Entropy and the Boltzmann Distribution

Rohan Hitchcock

16 August 2023

The concept of entropy first arose in thermodynamics. The modern expression of entropy was introduced by Ludwig Boltzmann in the proof of what is now known as *Boltzmann's H-theorem* [Bol72][1], where it was shown that the quantity we now call entropy can only increase as the system evolves over time. This showed that some processes in thermodynamics are irreversible, albeit under assumptions which were controversial at the time.

Later, in 1948, Claude Shannon introduced entropy to the world of information theory in the paper [Sha48]. In this work entropy is taken to be a measure of the 'amount of uncertainty' inherent to a probability distribution. More specifically, given a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a random variable $X : \Omega \to \mathbb{R}$, we want to be able to quantify how uncertain we are about the the value of $X(\omega)$ when we do not know $\omega \in \Omega$. This is equivalent to quantifying the amount of information about ω that we gain from knowing $X(\omega)$. From an axiomatic description of such a measurement, Shannon derives an essentially unique definition of this measurement in the case that X is a discrete random variable and identifies that it is formally identical to the expression of entropy in statistical mechanics, which by this time had become widely accepted.

In this note we approach entropy from Shannon's point of view. In the highly influential work [Jay57a; Jay57b] it is argued that the conceptions of entropy of Boltzmann and of Shannon are not just formally the same, but conceptually identical. Readers are invited to consider statistical mechanics as a form of statistical inference and in the point of view of [Jay57a; Jay57b] entropy, or more precisely entropy maximisation, plays the role of a model selection method. It is argued that choosing the distribution on phase space which maximises the entropy of the system amounts to choosing the distribution which is "maximally noncommittal with regard to missing information" [Jay57a, p. 620].

1 Entropy of a discrete random variable

We will now derive axioms for a measure of uncertainty in a discrete random variable. Having done so we will prove that there exists a unique, up to a change of units, way of measuring this uncertainty. This is [Sha48, Theorem 2].

Before doing so we will elaborate on a comment made in the introduction: that a measurement of uncertainty in a random variable can equivalently be thought of as a measurement of the information inherent to a random variable. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and recall that we can think of Ω as being the set of possible outcomes of some random experiment. We are highly uncertain about a (discrete) random variable $X : \Omega \to \mathbb{R}$ if it takes many different values with similar probabilities. That is, for any[2] $x \in \operatorname{supp} X$ we have that $\mathbf{P}(X = x)$ is reasonably small. This precisely means that if we know $X(\omega) = x$, the region of Ω in which ω must reside $X^{-1}(x) = \{\omega \in \Omega \mid X(\omega) = x\}$

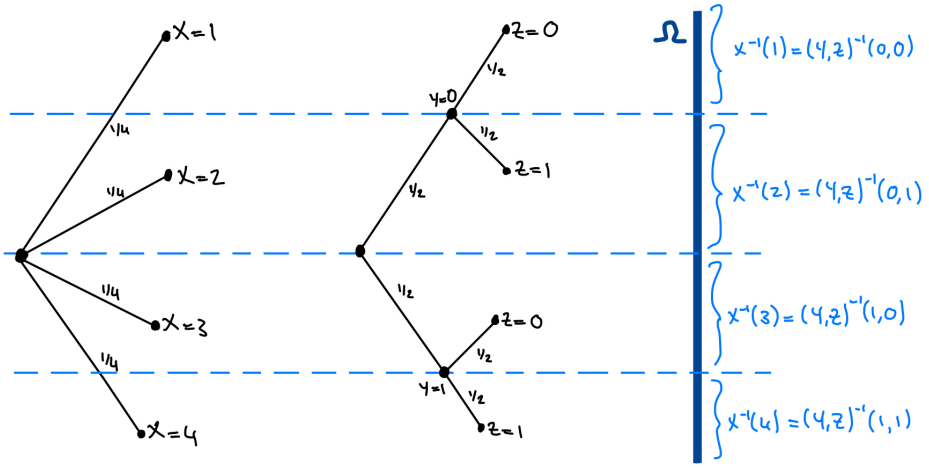
[1] An English translation is available in [Bol03].

[2] For any $x \in \mathbb{R}$ with $\mathbf{P}(X = x) > 0$.

is a small part of Ω as measured by $\mathbf{P}$. In other words, for a random variable with a high amount of uncertainty knowing that $X = x$ tells as a lot about the outcome of the random experiment, compared to a random variable with a low amount of uncertainty.

Now let us consider axioms for this measurement of uncertainty. If X is a discrete random variable let $H(X)$ denote the amount of uncertainty we have about X:

- If a random variable X takes many different values each with similar probabilities then there is more uncertainty in X compared to a random variable Y taking fewer values. Precisely, if X takes N values each with probability $1/N$ and Y takes M values each with probability $1/M$, if $M < N$ then $H(Y) < H(X)$.
- We should not be able to change the amount of uncertainty by changing how we encode random variables. By way of an example, suppose X takes the values $\{1, 2, 3, 4\}$ each with probability $1/4$. This could be encoded as two random variables: a random variable Y which takes the value 0 when $X \in \{1, 2\}$ and the value 1 when $X \in \{3, 4\}$, and a random variable Z conditionally dependent on Y. When $Y = 0$ then $Z = 0$ when $X = 1$ and $Z = 1$ when $X = 2$, and likewise for $Y = 1$:



The expression we arrive at is that

$$H(X) = H(Y) + \mathbf{P}(Y = 0)H(Z|Y = 0) + \mathbf{P}(Y = 1)H(Z|Y = 1)$$

where $H(Z|Y = i)$ the uncertainty in the conditional distribution of Z given $Y = i$, which in this case is the uniform distribution on two choices.

It turns out that these two assumptions, along with a continuity assumption, are all we need to uniquely specify H. Before proceeding to the proof of this theorem we need the following result:

Lemma 1 ([Sha48, Appendix 2])**.** *Let $A : \mathbb{N}_{>0} \to \mathbb{R}$ be a monotonic increasing function satisfying*

$$A(n) + A(m) = A(nm) \tag{1.1}$$

for all $n, m \in \mathbb{N}_{>0}$. Then $A(n) = K \log(n)$ for all n, for some fixed $K > 0$.

Proof. See Section 4. □

Let $\Delta^N = \{(p_1, \ldots, p_N) \in (0, 1)^N \mid \sum_{i=1}^N p_i = 1\}$ be the set of all probability distributions whose support has cardinality N and $\Delta = \bigcup_{N=1}^{\infty} \Delta^N$ the set of all probability distributions with finite support.

Theorem 2 ([Sha48, Theorem 2]). *Let $H : \Delta \to \mathbb{R}$ be a function satisfying the following three properties:*

(1) H is continuous.

(2) The function $A(N) = H(\frac{1}{N}, \ldots, \frac{1}{N})$ is a monotonic increasing function of N.

(3) Given $(p_1, \ldots, p_N) \in \Delta$ we can consider the distribution obtained by grouping the first n_1 events together, the next n_2, and so on up to n_M where $n_1 + n_2 + \cdots + n_M = N$. This distribution is given by $(q_1, \ldots, q_M) \in \Delta$, where $q_1 = p_1 + \cdots + p_{n_1}$, $q_2 = p_{n_1+1} + \cdots + p_{n_1+n_2}$ and so on. Then

$$H(p_1, \ldots, p_N) = H(q_1, \ldots, q_M) + q_1 H(\tfrac{p_1}{q_1}, \ldots, \tfrac{p_{n_1}}{q_1}) + q_2 H(\tfrac{p_{n_1+1}}{q_2}, \ldots, \tfrac{p_{n_1+n_2}}{q_2}) + \cdots .$$

Then H is necessarily of the form

$$H(p_1, \ldots, p_N) = -K \sum_{i=1}^{N} p_i \log(p_i) \tag{1.2}$$

for all $(p_1, \ldots, p_N) \in \Delta$, where $K > 0$ is fixed.

Proof. Suppose that H satisfies the above conditions $(1-3)$. Since H is continuous it suffices to show that (1.2) holds when $p_1, \ldots, p_N$ are rational numbers. Let $n_1, \ldots, n_N \in \mathbb{N}$ and set $p_i = n_i / \sum_{i=1}^{N} n_i$. Due to (3), we have that

$$H(p_1, \ldots, p_N) + \sum_{i=1}^{N} p_i A(n_i) = A(\textstyle\sum_{i=1}^{N} n_i) \ . \tag{1.3}$$

To explain how this follows from (3) consider the distribution on $\sum_{i=1}^{N} n_i$ choices, where each choice is equally likely. The entropy of this distribution is $A(\sum_{i=1}^{N} n_i)$. Our original distribution $(p_1, \ldots, p_N)$ is obtained by grouping the first n_1 events together into an event with probability $p_1 = n_1 / \sum_{i=1}^{N} n_i$, the next n_2 into an event with probability p_2, and so on. Applying assumption (3) yields (1.3).

Hence it suffices to determine $A(n)$ for all integers n. Let $n, m \in \mathbb{N}$ and consider the uniform distribution on nm choices, whose entropy is given by $A(nm)$. By grouping these choices into n events of equal probability $1/m$ and applying (3) again we find that A must satisfy

$$A(nm) = A(n) + A(m) \ .$$

By Lemma 1 it must be the case that $A(n) = K \log n$ for all $n \in \mathbb{N}$, for some fixed $K > 0$.

Then from (1.3) we find that

$$\begin{aligned} H(p_1, \ldots, p_N) &= K \log(\textstyle\sum_{j=1}^{N} n_j) - K \sum_{i=1}^{N} p_i \log(n_i) \\ &= K \sum_{i=1}^{N} p_i \log(\textstyle\sum_{j=1}^{N} n_j) - K \sum_{i=1}^{N} p_i \log(n_i) \\ &= K \sum_{i=1}^{N} p_i \log(\textstyle\sum_{j=1}^{N} n_j / n_i) \\ &= -K \sum_{i=1}^{N} p_i \log(p_i) \end{aligned}$$

as required. It is straightforward to check that the expression (1.2) does indeed satisfy conditions $(1-3)$. □

2 Entropy maximisation

Suppose we have a set of candidate distributions for which to model our system, all of which are compatible with our partial knowledge the system. It is argued in [Jay57a] that, since entropy is a measurement of the uncertainty inherent to a probability distribution, we should choose the one with maximal entropy. Choosing a distribution which does not maximise entropy, it is argued, amounts to introducing additional assumptions which are not based on our knowledge of the system. In this section we find the maximal entropy distribution in the case we know the expected value of some function.

Consider Δ^N as probability distributions on the N-set $\mathcal{X} = \{x_1, \dots, x_N\}$, meaning we identify $\boldsymbol{p} = (p_1, \dots, p_N) \in \Delta^N$ with a random variable $X_{\boldsymbol{p}}$ such that $X_{\boldsymbol{p}} = x_i$ with probability p_i for all $i = 1, \dots, N$.

Theorem 3. *For some function $f : \mathcal{X} \to \mathbb{R}$ not identically zero and $F \in \mathbb{R}$ we consider the distributions $\boldsymbol{p} = (p_1, \dots, p_N) \in \Delta N$ such that $\mathbf{E}f(X_{\boldsymbol{p}}) = F$. That is*

$$F = \sum_{i=1}^{N} p_i f(x_i) \ . \tag{2.1}$$

Of the distributions satisfying (2.1), the one with maximal entropy is the Boltzmann distribution on N-symbols. That is, for each $i = 1, \dots, N$

$$p_i = \frac{1}{Z} e^{-\beta f(x_i)} \qquad \text{where } Z = \sum_{j=1}^{N} e^{-\beta f(x_j)}$$

for some β.

Proof. We wish to maximise the entropy

$$H(\boldsymbol{p}) = -\sum_{i=1}^{N} p_i \log(p_i)$$

subject to the constraints

$$\sum_{i=1}^{N} p_i = 1 \qquad \text{and} \qquad \sum_{i=1}^{N} p_i f(x_i) = F \ .$$

Since f is not identically zero we can use the method of Lagrange multipliers. Set

$$L = -\sum_{i=1}^{N} p_i \log p_i + \alpha(1 - \sum_{i=1}^{N} p_i) + \beta(F - \sum_{i=1}^{N} p_i f(x_i))$$

for some $\alpha, \beta \in \mathbb{R}$. We have

$$\frac{\partial L}{\partial p_i} = -\log(p_i) - 1 - \alpha - \beta f(x_i)$$

and setting $\frac{\partial L}{\partial p_i} = 0$ and solving for p_i gives

$$p_i = \frac{e^{-\beta f(x_i)}}{e^{1+\alpha}}$$

Set $Z = e^{1+\alpha}$. Since $\sum_{i=1} p_i = 1$ we have that

$$Z = \sum_{j=1}^{N} e^{-\beta f(x_j)}$$

□

Theorem 3 can be extended to the case of multiple constraints. If we have $f_1, \dots, f_K : \mathcal{X} \to \mathbb{R}$ and $F_1, \dots, F_K \in \mathbb{R}$ such that

$$F_k = \sum_{i=1}^{K} p_i f_k(x_i) \qquad \text{for all } k = 1, \dots, K$$

then one can show using the same method that the choice of $p_1, \dots, p_N$ with maximal entropy is

$$p_i = \frac{1}{Z} e^{-\sum_{k=1}^{K} \beta_k f_k(x_i)} \qquad \text{where } Z = \sum_{j=1}^{N} e^{-\sum_{k=1}^{K} \beta_k f_k(x_j)}$$

for some $\beta_1, \dots, \beta_K$.

3 Continuous distributions

Up until now we have been working only with probability distributions that have finite support. Note that the entropy of a discrete random variable X is the expectation $\mathbf{E} \log(p(X))$ where $p(x)$ is the probability mass function of X. This suggests the following definition:

Definition 4. Let Y be a continuous random variable and $p(y)$ its probability density function. The *differential entropy* of Y is

$$H(Y) = \mathbf{E} \log(p(Y)) = \int_{\mathcal{Y}} p(y) \log(p(y)) dy \ .$$

Differential entropy was also defined in [Sha48, Section 20], however unlike the discrete case no justification was given. An alternative is looking at the *relative entropy* to some fixed distribution $\varphi(y)$, also known as as *Kullback-Leibler divergence*:

$$\mathrm{KL}(Y|\varphi) = \int_{\mathcal{Y}} p(y) \log(p(y)/\varphi(y)) dy$$

From the relative entropy one can define the Boltzmann distribution in the continuous case in the obvious way:

$$p(y) = \frac{1}{Z} e^{-\beta f(y)} \qquad Z = \int_{\mathcal{Y}} e^{-\beta f(y)} \varphi(y) dy$$

Justification of these definitions will be deferred to a future note.

4 Proof Lemma 1

We now prove Lemma 1, following the proof in [Sha48, Appendix 2]. First note that by substituting $n = m = 1$ into (1.1) we have $A(1) = 0$, and so since A is monotonically increasing $A(n) > 0$ for any $n \neq 1$.

Let $t, s \in \mathbb{N}_{>0}$ where $s > t > 1$. We first aim to show that $\frac{A(s)}{A(t)} = \frac{\log(s)}{\log(t)}$. Let $n \in \mathbb{N}$ be arbitrary and choose $m \in \mathbb{N}$ sufficiently large so that

$$s^m \leq t^{2n} < s^{m+1} \tag{4.1}$$

Taking logarithms and then dividing by $2n\log(s)$ gives

$$\frac{m}{2n} \leq \frac{\log(t)}{\log(s)} < \frac{m}{2n} + \frac{1}{2n} \ . \tag{4.2}$$

Now, by (1.1) we have that $A(s^m) = mA(s)$ and $A(t^{2n}) = 2nA(t)$ and so by applying A to (4.1) and dividing by $2nA(s)$ we find

$$\frac{m}{2n} \leq \frac{A(t)}{A(s)} < \frac{m}{2n} + \frac{1}{2n} \tag{4.3}$$

where we have also used the fact that A is monotonically increasing. From (4.2) and (4.3) we have

$$\left|\frac{\log(t)}{\log(s)} - \frac{m}{2n}\right| < \frac{1}{2n} \qquad \text{and} \qquad \left|\frac{A(t)}{A(s)} - \frac{m}{2n}\right| < \frac{1}{2n}$$

and so

$$\left|\frac{\log(t)}{\log(s)} - \frac{A(t)}{A(s)}\right| \leq \left|\frac{\log(t)}{\log(s)} - \frac{m}{2n}\right| + \left|\frac{\log(t)}{\log(s)} - \frac{m}{2n}\right| < \frac{1}{n} \ .$$

Since this holds for all n we have $\frac{\log(t)}{\log(s)} = \frac{A(t)}{A(s)}$.

To complete the proof, let K_t and K_s be such that $A(s) = K_s \log(s)$ and $A(t) = K_t \log(t)$. From $\frac{\log(t)}{\log(s)} = \frac{A(t)}{A(s)}$ we find $K_t = K_s$. This holds for any $s > t > 1$ and so we have $A(n) = K\log(n)$ for any $n > 1$, for some fixed $K \in \mathbb{R}$. That $K > 0$ follows from A being monotonically increasing, and we have already seen that $A(1) = K\log(1) = 0$. $\square$

References

[Bol03] Ludwig Boltzmann. 'Further Studies on the Thermal Equilibrium of Gas Molecules'. Stephen G. Brush and Nancy S. Hall. *The Kinetic Theory of Gases: An Anthology of Classic Papers with Historical Commentary*. Imperial College Press, 2003, pp. 262–349.

[Bol72] Ludwig Boltzmann. 'Weitere Studien Über Das Wärmegleichgewicht Unter Gasmolekülen'. *Sitzungsberichte Akad. Wiss.* 66 (1872), pp. 275–370.

[Jay57a] E. T. Jaynes. 'Information Theory and Statistical Mechanics'. *Physical Review* 106.4 (15 May 1957), pp. 620–630.

[Jay57b] E. T. Jaynes. 'Information Theory and Statistical Mechanics. II'. *Physical Review* 108.2 (15 Oct. 1957), pp. 171–190.

[Sha48] C. E. Shannon. 'A Mathematical Theory of Communication'. *The Bell System Technical Journal* 27.3 (1948), pp. 379–423.