1

Introduction

1.1 Statistical physics

Statistical physics is about systems composed of many parts.

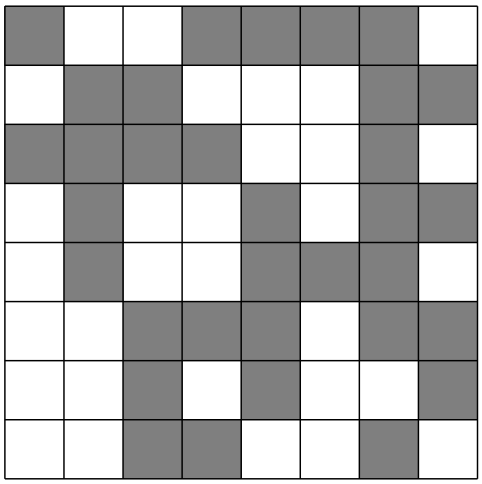
Examples:

- atoms or molecules in a gas, liquid, or solid;
- electrons in a metal, semiconductor, or plasma;
- quanta in quantum £elds, particularly photons in electromagnetic £elds and phonons in sound.
- individuals in populations, particularly evolution (changes in gene frequencies in populations), the spread of disease, social interactions;
- species in an ecosystem;
- computers in a network;
- agents in a market, such as a stock market;
- swarms of insects, such as ants.

The techniques for studying these systems are based largely on combinatorics and probability theory, hence the name *statistical* physics.

1.2 Percolation

Imagine coloring in the squares on a square lattice at random with probability p:

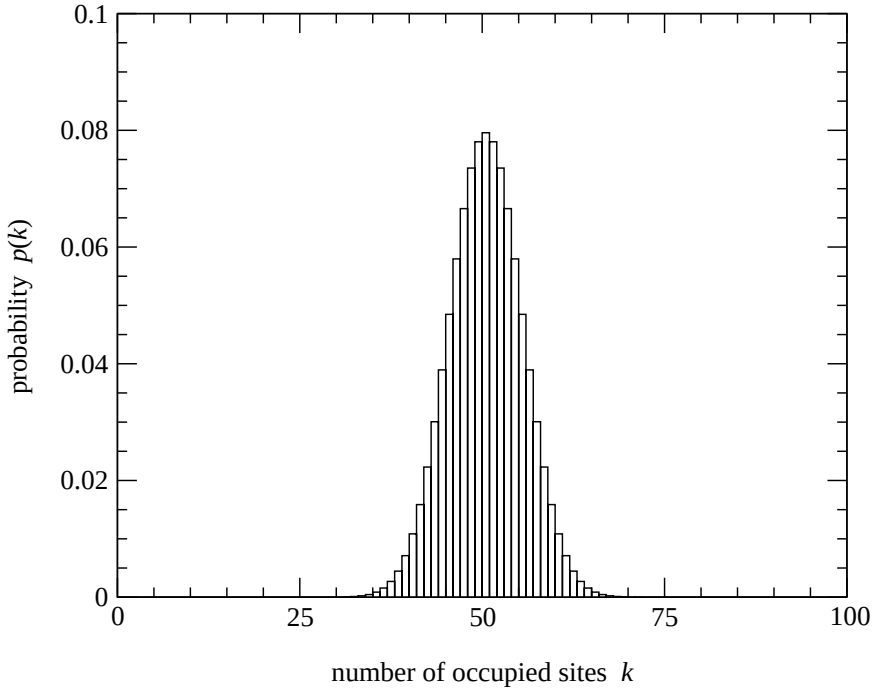


This simple system is called **site percolation** on the square lattice.

How many sites are colored? Suppose there are N sites in total. Then the probability of there being k of them colored in is

$$p_k = \binom{N}{k} p^k (1-p)^{N-k}. \tag{1.1}$$

This is the **binomial distribution**. For a square lattice of $N = 10 \times 10 = 100$, for example, with $p = \frac{1}{2}$, it looks like this:



The mean of this distribution is

$$\langle k\rangle = \sum_{k=0}^{N} k p_k = \sum_{k=0}^{N} k \binom{N}{k} p^k q^{N-k}, \tag{1.2}$$

where $q = 1 - p$. But this is equal to

$$\langle k\rangle = p\frac{\partial}{\partial p}\sum_{k=0}^{N}\binom{N}{k}p^k q^{N-k} = \frac{\partial}{\partial p}(p+q)^N = pN. \tag{1.3}$$

So the mean is right at the average occupation probability, as we might expect. The mean square of the distribution is

$$\langle k^2\rangle = \sum_{k=0}^{N} k^2 \binom{N}{k} p^k q^{N-k} = p\frac{\partial}{\partial p}p\frac{\partial}{\partial p}(p+q)^N = pN + p^2N(N-1). \tag{1.4}$$

So the variance is

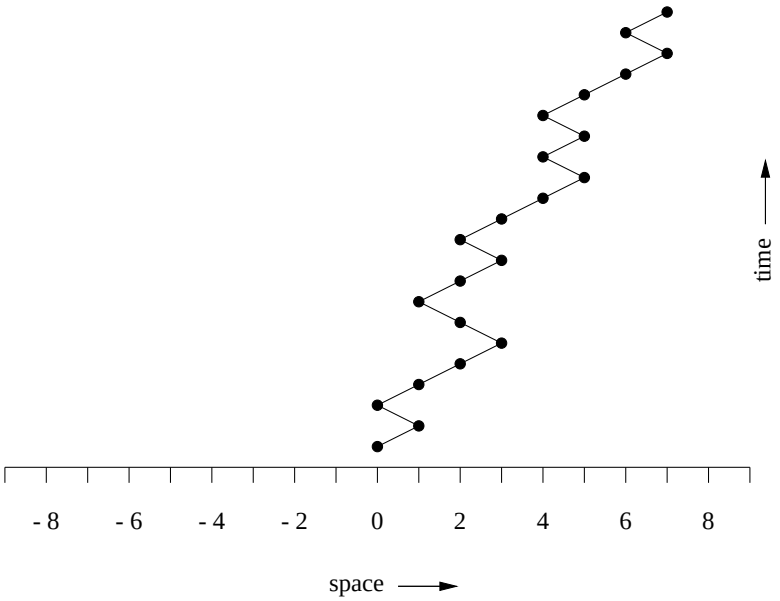
$$\sigma_k^2 = \langle k^2\rangle - \langle k\rangle^2 = pN + p^2N(N-1) - p^2N^2 = pN - p^2N = p(1-p)N. \tag{1.5}$$

So the standard deviation goes as $\sigma_k \sim \sqrt{N}$, and hence the distribution gets narrower, as a fraction of N as N becomes large. Thus as N becomes large, we can predict the value of k/N with better and better accuracy.

> Some values of k are statistically more likely than others—sometimes *much* more likely. Knowing nothing else about this system, we can make a prediction about what the value of k is likely to be here.

1.3 Random walk

Here's another example, the random walk. Consider a walker on a straight line who takes one step every second with probability p of going to the right and probability $q = 1 - p$ of going to the left:



The position $x(N)$ of the walker after N steps is given by

$$x(N) = r(N) - l(N) = 2r(N) - N, \tag{1.6}$$

where $r(N)$ is the number of steps to the right and $l(N) = N - r(N)$ the number to the left.

What is the probability distribution of $x(N)$? The probability of taking r steps to the right and $N - r$ to the left is

$$p_r = \binom{N}{r} p^r q^{N-r} \tag{1.7}$$

which is just the binomial distribution again. Thus we know immediately that the average value of $x(N)$ is

$$\langle x(N) \rangle = 2pN - N = N(p - q), \tag{1.8}$$

and the variance is

$$\begin{aligned} \sigma_x^2 &= \langle x^2 \rangle - \langle x \rangle^2 = 4\langle r^2 \rangle - 4N\langle r \rangle + N^2 - N^2(p-q)^2 \\ &= 4pN + 4p^2N^2 - 4p^2N - 4pN^2 + N^2 - 4p^2N^2 + 4pN^2 - N^2 \\ &= 4pqN. \end{aligned} \tag{1.9}$$

Thus the random walk, which is a sum of independent random variables ± 1, gives a binomial distribution in x. This is a special case of the **central limit theorem**.

1.4 Entropy

The simple results of the previous section are a particular example of a general concept:

- A **microstate** is one of the individual states of our system, such as a particular set of occupied sites on the percolation lattice, or a particular path taken by the random walker.

- A **macrostate** is a state of the system de£ned by some large scale property of the system, such as the number of occupied sites, or the distance traveled by the walker.

In general there are many microstates which can correspond to a given macrostate. For example, there are usually many paths the walker can take which will result in it traveling a distance x.

Let $\Omega(x)$ be the number of microstates corresponding to macrostate x. Then the most likely value of x is the one which maximizes $\Omega(x)$. Conventionally, in fact, one maximizes

$$S(x) = k \log \Omega(x), \tag{1.10}$$

which is called the **entropy**. (Strictly, it's the **microcanonical entropy**—we'll come to that.)

From the calculations above, we know that the width of the peak in $\Omega(x)$ gets narrower as N gets larger, so that in the limit of large N maximizing the entropy gives a very good estimate of the value of x.

A real-world example: Here is a picture of my of£ce at the Santa Fe Institute.

It's messy. Why? Because there are many microstates of my of£ce—many places I could put each paper and book for example—but most of them correspond to what we would de£ne as "messy" and very few to what we would de£ne as "tidy." Messy and tidy are the macrostates in this case, and the of£ce is messy because $\Omega(\text{messy}) \gg \Omega(\text{tidy})$. Of course, I could tidy up my of£ce (something I do about once or twice a decade), and so lower the entropy by moving from the messy macrostate to the tidy one. But this requires **work** (and work of a particularly unattractive kind too).

> Macrostates with high entropy are more likely than ones with low entropy. This allows us to predict which macrostates are likely to occur. It also means that most systems are in high-entropy states. Their entropy can be lowered, but only by doing work.

1.5 More general forms for entropy

The de£nition we gave for entropy above is correct, but limited. In particular, it makes two assumptions:

1. that all the microstates corresponding to a given macrostate are equally likely to occur;
2. that the macrostate is speci£ed by the value of a quantity which is measurable individually for each microstate (number of occupied sites, distance traveled by a walker).

In general neither of these things is true. It is of course perfectly possible to have systems in which microstates occur with different probabilities—even ones which correspond to the same macrostate. Also, more general ways of specifying the macrostate are possible. Indeed, the most general way to specify the macrostate is simply to state the set of probabilities $\{p_i\}$ that the system will be found in a given microstate. From this set, any other macroscopic variable, e.g., most probable microstate as above, can be calculated trivially.

Consider then a system which can be in any one of N microstates denoted by $i = 1 \ldots N$. Imagine in fact that we have a large number $M \gg N$ of copies of this system—a so-called **ensemble**—and that we measure each one to £nd out what microstate it is in. Let n_i be the number of systems found to be in the ith microstate. Then the number of ways of getting a particular set of values $\{n_i\}$—the number of microstates corresponding to this macrostate—is given by the multinomial distribution

$$\Omega(\{n_i\}) = \frac{M!}{n_1! n_2! \ldots n_N!}. \tag{1.11}$$

Then the most likely macrostate is the one which corresponds to the maximum of this quantity, or equivalently to the maximum of the entropy

$$S = \frac{1}{M} \log \Omega = \frac{1}{M} \Big[\log M! - \sum_{i=1}^{N} \log n_i! \Big]. \tag{1.12}$$

We make use of Sterling's approximation

$$\log k! \simeq k \log k - k, \tag{1.13}$$

giving

$$\begin{aligned} S &= \frac{1}{M} \Big[M \log M - M - \sum_{i=1}^{N} n_i \log n_i + \sum_{i=1}^{N} n_i \Big] = - \sum_i \frac{n_i}{M} \log \frac{n_i}{M} \\ &= - \sum_i p_i \log p_i, \end{aligned} \tag{1.14}$$

where

$$p_i = \frac{n_i}{M}. \tag{1.15}$$

Note that in this formulation the macrostate can only be de£ned with respect to the entire ensemble. Also, note the minus sign.

There is sometimes a constant k given in front of the de£nition of the entropy thus:

$$S = -k \sum_i p_i \log p_i. \tag{1.16}$$

Of course, this constant makes no difference to where the maximum of the entropy is. In traditional statistical mechanics, $k = 1.3807 \times 10^{-23}$ JK^{-1}, for reasons which are rooted in the obscure and often nonsensical history of physics.

> Equation (1.16) is perhaps the most important equation in statistical physics. It gives the Gibbs entropy for an ensemble. The Gibbs entropy is the quantity which is maximized in order to £nd the most probable macrostate of the ensemble, which corresponds to a set of values $\{p_i\}$.

1.6 Examples of the use of the Gibbs entropy

To make use of the Gibbs entropy, one usually speci£es the system of interest and any relevant constraints on the probabilities p_i, and then maximizes the entropy to £nd the most probable set $\{p_i\}$ subject to those constraints. Here are some examples.

1.6.1 The microcanonical ensemble

Consider again systems like the simple ones at the beginning of this lecture in which all microstates i are equally likely, and a macrostate m corresponds to a speci£c set of microstates. Then the constraints on p_i are simple:

$$p_i = \begin{cases} \Omega_m^{-1} & \text{if state } i \text{ belongs to macrostate } m \\ 0 & \text{otherwise.} \end{cases} \tag{1.17}$$

Thus

$$S_m = -\sum_{i \in m} \frac{1}{\Omega_m} \log \frac{1}{\Omega_m} = \log \Omega_m, \tag{1.18}$$

exactly as we de£ned it before.

In fact, if we don't restrict all p_i to be equal, we £nd that $p_i =$ constant maximizes S anyway—the uniform probability distribution maximizes the entropy with or without the constraint.

1.6.2 The canonical ensemble

A more realistic type of constraint on a system is a constraint on the average value of some observable quantity E. In almost all experiments that we do on systems we don't simply measure a quantity once, we measure it repeatedly. The universal assumption one makes, which is almost entirely unproven, and probably wrong except in all the cases that matter, is the **ergodic hypothesis**:

The average of a large number of measurements on the same system will be the same as the average of measurements on an ensemble of different and independent systems.

This means that the average of our measurements, which is the thing one almost always calculates, is

$$\langle E \rangle = \sum_i p_i E_i. \tag{1.19}$$

Suppose we have measured $\langle E \rangle$, and we want to know what the most likely probability distribution over microstates is. Then we should maximize Eq. (1.16) subject to the constraint (1.19), as well as the obvious sum rule

$$\Sigma = \sum_i p_i = 1. \tag{1.20}$$

We can do the maximization using the method of Lagrange multipliers:

$$\frac{\partial S}{\partial p_i} - \alpha \frac{\partial \Sigma}{\partial p_i} - \beta \frac{\partial \langle E \rangle}{\partial p_i} = 0 \qquad \text{for all } i, \tag{1.21}$$

which gives us

$$\log p_i - 1 - \alpha - \beta E_i = 0. \tag{1.22}$$

Or equivalently

$$p_i = \frac{\mathrm{e}^{-\beta E_i}}{Z}, \tag{1.23}$$

where Z is a normalization coef£cient which ensures that Eq. (1.20) is satis£ed. Z's value is

$$Z = \sum_i \mathrm{e}^{-\beta E_i}, \tag{1.24}$$

and it has a special name: it's called the **partition function**.

The Lagrange multiplier β is given in terms of $\langle E \rangle$ by substituting Eq. (1.23) back into Eq. (1.19). Alternatively, in some cases one actually speci£es β and then calculates $\langle E \rangle$ from Eqs. (1.19) and (1.23). For example, in classical equilibrium statistical mechanics $\beta = (kT)^{-1}$, where T is the temperature of the system, k is the Boltzmann constant de£ned in Section 1.5, and the observable E is, in this case, the total internal energy of the system.

Since it is by far the most common approach to measure the average of an observable quantity as in Eq. (1.19), the distribution (1.23) applies to a huge variety of different systems. This distribution is called the **Boltzmann distribution**.

Once we have the distribution of probabilities p_i we can use it to predict other things. For example, the variance of E immediately follows from

$$\sigma_E^2 = \langle E^2 \rangle - \langle E \rangle^2 = \frac{\sum_i \mathrm{e}^{-\beta E_i} E_i^2}{\sum_i \mathrm{e}^{-\beta E_i}} - \langle E \rangle^2. \tag{1.25}$$

1.6.3 Information theory

Consider a communication channel—a letter sent through the mail for example, or a page of a book, or an email message. Suppose there are N different possible messages that can be sent, and suppose that message i is sent with probability p_i. How much information is received per message sent?

Imagine receiving a large number M of messages. The distribution p_i de£nes the numbers n_i of messages of each type received. The information contained in them is only in their order. How many orders are there? There are

$$\Omega(\{n_i\}) = \frac{M!}{n_1!n_2!\ldots n_N!}. \tag{1.26}$$

Thus Ω is a measure of the information sent, as is its logarithm:

$$S = -\sum_i p_i \log p_i. \tag{1.27}$$

This is the **Shannon information** or **Shannon entropy** of a message. If the logarithms are taken base 2, then the units of information are **bits**.

For example, suppose that our messages are just single letters. Here are the frequencies of the 26 alphabetic letters in the 1.2 million characters of Herman Melville's dreary and frankly odious novel *Moby Dick*:

letter	frequency	percentage	letter	frequency	percentage
A	75982	8.16583	N	64146	6.89381
B	16489	1.77208	O	67654	7.27082
C	22036	2.36822	P	17507	1.88149
D	37387	4.01800	Q	1510	0.16228
E	114225	12.27580	R	50781	5.45746
F	20358	2.18789	S	62704	6.73884
G	20334	2.18531	T	85998	9.24226
H	61366	6.59504	U	25967	2.79069
I	64146	6.89381	V	8429	0.90587
J	1046	0.11241	W	21617	2.32319
K	7888	0.84773	X	1199	0.12886
L	41861	4.49883	Y	16462	1.76918
M	22765	2.44657	Z	630	0.06771

Feeding these probabilities into Shannon's formula, we £nd that the entropy per letter of *Moby Dick* is:

$$S_{\rm moby} = 4.178 \text{ bits per letter}. \tag{1.28}$$

Note that a simple ASCII £le containing the text of the book uses 8 bits per letter. Thus it is immediately clear that it should be possible to compress *Moby Dick* by about a factor of two. In fact the Unix program `gzip` can compress *Moby Dick* from 1202863 characters to 489159, which is somewhat better than a factor of two—better than Shannon's information theory predicts. Exercise: Why is this?