1

Entropy and information theory

Some notes on the connections between statistical mechanics and information theory.

1.1 Shannon entropy

Consider a system that can be in any one of N microstates denoted by $i = 1 \dots N$. 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 find 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

$$\Omega(\{n_i\}) = \frac{M!}{n_1! n_2! \dots n_N!}. \tag{1.1}$$

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

$$S = \frac{1}{M} \log \Omega = \frac{1}{M} \biggl[\log M! - \sum_{i=1}^{N} \log n_i! \biggr]. \tag{1.2}$$

We make use of Sterling's approximation

$$\log k! \simeq k \log k - k, \tag{1.3}$$

giving

$$\begin{aligned} S &= \frac{1}{M} \biggl[M \log M - M - \sum_{i=1}^{N} n_i \log n_i + \sum_{i=1}^{N} n_i \biggr] = - \sum_i \frac{n_i}{M} \log \frac{n_i}{M} \\ &= - \sum_i p_i \log p_i, \end{aligned} \tag{1.4}$$

where

$$p_i = \frac{n_i}{M}. \tag{1.5}$$

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

There is sometimes a constant k given in front of the definition of the entropy thus:

$$S = -k \sum_i p_i \log p_i. \tag{1.6}$$

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

Equation (1.6) 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 find the most probable macrostate of the ensemble, which corresponds to a set of values $\{p_i\}$.

1.2 Examples of the use of the Gibbs entropy

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

1.2.1 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 specific set of microstates. Then the constraints on p_i are simple:

$$p_i = \begin{cases} \Omega_m^{-1} & \text{if state } i \text{ belongs to macrostate } m \\ 0 & \text{otherwise.} \end{cases} \tag{1.7}$$

Thus

$$S_m = -\sum_{i \in m} \frac{1}{\Omega_m} \log \frac{1}{\Omega_m} = \log \Omega_m, \tag{1.8}$$

exactly as we defined it before.

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

1.2.2 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

$$\langle E \rangle = \sum_i p_i E_i. \tag{1.9}$$

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. (1.6) subject to the constraint (1.9), as well as the obvious sum rule

$$\Sigma = \sum_i p_i = 1. \tag{1.10}$$

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

$$\frac{\partial S}{\partial p_i} - \alpha \frac{\partial \Sigma}{\partial p_i} - \beta \frac{\partial \langle E \rangle}{\partial p_i} = 0 \qquad \text{for all } i, \tag{1.11}$$

which gives us

$$\log p_i - 1 - \alpha - \beta E_i = 0. \tag{1.12}$$

Or equivalently

$$p_i = \frac{\mathrm{e}^{-\beta E_i}}{Z}, \tag{1.13}$$

where Z is a normalization coefficient which ensures that Eq. (1.10) is satisfied. Z's value is

$$Z = \sum_i \mathrm{e}^{-\beta E_i}, \tag{1.14}$$

and it is called the **partition function**.

The Lagrange multiplier β is given in terms of $\langle E \rangle$ by substituting Eq. (1.13) back into Eq. (1.9). Alternatively, in some cases one actually specifies β and then calculates $\langle E \rangle$ from Eqs. (1.9) and (1.13). For example, in classical equilibrium statistical mechanics $\beta = (kT)^{-1}$, where T is the temperature of the system, k is the Boltzmann constant defined in Section 1.1, 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. (1.9), the distribution (1.13) 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

$$\sigma_E^2 = \langle E^2 \rangle - \langle E \rangle^2 = \frac{\sum_i \mathrm{e}^{-\beta E_i} E_i^2}{\sum_i \mathrm{e}^{-\beta E_i}} - \langle E \rangle^2. \tag{1.15}$$

1.2.3 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 defines 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

$$\Omega(\{n_i\}) = \frac{M!}{n_1!n_2!\ldots n_N!}. \tag{1.16}$$

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

$$S = -\sum_i p_i \log p_i. \tag{1.17}$$

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**.

The maximum information per message is achieved when all messages are equally likely, in which case we have $S = \log N$. For example, the English language, in which the "messages" are just single letters, would have an information content of $\log_2 26 = 4.7$ bits per letter. In fact, however, not all letters are equally likely, so the information content is less than this. Here are the frequencies of the 26 alphabetic letters in the 1.2 million characters of the 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

Exercise: estimate the entropy per character of the text of *Moby Dick*.