

W. W. HANSEN LABORATORIES OF PHYSICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

MICROWAVE LABORATORY
HIGH ENERGY PHYSICS LABORATORY

DAvenport 3-2441

June 17, 1957

Mr. Hugh Everett, III
Institute for Defense Analyses
The Pentagon
Washington, D. C.

Dear Mr. Everett:

Thank you for your letter of June 11, 1957, concerning my paper, "Information Theory and Statistical Mechanics." I appreciate very much your taking the trouble to write in such detail, because when developing a new theory it is always a good idea to try to understand it from as many different viewpoints as possible, and your letter has shown me how to add one more to my "arsenal" of different ways to explain the entropy principle. I am sending you, under separate cover, a preprint copy of the second paper in this series, which was submitted to the Phys. Rev. a couple of months ago.

Before turning to the specific points raised in your letter, a little historical review of this work might be of interest to you. I "discovered" Information Theory eight years ago, while in Princeton, and immediately felt that at last I understood what statistical mechanics was all about. It seemed to me at the time that every physicist who read Shannon's two papers would see at once that this had great implications for the proper interpretation of statistical mechanics, and that Information Theory was going to be the thing which would make the subjective view of probability once more respectable. Unfortunately, this did not happen, and physicists have tended to avoid information theory. I think the reason for it is that the subject appears so sensational at first; one has the impression that he is getting something for nothing.

For the past seven years, I have been working on this theory with students at Stanford, and giving quite a few seminar talks. No attempt at publication was made until I felt that I had several new results to communicate, rather than merely an exposition of a viewpoint.

The strange thing about the information principle is that the difficulties are not mathematical, but conceptual. The mathematics is very elementary, but there is the greatest difficulty in finding the proper words to convey its meaning.

The trouble here is just one well recognized in semantics, that bad habits of language can influence our thought processes, and can even prevent us from seeing very simple things. The difficulties which were first encountered in trying to explain the principle of Relativity and the Complementarity principle, illustrate this very well, and I think that statistical mechanics is long overdue for a similar house-cleaning in habits of expression. In particular, the terms, "assumption" and "a-priori probability" carry so many different and conflicting meanings that one really should avoid them if ideas are to be put across with reasonable fidelity. The objections which you raise show that I did not do this sufficiently well in my paper.

In a purely utilitarian sense, discussions of the nice-ties of just what "assumptions" are in the theory, are unimportant, because it is clear that the reliability of the predictions comes from the principle of macroscopic uniformity and not from our having used the "correct" information measure. J. M. Richardson tells me that he has a rather general theorem on this. However, the questions of principle are really the most interesting, and they may lead to more general theories, so they should be discussed very carefully.

The fact that $\sum p_i \log(p_i/u_i)$ satisfies simpler inequalities than $\sum p_i \log p_i$, is something which one finds immediately when one tries to work with this theory. I learned it from translating the chapter of Gibbs, on maximum and minimum properties, into modern notation. In fact, all of his inequalities make use of it. Also, the fact that the entropy of a continuous distribution is an information measure relative to a certain standard weighting, was pointed out by Shannon in his second paper. In application to classical statistical mechanics, Liouville's theorem and the more general invariance of phase volume under arbitrary canonical transformations, give at least a strong hint about the reasonable standard weighting to adopt, but I think this is still a little forced, and there must be some new way to handle this. At present I regard the justification for equal weighting of equal phase volumes to lie principally in the fact that it goes over into uniform weighting of different quantum states, when one makes the transition to quantum theory. The justification for the latter is discussed below.

For the situation treated in my first paper, however, I think that the procedure I gave is the only reasonable one to follow. Consider first the case where the only available information consists of the possible values of x , and certain averages. Here there can be no question about introducing a special weighting u_i ; that would, again amount to arbitrarily

Is p what we know or available?

assuming something which was not given in the statement of the problem. You may claim that nevertheless there exists some "correct" nonuniform weighting, but I will reply, "If we don't know this "correct" weighting, what else can we do?" This theory is not one where we allow ourselves the luxury of refusing to work on a problem merely because we don't have the type of information we would like; we must do the best we can with what we have. I prefer to put this in stronger form: to assert that some nonuniform weighting exists when the available information gives no evidence for it, is logically equivalent to asserting that there is vegetation on the moon, but it is all on the other side, where we cannot see it. On the other hand, as I will show below, whenever the available information does tell us something about the weighting to adopt, my theory as formulated automatically takes this into account.

Whether one should call the absence of a weighting function an "assumption of equal a-priori probabilities," is just a question of semantics. It is the customary habit of language which has been developed, but I think it is totally misleading, and in fact prevents us from seeing how simple the problem really is. There are two reasons for this.

In the first place, such a manner of speaking gives one the impression that there are certain a-priori probabilities existing before knowledge of the averages is introduced, which are then modified by incorporating this additional information into the problem, thus becoming the a-posteriori (maximum-entropy) probabilities. If this were the case, however, the application of the information principle would be a special case of the application of Bayes' theorem. But any attempt to interpret it in this way fails, because the likelihood ratio then involves probabilities conditional on mutually contradictory hypotheses, as you easily see upon writing out Bayes' theorem for this case. This shows that we should interpret the theory as follows. The average values represent part of the initial information, so that the probabilities resulting from application of the entropy principle are themselves the a-priori probabilities; the information principle merely tells us the consistent way of finding this assignment. Stated differently, the information principle is just a tool which helps us to translate "raw" information into numerical values of a-priori probabilities. As time goes on, this probability assignment may then undergo modifications representing our knowledge of the equations of motion. This gives a theory of "irreversible processes," as shown in my second paper.

For the second reason, note that if we agree to abide by the usual mathematical rules of probability theory, in particular that probabilities are to be normalized to unity ($\sum p_i = 1$), the rules of the game are quite unambiguous. We simply enumerate the different, mutually exclusive possibilities, then assign

probabilities to them in such a way that the entropy is maximized subject to whatever is known. The fact that the states are mutually exclusive is required by the mathematical rules, because the formula $p(A \text{ or } B) = p(A) + p(B)$ holds only when A and B are mutually exclusive. This in turn implies that we do not count the same state twice; each state must be counted once and only once. It is misleading to call this an assumption of equal a-priori probabilities, because to call it an assumption implies that we were free to do something different. Not only the above argument, but also simple common sense denies us this freedom. It makes sense to say, "The system may be in state A or it may be in state B." It makes no sense to say, "The system may be in state A or it may be in state A." But if we count state A twice in our enumeration, and then continue to normalize our probabilities to unity, our equations would have just the logical content of the latter statement. For this reason, I claim that the theory contains no assumption at this point.

One may, of course, consider the problem of inference in other cases than the one in my paper, for example where one has additional information not expressible in the form of enumeration or average values. The problem of maximizing the entropy subject to the available information might then take a quite different mathematical form than in my paper, and/or new principles might have to be discovered. However, this is not the case if the additional information is of a type which leads merely to a different weighting function than the uniform one, and in fact the theory as I described it includes this possibility.

Let me now try to explain this more carefully. You claim that my theory is only a special case of your theory, with one particular information measure. I can, with equal justice, claim that your theory is a special case of mine. In Section 5 of my first paper, it is shown how the probability distribution which represents our state of knowledge is the same whether the given information consists of the value of $\langle f_r(x) \rangle$ or its statistically conjugate quantity λ_r . In the latter case, one is always free, however, to interpret the theory in the way you suggested. If I know λ_r rather than $\langle f_r(x) \rangle$, I can always define $u_i = \exp [-\lambda_r f_r(x_i)]$, and say that I am minimizing, not $\sum p_i \log p_i$ of some larger system (including the "heat bath"), but rather $\sum p_i \log(p_i/u_i)$ of the small system of interest. Thus, for example, if the available information consists of the temperature and $\langle f(x) \rangle$, where $f(x)$

is not the energy, my theory gives the result of using an information measure defined relative to the Boltzmann distribution (which in this case is the stationary one), and minimized subject to the constraint represented by the value of $\langle f(x) \rangle$. You see that this comes about automatically, whenever the available information is of the type which justifies any such nonuniform weighting, and thus it cannot be considered as an extension of the theory or as anything which has to be taken into account separately.

Your two-state Markov process with the stationary distribution $p_1^* = 1/3$, $p_2^* = 2/3$, is a very nice example of this. The fact that the entropy relative to the stationary distribution never decreases is a very useful property, which has been used in thermodynamics for a long time, but it is not a property which has any reasonable interpretation in terms of information. Here the true entropy, $-\sum p_i \log p_i$, does not always increase, but may go either up or down depending on the initial state. This model is, in fact, just the one describing an atom with two energy levels, separated by $\delta E = E_1 - E_2$, which eventually comes to thermal equilibrium with a heat bath of a temperature such that $\exp(\delta E/kT) = 2$. If we start out in the state $p_1 = p_2 = 1/2$, corresponding to "infinite initial temperature," then the system cools down, and its entropy decreases. In this case, we obviously gain, rather than lose information about the state of the system by allowing the stochastic process to take place. There is nothing at all contradictory about this - in fact, if you prove that the entropy of a system can never decrease, then you have proved too much, for the fact is that eggs do manage, somehow, to cool off when placed in a refrigerator. In any reasonable measure of information, we obviously know more about the state of a cold egg than a hot one, since for every possible quantum state of the former, there will be something like $(10^{10})^{20}$ possible states of the latter.

Now let us see what your conditional entropy means in this model. Your weighting function is $p_i^* = (\text{const.}) \cdot \exp[-E_i/kT]$, and the conditional entropy reduces to

$$-\sum p_i \log(p_i/p_i^*) = S - \frac{U}{kT} = \Psi$$

This is just the Planck thermodynamic potential, which reaches its maximum value for thermal equilibrium, a criterion which physical chemists have been using for over 50 years.

Let me now turn to your remark, "If you try to make predictions about this example using a minimum $\sum p_i \log p_i$, you will make worse predictions than I, who use $\sum p_i \log(p_i/p_i^*)$, since I take into account the known fact that this system is not equally likely to be in any of its states." I think the answer

is clear from the above; I will use uniform weighting, minimize $\sum p_i \log p_i$, but I will also incorporate into my equations the fact that the quantity statistically conjugate to the energy is known. You will use nonuniform weighting, minimize $\sum p_i \log(p_i/p_i^*)$, and in so doing, consider only the other constraints. We will, in fact, be writing down exactly the same probability distributions if you allow me to use all the information which you used. If we base our subjective probability assignments on the same data, we will end up by making the same predictions, with possibly a little difference in the philosophy of how we got them.

I am not sure whether I have done a good job of exposition in this letter, and if there remain any doubts in your mind, I will be happy to try again. Finally, let me thank you once more for starting me on the train of thought which showed the equivalence of these two points of view.

Very truly yours,

E.T. Jaynes

E. T. Jaynes
Associate Professor
Microwave Laboratory

ETJ/nmm