7 Entropy arguments

This section is slightly different from previous ones in that we shall need to develop a little bit of theory, and only after that will the arguments be very short. However, the theory consists of a few basic statements, and what really matters is not the proofs of those statements, but how to use them. To put it another way, I recommend treating those statements more like axioms than lemmas. To encourage this, I shall do so myself, but just to give you something to hold on to, which makes some of the axioms more intuitive, you should think of the entropy $H[X]$ of a discrete random variable X as a real number that measures the "information content" of X. Roughly speaking, this is how many bits of information you gain, on average, if you find out the value of X, or equivalently, the expected number of bits needed to specify X.

7.1 The Khinchin axioms for entropy and some simple consequences

Entropy has the following properties, which are called the Khinchin (or Shannon-Khinchin) axioms. (I got some of these from a set of lecture notes by Cosma Shalizi, which I recommend for further discussion of the pros and cons of an axiomatic approach.)

0. *Normalization.* If X takes the values 0 and 1, each with probability $1/2$, then $H[X] = 1$.

1. *Invariance.* $H[X]$ depends only on the probability distribution of X. That is, if $Y = f(X)$ for a function f that is bijective on the values taken by X, then $H[Y] = H[X]$.

2. *Maximality.* If X takes at most k distinct values, then $H[X]$ is maximized when X takes each value with equal probability $1/k$.

3. *Extensibility.* If X is a random variable that takes values in a finite set A and Y is a random variable that takes values in a set B with $A \subset B$, and if $\mathbb{P}[X = a] = \mathbb{P}[Y = a]$ for every $a \in A$ (and hence $\mathbb{P}[Y = b] = 0$ for every $b \in B \setminus A$), then $H[Y] = H[X]$.

4. *Additivity.* For any two random variables X and Y, $H[X, Y] = H[X] + H[Y|X]$, where
$$H[Y|X] = \sum_x \mathbb{P}[X = x] H[Y|X = x].$$

5. *Continuity.* $H[X]$ depends continuously on the probabilities $\mathbb{P}[X = x]$.

Axiom 0 is not really one of Khinchin's axioms, but the remaining axioms determine H only up to a multiplicative constant so it is there to fix that constant to a convenient value. Axioms 1-3 and 5 are rather basic properties of a kind that one might expect, but axiom 4 needs more comment. The quantity $H[X, Y]$ is simply the entropy (whatever that will turn out to mean) of the joint random variable (X, Y). The quantity $H[Y|X]$ is called the

conditional entropy of Y given X: it is the average entropy of Y given the value of X. Note that from its definition and the fact that H takes non-negative values (which implies that $H[Y|X = x]$ is non-negative for each x), it follows that $H[Y|X]$ is non-negative.

We now prove a sequence of lemmas, most of them very simple.

Lemma 7.1. *If X and Y are independent, then $H[Y|X] = H[Y]$ and $H[X, Y] = H[X] + H[Y]$.*

Proof. For each x the distribution of Y given that $X = x$ is the same as the distribution of Y, so $H[Y|X = x] = H[Y]$ for every x, by the invariance axiom. (The reason that axiom is needed is that strictly speaking the random variable $Y|X = x$ takes values of the form (x, y), where y is a value taken by Y.) It follows that

$$H[Y|X] = \sum_x \mathbb{P}[X = x]H[Y|X = x] = \sum_x \mathbb{P}[X = x]H[Y] = H[Y].$$

The second statement then follows from the additivity axiom. □

Lemma 7.2. *If X takes just one value, then $H[X] = 0$.*

Proof. $H[X, X] = H[X]$, by the invariance axiom. But X and (X, X) are independent, so $H[X, X] = 2H[X]$, by Lemma 7.1. □

And another.

Lemma 7.3. *Let $A \subset B$, let X be uniformly distributed on A, and let Y be uniformly distributed on B. Then $H[X] \leq H[Y]$, with equality if and only if $A = B$.*

Proof. By extensibility, $H[X]$ is not affected if we regard it as taking values in B. The inequality then follows from the maximality axiom.

Suppose now that $|A| = r$, $|B| = s$, and $r < s$. If $r = 1$, then the result follows from Lemma 7.2, the normalization axiom, and what we have just proved.

Otherwise, denote by X^n the A^n-valued random variable given by n independent copies of X, and similarly for Y. Then for any n, Lemma 7.1 and induction imply that $H[X^n] = nH[X]$ and $H[Y^n] = nH[Y]$.

Now choose n such that $r^n \leq s^{n-1}$. Then $|A^n| \leq |B^{n-1}|$, so

$$nH[X] = H[X^n] \leq H[Y^{n-1}] = (n - 1)H[Y],$$

where the inequality follows from what we have just proved (together with the invariance axiom). Since $H[X] \geq 1$ (again by what we proved above), it follows that $H[X] < H[Y]$. □

And another.

Lemma 7.4. *Let X be a random variable and let $Y = f(X)$ for some function f. Then $H[Y] \leq H[X]$.*

Proof. By the invariance axiom, $H[X] = H[X, Y]$, since there is a bijection between values x taken by X and values $(x, f(x))$ taken by (X, Y). Therefore, by the additivity axiom, $H[X] = H[Y] + H[X|Y]$. □

In an earlier version of these notes, I assumed that H was non-negative, having failed to see a proof of non-negativity from the axioms. However, Sean Eberhard (a postdoc at Cambridge) pointed out to me the following argument.

Lemma 7.5. *$H[X] \geq 0$ for every discrete random variable X that takes values in a finite set A.*

Proof. First let us suppose that there exists n such that $p_a = \mathbb{P}[X = a]$ is a multiple of n^{-1} for every $a \in A$. Let Y be uniformly distributed on $[n]$, let $(E_a : a \in A)$ be a partition of $[n]$ such that $|E_a| = p_a n$ for each $a \in A$, and let Z be the random variable where $Z = a$ if $Y \in E_a$. Then Z and X are identically distributed, so $H[Z] = H[X]$, by invariance.

Now $H[Y, Z] = H[Z] + H[Y|Z]$, by additivity. Also, $H[Y, Z] = H[Y]$ by Lemma 7.4, since (Y, Z) depends only on Y. And finally, for each $a \in A$, $H[Y|Z = a]$ is uniformly distributed on a set of size at most n, so by Lemma 7.3 it follows that $H[Y|Z = a] \leq H[Y]$. This implies that $H[Y|Z] \leq H[Y, Z]$, and therefore that $H[X] = H[Z] \geq 0$.

In the general case, since we can approximate the probabilities p_a arbitrarily closely by multiples of n^{-1} for a suitably large n, we can apply the continuity axiom to obtain the same conclusion. □

Here is a slightly less simple lemma.

Lemma 7.6. *Let X be a random variable that takes at least two values with non-zero probability. Then $H[X] > 0$.*

Proof. Let A be the set of values taken by X, let $\alpha = \max_{a \in A} \mathbb{P}[X = a]$, and for each n denote by X^n the A^n-valued random variable that is given by n independent copies of X. Then the maximum probability of any value taken by X^n is α^n. Since $\alpha < 1$, for any $\epsilon > 0$ there exists n such that $\alpha^n < \epsilon$. It follows that we can partition A^n into two sets E and F, each of which has probability between $1/2 - \epsilon$ and $1/2 + \epsilon$. Now let Y be a random variable that takes the value 0 if $X^n \in E$ and 1 if $X^n \in F$. Then $H[X^n] = nH[X]$, by Lemma 7.1 (and induction), and also $H[X^n] = H[Y] + H[X^n|Y] \geq H[Y]$. But $H[Y] > 0$ for sufficiently small ϵ, by the normalization axiom and continuity. It follows that $H[X^n] > 0$ and therefore that $H[X] > 0$. □

We end this sequence of lemmas with a result that is often useful. It is sometimes known as the *chain rule* for entropy.

Lemma 7.7. *Let $X_1, \dots, X_k$ be random variables taking values in a set A. Then*

$$H[X_1, \dots, X_k] = H[X_1] + H[X_2|X_1] + H[X_3|X_1, X_2] + \cdots + H[X_k|X_1, \dots, X_{k-1}].$$

Proof. By additivity,

$$H[X_1, \dots, X_k] = H[X_1, \dots, X_{k-1}] + H[X_k|X_1, \dots, X_{k-1}].$$

The result therefore follows by induction, with the additivity axiom as the base case. □

7.2 The number of paths of length 3 in a bipartite graph

Let us now consider the following problem. Suppose that G is a bipartite graph with finite vertex sets A and B and density α. (The density is defined to be the number of edges divided by $|A||B|$.) A *labelled* $P3$ is a quadruple (x_1, y_1, x_2, y_2) such that $x_1, x_2 \in A$, $y_1, y_2 \in B$, and x_1y_1, y_1x_2, and x_2y_2 are all edges of G. In other words, it is a path of length 3 in the graph, but we allow degeneracies such as $x_1 = x_2$.

How many labelled $P3$s must a bipartite graph with density α contain? If G is bi-regular, meaning that every vertex in A has degree $\alpha|B|$ and every vertex in B has degree $\alpha|A|$, then there are $\alpha^3|A|^2|B|^2$, since we can choose x_1 in $|A|$ ways, then y_1 in $\alpha|B|$ ways, then x_2 in $\alpha|A|$ ways, and finally y_2 in $\alpha|B|$ ways. We shall now show that this is the smallest number of labelled $P3$s that there can be. The proof will assume that entropy exists – that is, that there is some H that satisfies the Khinchin axioms. Later we shall see that this assumption is valid, and that will put in the final piece of the jigsaw.

Before you read the proof, I strongly recommend you try to prove the result for yourself by elementary means, since it looks as though it ought to be possible (given that the result is true), but turns out to be surprisingly tricky, and if you haven't experienced the difficulty, then you won't appreciate the power of the entropy approach.

How does one use entropy to prove results in combinatorics? Part of the answer lies in axiom 2. Suppose that X is uniformly distributed on a set of size k, and Y is a random variable with $H[Y] \geq H[X]$. then axiom 2 implies that Y takes at least k different values. Therefore, if we want to prove that a set A has size at least k, one way of doing it is to find a random variable that takes values in A and has entropy at least $H[X]$.

At first this may seem a very strange approach, since by axiom 2 we know that if there is such a random variable, then a random variable that is uniformly distributed on A will also work. If we define $f(n)$ to be the entropy of a random variable that is uniformly distributed on a set of size n, then all we seem to be doing is replacing the cardinality of a set by f of that cardinality, which doesn't look as though it will achieve anything.

However, there is a flaw in that criticism, which is that it might in principle be easier to obtain a lower bound for the entropy of a carefully chosen distribution on a set A (given certain assumptions about A) than it is to find a lower bound on the cardinality of A. And indeed, this turns out to be the case in many interesting situations, including the one at hand.

We wish to obtain a lower bound for the number of labelled $P3$s in a bipartite graph G of density α, and to do so we shall obtain a lower bound for the entropy of the following distribution on the set of labelled $P3$s, which is *not* uniform (except when the graph is regular). We choose an edge x_1y_1 (with $x \in A$ and $y \in B$ uniformly at random, then a vertex x_2 uniformly from the neighbours of y_1, and then a vertex y_2 uniformly from the neighbours of x_2.

Let X_1, Y_1, X_2, and Y_2 be the distributions of x_1, y_1, x_2, and y_2, respectively. We now wish to say something about the entropy $H[X_1, Y_1, X_2, Y_2]$. The chain rule (Lemma 7.7) tells us that it is equal to

$$H[X_1, Y_1] + H[X_2|X_1, Y_1] + H[Y_2|X_1, Y_1, X_2].$$

Now

$$H[X_2|X_1,Y_1] = \sum_{a\in A}\sum_{b\in B}\mathbb{P}[X_1=a,Y_1=b]H[X_2|X_1=a,Y_1=b].$$

But for each fixed b, the distributions of X_1 and X_2 given that $Y_1 = b$ are independent: the way we choose x_2 once we have chosen y_1 depends entirely on y_1 and not on how y_1 was obtained. Therefore, this simplifies to

$$\begin{aligned}\sum_{a\in A}\sum_{b\in B}\mathbb{P}[X_1=a,Y_1=b]H[X_2|Y_1=b] &= \sum_{b\in B}\mathbb{P}[Y_1=b]H[X_2|Y_1=b]\\ &= H[X_2|Y_1].\end{aligned}$$

In a similar way, if we know the value of X_2, then the distribution of Y_2 is independent of the values of X_1 and Y_1, so

$$H[Y_2|X_1,Y_1,X_2] = H[Y_2|X_2].$$

So we are interested in finding a lower bound for

$$H[X_1,Y_1] + H[X_2|Y_1] + H[Y_2|X_2].$$

By the additivity axiom, we can write this as

$$H[X_1,Y_1] + H[Y_1,X_2] + H[X_2,Y_2] - H[Y_1] - H[X_2].$$

Notice that the first three terms are the entropies of the distributions of the three edges of the random labelled $P3$. We now make an important observation.

Lemma 7.8. *Given a random labelled $P3$ from the distribution defined above, the three edges are all uniformly distributed over all edges.*

Proof. The first edge is uniformly distributed by the definition of the distribution. Now the number of edges is $\alpha|A||B|$ and the number of edges x_1y_1 with $y_1 = b$ is $d(b)$ (the degree of b), so the probability that $Y_1 = b$ is $d(b)/\alpha|A||B|$, and the probability that $X_2 = a$ given that $Y_1 = b$ is 0 if ab is not an edge and $d(b)^{-1}$ if ab is an edge. So the probability that $(X_2,Y_1) = (a,b)$ is $1/\alpha|A||B|$ whenever ab is an edge, which is another way of saying that X_2Y_1 is uniformly distributed over all edges. And once we know that, then the same proof shows that X_2Y_2 is uniformly distributed. □

We also know that $H[Y_1]$ and $H[X_2]$ are at most as big as they would be if Y_1 and X_2 were uniformly distributed. So if we let X be uniformly distributed over A, Y be uniformly distributed over B, and E be uniformly distributed over all edges, then a lower bound for the entropy of (X_1,Y_1,X_2,Y_2) is

$$3H[E] - H[X] - H[Y].$$

Consider now the random variable $(X_1, Y_1, X_2, Y_2, X, Y)$, where (X_1, Y_1, X_2, Y_2) is as before, X is a random element of A, and Y is a random element of B, with X and Y independent of each other and of (X_1, Y_1, X_2, Y_2). By the above bound and Lemma 7.1 this random variable has entropy at least $3H[E]$, which is the entropy of the uniform distribution over all triples of edges (by Lemma 7.1). From this and Lemma 7.3, it follows that $|A||B|$ times the number of labelled $P3$s is at least the cube of the number of edges, which is $\alpha^3|A|^3|B|^3$, and from this we get that the number of labelled $P3$s is at least $\alpha^3|A|^2|B|^2$, as required.

The statement just proved is a special case of the following famous conjecture of Sidorenko.

Conjecture 7.9. *Let G be a bipartite graph with finite vertex sets X and Y and density α. Let H be another bipartite graph with vertex sets A and B and let ϕ be a random function that takes A to X and B to Y. Then the probability that $\phi(a)\phi(b)$ is an edge of G for every edge ab of H is at least $\alpha^{|E(H)|}$.*

In rough terms, this can be thought of as saying that if you want to minimize the number of copies of H in a bipartite graph of density α, then you cannot do better than to pick the edges of the bipartite graph independently at random with probability α. The conjecture has been proved for several classes of bipartite graphs, some by entropy methods, but in general it remains stubbornly open.

7.3 The formula for entropy

I once wrote a blog post about the above proof, in which I did things differently. There I defined entropy in a more usual way – by writing down a formula for it – and then I *calculated* entropies, or gave bounds in the form of specific numbers. When I started this section, I decided to try to do it axiomatically, but I wasn't sure how successful I would be. Having completed the exercise, I am now completely convinced that it is the right thing to do, as it makes it much clearer that the proof is comparing the entropy of one distribution with the entropy of another, rather than merely obtaining a numerical lower bound that gives the desired answer. Also, the proof in the blog post used Jensen's inequality, whereas this proof used the simpler maximality axiom.

So when I now give the formula, I recommend that you resist the temptation to latch on to it and use it the whole time. It should be a last resort – your proofs will be clearer if you can avoid it. It's a little like helping a much younger mathematician to get out of the habit of replacing $\sqrt{2}$ by 1.414.... At a certain age, one feels the need to experience the square root of 2 as a number with a decimal expansion, but with experience one comes to realize that what really matters is the "axioms for $\sqrt{2}$" which are

1. $\sqrt{2} > 0$.
2. $(\sqrt{2})^2 = 2$.

Of course, sometimes we use facts such as that $\sqrt{2} > 1$, but those can be deduced from the above properties and the ordered-field axioms that the reals satisfy.

With those remarks out of the way, here's the formula. If X is a discrete random variable taking values in a set A, then, writing p_a for $\mathbb{P}[X = a]$, we have

$$H[X] = \sum_{a \in A} p_a \log(1/p_a),$$

where the logarithm is to base 2. People often write this as $-\sum_{a \in A} p_a \log(p_a)$, a practice I don't like because it has a kind of clever-clever "You thought I was negative but I'm not!" aspect to it.

A quick example to help with orientation: if X is uniformly distributed over a set A of size n, then the formula tells us that $H[X] = \sum_{a \in A} n^{-1} \log n = \log n$. In particular, if $n = 2^k$, then the entropy is k, which reflects the intuitive idea that we need k bits of information to specify an element of A.

It is not hard to prove that this function satisfies the axioms given earlier. Normalization, invariance, extensibility and continuity are obvious. Maximality is a simple consequence of Jensen's inequality: the function $\log x$ is concave, so if A is any finite set, then for any random variable X taking values in A, if we write p_a for $\mathbb{P}[X = a]$, then the p_a are non-negative and sum to 1, so we have

$$\sum_{a \in A} p_a \log(1/p_a) \leq \log(\sum_{a \in A} p_a/p_a) = \log(|A|),$$

which is the entropy of a uniformly distributed random variable taking values in A.

As for the additivity axiom, let X take values in A and Y take values in B and let p_a, p_b and p_{ab} have obvious meanings (in particular, $p_{ab} = \mathbb{P}[X = a, Y = b]$). Then

$$H[X, Y] = \sum_{a \in A} \sum_{b \in B} p_{ab} \log(1/p_{ab}).$$

Now $p_{ab} = p_a \mathbb{P}[Y = b | X = a]$, so the right-hand side equals

$$\sum_{a \in A} \sum_{b \in B} \Big(p_{ab} (\log(1/p_a) + \log(1/\mathbb{P}[Y = b | X = a]) \Big).$$

But

$$\sum_{a \in A} \sum_{b \in B} p_{ab} \log(1/p_a) = \sum_{a \in A} p_a \log(1/p_a) = H[X],$$

and

$$\begin{aligned}
\sum_{a \in A} \sum_{b \in B} p_{ab} \log(1/\mathbb{P}[Y = b | X = a]) &= \sum_{a \in A} p_a \sum_{b \in B} \mathbb{P}[Y = b | X = a] \log(1/\mathbb{P}[Y = b | X = a]) \\
&= \sum_{a \in A} p_a H[Y | X = a] \\
&= \mathbb{E}_{a \in A} H[Y | X = a] \\
&= H[Y | X].
\end{aligned}$$

Thus, $H[X,Y] = H[X] + H[Y|X]$.

This proves that there is a function that satisfies the entropy axioms, and that completes the proof of the lower bound for the number of labelled $P3$s.

7.3.1 The axioms uniquely determine the formula.

It turns out that the formula we have given for entropy is the only one that satisfies the entropy axioms. To show this, we begin by working out $H[X]$ when X is uniformly distributed. (Here H is any function that satisfies the axioms for entropy. Logarithms are to base 2 throughout.)

Lemma 7.10. *If X is uniformly distributed on a set of size 2^k, then $H[X] = k$.*

Proof. Let Y be uniformly distributed on a set of size 2. Then $H[Y] = 1$ by the normalization axiom, which implies that $H[Y^k] = k$ by Lemma 7.1 and induction. Since Y^k is uniformly distributed on a set of size 2^k, $H[X] = k$ as well, by invariance. □

Lemma 7.11. *If X is uniformly distributed on a set of size n, then $H[X] = \log n$.*

Proof. For each r, X^r is uniformly distributed on a set of size n^r, and $H[X^r] = rH[X]$. Therefore, if $2^k \le n^r \le 2^{k+1}$, then $k \le rH[X] \le k+1$, by Lemma 7.3. It follows that $k/r \le H[X] \le (k+1)/r$ whenever $k/r \le \log n \le (k+1)/r$. This implies that $H[X] = \log n$ as claimed. □

Corollary 7.12. *Let X take values in a finite set A, with $p_a = \mathbb{P}[X = a]$. Then $H[X] = \sum_a p_a \log\left(\frac{1}{p_a}\right)$.*

Proof. First let us assume that there exists n such that p_a is a multiple of n^{-1} for every $a \in A$. As in the proof of Lemma 7.5 let Y be uniformly distributed on $[n]$, let Y be partitioned into sets E_a, one for each $a \in A$, with $|E_a| = p_a n$, and let us assume that $X = a$ if and only if $Y \in E_a$ (which by the invariance axiom loses no generality).

Then by additivity, $H[Y] = H[X] + H[Y|X]$. But by Lemma 7.11 $H[Y] = \log n$, and

$$H[Y|X] = \sum_a p_a H[Y|X=a] = \sum_a p_a H[Y|Y \in E_a] = \sum_a p_a \log(p_a n),$$

where for the last equality we again applied Lemma 7.11. It follows that

$$H[X] = \log n - \sum_a p_a(\log p_a + \log n) = \sum_a p_a \log\left(\frac{1}{p_a}\right).$$

As in the proof of of Lemma 7.5 we obtain the general case by applying the continuity axiom. □