Information theory

15

In this chapter we are going to examine the concept of information and relate it to thermodynamic entropy. At first sight, this seems a slightly crazy thing to do. What on earth do something to do with heat engines and something to do with bits and bytes have in common? It turns out that there is a very deep connection between these two concepts. To understand why, we begin our account by trying to formulate one definition of information.

15.1 Information and Shannon entropy

Consider the following three true statements about Isaac Newton (1643–1727) and his birthday.[1]

(1) Isaac Newton's birthday falls on a particular day of the year.

(2) Isaac Newton's birthday falls in the second half of the year

(3) Isaac Newton's birthday falls on the 25th of a month.

The first statement has, by any sensible measure, no information content. *All* birthdays fall on a particular day of the year. The second statement has more information content: at least we now know which half of the year his birthday is. The third statement is much more specific and has the greatest information content.

How do we quantify information content? Well, one property we could notice is that the greater the probability of the statement being true *in the absence of any prior information*, the less the information content of the statement. Thus if you knew no prior information about Newton's birthday, then you would say that statement 1 has probability $P_1 = 1$, statement 2 has probability $P_2 = \frac{1}{2}$, and statement 3 has probability[2] $P_3 = \frac{12}{365}$; so as the probability decreases, the information content increases. Moreover, since the useful statements 2 and 3 are independent, then if you are given statements 2 and 3 together, their information contents should *add*. Moreover, the probability of statements 2 and 3 *both* being true, in the absence of prior information, is $P_2 \times P_3 = \frac{6}{365}$. Since the probability of two independent statements being true is the *product* of their individual probabilities, and since it is natural to assume that information content is *additive*, one is motivated to adopt the definition of information which was proposed by Claude Shannon (1916–2001) as follows:

[1] The statements take as prior information that Newton was born in 1643 and that the dates are expressed according to the calendar which was used in his day. The Gregorian calendar was not adopted in England until 1742.

[2] We are using the fact that 1643 was not a leap year!

The **information** content Q of a statement is defined by

$$Q = -k \log P, \tag{15.1}$$

where P is the probability of the statement and k is a positive constant.[3] If we use $\log_2$ (log to the base 2) for the logarithm in this expression and also $k = 1$, then the information Q is measured in **bits**. If instead we use $\ln \equiv \log_\mathrm{e}$ and choose $k = k_\mathrm{B}$, then we have a definition which, as we shall see, will match what we have found in thermodynamics. In this chapter, we will stick with the former convention since bits are a useful quantity with which to think about information.

Thus, if we have a set of statements with probability P_i, with corresponding information $Q_i = -k \log P_i$, then the average information content S is given by

$$S = \langle Q \rangle = \sum_i Q_i P_i = -k \sum_i P_i \log P_i. \tag{15.2}$$

The average information is called the **Shannon entropy**.

Example 15.1

- A fair die produces outcomes 1, 2, 3, 4, 5 and 6 with probabilities $\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$. The information associated with each outcome is $Q = -k \log \frac{1}{6} = k \log 6$ and the average information content is then $S = k \log 6$. Taking $k = 1$ and using log to the base 2 gives a Shannon entropy of 2.58 bits.
- A biased die produces outcomes 1, 2, 3, 4, 5 and 6 with probabilities $\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{2}$. The information contents associated with the outcomes are $k \log 10$,$k \log 10$, $k \log 10$, $k \log 10$, $k \log 10$ and $k \log 2$. (These are 3.32, 3.32, 3.32, 3.32, 3.32 and 1 bit respectively.) If we take $k = 1$ again, the Shannon entropy is then $S = k(5 \times \frac{1}{10} \log 10 + \frac{1}{2} \log 2) = k(\log \sqrt{20})$ (this is 2.16 bits). This Shannon entropy is smaller than in the case of the fair die.

The Shannon entropy quantifies how much information we gain, on average, following a measurement of a particular quantity. (Another way of looking at it is to say the Shannon entropy quantifies the amount of *uncertainty* we have about a quantity *before* we measure it.) To make these ideas more concrete, let us study a simple example in which there are only two possible outcomes of a particular random process (such as the tossing of a coin, or asking the question ‘will it rain tomorrow?’).

[3] We need k to be a positive constant so that as P goes up, Q goes down.

Example 15.2

What is the Shannon entropy for a Bernoulli[4] trial (a two-outcome random variable) with probabilities P and $1-P$ of the two outcomes?
Solution:

$$S = -\sum_i P_i \log P_i = -P \log P - (1-P)\log(1-P), \qquad (15.3)$$

where we have set $k=1$. This behaviour is sketched in Fig. 15.1. The Shannon entropy has a maximum when $p = \frac{1}{2}$ (greatest uncertainty about the outcome, or greatest information gained, 1 bit, following a trial) and a minimum when $p = 0$ or 1 (least uncertainty about the outcome, or least information gained, 0 bit, following a trial).

The information associated with each of the two possible outcomes is also shown in Fig. 15.1 as dashed lines. The information associated with the outcome having probability P is given by $Q_1 = -\log_2 P$ and decreases as P increases. Clearly when this outcome is very unlikely (P small) the information associated with getting that outcome is very large (Q_1 is many bits of information). However, such an outcome doesn't happen very often so it doesn't contribute much to the average information (i.e. to the Shannon entropy, the solid line in Fig. 15.1). When this outcome is almost certain (P almost 1) it contributes a lot to the average information but has very little information content. For the other outcome, with probability $1-P$, $Q_2 = -\log_2(1-P)$ and the behaviour is simply a mirror image of this. The maximum average information is when $P = 1-P = \frac{1}{2}$ and both outcomes have 1 bit of information associated with them.

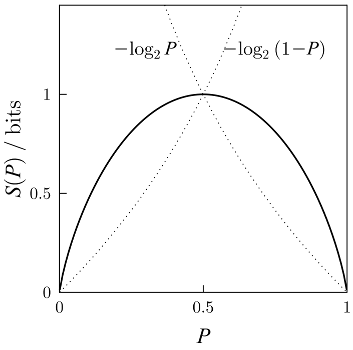

Fig. 15.1 The Shannon entropy of a Bernoulli trial (a two-outcome random variable) with probabilities P and $1-P$ of the two outcomes. The units are chosen so that the Shannon entropy is in bits. Also shown is the information associated with each outcome (dashed lines).

15.2 Information and thermodynamics

Remarkably, the formula for Shannon entropy in eqn 15.2 is identical (apart from whether you take your constant as k or k_{B}) to Gibbs' expression for thermodynamic entropy in eqn 14.48. This gives us a useful perspective on what thermodynamic entropy is. It is a measure of our uncertainty of a system, based upon our limited knowledge of its properties and ignorance about which of its microstates it is in. In making inferences on the basis of partial information, we can assign probabilities on the basis that we maximize entropy subject to the constraints provided by what is known about the system. This is exactly what we did in Example 14.7, when we maximized the Gibbs entropy of an isolated system subject to the constraint that the total energy U was constant; hey presto, we found that we recovered the Boltzmann probability distribution. With this viewpoint, one can begin to understand thermodynamics from an information theory viewpoint.

[4] James Bernoulli (1654–1705).

However, not only does information theory apply to physical systems, but as pointed out by Rolf Landauer (1927–1999), information itself is a physical quantity. Imagine a physical computing device which has stored N bits of information and is connected to a thermal reservoir of temperature T. The bits can be either one or zero. Now we decide to physically erase that information. Erasure must be irreversible. There must be no vestige of the original stored information left in the erased state of the system. Let us erase the information by resetting all the bits to zero.[5] Then this irreversible process reduces the number of states of the system by $\ln 2^N$ and hence the entropy of the system goes down by $Nk_\mathrm{B}\ln 2$, or $k_\mathrm{B}\ln 2$ per bit. For the total entropy of the Universe not to decrease, the entropy of the surroundings must go up by $k_\mathrm{B}\ln 2$ per bit and so we must dissipate heat in the surroundings equal to $k_\mathrm{B}T\ln 2$ per bit erased.

This connection between entropy and information helps us in our understanding of Maxwell's demon discussed in Section 14.7. By performing computations about molecules and their velocities, the demon has to store information. Each bit of information is associated with entropy, as becomes clear when the demon has to free up some space on its hard disk to continue computing. The process of erasing one bit of information gives rise to an increase of entropy of $k_\mathrm{B}\ln 2$. If Maxwell's demon reverses the Joule expansion of 1 mole of gas, it might therefore seem like it has decreased the entropy of the Universe by $N_\mathrm{A}k_\mathrm{B}\ln 2 = R\ln 2$, but it will have had to store at least N_A bits of information to do this. Assuming that Maxwell's demons only have on-board a storage capacity of a few hundred gigabytes, which is much less than N_A bits, the demon will have had to erase its disk many many times in the process of its operation, thus leading to an increase in entropy of the Universe which at least equals, and probably outweighs, the decrease of entropy of the Universe it was aiming to achieve.

If the demon is somehow fitted with a vast on-board memory so that it doesn't have to erase its memory to do the computation, then the increase in entropy of the Universe can be delayed until the demon needs to free up some memory space. Eventually, one supposes, as the demon begins to age and becomes forgetful, the Universe will reclaim all that entropy!

15.3 Data compression

Information must be stored, or sometimes transmitted from one place to another. It is therefore useful if it can be compressed down to its minimum possible size. This really begs the question what the actual irreducible amount of *real* information in a particular block of data really is; many messages, political speeches, and even sometimes book chapters, contain large amounts of extraneous padding that is not really needed. Of course, when we compress a file down on a computer we often get something which is unreadable to human beings. The English

[5] We could equally well reset the bits to one.

language has various quirks, such as when you see a letter ‘q’ it is almost always followed by a ‘u’, so is that second ‘u’ really needed when you know it is coming? A good data compression algorithm will get rid of extra things like that, plus much more besides. Hence, the question of how many bits are in a given source of data seems like a useful question for computer scientists to attempt to answer; in fact we will see it has implications for physics!

We will here not prove **Shannon’s noiseless channel coding theorem**, but motivate it and then state it.

Example 15.3

Let us consider the simplest case in which our data are stored in the form of the binary digits ‘0’ and ‘1’. Let us further suppose that the data contain ‘0’ with probability P and ‘1’ with probability $1 - P$. If $P = \frac{1}{2}$ then our data cannot really be compressed, as each bit of data contains real information. Let us now suppose that $P = 0.9$ so that the data contain more 0’s than 1’s. In this case, the data contain less information, and it is not hard to find a way of taking advantage of this. For example, let us read the data into our compression algorithm in pairs of bits, rather than one bit at a time, and make the following transformations:

$$\begin{array}{lcl} 00 & \rightarrow & 0 \\ 10 & \rightarrow & 10 \\ 01 & \rightarrow & 110 \\ 11 & \rightarrow & 1110 \end{array}$$

In each of the transformations, we end on a single ‘0’, which lets the decompression algorithm know that it can start reading the next sequence. Now, of course, although the pair of symbols ‘00’ have been compressed to ‘0’, saving a bit, the pair of symbols ‘01’ has been enlarged to ‘110’ and ‘11’ has been even more enlarged to ‘1110’, costing 1 extra or 2 extra bits respectively. However, ‘00’ is very likely to occur (probability 0.81) while ‘01’ and ‘11’ are much less likely to occur (probabilities 0.09 and 0.01 respectively), so overall we save bits using this compression scheme.

This example gives us a clue as to how to compress data more generally. The aim is to identify in a sequence of data what the typical sequences are and then efficiently code only those. When the amount of data becomes very large, then anything other than these typical sequences is very unlikely to occur. Because there are fewer typical sequences than there are sequences in general, a saving can be made. Hence, let us divide up some data into sequences of length n. Assuming the elements in the data do not depend on each other, then the probability of finding a sequence $x_1, x_2, \ldots, x_n$ is

$$P(x_1, x_2, \ldots, x_n) = P(x_1)P(x_2)\ldots P(x_n) \approx P^{nP}(1-P)^{n(1-P)}, \quad (15.4)$$

for typical sequences. Taking logarithms to base 2 of both sides gives

$$-\log_2 P(x_1, x_2, \ldots, x_n) \approx -nP\log_2 P - n(1-P)\log_2(1-P) = nS, \quad (15.5)$$

where S is the entropy for a Bernoulli trial with probability P. Hence

$$P(x_1, x_2, \ldots, x_n) \approx \frac{1}{2^{nS}}. \quad (15.6)$$

This shows that there are at most only 2^{nS} typical sequences and hence it only requires nS bits to code them. As n becomes larger, and the typical sequences become longer, the possibility of this scheme failing becomes smaller and smaller.

A compression algorithm will take a typical sequence of n terms $x_1, x_2, \ldots, x_n$ and turn them into a string of length nR. Hence, the smaller R is, the greater the compression. Shannon's noiseless channel coding theorem states that if we have a source of information with entropy S, and if $R > S$, then there exists a reliable compression scheme of compression factor R. Conversely, if $R < S$ then any compression scheme will not be reliable. Thus the entropy S sets the ultimate compression limit on a set of data.

15.4 Quantum information

This section shows how the concept of information can be extended to quantum systems and assumes familiarity with the main results of quantum mechanics.

In this chapter we have seen that in classical systems the information content is connected with the probability. In quantum systems, these probabilities are replaced by **density matrices**. A density matrix is used to describe the statistical state of a quantum system, as can arise for a quantum system in thermal equilibrium at finite temperature. A summary of the main results concerning density matrices is given in the box on page 159.

For quantum systems, the information is represented by the operator $-k\log\rho$, where ρ is the density matrix; as before we take $k = 1$. Hence the average information, or entropy, would be $\langle -\log\rho\rangle$. This leads to the definition of the **von Neumann entropy** S as[6]

$$S(\rho) = -\mathrm{Tr}(\rho\log\rho). \quad (15.7)$$

If the eigenvalues of ρ are $\lambda_1, \lambda_2 \ldots$, then the von Neumann entropy becomes

$$S(\rho) = -\sum_i \lambda_i \log\lambda_i, \quad (15.8)$$

which looks like the Shannon entropy.

[6] The operator Tr means the **trace** of the following matrix, i.e. the sum of the diagonal elements.

The density matrix:

- If a quantum system is in one of a number of states $|\psi_i\rangle$ with probability P_i, then the density matrix ρ for the system is defined by
$$\rho = \sum_i P_i |\psi_i\rangle\langle\psi_i|. \qquad (15.9)$$
- As an example, think of a three-state system and think of $|\psi_1\rangle$ as a column vector $\begin{pmatrix}1\\0\\0\end{pmatrix}$, and hence $\langle\psi_1|$ as a row vector $(1, 0, 0)$, and similarly for $|\psi_2\rangle$, $\langle\psi_2|$, $|\psi_3\rangle$ and $\langle\psi_3|$. Then
$$\begin{aligned}\rho &= P_1\begin{pmatrix}1&0&0\\0&0&0\\0&0&0\end{pmatrix} + P_2\begin{pmatrix}0&0&0\\0&1&0\\0&0&0\end{pmatrix} + P_3\begin{pmatrix}0&0&0\\0&0&0\\0&0&1\end{pmatrix}\\ &= \begin{pmatrix}P_1&0&0\\0&P_2&0\\0&0&P_3\end{pmatrix} \qquad (15.10)\end{aligned}$$
This form of the density matrix looks very simple, but this is only because we have expressed it in a very simple basis.
- If $P_j \neq 0$ and $P_{i\neq j} = 0$, then the system is said to be in a **pure state** and ρ can be written in the simple form
$$\rho = |\psi_j\rangle\langle\psi_j|. \qquad (15.11)$$
Otherwise, it is said to be in a **mixed state**.
- One can show that the expectation value $\langle\hat{A}\rangle$ of a quantum mechanical operator $\hat{A}$ is equal to
$$\langle\hat{A}\rangle = \mathrm{Tr}(\hat{A}\rho). \qquad (15.12)$$
- One can also prove that
$$\mathrm{Tr}\rho = 1, \qquad (15.13)$$
where $\mathrm{Tr}\rho$ means the trace of the density matrix. This expresses the fact that the sum of the probabilities must equal unity, and is in fact a special case of eqn 15.12 setting $\hat{A} = 1$.
- One can also show that $\mathrm{Tr}\rho^2 \leq 1$ with equality if and only if the state is pure.
- For a system in thermal equilibrium at temperature T, P_i is given by the Boltzmann factor $\mathrm{e}^{-\beta E_i}$ where E_i is an eigenvalue of the Hamiltonian $\hat{H}$. The **thermal density matrix** ρ_{th} is
$$\rho_{\mathrm{th}} = \sum_i \mathrm{e}^{-\beta E_i}|\psi_i\rangle\langle\psi_i| = \exp(-\beta\hat{H}). \qquad (15.14)$$

Example 15.4

Show that the entropy of a pure state[7] is zero. How can you maximize the entropy?

Solution:

(i) As shown in the box on page 159, the trace of the density matrix is equal to one ($\mathrm{Tr}\rho = 1$), and hence

$$\sum \lambda_i = 1. \tag{15.15}$$

For a pure state only one eigenvalue will be one and all the other eigenvalues will be zero, and hence[8] $S(\rho) = 0$, i.e. the entropy of a pure state is zero. This is not surprising, since for a pure state there is no 'uncertainty' about the state of the system.

(ii) The entropy is maximized when $\lambda_i = 1/n$ for all i, where n is the dimension of the density matrix. In this case, the entropy is $S(\rho) = n \times (-\frac{1}{n}\log\frac{1}{n}) = \log n$. This corresponds to there being maximal uncertainty in its precise state.

Classical information is made up only of sequences of 0's and 1's (in a sense, all information can be broken down into a series of 'yes/no' questions). Quantum information is comprised of quantum bits (known as **qubits**), which are two-level quantum systems which can be represented by linear combinations[9] of the states $|0\rangle$ and $|1\rangle$. Quantum mechanical states can also be *entangled* with each other. The phenomenon of **entanglement**[10] has no classical counterpart. Quantum information therefore also contains entangled superpositions such as $(|01\rangle+|10\rangle)/\sqrt{2}$. Here the quantum states of two objects must be described with reference to each other; measurement of the first bit in the sequence to be a 0 forces the second bit to be 1; if the measurement of the first bit gives a 1, the second bit has to be 0; these correlations persist in an entangled quantum system even if the individual objects encoding each bit are spatially separated. Entangled systems cannot be described by pure states of the individual subsystems, and this is where entropy plays a rôle, as a quantifier of the degree of mixing of states. If the overall system is pure, the entropy of its subsystems can be used to measure its degree of entanglement with the other subsystems.[11]

In this text we do not have space to provide many details about the subject of quantum information, which is a rapidly developing area of current research. Suffice to say that the processing of information in quantum mechanical systems has some intriguing facets which are not present in the study of classical information. Entanglement of bits is just one example. As another example, the **no-cloning theorem** states that it is impossible to make a copy of non-orthogonal quantum mechanical states (for classical systems, there is no physical mechanism to stop you copying information, only copyright laws). All of these features lead to the very rich structure of quantum information theory.

[7] A pure state is defined in the box on page 159.

[8] Note that we take $0 \ln 0 = 0$.

[9] An arbitary qubit can be written as $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ where $|\alpha|^2 + |\beta|^2 = 1$.

[10] Einstein called entanglement 'spooky action at a distance', and used it to argue against the Copenhagen interpretation of quantum mechanics and show that quantum mechanics is incomplete.

[11] It turns out that a unitary operator, such as the time-evolution operator, acting on a state leave the entropy unchanged. This is akin to our results in thermodynamics that reversibility is connected with the preservation of entropy.

Chapter summary

- The information Q is given by $Q = -\ln P$ where P is the probability.
- The entropy is the average information $S = \langle Q \rangle = -\sum_i P_i \log P_i$.
- The quantum mechanical generalization of this is the von Neumann entropy given by $S(\rho) = -\mathrm{Tr}(\rho \log \rho)$ where ρ is the density matrix.

Further reading

The results which we have stated in this chapter concerning Shannon's coding theorems, and which we considered only for the case of Bernoulli trials, i.e. for binary outputs, can be proved for the general case. Shannon also studied communication over noisy channels in which the presence of noise randomly flips bits with a certain probability. In this case it is also possible to show how much information can be reliably transmitted using such a channel (essentially how many times you have to 'repeat' the message to get yourself 'heard', though actually this is done using error-correcting codes). Further information may be found in Feynman (1996) and Mackay (2003). An excellent account of the problem of Maxwell's demon may be found in Leff and Rex (2003). Quantum information theory has become a very hot research topic in the last few years and an excellent introduction is Nielsen and Chuang (2000).

Exercises

(15.1) In a typical microchip, a bit is stored by a 5 fF capacitor using a voltage of 3 V. Calculate the energy stored in eV per bit and compare this with the minimum heat dissipation by erasure, which is $k_\mathrm{B}T \ln 2$ per bit, at room temperature.

(15.2) A particular logic gate takes two binary inputs A and B and has two binary outputs A' and B'. Its truth table is

A	B	A'	B'
0	0	1	1
0	1	1	0
1	0	0	1
1	1	0	0

and this is produced by $A' = \mathrm{NOT}\,A$ and $B' = \mathrm{NOT}\,B$. The input has a Shannon entropy of 2 bits. Show that the output has a Shannon entropy of 2 bits.

A second logic gate has a truth table given by

A	B	A'	B'
0	0	0	0
0	1	1	0
1	0	1	0
1	1	1	1

This can be achieved using $A' = A\,\mathrm{OR}\,B$ and $B' = A\,\mathrm{AND}\,B$. Show that the output now has an entropy of $\frac{3}{2}$ bits. What is the difference between the two logic gates?

(15.3) Maximize the Shannon entropy $S = -k\sum_i P_i \log P_i$ subject to the constraints that $\sum P_i = 1$ and $\langle f(x) \rangle = \sum P_i f(x_i)$ and show that

$$P_i = \frac{1}{Z(\beta)} \mathrm{e}^{-\beta f(x_i)}, \qquad (15.16)$$

$$Z(\beta) = \sum \mathrm{e}^{-\beta f(x_i)}, \qquad (15.17)$$

$$\langle f(x) \rangle = -\frac{\mathrm{d}}{\mathrm{d}\beta} \ln Z(\beta). \qquad (15.18)$$

(15.4) Noise in a communication channel flips bits at random with probability P. Argue that the entropy associated with this process is

$$S = -P\log P - (1-P)\log(1-P). \qquad (15.19)$$

It turns out that the rate R at which we can pass information along this noisy channel is $1-S$. (This is an application of Shannon's noisy channel coding theorem, and a nice proof of this theorem is given on page 548 of Nielsen and Chuang (2000).)

(15.5) (a) The **relative entropy** measures the closeness of two probability distributions P and Q and is defined by

$$S(P||Q) = \sum P_i \log\left(\frac{P_i}{Q_i}\right) = -S_p - \sum P_i \log Q_i, \qquad (15.20)$$

where $S_p = -\sum P_i \log P_i$. Show that $S(P||Q) \geq 0$ with equality if and only if $P_i = Q_i$ for all i.

(b) If i takes N values with probability P_i, then show that

$$S(P||Q) = -S_P + \log N \qquad (15.21)$$

where $Q_i = 1/N$ for all i. Hence show that

$$S_P \leq \log N \qquad (15.22)$$

with equality if and only if P_i is uniformly distributed between all N outcomes.