# 1 Entropy

Before we define entropy, let's review briefly what we've learned from our considerations of the microstates and macrostates of large systems.

- For a single isolated system, we want to count the number of microstates of the whole system for any given macrostate, which we have generally characterized by the fixed energy. We generally assume all microstates are equally probable.

- We now know how to do the microstate counting for all three of our toy systems: the ideal gas, the Einstein crystal, and the two-state paramagnet.

- For two systems in weak thermal contact (but otherwise isolated), the (overwhelmingly) most probable division of energy, which we have taken to characterize all the possible macrostates, is the one with the greatest number of microstates of the combined system. So, if we want to predict the mean division of energy in equilibrium, which is an observable macroscopic quantity, we have to find the macrostate that maximizes the number of microstates, or equivalently, the macrostate with the greatest multiplicity.

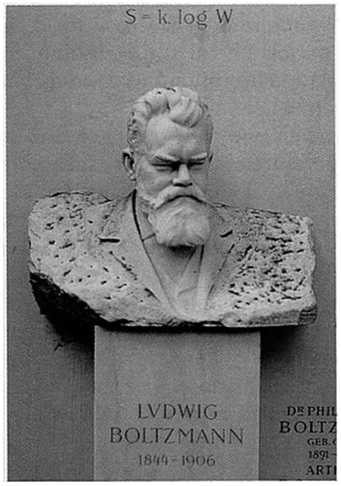
- The problem of predicting the equilibrium properties of two systems in weak contact is prototypical in thermal physics, so understanding first how to discover the equilibrium macrostate is critical.

In thermodynamics one learns that the entropy is a quantity that is maximized in equilibrium. By that it is meant that if one takes a constrained system, or equivalently two separated systems, and removes the constraint, e.g., whatever separates the originally separated systems, the equilibrium state that results is the state of the unconstrained, or combined, system that has the greatest entropy. This sounds very much like what we have concluded regarding the multiplicity of the macrostate with the greatest likelihood. Indeed, it is tempting to say that multiplicity and entropy are one and the same. Conceptually, that is the case, but in order to obtain a quantitative mapping between multiplicity and thermodynamic entropy, it is necessary to take the natural logarithm and multiply by Boltzmann's constant:

$$S = k \ln \Omega \,. \tag{1}$$

The proof of that requires some experience with the concept of entropy in thermodynamics, which we haven't yet acquired, so we won't do that now. The reason for the factor of Boltzmann's constant is simply to endow the entropy with the same dimensions it has in thermodynamics, which, in turn, arise from the way temperature is defined.

It's interesting to note that Boltzmann was so proud of his discovery of the statistical formula for entropy, that he requested that it be engraved on his tombstone:



## 1.1 But what is entropy, really?

In a sense, the definition, and it really is just a definition, given in (1) is all we need, since it gives us a way to calculate $S$, and it exposes its underlying probabilistic meaning. Unfortunately, it's difficult to express its meaning in succinct "everyday" terms. How you explain it to your mother in a few sentences? Nevertheless, it may be useful both to try to develop a deeper understanding and to place the concept into the broader context that has emerged since Boltzmann's discovery.

First, let's note that entropy is often described as a measure of uncertainty, of disorder, or of diversity. Those descriptions make some sense in the context we have established, since macrostates with a greater number of corresponding microstates than other macrostates do leave one with less certainty about exactly which microstate the system is in, and the corresponding collection of microstates exhibits greater disorder and diversity.

The uncertainty interpretation is readily made evident by the game of 20 questions. Suppose I say that I have chosen a number from 1 to 32, and you are to try to discover that number by asking me a series of questions to which I respond with either "yes" or "no." What is the smallest number of questions that is guaranteed to give you the number?

The probabilistic nature of entropy was made even clearer by the work of Claude Shannon, who derived an expression for the entropy of an arbitrary probability distribution in the context of his studies on communication theory at Bell Labs. Shannon's expression is more general than the one presented in (1), but it reduces to that in the case of a uniform probability distribution, and it coincides with a more general definition of entropy that is also used in statistical mechanics:

$$S = -k \sum_i p_i \ln p_i \,, \tag{2}$$

where $i$ indexes some event that occurs with probability $p_i$, and $k$ is an arbitrary constant, Boltzmann's constant in the context of statistical mechanics. In this form, we see that entropy is really a property of probability distributions.

To see that it maps nicely onto what we have done so far, think of any given macrostate as a specification of the probability distribution of the microstates belonging to it. It's a rather trivial probability distribution, since we've assumed the microstates to be equally likely, but it assigns equal probabilities to the allowed microstates and zero probability to each of the other microstates that are incompatible with that macrostate.

Exercise: show that when all probabilities are equal, Shannon's definition (2) reduces to Boltzmann's (1).

### 1.1.1 Shannon's theorem

It's interesting to take a look at Shannon's theorem for the insights it offers into the degree to which the choice of the definition of entropy is really a natural one. The hypotheses are remarkably minimal.

The context is this. We have some collection of $n$ possible "events" (states of a system, measured values of a discrete variable, or whatever), and there is assigned to those events a corresponding collection of probabilities $p_i$. We'll suppose the set of possibilities is exhaustive—there are no others in the context of interest. And we'll suppose the events to be mutually exclusive—if the system is in state 3, it is surely not in state 12.

It's useful to imagine that the probabilities are intended to represent our state of knowledge of the probable outcomes of measurements in a way that takes into account information we already know but is otherwise as unbiased as possible. That is, the goal is to choose a probability distribution that is as uncertain as possible within the constraints of prior knowledge, which is assumed to be inadequate to determine outcomes with certainty. In that light, having a measure of the uncertainty of a probability distribution is seen to be of great utility.

Shannon's theorem then demonstrates that the form of an entropy (uncertainty) function that satisfies certain desirable requirements is given by (2) and is unique, apart from the choice of the constant $k$, which is usually chosen to be 1 in probability theory and Boltzmann's constant in statistical mechanics.

Let's denote the entropy function by $H(p_1, \ldots, p_n)$.

**Theorem.** *Suppose*

1. *$H$ is a continuous function of the $p_i$. This simply assures that small changes in the probabilities don't result in big jumps in $H$.*

2. *If all of the $p_i$ are equal, then*

$$A(n) = H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) \tag{3}$$

   *is a monotonically increasing function of $n$. This enforces the intuitive expectation that a larger collection of equally probably events has more uncertainty than a smaller one. That is just the concept we have been using in our treatment of macrostate multiplicities.*

3. *The value of $H$ is independent of the grouping of the events.*

*Then*

$$H(p_1, \ldots, p_n) = -k \sum_{i=1}^{n} p_i \ln p_i \,. \tag{4}$$

The last part of the hypothesis requires a bit of elaboration. The idea is that we can group the events into composite events in a variety of ways, obtaining the entropy in terms of the hierarchy of probability distributions for the groups and for the individual events within the groups. A simple pictorial example is this:

This is what is done in the 20 questions game: the first question groups the possible events, the possible values of the chosen number, into two groups with about equal probabilities. Once the group has been determined, one must ask further questions, breaking down each group in steps to get to the answer. At each step, there is an associated probability distribution and an entropy for the groups chosen at that step. There are many ways of choosing the groupings, and third hypothesis requires the total entropy to be the same, regardless of the way the groups are chosen.

Another example is the determination of the identity of a card drawn at random from a deck of playing cards. One could calculate the entropy of the overall distribution, or one could calculate that of the distribution of the four suits, with subsequent determination of the entropy of the probability distribution for cards within each suit.

Stated generally, we might group the first $j$ events into a subset whose probability is

$$w_1 = p_1 + \ldots + p_j \;, \tag{5}$$

and make similar groupings of the other events, a total of $r$ altogether. Then there is an entropy function associated with the probability distribution defined by the $w_i$, and for each of the groups, there is another entropy that must be added to that of the groups. But the groups don't necessarily occur with equal probability, so the entropies of the groups must be added with weights equal to the probabilities of the groups:

$$\begin{aligned} H(p_1, \ldots, p_n) = H(w_1, \ldots, w_r) + w_1 H(p_1|w_1, \ldots, p_j|w_1) \\ + w_2 H(p_{j+1}|w_2, \ldots, p_{j+k}|w_2) + \ldots \;, \end{aligned} \tag{6}$$

where $p_i|w_l$ is the probabality of event $i$ given that some event in group $w_l$ has occurred.

The significance of that requirement is that the knowledge we gain ultimately from either inquiry on the individual events one-by-one or from a hierarchical inquiry on groupings of them, followed by inquiry within the groups is the same. The 20 questions game should make this clear: you're welcome to ask in succession whether the chosen number is 1, 2, and so on, or to ask whether the number is in either of two groups, and so on. No matter how you ask the questions, you get to the same state of knowledge in the end. Therefore, the uncertainty at the beginning must have been the same, and the third hypothesis forces the entropy function to reflect that.

*Proof.* I won't present a complete proof, but this should make the conclusion pleasingly plausible. This proof is closely similar to one favored by E. T. Jaynes, rather than Shannon's own proof.

The first hypothesis, continuity, allows us to discretize the calculation. That is, we can work with probabilities that are rational numbers:

$$p_i = \frac{n_i}{\sum_{i=1}^{n} n_i}\,. \tag{7}$$

Once the set $\{n_i\}$ has been chosen, we can regard that choice, which gives the probabilities of the events $i$ as a grouping of some set of "subevents" that are all equally likely, with their total number being $\sum n_i$. Within each group, the entropy function is that of equally probably events, $A(n_i)$. Thus, the overall entropy for the complete collection of subevents is, by (6),

$$A\left(\sum n_i\right) = H(p_1,\ldots,p_n) + \sum_{i=1}^{n} p_i A(n_i)\,. \tag{8}$$

As a specific example, we might choose $n_1 = 3$, $n_2 = 4$, and $n_3 = 2$, for which the entropy calculation would look like:

$$A(9) = H\left(\frac{3}{9},\frac{4}{9},\frac{2}{9}\right) + \frac{3}{9}A(3) + \frac{4}{9}A(4) + \frac{2}{9}A(2)\,. \tag{9}$$

Now suppose all $n$ of the $n_i$ are chosen to be the same value $m$, in which case their sum is $mn$. The the entropy calculation simplifies to:

$$A(mn) = A(m) + A(n)\,. \tag{10}$$

Here's where we depart from rigor. We'll imagine $m$ and $n$ to be continuous, rather than discrete variables; let's rename them $x$ and $y$. Then we have

$$A(xy) = A(x) + A(y)\,. \tag{11}$$

Now differentiate this separately with respect to $x$ and $y$ to get

$$y\frac{dA(xy)}{d(xy)} = \frac{dA(x)}{dx} \quad \text{and} \quad x\frac{dA(xy)}{d(xy)} = \frac{dA(y)}{dy}\,. \tag{12}$$

Solve both of those for $dA(xy)/d(xy)$ and equate the results to get

$$x\frac{dA(x)}{dx} = y\frac{dA(y)}{dy}\,. \tag{13}$$

Since this holds for arbitrary values of the independent variables $x$ and $y$, both sides of this equation must equal the same constant, which we'll call $k$:

$$\frac{dA(x)}{dx} = \frac{k}{x}\,. \tag{14}$$

Integration of this gives

$$A(x) = k\ln x + C\,, \tag{15}$$

where $C$ is an integration constant. Substituting this back into (10), with $x$ and $y$ replacing $m$ and $n$, we get

$$k\ln(xy) + C = k\ln x + C + k\ln y + C\,, \tag{16}$$

which requires that $C = 0$. Thus, the entropy function for equal probabilities is

$$A(n) = k\ln n\,. \tag{17}$$

This is just the Boltzmann form.

To complete the proof, we just substitute this result back into the general form (8) to get

$$k\ln\left(\sum_{i=1}^{n} n_i\right) = H(p_1,\ldots,p_n) + k\sum_{i=1}^{n} p_i\ln n_i\,, \tag{18}$$

which leads to

$$\begin{aligned} H(p_1,\ldots,p_n) &= k\left[\ln\left(\sum_{i=1}^{n} n_i\right) - \sum_{i=1}^{n} p_i\ln n_i\right] \\ &= -k\sum_{i=1}^{n} p_i\ln p_i\,. \end{aligned} \tag{19}$$

□