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.

# 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 . /\* monotonicity axiom \*/

**2 )** . /\* additivity axiom \*/

**3 )**

/\* grouping axiom \*/

**4 )** is a continuous function of . /\* continuity axiom \*/

# Constructing the Uncertainty Measure using the four axioms

**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 monotonosly 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 )**

This is an immediate consequence of ( c ) and the continuity axiom 4).

This is because if by the continuity axiom

where is a rational number ()

**( e )**

We again proceed by induction. We have already established the validity of the above formula for and . If , we use the grouping axiom which yields

.

Assuming the formula valid for positive integers up to , we obtain:

We shall assume and all logarithms with base .The units of are called *bits*. One bit of uncertainty is associated with the toss of unbiased coin. Biasing the coin tends to decrease the uncertainty as shown on the Figure below showing .

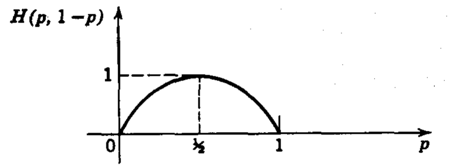


Figure :

# Constructing uncertainty measure based on the third Cauchy’s functional equation

An alternative approach to the construction of an uncertainty measure involves a set of axioms for the function .

If and are independent events, the occurrence of should not affect the odds that will occur.

If , then so that the uncertainty associated with is .

The statement that A has occurred removes an amount of uncertainty , leaving an uncertainty because of the independence of and . Thus we might require that

, , .

Two other reasonable requirements on are that be a decreasing function of p, and that h be continuous. The only function satisfying these three requirements is of the form . For details on the solution see the Appendix.

## 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 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].

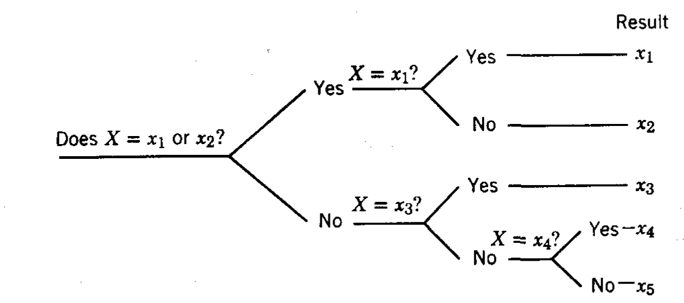
# Three Interpretations of the Uncertainty Function

## First Interpretation of the Uncertainty Function

The uncertainty function is a weighted average of the numbers , where the weights are the probabilities of the various values of . This suggests that may be interpreted as the **expectation** of a r.v. which assumes the value with probability .

## Second Interpretation of the Uncertainty Function

Another interpretation of is closely related to the construction of codes. Suppose takes on five values with probabilities 0.3, 0.2, 0.2, 0.15 and 0.15, respectively.

Suppose that the value of is to be revealed to us by someone who cannot communicate except by means of the words “yes” and “no”.

So we try to arrive at the correct value by a sequence of “yes” and “no” questions as shown on the Figure above.

Then the average number of questions is :

*Noiseless Coding Theorem*: The average number of “yes” or “no” questions needed to specify the value of can never be less than the uncertainty of .

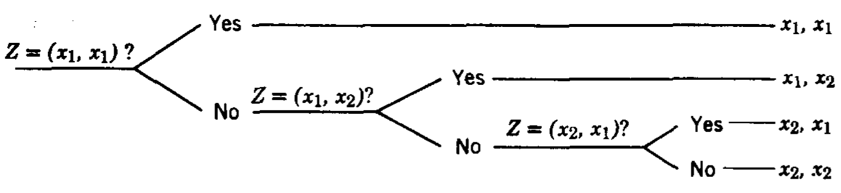
In this example:

Further , no scheme can be devised which on average uses fewer questions than the one described above.

**Note 1**: we can improve the efficiency of our guesswork if we assume that the experiment associated with is performed independently n times and that we are allowed to wait until all n observations are recorded and then guess all n results simultaneously.

Example: Let the r.v. take on two values, and with probabilities and respectively. Thus, one question is needed to specify the value of . If, however, we are allowed to make guesses about the outcome of a joint experiment involving two independent observations of , we may use the scheme shown on the Figure below.

This scheme uses questions on the average or questions *per value* of .



**Note** **2**: It can be shown that by making guesses about longer and longer blocks, the average number of questions *per value* of can be made to approach as closely as desired. Thus we arrive at the second interpretation of the uncertainty measure:

= the minimum average number of “yes” or “no” questions required to determine the result of one observation of .

## Third Interpretation of the Uncertainty Function

There is a third interpretation of uncertainty which is related to the asymptotic behavior of a sequence of independent, identically distributed r.v.’s.

Let be a r.v. taking on the values with probabilities respectively. Suppose that the experiment associated with is performed independently times. In other words, we consider a sequence of independent, identically distributed r.v.’s, each having the same distribution as . Let be the number of times that the symbol occurs in the sequence ; then has a binomial distribution with parameters and . Given choose any positive number such that ; We fix and for the rest of the argument.

//TODO: finish the section on the interpretations of the uncertainty function

# 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.

We have already defined axioms based on which we have constructed a measure of uncertainty , increasing when uncertainty is higher, and it is additive for independent sources of uncertainty:

(3)

where is positive constant. We will denote this quantity as the *entropy of the probability distribution* .

We want to solve our problem (1); in making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to the partial knowledge we are given.

This is the **only unbiased assignment we can make**; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have.

To maximize (3) subject to the constraints (1) and (2) we introduce Lagrangian multipliers and :

where (4)

subject to :

and (5)

In matrix notation (5) can be rewritten as

(6)

where is the dimensional unit row vector and .

The Karush-Kuhn-Tucker condition for (4) subject to (6) is:

(7)

Here is the dimensional unit column vector, and is the dimensional zero column vector.

( for details on the KKT condition and the Lagrange multipliers see for instance *Chapter 6* of [7] )

Since (7) becomes

(8)

where

(9)

(9) in (5) leads to

; (10)

(11)

is known as the partition function of the max entropy distribution.

The maximum entropy estimation described in (1) – (11) can be generalized for a linear combination of functions :

(12)

(13)

Then the maximum-entropy probability distribution is given by

(14)

(15)

(16)

(17)

where the constant in (3) has been set equal to unity. The variance of the distribution of is found to be

(18)

Proof:

//TODO: finish the section on max entropy estimates

# 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)

[5] [Note on the Functional Equations , Joseph Milkman, Proceedings of the AMS, 1950](https://www.jstor.org/stable/2032319)

[6] [Introduction to Functional Equations, Costas Eftimiou, Chapter 5: Cauchy’s Equations, 2010](https://github.com/dimitarpg13/information_theory_and_statistical_mechanics/blob/main/literature/books/Introduction_to_functional_equations_Efthimiou_2010.pdf)

[7] [Optimization: Principles and Algorithms, Michel Bierlaire, EPFL, 2018](https://github.com/dimitarpg13/optimization_classification_regression/blob/main/literature/books/OptimizationPrinciplesAlgorithms_Bierlaire_EPFL_2018.pdf)

# Appendix

## The First Cauchy Equation

**Problem Statement**:

Let be a continuous function of a continuous real variable that satisfies the functional relation

, (A1.1)

Find all functions that satisfy the above conditions.

Eq. (A1.1) is known as the *first Cauchy functional equation* of the *linear Cauchy functional equation*.

Solution:

Setting in the defining equation, we see that

Also, if ,

. (A1.2)

We now notice that for any natural number ,

(A1.3)

This is verified by induction. For , the equation above is true by definition. Assuming it is true for some ,

, (A1.4)

we see that it is true for :

. (A1.5)

This proves the statement; that is, (A1.3) is true for all natural numbers.

Setting in (A1.3)

(A1.6)

If moreover, , with ,

. (A1.7)

From this equation and (A1.2), we conclude that

(A1.8)

In other words, so far, we’ve proved that

, (A1.9)

When ,

,

where we set .

Now, let . There is a sequence of rational numbers such that

.

For the terms of the sequence, the function gives

,

and since it is continuous

.

## The Second Cauchy Equation

**Problem Statement**:

Let be a continuous function of a real variable that satisfies the functional relation

(A2.1)

Find all functions that satisfy the above conditions are not identically zero.

Eq. (A2.1) is the *second Cauchy functional equation* or the *exponential Cauchy functional equation*.

Solution 1: First, we will show that is positive for any real number . From the defining relation, we see that , and therefore . Let be a real number for which . Then , i.e. the function would vanish identically. If there are solution which do not vanish identically, then they cannot vanish at any point.

Since , the function is well-defined. Then the defining Eq. (A2.1) can be written in terms of the function as:

,

which has the unique continuous solution . From this the function f is found to be

where we set .

If we want to avoid using Eq. (A1.3) as a solution for (A1.1) there is an alternative solution:

Solution 2:

Similarly with Solution 1 we prove that .Setting in the defining equation, we see that

,

and therefore we must have .

Also, if ,

(A2.2)

We now notice that for any natural number

(A2.3)

This is verified by induction. For , Eq (A2.3) is true by definition. Assuming, that it is true for some

,

we see that it is true for :

//TODO: finish the second on the second Cauchy Equation

## The Third Cauchy Equation

**Problem Statement**:

Let be a continuous function of a continuous real variable that satisfies the functional relation

(A.3)

Find all functions that satisfy the above conditions.

Eq. (A3.1) is known as the *third Cauchy functional equation* or the *logarithmic Cauchy functional equation*.

**Solution 1**: Since the independent variable takes values in , the change of variable leads to a variable taking values in . Then the function is a function , satisfying the relation

or

.

The solution to this functional relation is which can be inverted to give .

//TODO: finish the section on the third Cauchy Equation