

ENTROPY AND PROBABILITY THEORY

TIM AUSTIN

These brief notes have been written to accompany an introductory talk about entropy, its rôle in probability theory and the underlying relations between probability and thermodynamics that it symptomatizes. The talk was intended for an audience of students from the sixth form (that is, the last two years of high-school) at Winchester College, UK.

1 The law of averages

We consider the following data.

Let Ω be some finite set and \mathbf{p} some fixed **probability distribution** on Ω : that is, writing the members of Ω as $\omega_1, \omega_2, \dots, \omega_m$ for some m , \mathbf{p} is a vector (p_1, p_2, \dots, p_m) with each $p_j \geq 0$ and $\sum_{j=1}^m p_j = 1$. The value p_j will be interpreted as specifying the probabilities with which some member of Ω will be chosen at random; we might suppose, for example, that we are conducting an experiment for which we know that the possible outcomes are $\omega_1, \omega_2, \dots, \omega_m$, and such that we have some theory telling us that the probability of outcome ω_j is p_j . We will refer to the set Ω as our **state space**, and its members as **states**.

Let us suppose that we make such a random choice of a state in Ω many times over, independently — perhaps we perform our experiment repeatedly, taking care in our set-up that previously-obtained results do not influence the next trial of the experiment — to give a sequence X_1, X_2, \dots, X_N of states in Ω , each selected at random according to the distribution of probabilities \mathbf{p} . We will refer to these as Ω -**random variables**, or as our **outcomes**.

Given this, we may consider the frequencies of different outcomes: let us write \mathbf{X} for the whole string of outcomes (X_1, X_2, \dots, X_N) , and define

$$q_j(\mathbf{X}) = \frac{\# i \leq N \text{ such that } X_i = \omega_j}{N}.$$

These are non-negative fractions with denominator N (or possibly some divisor of it, if we convert to the simplest possible form), and one can easily see that these fractions sum to 1. This means that, if we write

$$\mathbf{q}(\mathbf{X}) = (q_1(\mathbf{X}), q_2(\mathbf{X}), \dots, q_m(\mathbf{X}))$$

for the vector of all these frequencies, then this $\mathbf{q}(\mathbf{X})$ is some new probability distribution on Ω — it satisfies the same conditions as we placed on \mathbf{p} above. This $\mathbf{q}(\mathbf{X})$, which captures the information contained in each of the separate frequencies, is called the **empirical distribution** of the sequence X_1, X_2, \dots, X_N . We must stress here that $\mathbf{q}(\mathbf{X})$ is, on the one hand, a new distribution of probabilities on the state space Ω , but on the other, *it is itself a random object*: the weights it assigns depend on the actual outcomes X_1, X_2, \dots, X_N .

We will be interested in these frequencies $q_j(\mathbf{X})$ in the ‘long-run’; that is, for a very large number of experiments N . What can we say about them in this case? Intuitively, if N is large, we expect each $q_j(\mathbf{X})$

to be close to p_j . Indeed, this is simply the colloquial ‘law of averages’: if we flip a coin with head-bias p very many times, then we expect the proportion of the outcomes that equal heads to be roughly p . Since we expect this to hold for each $j = 1, 2, \dots, m$, we can say instead that for N large we expect the empirical distribution $\mathbf{q}(\mathbf{X})$ to be close to \mathbf{p} .

Our main goal in this talk will be to formulate a precise mathematical version of this ‘closeness’, and then to prove it. The end result — one of several versions of the **Law of Large Numbers** — is a central pillar of probability theory, and worth knowing for that reason alone. However, we will choose to do more: our proof will rely on tools motivated by a more-than-superficial analogy with thermodynamics — in particular, a suitable notion of ‘entropy’ — and so may go some way towards throwing a light on that subject also.

2 Precise formulation: the Law of Large Numbers

It will help to increase our level of abstraction just a little more: having already introduced Ω , let us now introduce also $\text{Pr}(\Omega)$, the set of all possible probability distributions on Ω — that is, of all possible vectors $\mathbf{r} = (r_1, r_2, \dots, r_m)$ satisfying the conditions that $r_j \geq 0$ for $j = 1, 2, \dots, m$ and $\sum_{j=1}^m r_j = 1$. Then the empirical distribution $\mathbf{q}(\mathbf{X})$ is also a member of $\text{Pr}(\Omega)$; a random member, since it depends on the random variables \mathbf{X} , as discussed above. We will be trying to prove that for N large, this random empirical distribution $\mathbf{q}(\mathbf{X})$ is ‘typically’ ‘close to’ \mathbf{p} as a member of $\text{Pr}(\Omega)$. Let us note, however, that $\mathbf{q}(\mathbf{X})$ cannot be just any member of $\text{Pr}(\Omega)$; for any given coordinate of a probability distribution in $\text{Pr}(\Omega)$ can take any value between 0 and 1, but the values $q_j(\mathbf{X})$ can all be written as rational numbers with denominator N . This means that $\mathbf{q}(\mathbf{X})$ must live in the subset of those $\mathbf{r} = (r_1, r_2, \dots, r_m)$ in $\text{Pr}(\Omega)$ with each r_j of the form k_j/N for some $k_j = 0, 1, 2, \dots, N$. We will write $\text{Pr}_N(\Omega)$ for this subset.

For our precise formulation, it remains to specify the meanings of ‘typically’ and ‘close to’ in the previous paragraph. The point is that there are two separate kinds of approximation at work here, and we will want to control both of them:

- On the one hand, we want $\mathbf{q}(\mathbf{X})$ to be close to \mathbf{p} ; we will take this to mean that we are given some fixed error tolerance $\varepsilon > 0$, and then each of the weights $q_j(\mathbf{X})$ is individually within ε of p_j , so that the maximum of the differences $\max_{j=1,2,\dots,m} |q_j(\mathbf{X}) - p_j|$ is less than ε ; let us introduce the notation $\|\mathbf{r} - \mathbf{p}\|$ for $\max_{j=1,2,\dots,m} |r_j - p_j|$, so the above says that $\|\mathbf{q}(\mathbf{X}) - \mathbf{p}\| < \varepsilon$. This is the ‘close to’ part.
- However, we must bear in mind that $\mathbf{q}(\mathbf{X})$ depends on the random variables $\mathbf{X} = (X_1, X_2, \dots, X_N)$, which could, in principle, take as values *any* sequence of states which are individually given positive weights by \mathbf{p} . Some such sequences will give an empirical distribution $\mathbf{q}(\mathbf{X})$ very far away from \mathbf{p} . What we want to show is that for N large, the *probability* of a sequences of outcomes X_1, X_2, \dots, X_N that gives such a distant $\mathbf{q}(\mathbf{X})$ will turn out to be very small: and so we will actually prove that, for a given error tolerance $\varepsilon > 0$, the probability

$$\mathbb{P}\left(X_1, X_2, \dots, X_N : \|\mathbf{q}(\mathbf{X}) - \mathbf{p}\| \geq \varepsilon\right)$$

tends to zero as N increases. This last, probabilistic statement is the ‘typically’ part. This is the Law of Large Numbers.

3 The proof

There are many proofs of the Law of Large Numbers, and we will not give the simplest possible. Instead, we will follow an approach that will then allow us to examine a link with an idea from a historically rather different area of science: entropy, in thermodynamics.

However, let us first describe our plan for proving our law of large numbers, and then execute it. We are going to introduce a real-valued function H defined on the set of all probability distributions $\text{Pr}(\Omega)$ which allows us to estimate the probabilities

$$\mathbb{P}\left(X_1, X_2, \dots, X_N : \|\mathbf{q}(\mathbf{X}) - \mathbf{p}\| \geq \varepsilon\right)$$

quite explicitly. The function H will have the property that $H(\mathbf{r}) \geq 0$ for any probability distribution \mathbf{r} on Ω , with $H(\mathbf{r}) > 0$ unless $\mathbf{r} = \mathbf{p}$, our original ‘background’ probability distribution, in which case $H(\mathbf{p}) = 0$.

Given the error tolerance $\varepsilon > 0$, consider the set U of all possible distributions \mathbf{r} with $\|\mathbf{r} - \mathbf{p}\| \geq \varepsilon$; let us refer to these as the **bad** distributions. We wish to estimate the probability that the empirical distribution $\mathbf{q}(\mathbf{X})$ falls in the set U ; that is, that it is bad.

We introduce an H which we will then prove can be used to estimate this probability according to the following recipe. Consider the values $H(\mathbf{r})$ as \mathbf{r} varies in this set U . None of these values is zero, by the second property of H above, since \mathbf{p} does *not* lie in U . In fact, more is true: we can choose some *fixed* $\alpha > 0$ that is less than all of these values $H(\mathbf{r})$. This last fact relies on just a little so-called ‘real analysis’; however, for the purposes of these notes I would invite you, with or without real analysis, to regard the existence of such a strictly positive α as ‘obvious’. Then whenever N is sufficiently big, the following is true:

$$\mathbb{P}\left(X_1, X_2, \dots, X_N : \|\mathbf{q}(\mathbf{X}) - \mathbf{p}\| \geq \varepsilon\right) < CN^t e^{-\alpha N}$$

for some fixed constants $C > 0$ and $t > 0$.

Since our $\alpha > 0$, C and t are fixed, as N increases the exponential decay $e^{-\alpha N}$ easily compensates for the growth of the polynomial CN^t , and so our probability tends to 0 (indeed, essentially at the rate $e^{-\alpha N}$: exponentials grow *overwhelmingly* faster than polynomials), which is what we wanted. We defer a proof of this to Appendix A. The value we were able to take for this α depended on the function H ; it is this latter which gives us the control we want.

Here, finally, is our formula for H :

$$H(\mathbf{r}) = \sum_{j=1}^m r_j \log r_j - \sum_{j=1}^m r_j \log p_j.$$

Of course, the exact form of this function also depends on \mathbf{p} , which we are assuming fixed for our problem. In the special case that \mathbf{p} is the uniform distribution — that is, if all the states in Ω are equally likely —

then each $p_j = 1/m$ and the above becomes

$$\begin{aligned}
H(\mathbf{r}) &= \sum_{j=1}^m r_j \log r_j - \log \frac{1}{m} \left(\sum_{j=1}^m r_j \right) \\
&= \log m + \sum_{j=1}^m r_j \log r_j \\
&= \log m - \left(- \sum_{j=1}^m r_j \log r_j \right)
\end{aligned}$$

(note that $\sum_{j=1}^m r_j \log r_j$ is *less* than or equal to zero, since $0 \leq r_j \leq 1$ for each j and so each term of the sum is individually less than or equal to 0); this is now a difference of a quantity depending only on \mathbf{p} and another depending only on \mathbf{r} . Later we will interpret it as a difference to two entropy values, but for now we merely adopt a name for H suggested by this fact: it is the **relative entropy of \mathbf{r} over \mathbf{p}** . It is an interesting exercise to show that this H has the properties described above; we sketch the details in Appendix B.

We will now prove our particular exponential estimate. Still there is a choice among different possible arguments; we shall take a fairly elementary approach, but more sophisticated techniques exist (and, in fact, are necessary for the study of more complex models).

For a given N , we want to estimate the probability that our sequence of outcomes $\mathbf{X} = (X_1, X_2, \dots, X_N)$ gives rise to a bad empirical distribution $\mathbf{q}(\mathbf{X})$ — that is, one lying in U , and so necessarily in the finite set $U \cap \text{Pr}_N(\Omega)$ (since certainly it must lie in $\text{Pr}_N(\Omega)$).

Let us first estimate the probability that it equals any one given such bad distribution \mathbf{r} : if $\mathbf{r} = (k_1/N, k_2/N, \dots, k_m/N)$, this is

$$\mathbb{P} \left(\begin{array}{l} k_1 \text{ of the outcomes } X_i \text{ are equal to } \omega_1 \\ \& k_2 \text{ of the outcomes } X_i \text{ are equal to } \omega_2 \\ \& \dots \end{array} \right),$$

and so it can be computed exactly, as follows. Since for any given fixed (that is, not random) string of possible states $\omega_{j_1}, \omega_{j_2}, \dots, \omega_{j_N}$, the probability of obtaining these as our random outcomes X_1, X_2, \dots, X_N is

$$\mathbb{P}(X_1 = \omega_{j_1}, X_2 = \omega_{j_2}, \dots, X_N = \omega_{j_N}) = \mathbb{P}(X_1 = \omega_{j_1}) \mathbb{P}(X_2 = \omega_{j_2}) \cdots \mathbb{P}(X_N = \omega_{j_N}) = p_{j_1} p_{j_2} \cdots p_{j_N}$$

(this is our crucial appeal to the independence of X_1, X_2, \dots, X_N), and now we can re-arrange the order of the factors in this product to obtain

$$p_1^{(\#\{i \leq N \text{ s.t. } j_i=1\})} p_2^{(\#\{i \leq N \text{ s.t. } j_i=2\})} \cdots p_m^{(\#\{i \leq N \text{ s.t. } j_i=m\})},$$

we see that for any one sequence of outcomes that gives the empirical distribution *bfr* above has probability $p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m}$. Therefore our overall probability of obtaining this empirical distribution \mathbf{r} is

$$\begin{aligned}
& (\# \text{ strings of states giving empirical distribution } (k_1/N, k_2/N, \dots, k_m/N)) \times p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m} \\
& = (\# \text{ ways to partition } N \text{ into } m \text{ subsets of sizes } k_1, k_2, \dots, k_m) \times p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m}.
\end{aligned}$$

Now, the initial factor in this expression is given by the **multinomial coefficient**: this is a direct extension of the binomial coefficient, and can be reduced to an explicit formula involving factorials by an argument

directly analogous with the binomial case. It is given by

$$\binom{N}{k_1, k_2, \dots, k_m} = \frac{N!}{k_1! k_2! \dots k_m!},$$

and so our above expression becomes

$$\binom{N}{k_1, k_2, \dots, k_m} p_1^{k_1} p_2^{k_2} \dots p_m^{k_m}.$$

We will call on the following estimate:

$$\binom{N}{k_1, k_2, \dots, k_m} = \frac{N!}{k_1! k_2! \dots k_m!} \leq CN^s \left(\frac{N^N}{k_1^{k_1} k_2^{k_2} \dots k_m^{k_m}} \right)$$

for some fixed constants $C > 0$ and $s \geq 1$. This follows from a standard result called Stirling's approximation (which estimates the individual factors $N!$, $k_1!$, \dots), or via other routes; a proof is given in Appendix C. Since $k_1 + k_2 + \dots + k_m = N$, the right-hand side of the above can now be re-arranged to

$$CN^s \left(\frac{k_1}{N} \right)^{-k_1} \left(\frac{k_2}{N} \right)^{-k_2} \dots \left(\frac{k_m}{N} \right)^{-k_m} = CN^s e^{-N \sum_{j=1}^m \frac{k_j}{N} \log \frac{k_j}{N}}.$$

Given this, it follows that our above probability for $\mathbf{q}(\mathbf{X})$ to equal \mathbf{r} is at most

$$\begin{aligned} CN^s e^{-N \sum_{j=1}^m \frac{k_j}{N} \log \frac{k_j}{N}} p_1^{k_1} p_2^{k_2} \dots p_m^{k_m} &= CN^s e^{-N \sum_{j=1}^m \frac{k_j}{N} \log \frac{k_j}{N}} e^{N \sum_{j=1}^m \frac{k_j}{N} \log p_j} \\ &= C e^{-N \left(\sum_{j=1}^m r_j \log r_j - \sum_{j=1}^m r_j \log p_j \right)} = e^{-NH(\mathbf{r})}. \end{aligned}$$

Suddenly our function H has appeared, and now we can see why: ignoring the initial factor CN^s (which, although it grows polynomially, is still ignorable compared with the exponential, as remarked earlier), our expression for the probability of a particular bad distribution \mathbf{r} arising as the empirical distribution is

$$e^{-N \left(\sum_{j=1}^m r_j \log r_j - \sum_{j=1}^m r_j \log p_j \right)},$$

where on the one hand the exponent $-N \sum_{j=1}^m r_j \log r_j$ (a *positive* number) arises from the (very large) number of possible strings \mathbf{X} of outcomes that could have given rise to this bad empirical distribution, via our earlier approximation for the multinomial coefficient, and on the other the exponent $N \sum_{j=1}^m r_j \log p_j$ (a negative number) arises from the (very small) individual probabilities of each such string of possible outcomes. These two factors are multiplied to give our approximation $e^{-NH(\mathbf{r})}$ for the probability of the bad distribution \mathbf{r} coming out as the empirical distribution in some way or other, and so the largeness of the first and smallness of the second are in competition. The point of the theorem is that the latter wins, and by an exponential margin given precisely by the relative entropy H : our overall expression is $e^{-NH(\mathbf{r})}$, and we know $H(\mathbf{r})$ is strictly positive for \mathbf{r} differing from \mathbf{p} (and so certainly for bad \mathbf{r}).

To complete the proof, it remains to estimate the overall probability of *any* bad distribution being the empirical distribution. We find this just by adding up the above estimates over all possible bad distributions \mathbf{r} .

$$\begin{aligned} &\mathbb{P} \left(X_1, X_2, \dots, X_N : \max_{j=1,2,\dots,m} |q_j(\mathbf{X}) - p_j| \geq \varepsilon \right) \\ &= \sum_{\text{bad } \mathbf{r} \in \text{Pr}_N(\Omega)} \mathbb{P} \left(X_1, X_2, \dots, X_N : \mathbf{q}(\mathbf{X}) = \mathbf{r} \right) \\ &\leq C \sum_{\text{bad } \mathbf{r} \in \text{Pr}_N(\Omega)} N^s e^{-NH(\mathbf{r})}. \end{aligned}$$

Now, the whole of $\text{Pr}_N(\Omega)$ contains fewer than N^m vectors (since this is the total number of vectors with each entry of the form k/N with $0 \leq k \leq N$): the number of possible strings of outcomes that can give rise to a given empirical distribution is large like an exponential (i.e., very very large), but the number of different possible empirical distributions is only large like a polynomial (still large, but much less so); so, referring to our recipe for choosing α , we must have $\alpha \leq H(\mathbf{r})$ for all bad \mathbf{r} , and hence $e^{-NH(\mathbf{r})} \leq e^{-\alpha N}$, for every bad \mathbf{r} . So the above sum is at most

$$CN^m \times N^s e^{-\alpha N} = CN^t e^{-\alpha N}$$

with $t = m + s$; this is of the required form. □

We end this section by remarking that, in some sense, the above estimate is as good as it can be:

EXERCISE: We can find an inequality for our probability going the other way: if $\beta > 0$ is any real number strictly *greater* than some of the values $H(\mathbf{r})$ for some bad probability distribution \mathbf{r} , then

$$\mathbb{P}\left(X_1, X_2, \dots, X_N : \max_{j=1,2,\dots,m} |q_j(\mathbf{X}) - p_j| > \varepsilon\right) > e^{-\beta N}$$

for N sufficiently large [Hint: consider the contribution to our probability from any one bad empirical distribution \mathbf{r}]. ◁

4 The link with thermodynamics

Having introduced the function H and used it to prove the Law of Large Numbers, let us finally address the underlying theme of this essay: the relationship between probability and thermodynamics.

We will base our discussion on an analogy between the probabilistic data we have considered above and the objects of thermodynamics. This is not so hard to motivate; it results from a slightly different, but also fairly intuitive, interpretation of our data instead of our previous ‘outcomes-of-repeated-experiments’ description. We can imagine now that Ω is the space of all possible states that a molecule of some gas in equilibrium can be in — presumably specifying its position in a container together with its velocity, and possibly some other relevant variables such as magnetic spin, if it has one — and the random variables X_1, X_2, \dots, X_N are the actual states that the N (a very large number) molecules in our gas are in at a given time.

Of course, there are technical issues here that we are ignoring: we wouldn’t choose a model for a real gas that allowed only a finite number of such possible states for the molecules, but rather a continuum of states, for example. Moreover, a real equilibrium thermodynamic system is not actually static: the molecules move around, and we would need to take into the time-variations of their states as well. However, the analogy is still suggestive for our toy model, and it turns out that the methods introduced above can indeed be extended to such more sophisticated models, but that’s a rather longer story.

To make one further leap, we can now interpret the empirical distribution $\mathbf{q}(\mathbf{X})$ — which, we recall, captures the overall proportion of the molecules that occupy each of the possible individual-molecule-states ω_j — as the macroscopic state of the whole system. It describes all the possible ‘macroscopic’ variables of the system (in classical thermodynamics, the pressure, temperature and volume), as these can be identified in the following way: given some statistic f for a single molecule — that is, some function f from Ω to \mathbb{R} — the corresponding macrostatistic is just the average of the values taken by f for the different molecule states X_i :

$$\frac{1}{N} \sum_{i \leq N} f(X_i);$$

however, this, in turn, is precisely $\sum_{j=1}^m q_j(\mathbf{X})f(\omega_j)$ (just by re-arranging the sum to take all the X_i s equal to ω_1 first, then all those equal to ω_2 , and so on), and hence it actually depends only on the empirical distribution (“macroscopic state”) $\mathbf{q}(\mathbf{X})$.

However, what has H to do with the entropy of thermodynamics? It is precisely the difference of two values which can themselves be considered as (absolute) entropies:

$$H(\mathbf{r}) = \sum_{j=1}^m r_j \log r_j - \sum_{j=1}^m r_j \log p_j = \left(- \sum_{j=1}^m r_j \log p_j \right) - \left(- \sum_{j=1}^m r_j \log r_j \right).$$

Now, this is slightly more complicated to analyze in the fully general case, and so we will restrict our discussion to the special case in which \mathbf{p} is the uniform distribution, so each p_j equals $1/m$. Then the above simplifies to

$$\log m - \left(- \sum_{j=1}^m r_j \log r_j \right),$$

as described previously.

The first term of this sum, $\log m$, is just the logarithm of the total number of possible states: up to a universal constant, this is precisely the formula Boltzmann originally proposed for his early far-sighted attempt at a microscopic description of the entropy of a particular macroscopic state in thermodynamics (which until then had only been treated as a mysterious, black-box quantity). The extra term $\sum_{j=1}^m r_j \log r_j$ is a competing entropy value depending on the probability distribution \mathbf{r} . When $\mathbf{r} = \mathbf{p}$, the uniform distribution, it compensates completely for the $\log m$ and leaves zero; if $\mathbf{r} \neq \mathbf{p}$, then the extra term falls short of $\log m$ (this is just the inequality for H that we proved before), and the gap gives an exponential approximation to the probability that the macroscopic state of the system is \mathbf{r} , given the assumption that for the individual molecules the different possible states are equally likely (that is, have the probabilities given by \mathbf{p}). The above calculation giving this exponentially small probability tells us that, when N is large, the system is overwhelmingly likely to be in that macroscopic state which maximizes its entropy.

With some other probability distribution \mathbf{p} (not uniform) a similar interpretation is possible, but we have to be more cautious, as H is chosen quite delicately to allow correctly for the different weights assigned by \mathbf{p} and \mathbf{r} .

Since its introduction to thermodynamics, various notions of entropy and relative entropy have proliferated in several areas of mathematics, particular probability theory and information theory, and new and more exotic ‘entropy functions’, which still act as estimates for probabilities but in more and more complicated situations, continue to appear all the time. Sadly, there seem to be few good sources for a common treatment of these different manifestations; introductions to information theory and coding (both topics that usually appear after the first year or two at university) may be the best place to start (although I know of no such books that either are so common or so excellent as to be worth recommending individually).

A Exponentials versus polynomials

Here we prove our claim concerning our initial form for our estimate on the probability of a bad empirical distribution, that exponential decay is overwhelmingly faster than any polynomial growth: that is, that whenever $\alpha > 0$ (however small) and $t \geq 0$ (however large) we have

$$n^t e^{-\alpha n} \rightarrow 0$$

as $n \rightarrow \infty$. There are many ways to prove this, all relying on one or other quite simple estimates; in fact, perhaps the most subtle aspect of this result is that it is *so* overwhelmingly true, and so can be proved by such crude estimates, that one might be misled by looking for an argument that is more delicate than necessary, when possibly no such ‘finer’ argument even exists.

Here we will use the following trick: letting $a_n = n^t e^{-\alpha n}$, we consider the successive quotients a_{n+1}/a_n :

$$\frac{a_{n+1}}{a_n} = \left(\frac{n+1}{n}\right)^t e^{-\alpha}.$$

Now, since $\alpha > 0$, $e^{-\alpha} < 1$. On the other hand, $\frac{n+1}{n} = 1 + \frac{1}{n} \rightarrow 1$ as $n \rightarrow \infty$, and so the same is true for $\left(\frac{n+1}{n}\right)^t$. Therefore, in particular, there is some fixed n_0 so that whenever $n \geq n_0$ we have $\left(\frac{n+1}{n}\right)^t < e^{\alpha/2}$, and therefore

$$\frac{a_{n+1}}{a_n} = \left(\frac{n+1}{n}\right)^t e^{-\alpha} < e^{-\alpha/2}.$$

Hence, for any $n > n_0$,

$$a_n = a_{n_0} \left(\frac{a_{n_0+1}}{a_{n_0}}\right) \left(\frac{a_{n_0+2}}{a_{n_0+1}}\right) \cdots \left(\frac{a_n}{a_{n-1}}\right) < a_{n_0} e^{-\alpha/2} \times e^{-\alpha/2} \times \cdots \times e^{-\alpha/2} = a_{n_0} e^{-(\alpha/2)(n-n_0)},$$

and this does decay exponentially (since the factor a_{n_0} is now fixed), as required.

EXERCISE: Modify the above argument to show that, in fact, whenever $\alpha_1 > 0$ is still slightly less than α — no matter how much less — we have

$$n^t e^{-\alpha n} < e^{-\alpha_1 n}$$

for all n sufficiently large. This says that any difference in the exponential decays $e^{-\alpha n}$ and $e^{-\alpha_1 n}$ with the second slower than the first is eventually enough to overcome the the polynomial factor n^t on the left. \triangleleft

B Inequalities for the relative entropy

Here we prove the basic property of the relative entropy function H , that $H(\mathbf{r}) \geq 0$ with equality if and only if $\mathbf{r} = \mathbf{p}$. This rests on a very powerful and general inequality called **Jensen’s inequality**.

The crucial observation here is that the function $f : [0, \infty) \rightarrow \mathbb{R}$ given by $f(x) = x \log x$ is *strictly convex*: that is, whenever we have real numbers $0 \leq x, y$ with $x \neq y$ and some θ with $0 \leq \theta \leq 1$, then

$$f(\theta x + (1 - \theta)y) \geq \theta f(x) + (1 - \theta)f(y),$$

with equality if and only if $x = y$. Diagrammatically, this is the picture we expect: looking at the points $(x, f(x))$ and $(y, f(y))$ on the graph of f , this inequality says that the line segment joining these points is entirely above the graph, and touches it only at the end-points of that line segments (try drawing a picture).

EXERCISE: Prove that f is convex [Hint: consider an inequality for f'']. \triangleleft

Now, once we know that f is convex, we may prove a slightly more general inequality than the one above by induction on m (EXERCISE): whenever $x_1, x_2, \dots, x_m \geq 0$ are *distinct* real numbers and $\theta_1, \theta_2, \dots, \theta_m \geq 0$ sum to 1 (equivalently, $(\theta_1, \theta_2, \dots, \theta_m)$ is a probability vector like those considered earlier), we have

$$f(\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m) \geq \theta_1 f(x_1) + \theta_2 f(x_2) + \cdots + \theta_m f(x_m),$$

with equality only if some one of the θ_j equals 1 and all of the others are 0. This is Jensen's inequality.

Given this, we can deduce our desired properties of H : given \mathbf{p} and \mathbf{r} , if we let $\theta_j = r_j$ and $x_j = p_j/r_j$, the above inequality becomes precisely what we wanted, and the conditions for equality show that we must have $\mathbf{r} = \mathbf{p}$ if $H(\mathbf{r}) = 0$.

C Counting partitions

Here we fill in the details of the estimate for the multinomial coefficient used during our main proof:

$$\binom{N}{k_1, k_2, \dots, k_m} = \frac{N!}{k_1! k_2! \dots k_m!} \leq CN^s \left(\frac{N^N}{k_1^{k_1} k_2^{k_2} \dots k_m^{k_m}} \right)$$

for some fixed $C > 0$ and $s \geq 1$ (not depending on k_1, k_2, \dots, k_m or N).

We will take quite a pedestrian route to this fact. It will follow from the following inequalities, themselves an ultra-crude version of Stirling's approximation: there is some fixed $s \geq 1$ and some constant $C \geq 1$ such that

$$\frac{n^n}{e^n} \leq n! \leq Cn^s \frac{n^n}{e^n}$$

for all n . In fact we will prove only the left-hand side of this, leave the right-hand side as an EXERCISE: although it looks more complicated, it follows by a slightly more subtle application of the same method.

We use the method of **integral comparison**, a popular, and surprisingly powerful, trick in elementary combinatorics, probability and number theory. By taking logarithms (noting that all of the above expressions are strictly positive) it suffices to show that

$$\log n! = \sum_{i=1}^n \log i \geq n \log n - n.$$

However, dropping the first term of the sum $\sum_{i=1}^n \log i$ since it is 0, we find then that this sum is equal precisely to the integral from 1 to n of the function f which takes the constant value $\log(i+1)$ on the interval $(i, i+1]$; that is, $f(x) = \log \lceil x \rceil$. This function f satisfies $f(x) \geq \log x$ for all $x \geq 1$, since $\lceil x \rceil \geq x$, and therefore

$$\sum_{i=2}^n \log i = \int_1^n f(x) dx \geq \int_1^n \log x dx = (x \log x - x)|_1^n = n \log n - n + 1 > n \log n - n,$$

and we are done. If you're not quite happy with this comparison, draw a picture of the graphs of f and \log for $x \geq 1$ and consider the areas under them.

(To prove the inequality in the other direction, think about how to use an integral comparison against $\log x$ of some other similar function f , this time less than or equal to $\log x$, which still has an integral equal to the desired sum $\sum_{i=1}^n \log i$).

DEPARTMENT OF MATHEMATICS
 UNIVERSITY OF CALIFORNIA AT LOS ANGELES, LOS ANGELES, CA 90095-1555, USA
 Email: timaustin@math.ucla.edu
 Web: <http://www.math.ucla.edu/~timaustin>