Notes on Information Theory and Statistical Mechanics by ET Jaynes

compiled by D. Gueorguiev, 12/31/24

# Introductory Notes

Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge and leads to a type of statistical inference – the *maximum entropy estimate*. The maximum entropy estimate is the least biased estimate possible on the given information – it is maximally noncommittal with regard to missing information. If one considers statistical mechanics as a form of statistical inference rather than as a physical theory, it is found that the usual computational rules, starting with the determination of the partition function, are an immediate consequence of the maximum-entropy principle.

In the resulting “subjective statistical mechanics”, the usual rules are thus justified independently of any physical argument, and in particular independently of experimental verification; whether or not the results agree with experiment, they still represent the best estimates that could have been made on the basis of the information available.

# Maximum-Entropy Estimates

The quantity is capable of assuming the discrete values . We are not given the corresponding probabilities ; all we know is the expectation value of the function :

(1)

On the basis of this information, what is the expectation value of the function ? At first glance, the problem seems insoluble because the given information is insufficient to determine the probabilities . (1) and the normalization condition

(2)

would have to be supplemented by more conditions before could be found.

The problem of specification of probabilities in case where little or no information is available was attempted to be resolved through a criterion of choice, in which one said that two events are to be assigned equal probabilities if there is no reason to think otherwise. However, except in cases where there is an evident element of symmetry that clearly renders the events “equally possible”, this assumption may appear just as arbitrary as any other that might be made. Furthermore, it has been very fertile in generating paradoxes in the case of continuously variable random quantities, since intuitive notions of “equally possible” are altered by a change of variables. Hence this way of resolving probabilistic distributions in problem has been abandoned as lacking constructive principle.

Again, our problem is to find a probability assignment which avoids bias, while agreeing with whatever information is given. Using information theory we can devise a unique, unambiguous criterion for the amount of uncertainty represented by a discrete probability distribution, which agrees with our intuitive notions that a broad distribution represents more uncertainty than does a sharply peaked one.

### Finding the explicit form of the uncertainty measure per Jaynes’ paper

This exposition is very similar to that in Shannon’s paper [3] and in Ash’s Information Theory [4].

Let us denote with the quantity which measures in a unique way the amount of uncertainty represented by this probability distribution. The three conditions on are

1. is a continuous function of the .
2. If all are equal, the quantity is a monotonic increasing function of .
3. The composition law: instead of giving the probabilities of the events directly, we might group the first of them together as a single event, and give its probability ; then the next possibilities are assigned the total probability . When this much has been specified, the amount of uncertainty as to the composite events is . Then we give the conditional probabilities of the ultimate events , given that the first composite event had occurred, the conditional probabilities for the second composite event and so on. We arrive ultimately at the same state of knowledge as if the has been given directly, therefore if our information measure is to be consistent, we must obtain the same ultimate uncertainty no matter how the choices were broken down in this way. Thus, we must have

(1)

The weight factor appears in the second term because the additional uncertainty is encountered only with probability . For example, .

From condition 1) , it is sufficient to determine for all rational values

with integers. Condition 2) implies that H is determined already from the quantities . We can regard a choice of one of the alternatives as a first step in the choice of one of

equally likely alternatives, the second step of which is also a choice between equally likely alternatives.

As an example, with , we might choose . For this case the composition law becomes

In general this could be written:

(2)

In particular, we could choose all equal to , which transforms (2) to

(3)

(3) is solved by setting:

(4)

where, by condition (2), .

The Proof that (4) is the only solution of (2) see Shannon’s paper [3] or Ash’ Information Theory [4].

More thorough analysis on the form of the Uncertainty Measure is developed in the next Section which follows the discussion in Shannon’s paper [3] and is derived rigorously in Ash’s Information Theory book [4].

## Axioms for the Uncertainty Measure

Question: Is it possible to find any quantity which measures in a unique way the amount of uncertainty represented by this probability distribution.

Suppose that a probabilistic experiment involves the observation of a discrete r.v. . Let take on a finite number of possible values with probabilities , respectively. We assume that all are strictly greater than zero and . We now attempt to arrive at a number that will measure the uncertainty associated with . We shall construct two functions and . The function will be defined on the interval .

will be interpreted as the uncertainty associated with an event with probability . Thus if the event has probability , we shall say that is the uncertainty associated with the event , or the uncertainty removed (or information conveyed) by revealing that has taken on the value in a given performance of the experiment. For each we shall define a function of the variables (we restrict the domain of by requiring all to be , and ). The function is to be interpreted as the average uncertainty associated with the events ; we require that . We abbreviate as or as . Thus is the average uncertainty removed by revealing the value of . The function is introduced merely as an aid to the intuition as it is helpful to think of as a weighted average of the numbers .

Now we proceed to impose requirements on the functions . The function will be referred as the “uncertainty of ”; the word “average” will be understood but not explicitly stated in the future discussions.

First, suppose that all values of are equally probable. We denote by the average uncertainty associated with equally likely outcomes, that is, . For example, would be the uncertainty associated with the toss of unbiased coin, while would be the uncertainty associated with picking a person at random in New York City. The expectation, of course, is that the uncertainty in the latter situation to be greater than that of the former. In fact, our first requirement on the uncertainty function is that should be a monotonically increasing function of .

With this requirement in mind, consider an experiment involving two independent variables and . Let take on the values with equal probability and let take on the values , also with equal probability. Thus the joint experiment involving and has equally likely outcomes, and therefore the average uncertainty about should not be affected because of the assumed independence of and . Hence, we expect that the average uncertainty associated with and together, minus the average uncertainty removed by revealing the value of , should yield the average uncertainty associated with . Revealing the value of removes, on the average, an amount of uncertainty equal to , and thus the second requirement on the uncertainty measure is that

,

At this point we remove the restriction of equally likely outcomes and turn to the general case. We divide the values of a r.v. into two groups, and , where consists of and consists of . We construct a compound experiment as follows: first we select one of the two groups, choosing group with probability and group with probability . Thus the probability of each group is the sum of the probabilities of the values in the group. If group is chosen, then we select with probability . Similarly, if group is chosen, then is selected with probability . This compound experiment is equivalent to the original experiment associated with . For if is the result of the compound experiment, the probability that is

Similarly, so that and have the same distribution. Before the compound experiment is performed, the average uncertainty associated with the outcome is . If we reveal which of the two groups and is selected, we remove on the average an amount of uncertainty . With probability , group is chosen and the remaining uncertainty is:

;

with probability , group is chosen, and the remaining uncertainty is

Thus *on average* the uncertainty remaining after the group is specified as

.

We expect that the average uncertainty about the compound experiment minus the average uncertainty removed by specifying the group equals the average uncertainty remaining after the group is specified. Hence, the third requirement we impose on the uncertainty function is that

.

As a numerical example of the above requirement, we may write

where and .

Finally, we require for mathematical convenience that be a continuous function of . That is, small changes in the probabilities of the values of corresponds to a small change in the uncertainty of .

To recap we formulated the following four conditions

1 ) is a monotonically increasing function of .

2 )

3 )

/\* grouping axiom \*/

4 ) is a continuous function of

**Statement**: The only function satisfying the four given axioms is

(1)

where is an arbitrary positive number, and the logarithm base is any number greater than 1.

Proof:

It is not difficult to verify that:

(1) satisfies 1)-4).

Changing the logarithm base in (1) is equivalent to changing the constant .

**( a )** We would like to prove that for all positive integers and . This is readily established by induction using 2).

**( b )** where C is a positive number.

First, let . We have by 2), hence . Obviously, there should not be any uncertainty in an experiment with only one possible outcome. Now let be a fixed positive integer greater than 1. If is an arbitrary positive integer, then the number lies somewhere between two powers of , that is, there is a nonnegative integer such that . It follows from 1) that , and this we have from (a) . The logarithm with base is a monotone increasing function and hence , from which we obtain . Since and are both between and we ca write

Since is fixed and is arbitrary, we may let to obtain

or where . Note that must be positive since and increases with .

**( c )** if is a rational number.

Let where and are positive integers. We consider

/\* by the grouping axiom \*/

Using (b) we have

.

Thus,

.

**( d )**

# References

[1] [Information Theory and Statistical Mechanics, E.T. Jaynes, Department of Physics, Stanford University, 1957](https://github.com/dimitarpg13/information_theory_and_statistical_mechanics/blob/main/literature/articles/Information_theory_and_statistical_mechanics_part1_Jaynes_1957.pdf)

[2] [Information Theory and Statistical Mechanics, E.T. Jaynes, Department of Physics, Stanford University II, 1957](https://github.com/dimitarpg13/information_theory_and_statistical_mechanics/blob/main/literature/articles/Information_theory_and_statistical_mechanics_part2_Jaynes_1957.pdf)

[3] [Mathematical Theory of Communication, Claude Shannon, 1957](https://github.com/dimitarpg13/information_theory_and_statistical_mechanics/blob/main/literature/articles/The_Mathematical_Theory_of_Communication_C_Shannon_bstj_1948_ver2.pdf)

[4] [Information Theory, Robert B. Ash, 1965](https://github.com/dimitarpg13/information_theory_and_statistical_mechanics/blob/main/literature/books/Information_Theory-Robert_Ash.pdf)