6 Entropy & the Boltzmann Law

$$S = k \log W$$

What Is Entropy?

Carved on the tombstone of Ludwig Boltzmann in the Zentralfriedhof (central cemetery) in Vienna is the inscription

$$S = k \log W. \tag{6.1}$$

This equation is the historical foundation of statistical mechanics. It connects the microscopic and macroscopic worlds. It defines the *entropy* S, a macroscopic quantity, in terms of the multiplicity W of the microscopic degrees of freedom of a system. For thermodynamics, $k = 1.380662 \times 10^{-23} \mathrm{J\,K^{-1}}$ is a quantity called Boltzmann's constant, and Boltzmann's inscription refers to the natural logarithm, $\log_e = \ln$.

In Chapters 2 and 3 we used simple models to illustrate that the composition of coin flips, the expansion of gases, the tendency of particles to mix, rubber elasticity, and heat flow can be predicted by the principle that systems tend toward their states of maximum multiplicity W. However, states that maximize W will also maximize W^2 or $15W^3+5$ or $k \ln W$, where k is any positive constant. Any monotonic function of W will have a maximum where W has a maximum. In particular, states that maximize W also maximize the entropy, $S = k \ln W$. Why does this quantity deserve special attention as a prediction principle, and why should it have this particular mathematical form?

In this chapter, we use a few general principles to show why the entropy must have this mathematical form. But first we switch our view of entropy from

a multiplicity perspective to a probability perspective that is more general. In the probability perspective, the entropy is given as

$$\frac{S}{k} = -\sum_{i=1}^{t} p_i \ln p_i. \tag{6.2}$$

Let's see how Equation (6.2) is related to Equation (6.1). Roll a t-sided die N times. The multiplicity of outcomes is given by Equation (1.18) (see page 12),

$$W = \frac{N!}{n_1! n_2! \ldots n_t!},$$

where n_i is the number of times that side i appears face up. Use Stirling's approximation $x! \approx (x/e)^x$ (page 56), and define the probabilities $p_i = n_i/N$, to convert Equation (1.18) to

$$\begin{aligned} W &= \frac{(N/e)^N}{(n_1/e)^{n_1}(n_2/e)^{n_2}\ldots(n_t/e)^{n_t}} \\ &= \frac{N^N}{n_1^{n_1} n_2^{n_2} \ldots n_t^{n_t}} = \frac{1}{p_1^{n_1} p_2^{n_2} \ldots p_t^{n_t}}. \end{aligned} \tag{6.3}$$

Take the logarithm and divide by N to get

$$\ln W = -\sum_{i=1}^{t} n_i \ln p_i \quad \Longrightarrow \quad \frac{1}{N} \ln W = -\sum_{i=1}^{t} p_i \ln p_i = \frac{S_N}{Nk}, \tag{6.4}$$

where S_N indicates that this is the entropy for N trials, and the entropy per trial is $S = S_N/N$. For this dice problem and the counting problems in Chapters 2 and 3, the two expressions for the entropy, Equations (6.2) and (6.1), are equivalent. The flattest distributions are those having maximum multiplicity W in the absence of constraints. For example, in N coin flips, the multiplicity $W = N!/[(n_H!)(N - n_H)!]$ is maximized when $n_H/N \approx n_T/N \approx 0.5$, that is, when the probabilities of heads and tails are as nearly equal as possible.

There are different types of entropy, depending on the degrees of freedom of the system. Examples 2.2 and 2.3 describe *translational* freedom due to the different positions of particles in space. In the next example we apply Equation (6.2) to the *rotational* or *orientational* entropy of *dipoles*. We show that flatter probability distributions have higher entropy than more peaked distributions.

EXAMPLE 6.1 Dipoles tend to orient randomly. Objects with distinguishable heads and tails such as magnets, chemically asymmetrical molecules, electrical dipoles with (+) charges at one end and (−) charges at the other, or even pencils with erasers at one end have rotational freedom as well as translational freedom. They can orient.

Spin a pencil on a table N times. Each time it stops, the pencil points in one of four possible directions: toward the quadrant facing north (n), east (e), south (s), or west (w). Count the number of times that the pencil points in each direction; label those numbers n_n, n_e, n_s, and n_w. Spinning a pencil and counting orientations is analogous to rolling a die with four sides labeled n, e, s or w. Each roll of that die determines the orientation of one pencil or

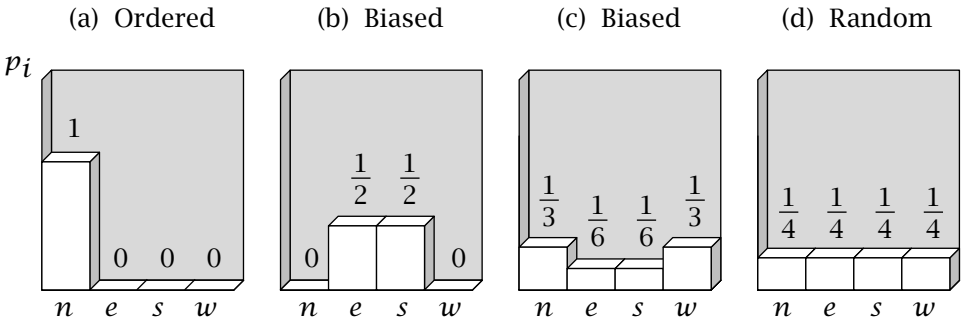

Figure 6.1 Spin a hundred pencils. Here are four (of a large number) of possible distributions of outcomes. (a) All pencils could point north (n). This is the most *ordered* distribution, $S/k = -1 \ln 1 = 0$. (b) Half the pencils could point east (e) and half could point south (s). This distribution has more entropy than (a), $S/k = -2(1/2 \ln 1/2) = 0.69$. (c) One-third of the pencils could point n, and one-third w, one-sixth e, and one-sixth s. This distribution has even more entropy, $S/k = -2(1/3 \ln 1/3 + 1/6 \ln 1/6) = 1.33$. (d) One-quarter of the pencils could point in each of the four possible directions. This is the distribution with highest entropy, $S/k = -4(1/4 \ln 1/4) = 1.39$.

dipole. N die rolls correspond to the orientations of N dipoles. The number of configurations for systems with N trials, distributed with any set of outcomes $\{n_1, n_2, \ldots, n_t\}$, where $N = \sum_{i=1}^{t} n_i$, is given by the multiplicity Equation (1.18): $W(n_1, n_2, \ldots, n_t) = N!/(n_1! n_2! \cdots n_t!)$. The number of different configurations of the system with a given composition n_n, n_e, n_s, and n_w is

$$W(N, n_n, n_e, n_s, n_w) = \frac{N!}{n_n! n_e! n_s! n_w!}.$$

The probabilities that a pencil points in each of the four directions are

$$p(n) = \frac{n_n}{N}, \qquad p(e) = \frac{n_e}{N}, \qquad p(s) = \frac{n_s}{N}, \qquad \text{and} \qquad p(w) = \frac{n_w}{N}.$$

Figure 6.1 shows some possible distributions of outcomes. Each distribution function satisfies the constraint that $p(n) + p(e) + p(s) + p(w) = 1$. You can compute the entropy per spin of the pencil of any of these distributions by using Equation (6.2), $S/k = -\sum_{i=1}^{t} p_i \ln p_i$. The absolute entropy is never negative, that is $S \geq 0$.

Flat distributions have high entropy. Peaked distributions have low entropy. When all pencils point in the same direction, the system is perfectly ordered and has the lowest possible entropy, $S = 0$. Entropy does not depend on being able to order the categories along an x-axis. For pencil orientations, there is no difference between the x-axis sequence *news* and *esnw*. To be in a state of low entropy, it does not matter which direction the pencils point in, just that they all point the same way. The flattest possible distribution has the highest possible entropy, which increases with the number of possible outcomes. In Figure 6.1 we have four states: the flattest distribution has $S/k = -4(1/4) \ln(1/4) = \ln 4 = 1.39$. In general, when there are t states, the flat distribution has entropy $S/k = \ln t$. Flatness in a distribution corresponds to *disorder* in a system.

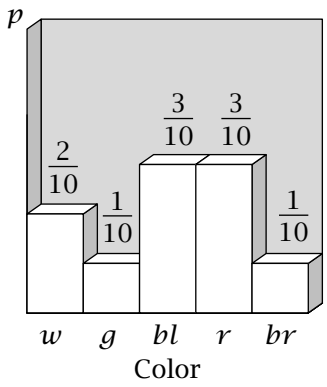

Figure 6.2 The entropy can be computed for *any* distribution function, even for colors of socks: white (w), green (g), black (bl), red (r), and brown (br).

The concept of entropy is broader than statistical thermodynamics. It is a property of any distribution function, as the next example shows.

EXAMPLE 6.2 Colors of socks. Suppose that on a given day, you sample 30 students and find the distribution of the colors of the socks they are wearing (Figure 6.2). The entropy of this distribution is

$$S/k = -0.2\ln 0.2 - 0.6\ln 0.3 - 0.2\ln 0.1 = 1.50.$$

For this example and Example 6.1, k should not be Boltzmann's constant. Boltzmann's constant is appropriate only when you need to put entropy into units that interconvert with energy, for thermodynamics and molecular science. For other types of probability distributions, k is chosen to suit the purposes at hand, so $k = 1$ would be simplest here. The entropy function just reports the relative flatness of a distribution function. The limiting cases are the most ordered, $S = 0$ (everybody wears the same color socks) and the most disordered, $S/k = \ln t = \ln 5 = 1.61$ (all five sock colors are equally likely).

Why should the entropy have the form of either Equation (6.1) or Equation (6.2)? Here is a simple justification. A deeper argument is given in 'Optional Material,' page 89.

The Simple Justification for $S = k \ln W$

Consider a thermodynamic system having two subsystems, A and B, with multiplicities W_A and W_B respectively. The multiplicity of the total system will be the product $W_{\text{total}} = W_A W_B$. Thermodynamics requires that entropies be *extensive*, meaning that the system entropy is the sum of subsystem entropies, $S_{\text{total}} = S_A + S_B$. The logarithm function satisfies this requirement. If $S_A = k \ln W_A$ and $S_B = k \ln W_B$, then $S_{\text{total}} = k \ln W_{\text{total}} = k \ln W_A W_B = k \ln W_A + k \ln W_B = S_A + S_B$. This argument illustrates why S should be a logarithmic function of W.

Let's use $S/k = -\sum_i p_i \ln p_i$ to derive the exponential distribution law, called the Boltzmann distribution law, that is at the center of statistical thermodynamics. The Boltzmann distribution law describes the energy distributions of atoms and molecules.

Underdetermined Distributions

In the rest of this chapter, we illustrate the principles that we need by concocting a class of problems involving die rolls and coin flips instead of molecules. How would you know if a die is biased? You could roll it N times and count the numbers of **1**'s, **2**'s, ..., **6**'s. If the probability distribution were perfectly flat, the die would not be biased. You could use the same test for the orientations of pencils or to determine whether atoms or molecules have biased spatial orientations or bond angle distributions. However the options available to molecules are usually so numerous that you could not possibly measure each one. In statistical mechanics you seldom have the luxury of knowing the full distribution, corresponding to all six numbers p_i for $i = 1, 2, 3, \ldots, 6$ on die rolls.

Therefore, as a prelude to statistical mechanics, let's concoct dice problems that are underdetermined in the same way as the problems of molecular

science. Suppose that you do not know the distribution of all six possible outcomes. Instead, you know only the total score (or equivalently, the average score per roll) on the N rolls. In thousands of rolls, the average score per roll of an unbiased die will be $3.5 = (1 + 2 + 3 + 4 + 5 + 6)/6$. If you observe that the average score is 3.5, it is evidence (but not proof)[1] that the distribution is unbiased. In that case, your best guess consistent with the evidence is that the distribution is flat. All outcomes are equally likely.

However, if you observe the average score per roll is 2.0, then you must conclude that every outcome from **1** to **6** is *not* equally likely. You know only that low numbers are somehow favored. This one piece of data—the total score—is not sufficient to predict all six unknowns of the full distribution function. So we aim to do the next best thing. We aim to predict the least biased distribution function that is consistent with the known measured score. This distribution is predicted by the maximum-entropy principle.

Maximum Entropy Predicts Flat Distributions When There Are No Constraints

The entropy function is used to predict probability distributions. Here we show that the tendency toward the maximum entropy is a tendency toward maximum flatness of a probability distribution function when there are no constraints. Roll an unbiased t-sided die many times. Because the probabilities must sum to one,

$$\sum_{i=1}^{t} p_i = 1 \quad \Longrightarrow \quad \sum_{i=1}^{t} dp_i = 0. \tag{6.5}$$

We seek the distribution, $(p_1, p_2, \ldots, p_t) = (p_1^*, p_2^*, \ldots, p_t^*)$, that causes the entropy function, $S(p_1, p_2, \ldots, p_t) = -k \sum_i p_i \ln p_i$ to be at its maximum possible value, subject to the normalization Equation (6.5). For this problem, let $k = 1$. To solve it by the Lagrange multiplier method (see Equation (5.35)), with multiplier α for constraint Equation (6.5), we want

$$\sum_{i=1}^{t} \left[\left(\frac{\partial S}{\partial p_i} \right)_{p_{j \neq i}} - \alpha \right] dp_i = 0. \tag{6.6}$$

Set the term inside the brackets equal to zero. The derivative of the entropy function gives $(\partial S/\partial p_i) = -1 - \ln p_i$ (since you are taking the derivative with respect to one particular p_i with all the other p's, $j \neq i$, held constant), so the solution is

$$-1 - \ln p_i - \alpha = 0 \quad \Longrightarrow \quad p_i^* = e^{(-1-\alpha)}. \tag{6.7}$$

To put this into a simpler form, divide Equation (6.7) by $\sum_i p_i^* = 1$ to get

$$\frac{p_i^*}{\sum_{i=1}^{t} p_i^*} = \frac{e^{(-1-\alpha)}}{t e^{(-1-\alpha)}} = \frac{1}{t}. \tag{6.8}$$

[1] For example, that score could also arise from 50% **2**'s and 50% **5**'s.

Maximizing the entropy predicts that when there is no bias, all outcomes are equally likely. However, what if there is bias? The next section shows how maximum entropy works in that case.

Maximum Entropy Predicts Exponential Distributions When There Are Constraints

Roll a die having t sides, with faces numbered $i = 1, 2, 3, \ldots, t$. You do not know the distribution of outcomes of each face, but you know the total score after N rolls. You want to predict the distribution function. When side i appears face up, the score is ε_i. The total score after N rolls will be $E = \sum_{i=1}^{t} \varepsilon_i n_i$, where n_i is the number of times that you observe face i. Let $p_i = n_i/N$ represent the fraction of the N rolls on which you observe face i. The average score per roll $\langle\varepsilon\rangle$ is

$$\langle\varepsilon\rangle = \frac{E}{N} = \sum_{i=1}^{t} p_i \varepsilon_i. \tag{6.9}$$

What is the expected distribution of outcomes $(p_1^*, p_2^*, \ldots, p_t^*)$ consistent with the observed average score $\langle\varepsilon\rangle$? We seek the distribution that maximizes the entropy, Equation (6.2), subject to two conditions: (1) that the probabilities sum to one and (2) that the average score agrees with the observed value $\langle\varepsilon\rangle$,

$$g(p_1, p_2, \ldots, p_t) = \sum_{i=1}^{t} p_i = 1 \quad \Longrightarrow \quad \sum_{i=1}^{t} dp_i = 0, \tag{6.10}$$

and

$$h(p_1, p_2, \ldots, p_t) = \langle\varepsilon\rangle = \sum_{i=1}^{t} p_i \varepsilon_i \quad \Longrightarrow \quad \sum_{i=1}^{t} \varepsilon_i dp_i = 0. \tag{6.11}$$

The solution is given by the method of Lagrange multipliers (pages 69–73):

$$\left(\frac{\partial S}{\partial p_i}\right) - \alpha\left(\frac{\partial g}{\partial p_i}\right) - \beta\left(\frac{\partial h}{\partial p_i}\right) = 0 \qquad \text{for } i = 1, 2, \ldots, t, \tag{6.12}$$

where α and β are the unknown multipliers. The partial derivatives are evaluated for each p_i:

$$\left(\frac{\partial S}{\partial p_i}\right) = -1 - \ln p_i, \quad \left(\frac{\partial g}{\partial p_i}\right) = 1, \quad \text{and} \left(\frac{\partial h}{\partial p_i}\right) = \varepsilon_i. \tag{6.13}$$

Substitute Equations (6.13) into Equation (6.12) to get t equations of the form

$$-1 - \ln p_i^* - \alpha - \beta\varepsilon_i = 0, \tag{6.14}$$

where the p_i^*'s are the values of p_i that maximize the entropy. Solve Equations (6.14) for each p_i^*:

$$p_i^* = e^{(-1-\alpha-\beta\varepsilon_i)}. \tag{6.15}$$

To eliminate α in Equation (6.15), use Equation (6.10) to divide both sides by one. The result is an **exponential distribution law**:

$$p_i^* = \frac{p_i^*}{\sum_{i=1}^{t} p_i^*} = \frac{e^{(-1-\alpha)}e^{-\beta\varepsilon_i}}{\sum_{i=1}^{t} e^{(-1-\alpha)}e^{-\beta\varepsilon_i}} = \frac{e^{-\beta\varepsilon_i}}{\sum_{i=1}^{t} e^{-\beta\varepsilon_i}}. \tag{6.16}$$

In statistical mechanics, this is called the **Boltzmann distribution law** and the quantity in the denominator is called the **partition function** q,

$$q = \sum_{i=1}^{t} e^{-\beta\varepsilon_i}. \tag{6.17}$$

Using Equations (6.11) and (6.16) you can express the average score per roll $\langle\varepsilon\rangle$ (Equation (6.9)) in terms of the distribution,

$$\langle\varepsilon\rangle = \sum_{i=1}^{t} \varepsilon_i p_i^* = \frac{1}{q}\sum_{i}^{t} \varepsilon_i e^{-\beta\varepsilon_i}. \tag{6.18}$$

The next two examples show how Equation (6.18) predicts all t of the p_i^*'s from the one known quantity, the average score.

EXAMPLE 6.3 Finding bias in dice by using the exponential distribution law. Here we illustrate how to predict the maximum entropy distribution when an average score is known. Suppose a die has $t = 6$ faces and the scores equal the face indices, $\varepsilon(i) = i$. Let $x = e^{-\beta}$. Then Equation (6.17) gives $q = x + x^2 + x^3 + x^4 + x^5 + x^6$, and Equation (6.16) gives

$$p_i^* = \frac{x^i}{\sum_{i=1}^{6} x^i} = \frac{x^i}{x + x^2 + x^3 + x^4 + x^5 + x^6}. \tag{6.19}$$

From the constraint Equation (6.18), you have

$$\langle\varepsilon\rangle = \sum_{i=1}^{6} i p_i^* = \frac{x + 2x^2 + 3x^3 + 4x^4 + 5x^5 + 6x^6}{x + x^2 + x^3 + x^4 + x^5 + x^6}. \tag{6.20}$$

You have a polynomial, Equation (6.20), that you must solve for the one unknown x (a method for solving polynomials like Equation (6.20) is given on page 55). You begin with knowledge of $\langle\varepsilon\rangle$. Compute the value x^* that solves Equation (6.20). Then substitute x^* into Equations (6.19) to give the distribution function $(p_1^*, p_2^*, \ldots, p_t^*)$.

For example, if you observe the average score $\langle\varepsilon\rangle = 3.5$, then $x = 1$ satisfies Equation (6.20), predicting $p_i^* = 1/6$ for all i, indicating that the die is unbiased and has a flat distribution (see Figure 6.3(a)).

If, instead, you observe the average score is $\langle\varepsilon\rangle = 3.0$, then $x = 0.84$ satisfies Equation (6.20), and you have $q = 0.84 + 0.84^2 + 0.84^3 + 0.84^4 + 0.84^5 + 0.84^6 = 3.41$. The probabilities are $p_1 = 0.84/3.41 = 0.25$, $p_2 = 0.84^2/3.41 = 0.21$, $p_3 = 0.84^3/3.41 = 0.17$, and so on, as shown in Figure 6.3(b).

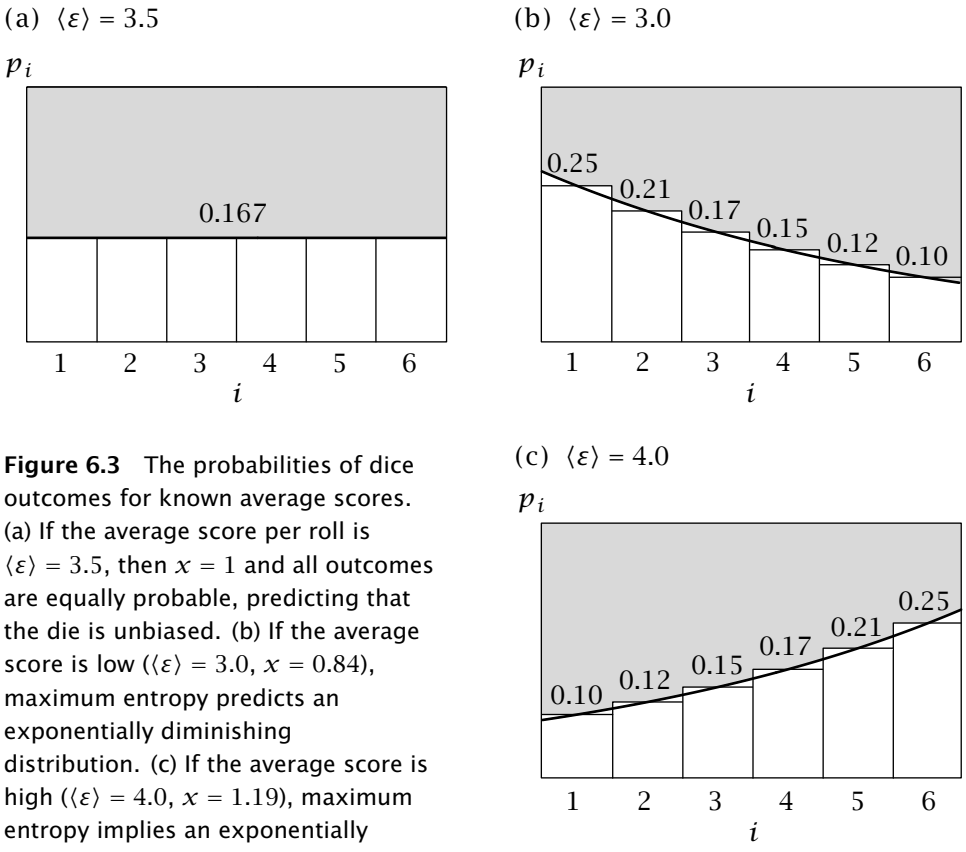

Figure 6.3 The probabilities of dice outcomes for known average scores. (a) If the average score per roll is $\langle\varepsilon\rangle = 3.5$, then $x = 1$ and all outcomes are equally probable, predicting that the die is unbiased. (b) If the average score is low ($\langle\varepsilon\rangle = 3.0$, $x = 0.84$), maximum entropy predicts an exponentially diminishing distribution. (c) If the average score is high ($\langle\varepsilon\rangle = 4.0$, $x = 1.19$), maximum entropy implies an exponentially increasing distribution.

If you observe $\langle\varepsilon\rangle < 3.5$, then the maximum-entropy principle predicts an exponentially decreasing distribution (see Figure 6.3(b)), with more **1**'s than **2**'s, more **2**'s than **3**'s, etc. If you observe $\langle\varepsilon\rangle > 3.5$, then maximum entropy predicts an exponentially increasing distribution (see Figure 6.3(c)): more **6**'s than **5**'s, more **5**'s than **4**'s, etc. For any given value of $\langle\varepsilon\rangle$, the exponential or flat distribution gives the most impartial distribution consistent with that score. The flat distribution is predicted either if the average score is 3.5 or if you have no information at all about the score.

EXAMPLE 6.4 Biased coins? The exponential distribution again. Let's determine a coin's bias. A coin is just a die with two sides, $t = 2$. Score tails $\varepsilon_T = 1$ and heads $\varepsilon_H = 2$. The average score per toss $\langle\varepsilon\rangle$ for an unbiased coin would be 1.5.

Again, to simplify, write the unknown Lagrange multiplier β in the form $x = e^{-\beta}$. In this notation, the partition function Equation (6.17) is $q = x + x^2$. According to Equation (6.16), the exponential distribution law for this two-state system is

$$p_T^* = \frac{x}{x + x^2} \qquad \text{and} \qquad p_H^* = \frac{x^2}{x + x^2}. \tag{6.21}$$

From the constraint Equation (6.18) you have

$$\langle \varepsilon \rangle = 1p_T^* + 2p_H^* = \frac{x + 2x^2}{x + x^2} = \frac{1 + 2x}{1 + x}.$$

Rearranging gives

$$x = \frac{\langle \varepsilon \rangle - 1}{2 - \langle \varepsilon \rangle}.$$

If you observe the average score to be $\langle \varepsilon \rangle = 1.5$, then $x = 1$, and Equation (6.21) gives $p_T^* = p_H^* = 1/2$. The coin is fair. If instead you observed $\langle \varepsilon \rangle = 1.2$, then $x = 1/4$, and you have $p_H^* = 1/5$ and $p_T^* = 4/5$.

There are two situations that will predict a flat distribution function. First, it will be flat if $\langle \varepsilon \rangle$ equals the value expected from a uniform distribution. For example, if you observe $\langle \varepsilon \rangle = 3.5$ in Example 6.3, maximum entropy predicts a flat distribution. Second, if there is no constraint at all, you expect a flat distribution. By the maximum-entropy principle, having no information is the same as expecting a flat distribution.

On page 84, we gave a simple rationalization for why S should be a logarithmic function of W. Now we give a deeper justification for the functional form, $S/k = -\sum_i p_i \ln p_i$. You might ask why entropy must be extensive, and why entropy, which we justified on thermodynamic grounds, also applies to a broad range of problems outside molecular science. Should $S = k \ln W$ also apply to interacting systems? The following section is intended to address questions such as these.

A Principle of Fair Apportionment Leads to the Function $-\sum p_i \ln p_i$

Optional Material

Here we derive the functional form of the entropy function, $S = -k \sum_i p_i \ln p_i$ from a Principle of Fair Apportionment. Coins and dice have intrinsic symmetries in their possible outcomes. In unbiased systems, heads is equivalent to tails, and every number on a die is equivalent to every other. The Principle of Fair Apportionment says that if there is such an intrinsic symmetry, and if there is no constraint or bias, then all outcomes will be observed with the same probability. That is, the system 'treats each outcome fairly' in comparison with every other outcome. The probabilities will tend to be apportioned between those outcomes in the most uniform possible way, if the number of trials is large enough. Throughout a long history, the idea that every outcome is equivalent has gone by various names. In the 1700s, Bernoulli called it the Principle of Insufficient Reason; in the 1920s, Keynes called it the Principle of Indifference [1].

However, the principle of the flat distribution, or maximum multiplicity, is incomplete. The Principle of Fair Apportionment needs a second clause that describes how probabilities are apportioned between the possible outcomes when there are constraints or biases. If die rolls give an average score that is convincingly different from 3.5 per roll, then the outcomes are not all equivalent and the probability distribution is not flat. The second clause that completes

Table 6.1 A possible distribution of outcomes for rolling a 30-sided die 1000 times. For example, a red **3** appears 18 times.

	j					
i	**1**	**2**	**3**	**4**	**5**	**6**
red	16	18	18	17	15	16
blue	25	35	45	55	65	75
green	30	28	32	30	50	30
white	40	42	38	40	40	50
black	20	25	30	20	30	25

the Principle of Fair Apportionment says that when there are independent constraints, the probabilities must satisfy the multiplication rule of probability theory, as we illustrate here with a system of multicolored dice.

A Multicolored Die Problem

Consider a 30-sided five-colored die. The sides are numbered **1** through **6** in each of the five different colors. That is, six sides are numbered **1** through **6** and are colored red, six more are numbered **1** through **6** but they are colored blue, six more are green, six more are white, and the remaining six are black. If the die is fair and unbiased, a blue **3** will appear $1/30$ of the time, for example, and a red color will appear $1/5$ of the time.

Roll the die N times. Count the number of appearances of each outcome and enter those numbers in a table in which the six columns represent the numerical outcomes and the five rows represent the color outcomes. Table 6.1 is an example of a possible result for $N = 1000$. To put the table into the form of probabilities, divide each entry by N. This is Table 6.2. For row $i = 1, 2, 3, \ldots, a$ and column $j = 1, 2, 3, \ldots, b$, call the normalized entry p_{ij} ($a = 5$ and $b = 6$ in this case). The sum of probabilities over all entries in Table 6.2 equals one,

$$\sum_{i=1}^{a} \sum_{j=1}^{b} p_{ij} = 1. \tag{6.22}$$

If there were many trials, and no bias (that is, if blue **3** appeared with the same frequency as green **5**, etc.), then the distribution would be flat and every probability would be $1/30$. The flat distribution is the one that apportions the outcomes most fairly between the 30 possible options, if there is no bias.

Now suppose that the system is biased, but that your knowledge of it is incomplete. Suppose you know only the sum along each row and the sum down each column (see Table 6.3). For tables larger than 2×2, the number of rows and columns will be less than the number of cells in the table, so that the probability distribution function will be *underdetermined*. For each row i the sum of the probabilities is

Table 6.2 The conversion of Table 6.1 to probabilities.

	j					
i	**1**	**2**	**3**	**4**	**5**	**6**
red	0.016	0.018	0.018	0.017	0.015	0.016
blue	0.025	0.035	0.045	0.055	0.065	0.075
green	0.030	0.028	0.032	0.030	0.050	0.030
white	0.040	0.042	0.038	0.040	0.040	0.050
black	0.020	0.025	0.030	0.020	0.030	0.025

Table 6.3 Suppose that you know only the row and column sums. In this case they are from Table 6.2. What is the best estimate of all the individual entries?

	j						
i	**1**	**2**	**3**	**4**	**5**	**6**	$u_i = \sum_{j=1}^{6} p_{ij}$
red	?	?	?	?	?	?	$u_1 = 0.10$
blue	?	?	?	?	?	?	$u_2 = 0.30$
green	?	?	?	?	?	?	$u_3 = 0.20$
white	?	?	?	?	?	?	$u_4 = 0.25$
black	?	?	?	?	?	?	$u_5 = 0.15$
$v_j = \sum_{i=1}^{5} p_{ij}$	v_1 0.131	v_2 0.148	v_3 0.163	v_4 0.162	v_5 0.200	v_6 0.196	

$$u_i = \sum_{j=1}^{b} p_{ij}, \tag{6.23}$$

where u_i represents the probability that a roll of the die will show a particular color, i = red, blue, green, white, or black. For example, u_2 is the probability of seeing a blue face with any number on it. Similarly, for each column j the sum of the probabilities is

$$v_j = \sum_{i=1}^{a} p_{ij}. \tag{6.24}$$

For example, v_4 is the probability of seeing a **4** of any color. If the die were unbiased, you would have $v_4 = 1/6$, but if the die were biased, v_4 might have a different value. Row and column sums constitute *constraints*, which are biases or knowledge that must be satisfied as you predict the individual table entries, the p_{ij}'s.

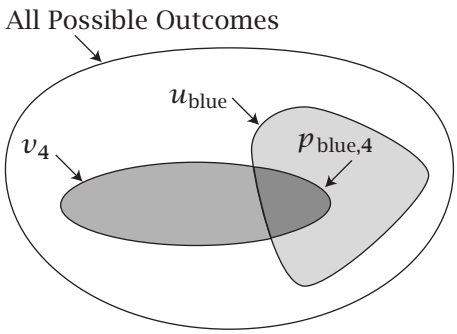

Figure 6.4 The darkest shaded region on this Venn Diagram represents a *blue* **4** outcome, the intersection of sets of all *blue* outcomes with all **4** outcomes.

How should you predict the p_{ij}'s if you know only the row and column sums? The rules of probability tell you exactly what to do. The joint probability p_{ij} represents the intersection of the set of number j and the set of color i (see Figure 6.4). Each p_{ij} is the product of u_i, the fraction of all possible outcomes that has the right color, and v_j, the fraction that has the right number (see Equation (1.6)),

$$p_{ij} = u_i v_j. \tag{6.25}$$

The multiplication rule was introduced in Chapter 1 (page 4). Table 6.4 shows this prediction: to get each entry p_{ij} in Table 6.4, multiply the row sum u_i by the column sum v_j from Table 6.3. With the multiplication rule, you can infer a lot of information (the full table) from a smaller amount (only the row and column sums). You have used only $a + b = 5 + 6 = 11$ known quantities to predict $a \times b = 30$ unknowns. The ability to make predictions for underdetermined systems is particularly valuable for molecular systems in which the number of entries in the table can be huge, even infinite, while the number of row and column sums might be only one or two.

However, if you compare Tables 6.4 and 6.2, you see that your incomplete information is not sufficient to give a perfect prediction of the true distribution function. Is there any alternative to the multiplication rule that would have given a better prediction? It can be proved that the multiplication rule is uniquely the only unbiased way to make consistent inferences about probabilities of the intersection of sets [2]. This is illustrated (but not proved) below, and is proved for a simple case in Problem 6 at the end of the chapter.

The Multiplication Rule Imparts the Least Bias

For simplicity, let's reduce our 5×6 problem to a 2×2 problem. Suppose you have a four-sided die, with sides red **1**, red **2**, blue **1**, and blue **2**. You do not know the individual outcomes themselves. You know only row and column sums: red and blue each appear half the time, **1** appears three-quarters of the time, and **2** appears one-quarter of the time (see Table 6.5). The data show no bias between red and blue, but show a clear bias between **1** and **2**. Now you want to fill in Table 6.5 with your best estimate for each outcome.

Because you know the row and column sums, you can express all the probabilities in terms of a single variable q, say the probability of a red **1**. Now Table 6.5 becomes Table 6.6.

Table 6.4 This table was created by using the multiplication rule Equation (6.26) and the row and column sums from Table 6.3. Compare with Table 6.2. The numbers are in good general agreement, including the large value of green 5 relative to the other green outcomes, and including the bias toward higher numbers in the blue series.

	j					
i	**1**	**2**	**3**	**4**	**5**	**6**
red	0.013	0.015	0.016	0.017	0.021	0.020
blue	0.040	0.044	0.049	0.050	0.062	0.059
green	0.026	0.030	0.033	0.033	0.041	0.039
white	0.033	0.037	0.041	0.042	0.052	0.049
black	0.020	0.022	0.024	0.025	0.031	0.029

Table 6.5 Assume the row and column constraints are known, but the probabilities of the individual outcomes are not, for this 2×2 case.

Color	**Number**		$u_i = \sum_{j=1}^{2} p_{ij}$
	1	**2**	
red	?	?	$u_1 = 1/2$
blue	?	?	$u_2 = 1/2$
$v_j = \sum_{i=1}^{2} p_{ij}$	$v_1 = 3/4$	$v_2 = 1/4$	

Table 6.6 The probabilities for Table 6.5 can be expressed in terms of a single variable q. All entries satisfy the row and column constraints.

Color	**Number**		
	1	**2**	
red	q	$1/2 - q$	$u_1 = 1/2$
blue	$3/4 - q$	$q - 1/4$	$u_2 = 1/2$
	$v_1 = 3/4$	$v_2 = 1/4$	

To ensure that each cell of Table 6.6 contains only a positive quantity (probabilities cannot be negative), the range of allowable values is from $q = 1/4$ to $q = 1/2$. At first glance, it would seem that you have four equations and four unknowns, so that you could solve directly for q. However, the four equations are not all independent, since $u_1 + u_2 = 1$ and $v_1 + v_2 = 1$. Tables 6.7(a), (b)

Table 6.7 These three tables are all consistent with the row and column constraints given in Table 6.5. Only (b) is unbiased. (b) is generated by the multiplication rule from the row and column constraints.

(a)

Color	Number 1	2
red	1/4	1/4
blue	1/2	0

(b)

Color	Number 1	2
red	3/8	1/8
blue	3/8	1/8

(c)

Color	Number 1	2
red	1/2	0
blue	1/4	1/4

and (c) show three possible probability distributions, all of which satisfy the given constraints.

Which of these tables is least biased? Notice that Tables 6.7(a) and (c) with $q = 1/4$ and $q = 1/2$ have unwarranted bias. The data tell us that there is a bias in favor of **1**'s relative to **2**'s but they give no evidence for a difference between the red and blue sides. Yet Tables 6.7(a) and (c) with $q = 1/4$ and $q = 1/2$ indicate that the red sides are different from the blue sides. These tables are perfectly consistent with the data, but they are not impartial toward each outcome. In short, Tables 6.7(a) and (c) are 'unfair' in their apportionment of probabilities to cells.

Only Table 6.7(b) with $q = 3/8$ offers a fair apportionment of the probabilities, and predicts that both colors are equivalent. This least-biased inference results from the multiplication rule. For example, the probability of a red **1** is the product of its row and column sums, $p(\text{red})p(\mathbf{1}) = (1/2)(3/4) = 3/8$. The multiplication rule applies because the constraints u_i and v_j are independent.

The Principle of Fair Apportionment of Outcomes

Equation (6.2) defines the entropy as $S/k = -\sum_i p_i \ln p_i$. Our aim in this section is to derive this expression by treating S as an unknown function of the p_i's using the requirement that S is maximal when the probabilities are fairly apportioned. Consider a table of probability cells indexed by labels $i = 1, 2, 3, \ldots, a$, and $j = 1, 2, 3, \ldots, b$. In typical dice problems, the score on face i is just equal to i. However, let us be more general here. Suppose you paste a score ε_i onto each face i. A score ε_{ij} is associated with each cell. Each cell represents an outcome having probability p_{ij}, the sum of which over all the cells equals one. There are many possible distributions of the p_{ij}'s, but only one distribution, $(p^*_{11}, p^*_{12}, p^*_{13}, \ldots, p^*_{ab})$, is apportioned fairly. If there are no constraints, the distribution will be flat and the probabilities uniform. If there are row and column constraints, all the probabilities will obey the multiplication rule. We seek a function of all the probabilities, $S(p_{11}, p_{12}, p_{13}, \ldots, p_{ab})$, which we will call the *entropy*. The entropy function is maximal, $S = S_{\max}$, when the probabilities are distributed according to fair apportionment.

Why do we need a function S to do this? The multiplication rule already tells us how to fill in the table if the row and column constraints are known. We need S for two reasons. First, we need a way to generalize to situations

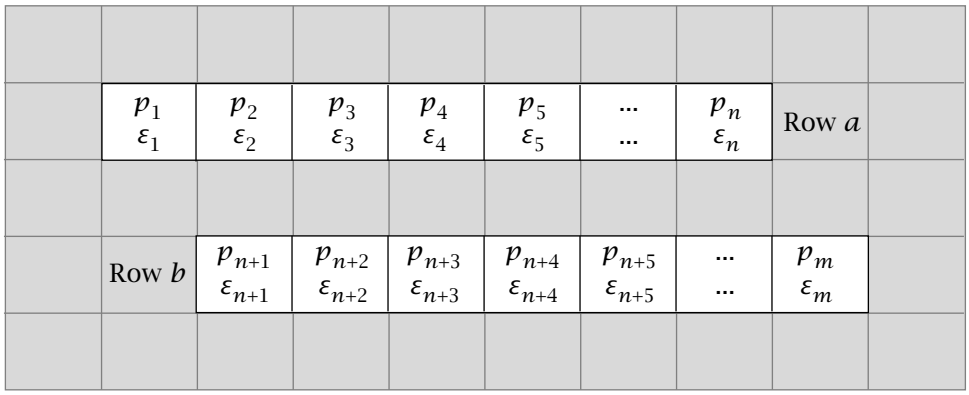

Figure 6.5 Two rows a and b in a grid of cells used to illustrate the notation for proving that entropy is extensive. Each row is subject to a constraint on the sum of probabilities and a constraint on the sum of the scores.

in which we have only a single constraint, or any set of constraints on only a part of the system. Second, we seek a measure of the driving force toward redistributing the probabilities when the constraints are changed.

The Principle of Fair Apportionment is all we need to specify the precise mathematical form that the entropy function must have. We will show this in two steps. First, we will show that the entropy must be *extensive*. That is, the function S that applies to the *whole system* must be a sum over *individual cell* functions s:

$$\begin{aligned} &S(p_{11}, p_{12}, p_{13}, \ldots, p_{ab}) \\ &\quad = s(p_{11}) + s(p_{12}) + s(p_{13}) + \cdots + s(p_{ab}). \end{aligned} \tag{6.26}$$

Second, we will show that the *only* function that is extensive and at its maximum satisfies the multiplication rule is the function $-\sum_{i=1}^{t} p_i \ln p_i$.

The Entropy Is Extensive

Focus on two collections of cells a and b chosen from within the full grid. To keep it simple, we can eliminate the index j if a is a row of cells with probabilities that we label $(p_1, p_2, \ldots, p_n)$, and b is a different row with probabilities that we label $(p_{n+1}, p_{n+2}, \ldots, p_m)$ (see Figure 6.5). Our aim is to determine how the entropy function for the combined system a plus b is related to the entropy functions for the individual rows.

The entropy S is a function that we aim to maximize subject to two conditions. Two constraints, g and h, apply to row a: one is the normalization on the sum of probabilities, and the other is a constraint on the average score,

$$\begin{aligned} g(p_1, p_2, \ldots, p_n) &= \sum_{i=1}^{n} p_i = u_a, \\ h(p_1, p_2, \ldots, p_n) &= \sum_{i=1}^{n} \varepsilon_i p_i = \langle \varepsilon \rangle_a. \end{aligned} \tag{6.27}$$

To find the maximum of a function subject to constraints, use the method of Lagrange multipliers. Multiply $(\partial g/\partial p_i) = 1$ and $(\partial h/\partial p_i) = \varepsilon_i$ by the corresponding Lagrange multipliers α_a and λ_a to get the extremum

$$\frac{\partial S(p_1, p_2, \ldots, p_n)}{\partial p_i} = \alpha_a + \lambda_a \varepsilon_i. \tag{6.28}$$

When this equation is satisfied, S is maximal subject to the two constraints. Express the entropy for row a as $S_a = S(p_1, p_2, \ldots, p_n)$. To find the total differential for the variation of S_a resulting from any change in the p's, sum Equation (6.28) over all dp_i's:

$$dS_a = \sum_{i=1}^{n} \left(\frac{\partial S_a}{\partial p_i} \right) dp_i = \sum_{i=1}^{n} (\alpha_a + \lambda_a \varepsilon_i)\, dp_i. \tag{6.29}$$

Equation (6.29) does not require any assumption about the form of $S_a = S(p_1, p_2, \ldots, p_n)$. It is just the definition of the total differential. Similarly, row b in a different part of the grid will be subject to constraints unrelated to the constraints on row a,

$$\alpha_b \sum_{i=n+1}^{m} dp_i = 0 \quad \text{and} \quad \lambda_b \sum_{i=n+1}^{m} \varepsilon_i dp_i = 0,$$

with Lagrange multipliers α_b and λ_b. For row b, index i now runs from $n+1$ to m, rather than from 1 to n, and we will use the corresponding shorthand notation $S_b = S(p_{n+1}, p_{n+2}, \ldots, p_m)$. The total differential for S_b is

$$dS_b = \sum_{i=n+1}^{m} \left(\frac{\partial S_b}{\partial p_i} \right) dp_i = \sum_{i=n+1}^{m} (\alpha_b + \lambda_b \varepsilon_i)\, dp_i. \tag{6.30}$$

Finally, use the function $S_{\text{total}} = S(p_1, p_2, \ldots, p_n, p_{n+1}, \ldots, p_m)$ to apportion probabilities over both rows a and b at the same time, given exactly the same two constraints on row a and the same two constraints on row b as above. The total differential for S_{total} is

$$\begin{aligned} dS_{\text{total}} &= \sum_{i=1}^{m} \left(\frac{\partial S_{\text{total}}}{\partial p_i} \right) dp_i \\ &= \sum_{i=1}^{n} (\alpha_a + \lambda_a \varepsilon_i) + \sum_{i=n+1}^{m} (\alpha_b + \lambda_b \varepsilon_i) = dS_a + dS_b, \end{aligned} \tag{6.31}$$

where the last equality comes from substitution of Equations (6.29) and (6.30). This says that if the entropy is to be a function having a maximum, subject to constraints on blocks of probability cells, then it must be *extensive*. The entropy of a system is a sum of subsystem entropies. The underlying premise is that the score quantities, $\langle \varepsilon \rangle_a$ and $\langle \varepsilon \rangle_b$, are extensive.

The way in which we divided the grid into rows a and b was completely arbitrary. We could have parsed the grid in many different ways, down to an individual cell at a time. Following this reasoning to its logical conclusion, and integrating the differentials, leads to Equation (6.26), which shows that the entropy of the whole grid is the sum of entropies of the individual cells.

The Form of the Entropy Function

To obtain the form of the entropy function, we use a table of probabilities with rows labeled by the index i and columns by the index j. Each row i is constrained to have an average score,

$$\sum_{j=1}^{b} \varepsilon_{ij} p_{ij} = \langle \varepsilon_i \rangle \quad \Longrightarrow \quad \lambda_i \sum_{j=1}^{b} \varepsilon_{ij} dp_{ij} = 0, \tag{6.32}$$

where λ_i is the Lagrange multiplier that enforces the score constraint for row i. Similarly, for column j,

$$\beta_j \sum_{i=1}^{a} \varepsilon_{ij} dp_{ij} = 0,$$

where β_j is the Lagrange multiplier that enforces the score constraint for column j. Finally,

$$\alpha \sum_{ij} dp_{ij} = 0$$

enforces the constraint that the probabilities must sum to one over the whole grid of cells. We use the extensivity property, Equation (6.26), to find the expression for maximizing the entropy subject to these three constraints,

$$\sum_{i=1}^{a} \sum_{j=1}^{b} \left[\left(\frac{\partial s(p_{ij})}{\partial p_{ij}} \right) - \lambda_i \varepsilon_{ij} - \beta_j \varepsilon_{ij} - \alpha \right] dp_{ij} = 0. \tag{6.33}$$

To solve Equation (6.33), each term in brackets must equal zero for each cell ij according to the Lagrange method (see Equation (5.35)).

Our aim is to deduce the functional form for $s(p_{ij})$ by determining how Equation (6.33) responds to changes in the row constraint u_i and column constraint v_j. To look at this dependence, let's simplify the notation. Focus on the term in the brackets, for one cell ij, and drop the subscripts ij for now. Changing the row or column sum u or v changes p, so express the dependence as $p(u, v)$.

Define the derivative $(\partial s_{ij}/\partial p_{ij}) = r_{ij}$, and express it as $r[p(u, v)]$. The quantities ε_{ij} differ for different cells in the grid, but they are fixed quantities and do not depend on u or v. Because the probabilities must sum to one over the whole grid no matter what changes are made in u and v, α is also independent of u and v. The value of the Lagrange multipliers $\lambda(u)$ for row i and $\beta(v)$ for column j will depend on the value of the constraints u and v (see Example 6.5). Collecting together these functional dependences, the bracketed term in Equation (6.33) can be expressed as

$$r[p(u, v)] = \lambda(u)\varepsilon + \beta(v)\varepsilon + \alpha. \tag{6.34}$$

Now impose the multiplication rule, $p = uv$. Let's see how r depends on u and v. This will lead us to specific requirements for the form of the entropy function. Take the derivative of Equation (6.34) with respect to v to get

$$\left(\frac{\partial r}{\partial v}\right) = \left(\frac{\partial r}{\partial p}\right)\left(\frac{\partial p}{\partial v}\right) = \left(\frac{\partial r}{\partial p}\right)\frac{p}{v} = \varepsilon\beta'(v). \tag{6.35}$$

where $\beta' = d\beta/dv$ and $(\partial p/\partial v) = u = p/v$. Now take the derivative of Equation (6.34) with respect to u instead of v to get

$$\left(\frac{\partial r}{\partial p}\right)\frac{p}{u} = \varepsilon\lambda'(u), \tag{6.36}$$

where $\lambda' = d\lambda/du$. Rearranging and combining Equations (6.35) and (6.36) gives

$$\left(\frac{\partial r}{\partial p}\right) = \frac{u\lambda'(u)\varepsilon}{p} = \frac{v\beta'(v)\varepsilon}{p}. \tag{6.37}$$

Notice that the quantity $u\lambda'(u)\varepsilon$ can only depend on (east–west) properties of the row i, particularly its sum u. Likewise, the quantity $v\beta'(v)\varepsilon$ can only depend on (north–south) properties of the column j, particularly its sum v. The only way that these two quantities can be equal in Equation (6.37) for any arbitrary values of u and v is if $u\lambda'(u)\varepsilon = v\beta'(v)\varepsilon =$ constant. Call this constant $-k$, and you have

$$\left(\frac{\partial r}{\partial p}\right) = \frac{-k}{p}. \tag{6.38}$$

Integrate Equation (6.38) to get

$$r(p) = -k\ln p + c_1,$$

where c_1 is the constant of integration. To get $s(p)$, integrate again:

$$\begin{aligned} s(p) &= \int r(p)\,dp = \int(-k\ln p + c_1)\,dp \\ &= -k(p\ln p - p) + c_1 p + c_2, \end{aligned} \tag{6.39}$$

where c_2 is another constant of integration. Summing over all cells ij gives the total entropy:

$$S = \sum_{i=1}^{a}\sum_{j=1}^{b} s(p_{ij}) = -k\sum_{ij} p_{ij}\ln p_{ij} + c_3\sum_{ij} p_{ij} + c_2, \tag{6.40}$$

where $c_3 = k + c_1$. Since $\sum_{ij} p_{ij} = 1$, the second term on the right side of Equation (6.40) sums to a constant, and Equation (6.40) becomes

$$S = -k\sum_{ij} p_{ij}\ln p_{ij} + \text{constant}. \tag{6.41}$$

If you define the entropy of a perfectly ordered state (one p_{ij} equals one, all others equal zero) to be zero, the constant in Equation (6.41) equals zero. If you choose the convention that the extremum of S should be a maximum rather than a minimum, k is a positive constant. If there is only a single cell index i, you have $S/k = -\sum_i p_i \ln p_i$.

Equation (6.41) is a most remarkable result. It says that you can uniquely define a function, the entropy, that at its maximum will assign probabilities to the

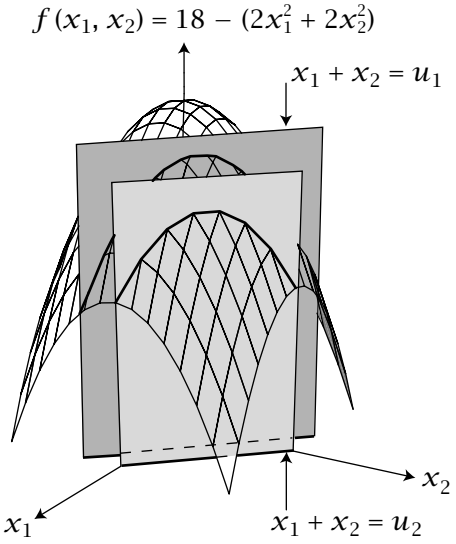

Figure 6.6 The optimum of $f(x_1, x_2)$ changes if a row sum $g = x_1 + x_2 = u$ changes from a value u_1 to u_2.

cells in accordance with the Fair Apportionment Principle: (1) when there are no constraints, all probabilities will be equal, and (2) when there are independent constraints, the probabilities will satisfy the multiplication rule. Because S is the only function that succeeds at this task when it is maximal, S must also be the only function that gives the deviations from the Fair Apportionment Principle when it is not maximal.

EXAMPLE 6.5 Illustrating the dependence $\lambda(u)$. A Lagrange multiplier can depend on the value of a row (or column) constraint. Consider a function $f(x_1, x_2) = 18 - (2x_1^2 + 2x_2^2)$ subject to the constraint $g(x_1, x_2) = x_1 + x_2 = u$ (see Figure 6.6). The extremum is found where

$$\left(\frac{\partial f}{\partial x_1}\right) = -4x_1 = \lambda, \quad \text{and} \quad \left(\frac{\partial f}{\partial x_2}\right) = -4x_2 = \lambda \quad \Longrightarrow \quad x_1 = x_2 = -\frac{\lambda}{4}. \tag{6.42}$$

Substitute Equation (6.42) into $g = x_1 + x_2 = u$ to find $\lambda(u) = -2u$.

Philosophical Foundations

For more than a century, entropy has been a controversial concept. One issue is whether entropy is a tool for making predictions based on the knowledge of an observer [1], or whether it is independent of an observer [3]. Both interpretations are useful under different circumstances. For dice and coin flip problems, constraints such as an average score define the limited knowledge that you have. You want to impart a minimum amount of bias in predicting the probability distribution that is consistent with that knowledge. Your prediction should be based on assuming you are maximally ignorant with respect to all else [1, 4]. Maximizing the function $-\sum_i p_i \ln p_i$ serves the role of making this prediction. Similar prediction problems arise in eliminating noise from spectroscopic signals or in reconstructing satellite photos. In these cases, the Principle we call Fair Apportionment describes how to reason and draw inferences

and make the least-biased predictions of probabilities in face of incomplete knowledge.

However, for other problems, notably those of statistical mechanics, entropy describes a force of nature. In those cases, constraints such as averages over a row of probabilities are not only numbers that are known to an observer, they are also physical constraints that are imposed upon a physical system from outside (see Chapter 10). In this case, the Principle of Fair Apportionment is a description of nature's symmetries. When there is an underlying symmetry in a problem (such as t outcomes that are equivalent), then Fair Apportionment says that external constraints are shared equally between all the possible states that the system can occupy (grid cells, in our earlier examples). In this case, entropy is more than a description about an observer. It describes a tendency of nature that is independent of an observer.

A second controversy involves the interpretation of probabilities. The two perspectives are the frequency interpretation and the 'inference' interpretation [1]. According to the frequency interpretation, a probability $p_A = n_A/N$ describes some event that can be repeated N times. However, the inference interpretation is broader. It says that probabilities are fractional quantities that describe a state of knowledge. In the inference view, the rules of probability are simply rules for making consistent inferences from limited information. The probability that it might rain tomorrow, or the probability that Elvis Presley was born on January 8, can be perfectly well defined, even though these events cannot be repeated N times. In the inference view, probabilities can take on different values depending on your state of knowledge. If you do not know when Elvis was born, you would say the probability is $1/365$ that January 8 was his birthday. That is a statement about your lack of knowledge. However, if you happen to know that was his birthday, the probability is one. According to the subjective interpretation, the rules of probability theory in Chapter 1 are laws for drawing consistent inferences from incomplete information, irrespective of whether or not the events are countable.

Historically, statistical mechanics has been framed in terms of the frequency interpretation. JW Gibbs (1839–1903), American mathematical physicist and a founder of chemical thermodynamics and statistical mechanics, framed statistical mechanics as a counting problem. He envisioned an imaginary collection of all possible outcomes, called an *ensemble*, which was countable and could be taken to the limit of a large number of imaginary repetitions. To be specific, for die rolls, if you want to know the probability of each of the six outcomes on a roll, the ensemble approach would be to imagine that you had rolled the die N times. You would then compute the number of sequences that you expect to observe. This number will depend on N. On the other hand, if the outcomes can instead be described in terms of probabilities, the quantity N never appears. For example, the grid of cells described in Table 6.2 describes probabilities that bear no trace of the information that $N = 1000$ die rolls were used to obtain Table 6.1.

These issues are largely philosophical fine points that have had little implication for the practice of statistical mechanics. We will sometimes prefer the counting strategy, and the use of $S/k = \ln W$, but $S/k = -\sum_i p_i \ln p_i$ will be more convenient in other cases.

Summary

The entropy $S(p_{11}, \ldots, p_{ij}, \ldots, p_{ab})$ is a function of a set of probabilities. The distribution of p_{ij}'s that cause S to be maximal is the distribution that most fairly apportions the constrained scores between the individual outcomes. That is, the probability distribution is flat if there are no constraints, and follows the multiplication rule of probability theory if there are independent constraints. If there is a constraint, such as the average score on die rolls, and if it is not equal to the value expected from a uniform distribution, then maximum entropy predicts an exponential distribution of the probabilities. In Chapter 10, this exponential function will define the Boltzmann distribution law. With this law you can predict thermodynamic and physical properties of atoms and molecules, and their averages and fluctuations. However, first we need the machinery of thermodynamics, the subject of the next three chapters.

Problems

1. Calculating the entropy of dipoles in a field. You have a solution of dipolar molecules with a positive charge at the head and a negative charge at the tail. When there is no electric field applied to the solution, the dipoles point north (n), east (e), west (w), or south (s) with equal probabilities. The probability distribution is shown in Figure 6.7(a). However when you apply a field to the solution, you now observe a different distribution, with more heads pointing north, as shown in Figure 6.7(b).

(a) What is the polarity of the applied field? (In which direction does the field have its most positive pole?)

(b) Calculate the entropy of the system in the absence of the field.

(c) Calculate the entropy of the system in the applied field.

(d) Does the system become more ordered or disordered when the field is applied?

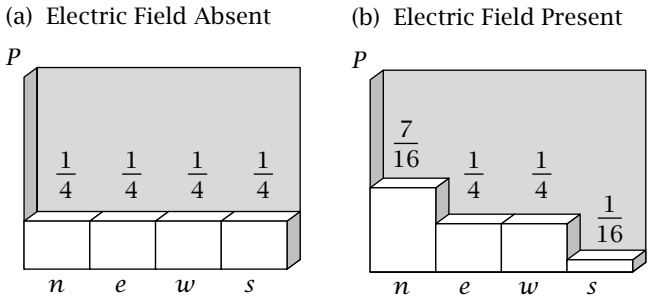

Figure 6.7

2. Comparing the entropy of peaked and flat distributions. Compute the entropies for the spatial concentration shown in Figures 6.8(a) and (b).

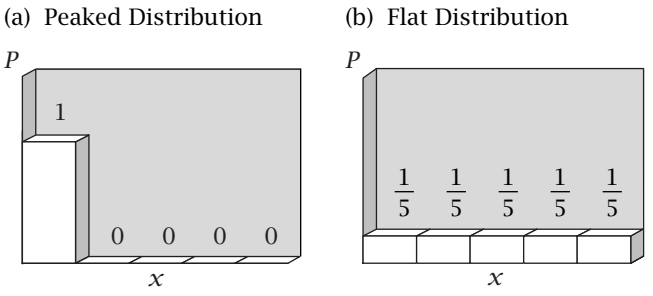

Figure 6.8

3. Comparing the entropy of two peaked distributions. Which of the two distributions shown in Figure 6.9 has the greater entropy?

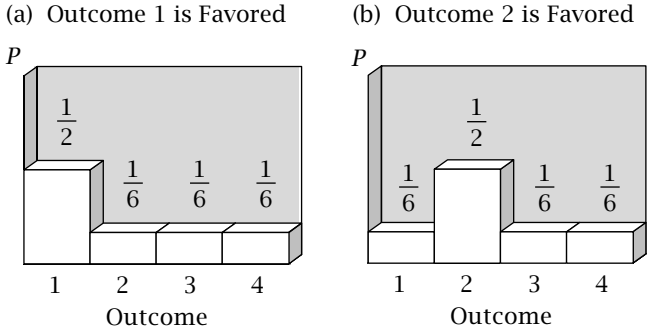

Figure 6.9

4. Calculating the entropy of mixing. Consider a lattice with N sites and n green particles. Consider another lattice, adjacent to the first, with M sites and m red particles. Assume that the green and red particles cannot switch lattices. This is state A.

(a) What is the total number of configurations W_A of the system in state A?

(b) Now assume that all $N + M$ sites are available to all the green and red particles. The particles remain distinguishable by their color. This is state B. Now what is the total number of configurations W_B of the system?

Now take $N = M$ and $n = m$ for the following two problems.

(c) Using Stirling's approximation on page 56, what is the ratio W_A/W_B?

(d) Which state, A or B, has the greatest entropy? Calculate the entropy difference given by

$$\Delta S = S_A - S_B = k \ln \left(\frac{W_A}{W_B} \right).$$

5. Proof of maximum entropy for two outcomes. Example 6.4 (page 88) is simple enough that you can prove the answer is correct even without using the maximum entropy principle. Show this.

6. Proving the multiplication rule maximizes the entropy. Show that if the entropy is defined by Equation (6.2), $S/k = -\sum_i p_i \ln p_i$, and if the 2×2 grids of Tables 6.4 through 6.7 have row sums u_1 and u_2, and column sums v_1 and v_2, then the function $p_{ij}(u_i, v_j)$ that maximizes the entropy is the multiplication rule $p_{ij} = u_i v_j$.

7. Other definitions of entropy do not satisfy the multiplication rule. In contrast to problem 6, show that if the entropy were defined by a least-squares criterion, as $S/k = \sum_{i=1}^{t} p_i^2$, the multiplication rule would not be satisfied when S is maximal.

8. Other definitions of entropy can predict the uniform distribution when there are no constraints. As in problem 7, assume that the definition of entropy was $S/k = \sum_{i=1}^{t} p_i^2$. Show that when there are no row or column constraints on the 2×2 grids of Tables 6.4 through 6.7, the definition will satisfactorily predict that the uniform distribution is an extremum of the entropy.

9. The maximum entropy distribution is Gaussian when the second moment is given. Prove that the probability distribution p_i that maximizes the entropy for die rolls subject to a constant value of the second moment $\langle i^2 \rangle$ is a Gaussian function. Use $\varepsilon_i = i$.

10. Maximum entropy for a three-sided die. You have a three-sided die, with numbers **1**, **2** and **3** on the sides. For a series of N dice rolls, you observe an average score $\langle \varepsilon \rangle$ per roll using the maximum entropy principle.

(a) Write expressions that show how to compute the relative probabilities of occurrence of the three sides, n_1^*/N, n_2^*/N, and n_3^*/N, if α is given.

(b) Compute n_1^*/N, n_2^*/N, and n_3^*/N if $\alpha = 2$.

(c) Compute n_1^*/N, n_2^*/N, and n_3^*/N, if $\alpha = 1.5$.

(d) Compute n_1^*/N, n_2^*/N, and n_3^*/N if $\alpha = 2.5$.

11. Maximum entropy in Las Vegas. You play a slot machine in Las Vegas. For every \$1 coin you insert, there are three outcomes: (1) you lose \$1, (2) you win \$1, so your profit is \$0, (3) you win \$5, so your profit is \$4. Suppose you find that your average expected profit over many trials is \$0 (what an optimist!). Find the maximum entropy distribution for the probabilities p_1, p_2, and p_3 of observing each of these three outcomes.

12. Flat distribution, high entropy. For four coin flips, for each distribution of probabilities, $(p_H, p_T) = (0,1), (1/4, 3/4), (1/2, 1/2), (3/4, 1/4), (1,0)$, compute W, and show that the flattest distribution has the highest multiplicity.

References

[1] ET Jaynes. *The Maximum Entropy Formalism*, RD Levine and M Tribus eds., MIT Press, Cambridge, 1979, page 15.

[2] RT Cox. *Am J Phys* **14**, 1 (1946).

[3] KG Denbigh and JS Denbigh. *Entropy in Relation to Incomplete Knowledge*. Cambridge University Press, Cambridge, 1985.

[4] ET Jaynes. *Phys Rev* **106**, 620 (1957).

Suggested Reading

ET Jaynes, *Phys Rev* **106**, 620 (1957). This is the classic article that introduces the maximum entropy approach to statistical thermodynamics. It is well-written and includes a scholarly historical introduction.

JN Kapur and HK Kesavan, *Entropy Optimization Principles with Applications*, Academic Press, Boston, 1992. Excellent overview of maximum entropy applied to many fields of science.

RD Levine and M Tribus, *The Maximum Entropy Formalism*, MIT Press, 1979. This is an advanced edited volume that discusses applications and other aspects of maximum entropy, including an extensive review by Jaynes.

The following three articles show how the maximum entropy principle follows from the multiplication rule. A brief overview is given in J Skilling, *Nature*, **309**, 748 (1984).

AK Livesay and J Skilling, *Acta Crystallographica*, **A41**, 113–122 (1985).

JE Shore and RW Johnson, *IEEE Trans Inform Theory* **26**, 26 (1980).

Y Tikochinski, NZ Tishby, RD Levine, *Phys Rev Lett* **52**, 1357 (1984).