendre transform (4.26) as follows:

$$I_\psi(\overline{f}) = \sup_{t\in\mathbb{R}} \left[t\overline{f} - \psi(t)\right]\,, \tag{4.27}$$

and we expect $I_\psi(\overline{f})$ to coincide with the convex envelope of $I(\overline{f})$. This procedure is useful whenever computing $\psi(t)$ is easier than directly estimate the probability distribution $P_N(\overline{f})$.

4.3.2 *Examples*

It is useful to gain some insight by considering a few examples.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Example 4.9 Consider the one-dimensional Ising model, without external magnetic field, cf. Sec. 2.5.1. To be precise we have $x_i = \sigma_i \in \{+1, -1\}$, and $P_N(\underline{\sigma}) = \exp[-\beta E(\underline{\sigma})]/Z$ the Boltzmann distribution with energy function

$$E(\underline{\sigma}) = -\sum_{i=1}^{N-1} \sigma_i \sigma_{i+1} \,. \tag{4.28}$$

We want to compute the large deviation properties of the magnetization

$$m(\underline{\sigma}) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i \,. \tag{4.29}$$

We know from Sec. 2.5.1, and from the symmetry of the energy function under spin reversal ($\sigma_i \to -\sigma_i$) that $\langle m(\underline{\sigma}) \rangle = 0$. In order to compute the probability of a large fluctuation of m, we apply the method described above. A little thought shows that $\psi(t) = \phi(\beta, t/\beta) - \phi(\beta, 0)$ where $\phi(\beta, B)$ is the free energy density of the model in an external magnetic field B, found in (2.63). We thus get

$$\psi(t) = \log \left(\frac{\cosh t + \sqrt{\sinh^2 t + e^{-4\beta}}}{1 + e^{-2\beta}} \right) . \tag{4.30}$$

One sees that $\psi(t)$ is convex and analytic for any $\beta < \infty$. We can apply Eq. (4.27) in order to obtain the rate function $I_\psi(m)$. In Fig. 4.3 we report the resulting function for several temperatures β. Notice that $I_\psi(m)$ is analytic and has strictly positive second derivative for any m and $\beta < \infty$, so that we expect $I(m) = I_\psi(m)$. This expectation is confirmed by Theorem 4.12 below.

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

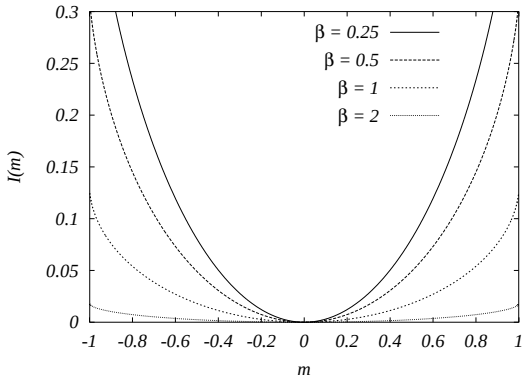

FIG. 4.3. Rate function for the magnetization of the one-dimensional Ising model. Notice that, as the temperature is lowered (β increased) the probability of large fluctuations increases.

{fig:largedev1dIsing}

Example 4.10 Consider a Markov chain $X_0, X_1, \ldots, X_i, \ldots$ taking values in a finite state space $\mathcal{X}$, as in the Example 2 of Sec. 1.3, and assume all the elements of the transition matrix $w(x \to y)$ to be strictly positive. Let us study the large deviation properties of the empirical average $\frac{1}{N}\sum_i f(X_i)$.

One can show that the limit moment generating function $\psi(t)$, cf. Eq. (4.24) exists, and can be computed using the following recipe. Define the 'tilted' transition probabilities as $w_t(x \to y) = w(x \to y)\exp[t\, f(y)]$. Let $\lambda(t)$ be the largest solution of the eigenvalue problem

$$\sum_{x\in\mathcal{X}} \phi^l_t(x)\; w_t(x \to y) = \lambda(t)\; \phi^l_t(y)\,. \tag{4.31}$$

The moment generating function is simply given by $\psi(t) = \log\lambda(t)$ (which is unique and positive because of Perron-Frobenius theorem).

Notice that Eq. (4.31) resembles the stationarity condition for a Markov chain with transition probabilities $w_t(x \to y)$. Unhappily the rates $w_t(x \to y)$ are not properly normalized ($\sum_y w_t(x \to y) \neq 1$). This point can be overcome as follows. Call $\phi^r_t(x)$ the right eigenvector of $w_t(x \to y)$ with eigenvalue $\lambda(t)$ and define:

$$\overline{w}_t(x \to y) \equiv \frac{1}{\lambda(t)\phi^r_t(x)}\; w_t(x \to y)\, \phi^r_t(y)\,. \tag{4.32}$$

We leave to the reader the exercise of showing that: (i) These rates are properly normalized; (ii) Eq. (4.31) is indeed the stationarity condition for the distribution $p_t(x) \propto \phi^l_t(x)\phi^r_t(x)$ with respect to the rates $\overline{w}_t(x \to y)$.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Example 4.11 Consider now the Curie-Weiss model without external field, cf. Sec. 2.5.2. As in Example 1, we take $x_i = \sigma_i \in \{+1, -1\}$ and $P_N(\underline{\sigma}) = \exp[-\beta E(\underline{\sigma})]/Z$, and we are interested in the large fluctuations of the global magnetization (4.29). The energy function is

$$E(\underline{\sigma}) = -\frac{1}{N} \sum_{(ij)} \sigma_i \sigma_j \,. \tag{4.33}$$

By repeating the arguments of Sec. 2.5.2, it is easy to show that, for any $-1 \le m_1 < m_2 \le 1$:

$$P_N\{m(\underline{\sigma}) \in [m_1, m_2]\} \doteq \frac{1}{Z_N(\beta)} \int_{m_1}^{m_2} dm \;\; e^{N\phi_{\rm mf}(m;\beta)} \,, \tag{4.34}$$

where $\phi_{\rm mf}(m;\beta) = \frac{\beta}{2} m^2 - \log[2\cosh(\beta m)]$. The large deviation property (4.23) holds, with:

$$I(m) = \phi_{\rm mf}(m^*;\beta) - \phi_{\rm mf}(m;\beta) \,. \tag{4.35}$$

and $m^*(\beta)$ is the largest solution of the Curie Weiss equation $m = \tanh(\beta m)$. The function $I(m)$ is represented in Fig. 4.4, left frame, for several values of the inverse temperature β. For $\beta < \beta_{\rm c} = 1$, $I(m)$ is convex and has its unique minimum in $m = 0$.

A new and interesting situation appears when $\beta > \beta_{\rm c}$. The function $I(m)$ is non convex, with two degenerate minima at $m = \pm m^*(\beta)$. In words, the system can be found in either of two well-distinguished 'states': the positive and negative magnetization states. There is no longer a *unique* typical value of the magnetization such that large fluctuations away from this value are exponentially rare.

Let us now look at what happens if the generating function approach is adopted. It is easy to realize that the limit (4.24) exists and is given by

$$\psi(t) = \sup_{m \in [-1,1]} [mt - I(m)] \,. \tag{4.36}$$

While at high temperature $\beta < 1$, $\psi(t)$ is convex and analytic, for $\beta > 1$ it develops a singularity at $t = 0$. In particular one has $\psi'(0+) = m^*(\beta) = -\psi'(0-)$. Compute now $I_\psi(m)$ using Eq. (4.27). A little thought shows that, for any $m \in [-m^*(\beta), m^*(\beta)]$ the supremum is achieved for $t = 0$, which yields $I_\psi(m) = 0$. Outside this interval, the supremum is achieved at the unique solution of $\psi'(t) = m$, and $I_\psi(m)$. As anticipated, $I_\psi(m)$ is the convex envelope of $I(m)$. In the range $(-m^*(\beta), m^*(\beta))$, an estimate of the magnetization fluctuations through the function $\doteq \exp(-N I_\psi(m))$ would *overestimate* the fluctuations.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

4.3.3 *The Gärtner-Ellis theorem*

The Gärtner-Ellis theorem has several formulations which usually require some technical definitions beforehand. Here we shall state it in a simplified (and somewhat weakened) form. We need only the definition of an **exposed point**: $x \in \mathbb{R}$ is an exposed point of the function $F : \mathbb{R} \to \mathbb{R}$ if there exists $t \in \mathbb{R}$ such that $ty - F(y) > tx - F(x)$ for any $y \neq x$. If, for instance, F is convex, a sufficient condition for x to be an exposed point is that F is twice differentiable at x with $F''(x) > 0$.

{thm:GE}

Theorem 4.12. (Gärtner-Ellis) *Consider a function $\overline{f}(\underline{x})$ (not necessarily of the form (4.22)) and assume that the moment generating function $\psi_N(t)$ defined in (4.24) exists and has a finite limit $\psi(t) = \lim_{N\to\infty} \psi_N(t)$ for any $t \in \mathbb{R}$. Define $I_\psi(\cdot)$ as the inverse Legendre transform of Eq. (4.27) and let $\mathcal{E}$ be the set of exposed points of $I_\psi(\cdot)$.*

1. *For any closed set $F \in \mathbb{R}$:*

$$\limsup_{N\to\infty} \frac{1}{N} \log P_N(\overline{f} \in F) \leq - \inf_{f\in F} I_\psi(f)\,. \tag{4.37}$$

2. *For any open set $G \in \mathbb{R}$:*

$$\limsup_{N\to\infty} \frac{1}{N} \log P_N(\overline{f} \in G) \geq - \inf_{f\in G\cap\mathcal{E}} I_\psi(f)\,. \tag{4.38}$$

3. *If moreover $\psi(t)$ is differentiable for any $t \in \mathbb{R}$, then the last statement holds true with the* inf *being taken over the whole set G (rather than over $G \cap \mathcal{E}$).*

Informally, the inverse Legendre transform (4.27) generically yields an upper bound on the probability of a large fluctuation of the macroscopic observable. This upper bound is tight unless a 'first order phase transition' occurs, corresponding to a discontinuity in the first derivative of $\psi(t)$.

It is worth mentioning that $\psi(t)$ can be non-analytic at a point t_* while its first derivative is continuous at t_*. This correspondsm in the statistical mechanics jargon, to a 'higher order' phase transition. As we shall see in the following Chapters, such phenomena have interesting probabilistic interpretations too.

4.3.4 *Typical sequences*

Let us get back to the concept of typical sequences, introduced in Section 4.2. More precisely, we want to investigate the large deviation of the probability itself, measured by $r(\underline{x}) = -\frac{1}{N} \log P(\underline{x})$. For independent random variables, the study of sect. 4.2.3 led to the concept of ε-typical sequences. What can one say about general sequences?

Let us compute the corresponding moment generating function (4.24):

$$\psi_N(t) = \frac{1}{N}\, \log \left\{ \sum_{\underline{x}} P_N(\underline{x})^{1-t} \right\}\,. \tag{4.39}$$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Without loss of generality, we can assume $P_N(\underline{x})$ to have the Boltzmann form:

$$P_N(\underline{x}) = \frac{1}{Z_N(\beta)} \exp\{-\beta E_N(\underline{x})\}\,, \tag{4.40}$$

with energy function $E_N(\underline{x})$. Inserting this into Eq. (4.39), we get

$$\psi_N(t) = \beta f_N(\beta) - \beta f_N(\beta(1-t))\,, \tag{4.41}$$

where $f_N(\beta) = -(1/N)\log Z_N(\beta)$ is the free energy density of the system with energy function $E_N(\underline{x})$ at inverse temperature β. Let us assume that the thermodynamic limit $f(\beta) = \lim_{N\to\infty} f_N(\beta)$ exists and is finite. It follows that the limiting generating function $\psi(t)$ exists and we can apply the Gärtner-Ellis theorem to compute the probability of a large fluctuation of $r(\underline{x})$. As long as $f(\beta)$ is analytic, large fluctuations are exponentially depressed and the asymptotic equipartition property of independent random variables is essentially recovered. On the other hand, if there is a phase transition at $\beta = \beta_c$, where the first derivative of $f(\beta)$ is discontinuous, then the likelihood $r(\underline{x})$ may take several distinct values with a non-vanishing probability. This is what happened with the magnetization in Example 3 above.

4.4 Gibbs free energy

{sec:Gibbs}

In the introduction to statistical physics of chapter 2, we assumed that the probability distribution of the configurations of a physical system is Boltzmann's distribution. It turns out that this distribution can be obtained from a variational principle. This is interesting, both as a matter of principle and in order to find approximation schemes.

Consider a system with a configuration space $\mathcal{X}$, and a real valued energy function $E(x)$ defined on this space. The Boltzmann distribution is $P_\beta(x) = \exp[-\beta(E(x) - F(\beta))]$, where $F(\beta)$, the 'free energy', is a function of the inverse temperature β defined by the fact that $\sum_{x\in\mathcal{X}} P_\beta(x) = 1$. Let us define the **Gibbs free energy** $G[P]$ (not to be confused with $F(\beta)$), which is a real valued functional over the space of probability distributions $P(x)$ on $\mathcal{X}$:

$$G[P] = \sum_{x\in\mathcal{X}} P(x)E(x) + \frac{1}{\beta}\sum_{x\in\mathcal{X}} P(x)\log P(x)\,. \tag{4.42}$$

It is easy to rewrite the Gibbs free energy in terms of the KL divergence between $P(x)$ and the Boltzmann distribution $P_\beta(x)$:

$$G[P] = \frac{1}{\beta} D(P||P_\beta) + F(\beta)\,, \tag{4.43}$$

This representation implies straightforwardly the following proposition (**Gibbs variational principle**):

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Proposition 4.13 *The Gibbs free energy $G[P]$ is a convex functional of $P(x)$, and it achieves its unique minimum on the Boltzmann distribution $P(x) = P_\beta(x)$. Moreover $G[P_\beta] = F(\beta)$, where $F(\beta)$ is the free energy.*

When the partition function of a system cannot be computed exactly, the above result suggests a general line of approach for estimating the free energy: one can minimize the Gibbs free energy in some restricted subspace of "trial probability distributions" $P(x)$. These trial distributions should be simple enough that $G[P]$ can be computed, but the restricted subspace should also contain distributions which are able to give a good approximation to the true behavior of the physical system. For each new physical system one will thus need to find a good restricted subspace.

Example 4.14 Consider a system with space of configurations $\mathcal{X} = \mathbb{R}$ and energy:

$$E(x) = \frac{1}{2} t\, x^2 + \frac{1}{4} x^4 \,, \tag{4.44}$$

with $t \in \mathbb{R}$. We ask the question of computing its free energy at temperature $\beta = 1$ as a function of t. With a slight abuse of notation, we are interested in

$$F(t) = -\log \left(\int dx\; e^{-E(x)} \right) . \tag{4.45}$$

The above integral cannot be computed in closed form and so we recur to the Gibbs variational principle. We consider the following family of trial probability distributions:

$$Q_a(x) = \frac{1}{\sqrt{2\pi a}}\, e^{-x^2/2a} \,. \tag{4.46}$$

It is easy to compute the corresponding Gibbs free energy for $\beta = 1$:

$$G[Q_a] = \frac{1}{2}\, ta + \frac{3}{4} a^2 - \frac{1}{2}\, (1 + \log 2\pi a) \equiv G(a,t) \,. \tag{4.47}$$

The Gibbs principle implies that $F(t) \leq \min_a G(a,t)$. In Fig. 4.5 we plot the optimal value of a, $a_{\rm opt}(t) = \arg\min_a G(a,t)$ and the corresponding estimate $G_{\rm opt}(t) = G(a_{\rm opt}(t), t)$.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

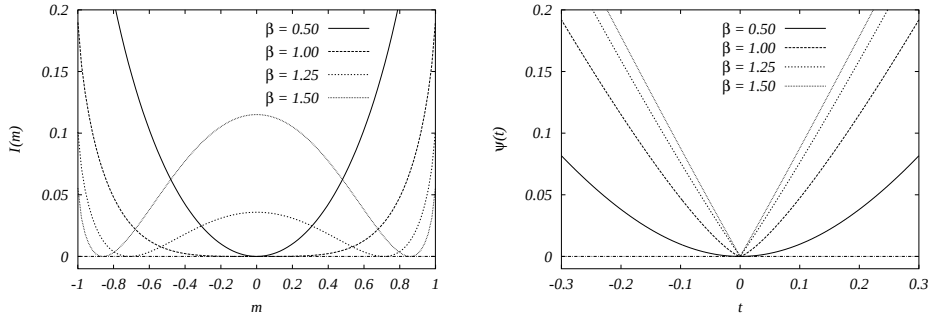

FIG. 4.4. The rate function for large fluctuations of the magnetization in the Curie-Weiss model (left) and the corresponding generating function (right).

{fig:largedevCW}

Example 4.15 Consider the same problem as above and the family of trials distributions:

$$Q_a(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-a)^2/2} \,. \tag{4.48}$$

We leave as an exercise for the reader the determination of the optimal value of $a_{\rm opt}$, and the corresponding upper bound on $F(t)$, cf. Fig. 4.5. Notice the peculiar phenomenon going on at $t_{\rm cr} = -3$. For $t > t_{\rm cr}$, we have $a_{\rm opt}(t) = 0$, while $G[Q_a]$ has two degenerate local minima $a = \pm a_{\rm opt}(t)$ for $t \leq t_{\rm cr}$.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

Example 4.16 Consider the Ising model on a d-dimensional lattice $\mathbb{L}$ of linear size L (i.e. $\mathbb{L} = [L]^d$), cf. Sec. 2.5. The energy function is (notice the change of normalization with respect to Sec. 2.5)

$$E(\underline{\sigma}) = -\sum_{(ij)} \sigma_i \sigma_j - B \sum_{i \in \mathbb{L}} \sigma_i \,. \tag{4.49}$$

For the sake of simplicity we assume **periodic boundary conditions**. This means that two sites $i = (i_1, \dots, i_d)$ and $j = (j_1, \dots, j_d)$ are considered nearest neighbors if, for some $l \in \{1, \dots, d\}$, $i_l - j_l = \pm 1 \pmod{L}$ and $i_{l'} = j_{l'}$ for any $l' \neq l$. The sum over (ij) in Eq. (4.49) runs over all nearest neighbors pairs in $\mathbb{L}$.

In order to obtain a variational estimate of the free energy $F(\beta)$ at inverse temperature β, we evaluate the Gibbs free energy on the following trial distribution:

$$Q_m(\underline{\sigma}) = \prod_{i \in \mathbb{L}} q_m(\sigma_i) \,, \tag{4.50}$$

with $q_m(+) = (1+m)/2$ and $q_m(-) = (1-m)/2$ and $m \in [-1, +1]$. Notice that, under $Q_m(\underline{\sigma})$, the σ_i's are i.i.d. random variables with expectation m.

It is easy to evaluate the Gibbs free energy on this distribution. If we define the per-site Gibbs free energy $g(m; \beta, B) \equiv G[Q_m]/L^d$, we get

$$g(m; \beta, B) = -\frac{1}{2}\, m^2 - B\, m + \frac{1}{\beta}\, \mathcal{H}((1+m)/2) \,. \tag{4.51}$$

Gibbs variational principle implies an upper bound on the free energy density $f(\beta) \leq \inf_m g(m; \beta, h)$. Notice that, apart from an additive constant, this expression (4.51) has the same form as the solution of the Curie-Weiss model, cf. Eq. (2.79). We refer therefore to Sec. 2.5.2 for a discussion of the optimization over m. This implies the following inequality:

$$f_d(\beta, h) \leq f_{\rm CW}(\beta, h) - \frac{1}{2} \,. \tag{4.52}$$

The relation between Gibbs free energy and Kullback-Leibler divergence in Eq. (4.43) implies a simple probabilistic interpretation of Gibbs variational principle. Imagine to prepare a large number $\mathcal{N}$ of copies of the same physical system. Each copy is described by the same energy function $E(\underline{x})$. Now consider the empirical distribution $P(\underline{x})$ of the $\mathcal{N}$ copies. Typically $P(\underline{x})$ will be close to the Bolzmann distribution $P_\beta(\underline{x})$. Sanov's theorem implies that the probability of an 'atypical' distribution is exponentially small in $\mathcal{N}$:

$$\mathbb{P}[P] \doteq \exp[-\mathcal{N}(G[P] - F(\beta))] \,. \tag{4.53}$$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

An illustration of this remark is provided by Exercise 4 of Sec. 4.2.

4.5 The Monte Carlo method

{sec:MonteCarlo}

The Monte Carlo method is an important generic tool which is common to probability theory, statistical physics and combinatorial optimization. In all of these fields, we are often confronted with the problem of sampling a configuration $\underline{x} \in \mathcal{X}^N$ (here we assume $\mathcal{X}$ to be a finite space) from a given distribution $P(\underline{x})$. This can be quite difficult when N is large, because there are too many configurations, because the typical configurations are exponentially rare and/or because the distribution $P(\underline{x})$ is specified by the Boltzmann formula with an unknown normalization (the partition function).

A general approach consists in constructing a Markov chain which is guaranteed to converge to the desired $P(\underline{x})$ and then simulating it on a computer. The computer is of course assumed to have access to some source of randomness: in practice pseudo-random number generators are used. If the chain is simulated for a long enough time, the final configuration has a distribution 'close' to $P(\underline{x})$. In practice, the Markov chain is defined by a set of transition rates $w(\underline{x} \to \underline{y})$ with $\underline{x}, \underline{y} \in \mathcal{X}^N$ which satisfy the following conditions.

1. The chain is **irreducible**, i.e. for any couple of configurations $\underline{x}$ and $\underline{y}$, there exists a path $(\underline{x}_0, \underline{x}_1, \dots \underline{x}_n)$ of length n, connecting $\underline{x}$ to $\underline{y}$ with non-zero probability. This means that $\underline{x}_0 = \underline{x}$, $\underline{x}_n = \underline{y}$ and $w(\underline{x}_i \to \underline{x}_{i+1}) > 0$ for $i = 0 \dots n-1$.
2. The chain is **aperiodic**: for any couple $\underline{x}$ and $\underline{y}$, there exists a positive integer $n(\underline{x}, \underline{y})$ such that, for any $n \geq n(\underline{x}, \underline{y})$ there exists a path of length n connecting $\underline{x}$ to $\underline{y}$ with non-zero probability. Notice that, for an irreducible chain, aperiodicity is easily enforced by allowing the configuration to remain unchanged with non-zero probability: $w(\underline{x} \to \underline{x}) > 0$.
3. The distribution $P(\underline{x})$ is **stationary** with respect to the probabilities $w(\underline{x} \to \underline{y})$:

$$\sum_{\underline{x}} P(\underline{x})\; w(\underline{x} \to \underline{y}) = P(\underline{y})\,. \tag{4.54}$$

   Sometimes a stronger condition (implying stationarity) is satisfied by the transition probabilities. For each couple of configurations $\underline{x}$, $\underline{y}$ such that either $w(\underline{x} \to \underline{y}) > 0$ or $w(\underline{y} \to \underline{x}) > 0$, one has

$$P(\underline{x})\, w(\underline{x} \to \underline{y}) = P(\underline{y})\, w(\underline{y} \to \underline{x})\,. \tag{4.55}$$

   This condition is referred to as **reversibility** or **detailed balance**.

The strategy of designing and simulating such a process in order to sample from $P(\underline{x})$ goes under the name of **dynamic Monte Carlo** method or **Monte Carlo Markov chain** method (hereafter we shall refer to it simply as Monte Carlo method). The theoretical basis for such an approach is provided by two classic theorems which we collect below.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

{thm:AsymptoticMarkov}

Theorem 4.17 *Assume the rates $w(\underline{x} \to \underline{y})$ to satisfy the hypotheses 1-3 above. Let $\underline{X}_0, \underline{X}_1, \ldots, \underline{X}_t, \ldots$ be random variables distributed according to the Markov chain with rates $w(\underline{x} \to \underline{y})$ and initial condition $\underline{X}_0 = \underline{x}_0$. Let $f : \mathcal{X}^N \to \mathbb{R}$ be any real valued function. Then*

1. *The probability distribution of X_t converges to the stationary one:*

$$\lim_{t\to\infty} \mathbb{P}[\underline{X}_t = \underline{x}] = P(\underline{x})\,. \tag{4.56}$$

2. *Time averages converge to averages over the stationary distribution*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} f(\underline{X}_s) = \sum_{\underline{x}} P(\underline{x}) f(\underline{x}) \qquad \textit{almost surely.} \tag{4.57}$$

The proof of this Theorem can be found in any textbook on Markov processes. Here we will illustrate it by considering two simple Monte Carlo algorithms which are frequently used in statistical mechanics (although they are by no means the most efficient ones).

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

Example 4.18 Consider a system of N Ising spins $\underline{\sigma} = (\sigma_1 \dots \sigma_N)$ with energy function $E(\underline{\sigma})$ and inverse temperature β. We are interested in sampling the Boltzmann distribution P_β. The **Metropolis algorithm** with random updatings is defined as follows. Call $\underline{\sigma}^{(i)}$ the configuration which coincides with $\underline{\sigma}$ but for the site i ($\sigma_i^{(i)} = -\sigma_i$), and let $\Delta E_i(\underline{\sigma}) \equiv E(\underline{\sigma}^{(i)}) - E(\underline{\sigma})$. At each step, an integer $i \in [N]$ is chosen randomly with flat probability distribution and the spin σ_i is flipped with probability

$$w_i(\underline{\sigma}) = \exp\{-\beta \max[\Delta E_i(\underline{\sigma}), 0]\} . \tag{4.58}$$

In formulae, the transition probabilities are given by

$$w(\underline{\sigma} \to \underline{\tau}) = \frac{1}{N} \sum_{i=1}^{N} w_i(\underline{\sigma})\, \delta(\underline{\tau}, \underline{\sigma}^{(i)}) \; + \; \left[1 - \frac{1}{N} \sum_{i=1}^{N} w_i(\underline{\sigma}) \right] \delta(\underline{\tau}, \underline{\sigma}) , \tag{4.59}$$

where $\delta(\underline{\sigma}, \underline{\tau}) = 1$ if $\underline{\sigma} \equiv \underline{\tau}$, and $= 0$ otherwise. It is easy to check that this definition satisfies both the irreducibility and the stationarity conditions for any energy function $E(\underline{\sigma})$ and inverse temperature $\beta < 1$. Furthermore, the chain satisfies the detailed balance condition:

$$P_\beta(\underline{\sigma})\, w_i(\underline{\sigma}) = P_\beta(\underline{\sigma}^{(i)})\, w_i(\underline{\sigma}^{(i)}) . \tag{4.60}$$

Whether the condition of aperiodicity is fulfilled depends on the energy. It is easy to construct systems for which it does not hold. Take for instance a single spin, $N = 1$, and let $E(\sigma) = 0$: the spin is flipped at each step and there is no way to have a transition from $\sigma = +1$ to $\sigma = -1$ in an even number of steps. (But this kind of pathology is easily cured modifying the algorithm as follows. At each step, with probability $1-\varepsilon$ a site i is chosen and a spin flip is proposed as above. With probability ε nothing is done, i.e. a null transition $\underline{\sigma} \to \underline{\sigma}$ is realized.)

Exercise 4.3 Variants of this chain can be obtained by changing the flipping probabilities (4.58). A popular choice consists in the **heath bath** algorithm (also referred to as **Glauber dynamics**):

$$w_i(\underline{\sigma}) = \frac{1}{2} \left[1 - \tanh\left(\frac{\beta \Delta E_i(\underline{\sigma})}{2} \right) \right] . \tag{4.61}$$

Prove irreducibility, aperiodicity and stationarity for these transition probabilities.

One of the reason of interest of the heath bath algorithm is that it can be easily generalized to any system whose configuration space has the form $\mathcal{X}^N$. In this algorithm one chooses a variable index i, fixes all the others variables, and

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

assign a new value to the i-th one according to its conditional distribution. A more precise description is provided by the following pseudocode. Recall that, given a vector $\underline{x} \in \mathcal{X}^N$, we denote by $\underline{x}_{\sim i}$, the $N-1$-dimensional vector obtained by removing the i-th component of $\underline{x}$.

Heat bath algorithm()

Input: A probability distribution $P(\underline{x})$ on the configuration space $\mathcal{X}^N$, and the number r of iterations.

Output: a sequence $\underline{x}^{(0)}, \underline{x}^{(1)}, \dots, \underline{x}^{(r)}$

1. Generate $\underline{x}^{(0)}$ uniformly at random in $\mathcal{X}^N$.
2. For $t = 1$ to $t = r$:
   2.1 Draw a uniformly random integer $i \in \{1, \dots, N\}$
   2.2 For each $z \in \mathcal{X}$, compute

$$P(X_i = z | \underline{X}_{\sim i} = \underline{x}^{(t-1)}_{\sim i} \underline{y}) = \frac{P(X_i = z, \underline{X}_{\sim i} = \underline{x}^{(t-1)}_{\sim i})}{\sum_{z' \in \mathcal{X}} P(X_i = z', \underline{X}_{\sim i} = \underline{x}^{(t-1)}_{\sim i})} \, . \tag{4.62}$$

   2.3 Set $x_j^{(t)} = x_j^{(t-1)}$ for each $j \neq i$, and $x_i^{(t)} = z$ where z is drawn from the distribution $P(X_i = z | \underline{X}_{\sim i} = \underline{x}^{(t-1)}_{\sim i} \underline{y})$.

Let us stress that this algorithm does only require to compute the probability $P(\underline{x})$ up to a multiplicative constant. If, for instance, $P(\underline{x})$ is given by Boltzmann law, cf. Sec. 2.1, it is enough to be able to compute the energy $E(\underline{x})$ of a configuration, and is instead not necessary to compute the partition function $Z(\beta)$.

This is a very general method for defining a Markov chain with the desired property. The proof is left as exercise.

Exercise 4.4 Assuming for simplicity that $\forall \underline{x}$, $P(\underline{x}) > 0$, prove irreducibility, aperiodicity and stationarity for the heat bath algorithm.

Theorem 4.17 confirms that the Monte Carlo method is indeed a viable approach for sampling from a given probability distribution. However, it does not provide any information concerning its computational efficiency. In order to discuss such an issue, it is convenient to assume that simulating a single step $\underline{X}_t \to \underline{X}_{t+1}$ of the Markov chain has a unitary time-cost. This assumption is a good one as long as sampling a new configuration requires a finite (fixed) number of computations and updating a finite (and N-independent) number of variables. This is the case in the two examples provided above, and we shall stick here to this simple scenario.

Computational efficiency reduces therefore to the question: how many step of the Markov chain should be simulated? Of course there is no unique answer to such a generic question. We shall limit ourselves to introduce two important figures of merit. The first concerns the following problem: how many steps should be simulated in order to produce a single configuration $\underline{x}$ which is distributed

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

approximately according to $P(\underline{x})$? In order to precise what is meant by "approximately" we have to introduce a notion distance among distributions $P_1(\cdot)$ and $P_2(\cdot)$ on $\mathcal{X}^N$. A widespread definition is given by the **variation distance**:

$$||P_1 - P_2|| = \frac{1}{2} \sum_{\underline{x} \in \mathcal{X}^N} |P_1(\underline{x}) - P_2(\underline{x})| \,. \tag{4.63}$$

Consider now a Markov chain satisfying the hypotheses 1-3 above with respect to a stationary distribution $P(\underline{x})$ and call $P_t(\underline{x}|\underline{x}_0)$ the distribution of $\underline{X}_t$ conditional to the initial condition $\underline{X}_0 = \underline{x}_0$. Let $d_{\underline{x}_0}(t) = ||P_t(\cdot|\underline{x}_0) - P(\cdot)||$ be the distance from the stationary distribution. The **mixing time** (or **variation threshold time**) is defined as

$$\tau_{\rm eq}(\varepsilon) = \min\{t > 0 \ : \ \sup_{\underline{x}_0} d_{x_0}(t) \le \varepsilon\} \,. \tag{4.64}$$

In this book we shall often refer informally to this quantity (or to some close relative) as the **equilibration time**. The number ε can be chosen arbitrarily, a change in ε implying usually a simple multiplicative change in in $\tau_{\rm eq}(\varepsilon)$. Because of this reason the convention $\varepsilon = 1/e$ is sometimes adopted.

Rather than producing a single configuration with the prescribed distribution, one is often interested in computing the expectation value of some observable $\mathcal{O}(\underline{x})$. In principle this can be done by averaging over many steps of the Markov chain as suggested by Eq. (4.57). It is therefore natural to pose the following question. Assume the initial condition $\underline{X}_0$ is distributed according to the stationary distribution $P(\underline{x})$. This can be obtained by simulating $\tau_{\rm eq}(\varepsilon)$ steps of the chain in a preliminary (equilibration) phase. We shall denote by $\langle\cdot\rangle$ the expectation with respect to the Markov chain with this initial condition. How many steps should we average over in order to get expectation values within some prescribed accuracy? In other words, we estimate $\sum P(\underline{x})\mathcal{O}(\underline{x}) \equiv \mathbb{E}_P\mathcal{O}$ by

$$\overline{\mathcal{O}}_T \equiv \frac{1}{T} \sum_{t=0}^{T-1} \mathcal{O}(\underline{X}_t) \,. \tag{4.65}$$

It is clear that $\langle \overline{\mathcal{O}}_T \rangle = \sum P(\underline{x})\mathcal{O}(\underline{x})$. Let us compute the variance of this estimator:

$$\mathrm{Var}(\overline{\mathcal{O}}_T) = \frac{1}{T^2} \sum_{s,t=0}^{T-1} \langle \mathcal{O}_s; \mathcal{O}_t \rangle = \frac{1}{T^2} \sum_{t=0}^{T-1} (T-t) \langle \mathcal{O}_0; \mathcal{O}_t \rangle \,, \tag{4.66}$$

where we used the notation $\mathcal{O}_t \equiv \mathcal{O}(\underline{X}_t)$. Let us introduce the **autocorrelation function** $C_{\mathcal{O}}(t-s) \equiv \frac{\langle \mathcal{O}_s; \mathcal{O}_t \rangle}{\langle \mathcal{O}_0; \mathcal{O}_0 \rangle}$, so that $\mathrm{Var}(\overline{\mathcal{O}}_T) = \frac{\langle \mathcal{O}_0; \mathcal{O}_0 \rangle}{T^2} \sum_{t=0}^{T-1} (T-t)\, C_{\mathcal{O}}(t)$. General results on Markov chain on finite state spaces imply that $C_{\mathcal{O}}(t)$ decreases exponentially as $t \to \infty$. Therefore, for large T, we have

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

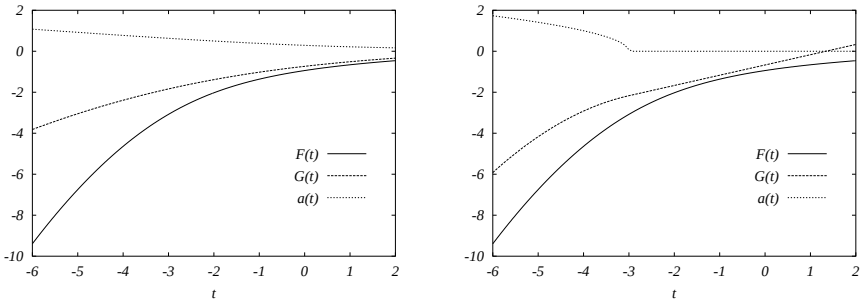

FIG. 4.5. Variational estimates of the free energy of the model (4.44). We use the trial distributions (4.46) on the left and (4.48) on the right.

{fig:variational_anh}

$$\mathrm{Var}(\overline{\mathcal{O}}_T) = \frac{\tau^{\mathcal{O}}_{\rm int}}{T}\,[\mathbb{E}_P\mathcal{O}^2 - (\mathbb{E}_P\mathcal{O})^2] + O(T^{-2})\,. \tag{4.67}$$

The **integrated autocorrelation time** $\tau^{\mathcal{O}}_{\rm int}$ is given by

$$\tau^{\mathcal{O}}_{\rm int} \equiv \sum_{t=0}^{\infty} C_{\mathcal{O}}(t)\,, \tag{4.68}$$

and provides a reference for estimating how long the Monte Carlo simulation should be run in order to get some prescribed accuracy. Equation (4.67) can be interpreted by saying that one statistically independent estimate of $\mathbb{E}_P\mathcal{O}$ is obtained every $\tau^{\mathcal{O}}_{\rm int}$ iterations.

Example 4.19 Consider the Curie-Weiss model, cf. Sec. 2.5.2, at inverse temperature β, and use the heath-bath algorithm of Example 2 in order to sample from the Boltzmann distribution. In Fig. **??** we reproduce the evolution of the global magnetization $m(\underline{\sigma})$ during three different simulations at inverse temperatures $\beta = 0.8,\, 1.0,\, 1.2$ for a model of $N = 150$ spin. In all cases we initialized the Markov chain by extracting a random configuration with flat probability.

A spectacular effect occurs at the lowest temperature, $\beta = 1.2$. Although the Boltzmann average of the global magnetization vanishes, $\langle m(\underline{\sigma})\rangle = 0$, the sign of the magnetization remains unchanged over extremely long time scales. It is clear that the equilibration time is at least as large as these scales. An order-of-magnitude estimate would be $\tau_{\rm eq} > 10^5$. Furthermore this equilibration time diverges exponentially at large N. Sampling from the Boltzmann distribution using the present algorithm becomes exceedingly difficult at low temperature.

4.6 Simulated annealing

{sec:SimulAnn}

As we mentioned in Sec. 3.5, any optimization problem can be 'embedded' in a statistical mechanics problem. The idea is to interpret the cost function $E(\underline{x})$, $\underline{x} \in$

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

$\mathcal{X}^N$ as the energy of a statistical mechanics system and consider the Boltzmann distribution $p_\beta(\underline{x}) = \exp[-\beta E(\underline{x})]/Z$. In the low temperature limit $\beta \to \infty$, the distribution concentrates over the minima of $E(\underline{x})$, and the original optimization setting is recovered.

Since the Monte Carlo method provides a general technique for sampling from the Boltzmann distribution, one may wonder whether it can be used, in the $\beta \to \infty$ limit, as an optimization technique. A simple minded approach would be to take $\beta = \infty$ at the outset. Such a straegy is generally referred to as **quench** in statistical physics and **greedy search** in combinatorial optimization, and is often bound to fail. Consider in fact the stationarity condition (4.54) and rewrite it using the Boltzmann formula

$$\sum_{\underline{x}} e^{-\beta\,[E(\underline{x})-E(\underline{y})]}\; w(\underline{x} \to \underline{y}) = 1\,. \tag{4.69}$$

Since all the terms on the left hand side are positive, any of them cannot be larger than one. This implies $0 \le w(\underline{x} \to \underline{y}) \le \exp\{-\beta\,[E(\underline{y}) - E(\underline{x})]\}$. Therefore, for any couple of configurations $\underline{x}$, $\underline{y}$, such that $E(\underline{y}) > E(\underline{x})$ we have $w(\underline{x} \to \underline{y}) \to 0$ in the $\beta \to \infty$ limit. In other words, the energy is always non-increasing along the trajectories of a zero-temperature Monte Carlo algorithm. As a consequence, the corresponding Markov chain is not irreducible, although it is irreducible at any $\beta < \infty$, and is not guaranteed to converge to the equilibrium distribution, i.e. to find a global minimum of $E(x)$.

Another simple minded approach would be to set β to some large but finite value. Although the Boltzmann distribution gives some weight to near-optimal configurations, the algorithm will visit, from time to time, also optimal configuratons which are the most probable one. How large should be β? How much time shall we wait before an optimal configuration is visited? We can assume without loss of generality that the minimum of the cost function (the ground state energy) is zero: $E_0 = 0$. A meaningful quantity to look at is the probability for $E(\underline{x}) = 0$ under the Boltzmann distribution at inverse temperature β. We can easily compute the logarithmic moment generating function of the energy:

$$\psi_N(t) = \frac{1}{N} \log \left[\sum_{\underline{x}} p_\beta(\underline{x})\, e^{tE(\underline{x})} \right] = \frac{1}{N}\, \log \left[\frac{\sum_x e^{-(\beta-t)E(x)}}{\sum_x e^{-\beta E(x)}} \right]. \tag{4.70}$$

This is given by $\psi_N(t) = \phi_N(\beta - t) - \phi_N(\beta)$, where $\phi_N(\beta)$ is the free entropy density at inverse temperature β. Clearly $p_\beta[E(x) = 0] = \exp[N\psi_N(-\infty)] = \exp\{N[\phi_N(\infty) - \phi_N(\beta)]\}$, and the average time to wait before visiting the optimal configuration is $1/p_\beta[E(x) = 0] = \exp[-N\psi_N(-\infty)]$.

‘‘Info Phys Comp’’ Draft: November 9, 2007 -- ‘‘Info Phys

Exercise 4.5 Assume that the cost function takes integer values $E = 0, 1, 2\dots$ and call $\mathcal{X}_E$ the set of cofigurations $\underline{x}$ such that $E(\underline{x}) = E$. You want the Monte Carlo trajectories to spend a fraction $(1-\varepsilon)$ of the time on optimal solutions. Show that the temperature must be chosen such that

$$\beta = \log\left(\frac{|\mathcal{X}_1|}{\varepsilon|\mathcal{X}_0|}\right) + \Theta(\varepsilon)\,. \tag{4.71}$$

In Section 2.4 we argued that, for many statistical mechanics models, the free entropy density has a finite thermodynamic limit $\phi(\beta) = \lim_{N\to\infty}\phi_N(\beta)$. In the following Chapters we will show that this is the case also for several interesting optimization problems. This implies that $p_\beta[E(x) = 0]$ vanishes in the $N \to \infty$ limit. In order to have a non-negligibile probability of hitting a solution of the optimization problem, β must be scaled with N in such a waythat $\beta \to \infty$ as $N \to \infty$. On the other hand, letting $\beta \to \infty$ we are going to face the reducibility problem mentioned above. Althouch the Markov chain is formally irreducible, its equilibration time will diverge as $\beta \to \infty$.

The idea of **simulated annealing** consists in letting β vary with time. More precisely one decides an **annealing schedule** $\{(\beta_1, n_1);\ (\beta_2, n_2); \dots (\beta_L, n_L)\}$, with inverse temperatures $\beta_i \in [0, \infty]$ and integers $n_i > 0$. The algorithm is initialized on a configutation $\underline{x}_0$ and executes n_1 Monte Carlo steps at temperature β_1, n_2 at temperature β_2, ..., n_L at temperature β_L. The final configuration of each cycle i (with $i = 1, \dots, L-1$) is used as initial configuration of the next cycle. Mathematically, such a process is a **time-dependent Markov chain**. The common wisdom about the simulated annealing algorithm is that varying the temperature with time should help avoiding the two problems encountered above. Usually one takes the β_i's to be an increasing sequence. In the first stages a small β should help equilibrating across the space of configurations $\mathcal{X}^N$. As the themperature is lowered the probability distribution concentrates on the lowest energy regions of this space. Finally, in the late stages, a large β forces the system to fix the few wrong details, and to find solution. Of course, this image is very simplistic. In the following Chapter we shall try to refine it by considering the application of simulated annealing to a variety of problems.

{app_sanov_ft}

4.7 Appendix: A physicist's approach to Sanov's theorem

Let us show how the formulas of Sanov's theorem can be obtained using the type of 'field theoretic' approach used in statistical physics. The theorem is easy to prove, the aim of this section is not so much to give a proof, but rather to show on a simple example a type of approach that is very common in physics, and which can be powerful. We shall not aim at a rigorous derivation.

The probability that the type of the sequence $x_1, \cdots, x_N$ be equal to $q(x)$ can be written as:

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

$$\mathbb{P}[q(x)] = \mathbb{E}\left\{ \prod_{x\in\mathcal{X}} \mathbb{I}\left(q(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x,x_i} \right) \right\}$$
$$= \sum_{x_1\cdots x_N} p(x_1)\cdots p(x_N)\, \mathbb{I}\left(q(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x,x_i} \right) . \qquad (4.72)$$

A typical approach in field theory is to introduce some auxiliary variables in order to enforce the constraint that $q(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x,x_i}$. For each $x \in \mathcal{X}$, one introduces a variable $\lambda(x)$, and uses the 'integral representation' of the constraint in the form:

$$\mathbb{I}\left(q(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x,x_i} \right) = \int_0^{2\pi} \frac{d\lambda(x)}{2\pi} \ \exp\left[i\lambda(x) \left(Nq(x) - \sum_{i=1}^{N} \delta_{x,x_i} \right) \right] . \qquad (4.73)$$

Dropping q-independent factors, we get:

$$\mathbb{P}[q(x)] = C \int \prod_{x\in\mathcal{X}} d\lambda(x) \ \exp\{NS[\lambda]\} ,$$

where C is a normalization constant, and the **action** S is given by:

$$S[\lambda] = i\sum_x \lambda(x)q(x) + \log\left[\sum_x p(x) e^{-i\lambda(x)} \right] \qquad (4.74)$$

In the large N limit, the integral in (4.74) can be evaluated with a saddle point method. The saddle point $\lambda(x) = \lambda^*(x)$ is found by solving the stationarity equations $\partial S/\partial\lambda(x) = 0$ for any $x \in \mathcal{X}$. One gets a family of solutions $-i\lambda(x) = C + \log(q(x)/p(x))$ with C arbitrary. The freedom in the choice of C comes from the fact that $\sum_x(\sum_i \delta_{x,x_i}) = N$ for any configuration $x_1 \ldots x_N$, and therefore one of the constraints is in fact useless. This freedom can be fixed arbitrarily: regardless of this choice, the action on the saddle point is

$$S[\lambda^*] = S_0 - \sum_x q(x) \log \frac{q(x)}{p(x)} , \qquad (4.75)$$

where S_0 is a q independent constant. One thus gets $P[q(x)] \doteq \exp[-ND(q||p)]$.

The reader who has never encountered this type of reasoning may wonder why use such an indirect approach. It turns out that it is a very common formalism in statistical physics, where similar methods are also applied, under the name 'field theory', to continuous $\mathcal{X}$ spaces (some implicit discretization is then usually assumed at intermediate steps, and the correct definition of a continuum limit is often not obvious). In particular the reader interested in the statistical physics approach to optimizations problems or information theory will often find this type of formalism in research papers. One of the advantages of this approach is

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys

that it provides a formal solution to a large variety of problems. The quantity to be computed is expressed in an integral form as in (4.74). In problems having a 'mean field' structure, the dimension of the space over which the integration is performed does not depend upon N. Therefore its leading exponential behavior at large N can be obtained by saddle point methods. The reader who wants to get some practice of this approach is invited to 'derive' in the same way the various theorems and corollaries of this chapter.

Notes

The theory of large deviations is exposed in the book of Dembo and Zeitouni (Dembo and Zeitouni, 1998), and its use in statistical physics can be found in Ellis's book (Ellis, 1985).

Markov chains on discrete state spaces are treated by Norris (Norris, 1997) A nice introduction to Monte Carlo methods in statistical physics is given in the lecture notes by Krauth (Krauth, 1998) and by Sokal (Sokal, 1996).

Simulated annealing was introduced by Kirkpatrick, Gelatt and Vecchi 1983 (Kirkpatrick, C. D. Gelatt and Vecchi, 1983). It is a completely "universal" optimization algorithm: it can be defined without reference to any particular problem. Beacause of this reason it ofteen overlooks important structures that may help solving the problem itself.

''Info Phys Comp'' Draft: November 9, 2007 -- ''Info Phys