WHERE DO WE GO FROM HERE?

E. T. Jaynes

Arthur Holly Compton Laboratory of Physics, Washington University, St. Louis, Missouri 63130

C. Ray Smith and W. T. Grandy, Jr. (eds.),
Maximum-Entropy and Bayesian Methods in Inverse Problems, 21–58.
© 1985 by D. Reidel Publishing Company

1. Introduction

With this meeting, we enter a new era for the Principle of Maximum Entropy. Gathered in one room are many people who have been working largely in isolation from each other, while thinking very similar thoughts. Each has discovered, in his own way and in his own context, that this principle solves real, nontrivial problems in a way that cannot be approached by other statistical methods.

But future progress will be more rapid if we can pool the experience gained thus far in many different applications, and recognize common problems still in need of solution. That is the main purpose of these meetings.

My own work has been concerned mostly with development of the general theory of irreversible thermodynamics, although attempts have been made also to point out other possible applications. The history of the principle of maximum entropy, as it applies to thermodynamics, has been told recently (Jaynes, 1978) at such great length that hardly anyone could wish to have it all told again here. But as applied to problems far removed from thermodynamics, a few recollections about early experiences may help to set the record straight and explain to present students why the method is only now coming into its own.

Section 2 of this paper recalls a little history, but very briefly because our concern today should be with the future. It appears, in retrospect, that the development of computer programs capable of dealing with dozens to thousands of simultaneous constraints was the key factor in establishing the power of this method, in a way that transcended all philosophical arguments.

Sections 3 through 6 deal with technical problems, first in generality, then as applied specifically to spectrum analysis and image reconstruction. In these discussions we try to formulate some of the common problems that need to be dealt with in the immediate future in connection with presently established applications. Finally, we speculate a little about new applications that might develop in the more distant future.

2. Setting the Record Straight

Of course, the maximum-entropy algorithm originated with Boltzmann and Gibbs. But in their writings they did not make its meaning crystal clear—Boltzmann because he was not very clear in his own mind, repeatedly changing his position (Klein, 1973), and Gibbs because his work was left unfinished (Jaynes, 1967). In Gibbs' "Heterogeneous Equilibrium" (1875-1878), in fact, we find a much clearer and deeper explanation of the properties of entropy than in his final work, Statistical Mechanics (1902).

Therefore, as far as the rationale of the method was concerned, Boltzmann and Gibbs left a kind of vacuum that was filled, as such vacua always are, by followers whose thinking had little in common with that of Boltzmann, and nothing in common with that of Gibbs. For 60 years after Gibbs, the "official" statistical mechanics of our textbooks stood on the premise that the canonical (maximum entropy) distribution was an actual

physical fact, the ultimate result of the mechanical equations of motion operating over long times, and that the fundamental unfinished goal of statistical mechanics was to prove this rigorously by "ergodic" theorems.

In the early 1950s—just when Enders Robinson was doing some of the first hand-run predictive deconvolutions of geophysical data at MIT—I was at Stanford, studying not only statistical mechanics but also the problem of detecting hidden regions of different dielectric or acoustical properties (nonmetallic land mines in Korea, crevasses in Greenland, brain tumors, flaws in structural materials, etc.).

In those days nobody even dreamed of being able to use digital signal processing hardware for such purposes although we did dream of analog calculation, recording the observed signal $s(x,y)$ as a function of the antenna position, and correlating it with those to be expected from every possible target by endless-loop tape recorders (which would have been so bulky and slow as to be almost useless).

A conceivable fast solution was to try to design the antenna so that its radiation pattern (or more accurately, the dot product $\mathbf{E}_t \cdot \mathbf{E}_r$ of transmitter and receiver fields) maximized the ratio (peak signal)^2/(mean square response to soil anomalies), that is, so that the antenna itself approximated a two-dimensional matched filter. But of course the equations defining the optimal matched filter contradicted Maxwell's equations or the acoustical wave equations; and even if the optimal antenna could have been realized, it would have been optimal only for soil of one particular autocorrelation function. Today, of course, all this sounds trivial; an electronic package the size of a matchbook could do digital processing of these signals, according to any algorithm we please, in real time.

It was realized that if we had the capability to deconvolve any antenna response pattern, the antenna design problem would be radically changed. As those familiar with the theory of integral equations will see at once, the only important antenna design parameter is then the size of the "window" that it opens up in Fourier transform space. That is, over how large a region (k_1,k_2) is the transform

$$F(k_1,k_2) = \int dx \int dy(\mathbf{E}_r \cdot \mathbf{E}_t) \exp[i(k_1x+k_2y)]$$

above the thermal noise level, thus delivering relevant information to the computer?

These theoretical conclusions were summed up in a classified report in 1954, but at the time the technology to exploit them did not exist and no actual hardware resulted. Still, this association was an important influence on my general thinking, making me see a close relationship between the rationale of these engineering problems and that of Gibbsian statistical mechanics.

To see why a relationship exists, let us ask: What determines our probability distribution for soil anomalies? Conventional thinking—then and

now—would suppose this is an actual physical fact, a "real physical property" of the soil, like density, chemical composition, etc., with different kinds of soil having different intrinsic autocorrelation functions. The more I held lumps of soil in my hand, the less I believed this.

The probability distribution for the soil dielectric constant that we use in our detection theory can be useful only to the extent that it tells us something about the one sample of soil that exists under our antenna. Whether variations in other samples do or do not follow this distribution in the frequency sense cannot be relevant to our problem. Indeed, if we had prior information telling us the actual condition of our sample of soil, the frequency distributions in other samples that we are not looking at would be completely irrelevant.

Realizing this seemed to me an important new insight into the nature of statistical inference. Orthodox thinking—then and now—wants us to define probabilities only as physical frequencies, and deplores any other criterion as not "objective." Yet when confronted with a (literally) dirty, objective real problem, common sense overrides orthodox teaching and tells us that to make the most reliable inferences about the special case before us, we ought to take into account all the information we have, whatever its nature; a rational person does not throw away cogent evidence merely because it does not fit into a preconceived pattern. But this means that our probabilities can be equal to frequencies only when (a) we actually have frequency data and (b) we have no other "prior information" beyond those frequencies. (We should add here that R. A. Fisher, the great proponent of frequency interpretations, eventually recognized this also, and in his final book (1956) he acknowledged that fiducial inference is valid only when we have no prior information; unfortunately, many writers of statistics textbooks still have not recognized this.)

The probability distributions that we use for inference must in general represent not merely frequencies, but our total state of knowledge, of which frequencies are only a part. Indeed, some of the most important real problems of inference are unrelated to frequencies in any "random experiment." The problem is that our information is incomplete; there is nothing "random" about it. But strict adherence to orthodox principles would deny us the use of probability theory in such problems.

So there is the connection: the only way known to set up a probability distribution that honestly represents a state of incomplete knowledge is to maximize the entropy, subject to all the information we have. Any other distribution would necessarily either assume information that we do not have, or contradict information that we do have. But that is just the procedure that Boltzmann and Gibbs advocated in statistical mechanics, long before Shannon showed us how to interpret it.

The principle itself was not new at all, but the realization that it applied as well to problems far removed from thermodynamics, and that these engineering problems, when finally solved correctly, would be seen as having the same rationale and the same algorithm as statistical mechanics, was so new and startling that it could not be conveyed to others. Attempts

to explain this to other physicists, engineers, mathematicians, and statisticians at Stanford and Berkeley, and to the Army Engineers, met with strictly zero success—with only one exception.

David Blackwell saw the point at once, and put his finger on the basic "fairness" property of the MAXENT algorithm: It does not allow you to assign zero probability to any situation unless your information really rules out that situation. Nobody else was able to free his mind from frequentist preconceptions; and so from this engineering episode I returned to statistical mechanics with nothing to show for it except a new appreciation of the generality of the MAXENT principle.

But before one can see this generality it is necessary to realize that it is a principle of reasoning, not a principle of mechanics, and thus free it from the supposed dependence on Hamiltonian equations of motion and ergodic theorems.

What I did (Jaynes, 1957) was only to suggest a different interpretation of Gibbs' "canonical ensemble" method. If we regard it as representing, not mechanical prediction from the equations of motion, but only the process of inference (make the best predictions you can from the information you have), then in that sense his algorithm can be justified in great generality as a principle of probability theory—essentially just the process of rational thinking—without any appeal to ergodicity.

To this day, our ergodist critics seem to have a hang-up over the distinction between mechanical prediction and inference; they continue to complain about the latter because it is not the former. Nevertheless, for thermal equilibrium (the only case where ergodic theorems could ever have applied) we obtain the same actual predictions that the ergodists wanted; and indeed, any other predictions would be in conflict with experiment. Ergo: the most that ergodic theorems could ever have accomplished is to confirm what we already know, namely that in equilibrium problems taking the equations of motion into account does not alter the predictions that we obtain directly by maximizing the entropy.

In fact, it is rather elementary that equations of motion can tell us only how probabilities change with time, and not what probabilities should be assigned initially. Therefore, in any problem where the given information (set of constraints) tells us only something about constants of the motion, the equations of motion can provide zero additional information about the state of the system. In John Burg's happy phrase, then, all the philosophical arguments "do not change a single number" in the calculations.

The pragmatic advantage of the inference viewpoint (which hardly needs pointing out to this audience) is that we then see Gibbs' method as something of far greater generality than the ergodists ever dreamed of. It applies equally well not only to nonequilibrium situations but to any problem of inference, in or out of physics, in which the information at hand can be described by enumerating a set of conceivable hypotheses ("prior information") and specifying whatever else we know ("data") in the form of constraints that narrow down the set of possibilities. In such applications, it has some easily proved optimality properties.

However, in my papers of 1957 I was still under the influence of textbooks that attributed the ergodic view to Gibbs; therefore I thought I was suggesting something new. Three years later, finally getting around to a really careful reading of Gibbs, I realized that I had been brainwashed. Gibbs' actual thinking was utterly different from that commonly attributed to him, and I had only rediscovered his original viewpoint! (There was, however, a technical advance in that we now had Shannon's theorems that filled a gap in Gibbs' argument.) In the trauma of this discovery, a book review I wrote just then (Jaynes, 1961) came down hard on two other poor souls who had only been victims of this same brainwashing, but who had had the misfortune to write books before realizing it. Expressions of sympathy would have been more appropriate.

For some years (1958-1964) I traveled around a circuit of talks at universities and industrial research laboratories (including those of three well known oil companies, at Dallas, Tulsa, and Bartlesville), pointing out the existence and generality of the MAXENT algorithm to hundreds of people who might have profited by using it in all kinds of applications. But the reaction was mostly negative. Few saw the point, and most (under the influence of conventional physical and/or statistical teaching) denied vehemently—even angrily—that useful predictions could come from a method that expressed only "subjective" probabilities, not "objective" frequencies.

On one of these occasions a member of the audience was moved to express these doubts in poetry. I do not recall his name, or even the year and place, but the poem has been preserved because I was in turn moved to compose a counter-poem and read them both to my students. They were rediscovered this year in some yellowed old course notes.

Here is the original, whose author is presumably just as happy to remain anonymous (if not, he may come forward and claim it):

MAXIMUM ENTROPY

There's a great new branch of science which is coming to the fore,
And if you learn its principles you'll need to learn no more,
For though others may be doubtful (which can make them fuss and fret)
You'll be certain your uncertainty's as large as it can get.

This is no toy of theory, to invite the critic's jeering
But a powerful new method used in modern engineering.
For uncertainty, which often in the past just gave us fidgets
Has proved to be a vital tool in mass-producing Widgets.

The procedure's very simple (it is due to Dr. Jaynes)
You just maximize the entropy, which doesn't take much brains.
Then you form a certain function which we designate by Z,
Differentiate its log by every lambda that you see.

And LO!—we see before us (there is nothing more to do)
All the laws of thermal physics, and decision theory too.
Possibilities are endless, no frontier is yet in sight,
And regardless of your ignorance, you'll always know you're right!

So, are you faced with problems you can barely understand?
Do you have to make decisions, though the facts are not at hand?
Perhaps you'd like to win a game you don't know how to play.
Just apply your lack of knowledge in a systematic way.

Handed a copy of this upon leaving the meeting, I composed a reply during my return flight:

MAXIMUM ENTROPY REVISITED

There's a fine old branch of science coming back now to the fore
And if you learn its principles you'll need to grope no more.
For though others may be blinded (which can make them twist and turn)
You'll be certain that you see whatever is there to discern.

This is no cut-and-try device, empirical ad hockery
But a principle of reasoning, of proven optimality.
For prior information, which empiricism spurns
Has proved to be the pivot point on which decision turns.

The procedure's very simple (it is due to Willard Gibbs)
First define your prior knowledge without telling any fibs
On this space of possibilities, a constraint is then applied
For every piece of data, until all are satisfied.

If the states you thought were possible, in setting up this game
And the real possibilities in Nature, are the same
Then LO! you see before you (there is nothing more to grind)
Reproducible connections that experimenters find.

But the principles of logic are the same in every field
And regardless of your ignorance, you'll always know they yield
What your information indicates; and (whether good or bad)
The best predictions one could make, from data that you had.

This reply tried to correct the historical record and the mistaken emphasis by pointing out the positive nature of the principle, which seemed obviously true to me—and obviously false to nearly everybody else.

Not quite everybody, however. During the early 1960s several rumors circulated, too vague to leave a trail to fact, but hinting guardedly at uses of maximum entropy in oil exploration. Finally, with the landmark paper of Burg (1967), it became public knowledge that a few people had indeed taken the trouble to work out the detailed algorithms and computer programs needed to try the method on some real problems far removed from thermodynamics. They found, of course, that it worked just as I had said it would (or else we would not be meeting here today).

From this point on, I shall not presume to recount a history of these applications, so much better known to those here who made it. But, for that goal of setting the record straight, perhaps other speakers will be willing to share with us some of their own recollections of how it all started.

3. General Remarks

Maximum entropy represents an entirely different kind of thinking from what has been taught in statistics courses for 50 years. Many of those trained in orthodox statistics find the difference so mind-wrenching that the rationale of MAXENT remains incomprehensible to them after repeated attempts at explanation. Yet beginning students see the point at once with no difficulty because MAXENT is just the natural, common-sense way in which anybody does think about his problems of inference—unless his mind has been warped by orthodox teaching. This difference will be less troublesome if we explain it first in very general terms.

ORTHODOX APPROACH. You are given a set of observed data $D_{obs} = \{d_1 \ldots d_m\}$ from which you are to decide whether some hypothesis H about the real world is true (or, put more cautiously, whether to act as if it were true). For example, H might be the statement that some unobserved quantity θ has a value in a specified interval ($a \leq \theta \leq b$); in fact, by imaginative use of language, almost any hypothesis could be stated in this form.

Given this problem, the first thing orthodox statistics does is to imbed the observed data set in a "sample space," which is an imaginary collection containing other data sets $\{D_1 \ldots D_N\}$ that one thinks might have been observed but were not. Then one introduces the probabilities

$$p(D_i \mid H), \qquad 1 \leq i \leq N, \tag{1}$$

that the data set D_i would be observed if H were true. This is called the "sampling distribution," and $p(D_i \mid H)$ is interpreted as the frequency with which the data set D_i would be observed in the long run if the measurement were made repeatedly with H constantly true.

When one asserts the long-run results of an arbitrarily long sequence of measurements that have not been performed, it would appear that he is drawing either on a vivid imagination or on a rather large hidden fund of prior knowledge about the phenomenon. If we are not told what that knowledge is and how it was obtained, we might be excused for doubting its

existence. But, for the sake of argument, let us suppress these doubts and accept the orthodox interpretation of $p(D_i|H)$ at its face value.

The sampling distribution is the only probability distribution we are allowed to use. For that reason, orthodox statistics is often called "sampling theory." In principle, the merits of any proposed method of data analysis are to be judged by its sampling properties. That is, in the long run, how often would it lead us to a correct conclusion, or how large would the average error of estimation be? But practice sometimes ignores precept (as when one insists on using an unbiased estimator $(n-1)^{-1}\ \Sigma(x_i-\bar{x})^2$ of variance, even though the biased estimator $n^{-1}\ \Sigma(x_i-\bar{x})^2$ would yield a smaller mean square error).

PURE MAXENT APPROACH. By contrast, in the pure maximum-entropy (noiseless Bayes) method, our reasoning format is almost the opposite. Instead of considering the class of all data sets $\{D_1 \ldots D_N\}$ consistent with a hypothesis H, we consider the class of all hypotheses $\{H_1 \ldots H_n\}$ consistent with the one data set D_{obs} that was actually observed. In addition we use prior information I that represents our knowledge (from physical law, usually expressed as combinatorial multiplicity factors W_i) of the possible ways in which Nature could have generated the various H_i. Out of the class C of hypotheses consistent with our data, we pick the one favored by the prior information I—which means, usually, having the greatest multiplicity W (that is, greatest entropy log W).

Now, which kind of reasoning do you and I use in our everyday problems of inference? An automobile driver must make a decision (stop, slow down, go through, speed up, turn) at every intersection, based on what he can see. He does not think about the class of all things he might have seen but doesn't see. He thinks about the class of all contingencies that are consistent with what he does see, and acts according to which ones, in that class, seem a priori most likely from his previous experience.

If you go to a doctor and tell him your symptoms, he does not start thinking about the class of all symptoms you might have had but don't have. He thinks about the class of all disorders that might cause the symptoms you do have. The first one he will test for is the one which, in that class, appears to be a priori most likely from your medical history.

The same common-sense reasoning format is used by a TV repairman, a detective, an analytical chemist—and a player of the game "Twenty Questions." Each successive piece of data that one obtains is a new constraint that, if cogent, restricts the possibilities permitted by our previous information. By Blackwell's principle, at any stage an honest description of what we know must take into account (that is, assign nonzero probability to) every possibility that is not ruled out by our prior information and data.

Clearly, this is the way any rational person does—and should—think about such problems of inference. Only a warped mentality could see anything peculiar or illogical about it. Yet for 25 years MAXENT methods have been rejected—in some cases attacked—by persons who did not comprehend this simple rationale.

Even when we come to the proof of the pudding and present our superior numerical results (such as those of Burg, Currie, Montgomery, Frieden, Skilling, Gull and Daniell, and Shore as presented or recalled at this meeting), the issue is not always resolved. MAXENT results are sometimes regarded with suspicion because those familiar only with orthodox methods and results simply cannot believe it is possible to extract so much detailed information from the data.

In fact, their instincts are quite correct. It is not possible to extract all that detail from the data alone, or else orthodox methods might have done so. MAXENT gives us more information only because we have put more information into it. In image reconstruction, MAXENT is taking into account not only the data, but also our prior information about the multiplicity of different scenes. This information is, as we shall see, of the highest relevance to the problem, but orthodox ideology does not recognize it because it does not consist of frequencies in any random experiment.

In the problems where pure MAXENT is appropriate (which include all of statistical mechanics, equilibrium or nonequilibrium) we are concerned, not with frequencies in any random experiment, but with rational thinking in a situation where our information is incomplete. We are trying to do the best reasoning we can about the real situation that exists here and now—and not about the long run in some other situations that exist only in a statistician's imagination. In such applications, a probability distribution is not an assertion about frequencies but only a means of describing our state of knowledge.

In orthodox thinking, a frequency is considered "objective" and therefore respectable, while a mere state of knowledge is "subjective" and unscientific. But in the real-world problems of inference faced by every engineer, scientist, economist, business man, or administrator, it is evidently his state of knowledge that determines the quality of the decisions he is able to make in situations that will never be repeated; and it is the frequencies that are figments of the imagination.

Interestingly, the sampling distribution that orthodox theory does allow us to use is nothing more than a way of describing our prior knowledge about the "noise" (measurement errors). Thus, orthodox thinking is in the curious position of holding it decent to use prior information about noise, but indecent to use prior information about the "signal" of interest. Yet one man's signal is another man's noise. This arbitrary rejection of part of the information is the reason orthodox methods are incapable of dealing with "generalized inverse" problems of the aforementioned kind, which arise constantly in the real world.

FULL BAYES METHOD. We have expounded two more or less opposite extremes of reasoning, each of which is appropriate in a certain class of problems. The orthodox sampling distribution $p(D_i|H)$ is described only loosely as "noise." Stated more carefully, it might represent two different things:

(A) The variability of the data to be expected if the experiment were repeated under seemingly identical conditions.
(B) The likely deviations of the actually observed data D_{obs} from the truth.

Orthodox statistics, failing to distinguish between meanings (A) and (B), makes a single sampling distribution $p(D_i|H)$ serve two different functions. But, as I realized while holding those lumps of soil, (A) and (B) are logically quite different things. Indeed, it is (B) that one needs to know in order to make sound inferences that make allowance for the unreliability of the data D_{obs} actually obtained. Given the probability distribution appropriate for (B), it may or may not give also the variability (A) of data on other measurements that are not made, but the question is obviously irrelevant to our inference from the data we do have.

However, if the sole information we have pertaining to question (B) is derived from knowledge of variability in sense (A), then one may confuse the meanings (A) and (B) without harm. If in addition there is no prior information about the quantities being estimated, orthodox methods will be at least useful, and perhaps (if we have sufficient statistics and no nuisance parameters) even optimal. Likewise, if we have relevant prior information but no appreciable noise, the pure MAXENT method will be appropriate and near-optimal.

But just as prior information can make an orthodox analysis invalid (a maximum likelihood estimate $\hat{\theta} = 8$ is useless and misleading if we know in advance that $\theta < 6$), so appreciable noise can make the pure MAXENT method inappropriate and misleading.

Fortunately, in most of our everyday problems of inference, noise is not an important factor, so the MAXENT reasoning format applies. But if that automobile driver at the intersection had to look out through an optically turbulent medium, and so saw something different every time he looked even though the factual situation was unchanging, then he would have a more complicated problem and the MAXENT reasoning would have to be modified to make allowances for the unreliability of his data. We all know how much more cautiously we drive when a heavy rainstorm converts the windshield into just that turbulent medium. We cannot be sure what lies before us and must first guess at that before trying to decide what action to take.

If we have both noise and prior information, neither of the above methods is adequate. But both are only limiting special cases of a more general method that applies in all cases known to the writer. Adding prior information capabilities to orthodox methods, or noise capabilities to MAXENT, we arrive in either case at the Bayes method, which is actually simpler conceptually and older historically than either of these special cases.

To define the Bayesian method, we introduce some rudimentary Boolean algebra. We denote various propositions by A, B, C, etc., "not A" by ~A, "A and B" by AB. Then the probability symbols $p(A|{\sim}B)$ and $p(AB|C)$ are read as "the probability that A is true, given that B is false," and "the probability that both A and B are true, given that C is true." Mathematically, then, probability theory consists of nothing but the sum and product rules

$$p(A\,|\,B) + p(\sim A\,|\,B) = 1 \tag{2a}$$

$$p(AB\,|\,C) = p(A\,|\,C)\,p(B\,|\,AC) \tag{2b}$$

and the unending stream of consequences that can be deduced from them.

All schools of thought accept these rules as mathematically correct. It is over their interpretation—their relationship to the real world—that the 150-year-old philosophical controversies swirl. Orthodox doctrine, as noted, interprets p as a frequency, and it is a trivial observation that frequencies do indeed combine according to these rules. But orthodoxy also takes a militant stand, declaring all other interpretations to be metaphysical nonsense. In some 35 years of perusing the literature, I have found no orthodox writer who has advanced a logical reason for this position; it is merely asserted. But there is a vast literature indicating that this position was ill-advised, in effect putting orthodox statistics in a kind of straitjacket that makes it incapable of dealing with the current real problems of science, engineering, and economics.

A broader—and, we believe, far more useful—view, interprets $p(A\,|\,B)$ as a measure of the degree of plausibility of A, on a (0,1) scale. Then the equations of probability theory are not merely rules for calculating frequencies; they are also rules for conducting inference. In fact, R. T. Cox showed many years ago (Jaynes, 1976, 1978) that any set of rules for conducting inference, in which we represent degrees of plausibility by real numbers, is necessarily either equivalent to Eqs. (2) or inconsistent (in the sense that with any other rules one can find two different methods of calculation, both obeying the rules, that yield different results).

Of course, since Gödel one does not expect probability theory to provide a proof of its own consistency. The rules (2) as derived by Cox refer to finite discrete sets, and when they are applied in problems of inference on such sets, no inconsistency has ever been found. But there are many "paradoxes" in the literature—the marginalization paradox, the ambiguity of posterior odds ratios, the Borel-Kolmogorov paradox, the nonconglomerability paradox, etc.—all of which are caused by trying to apply these rules directly and indiscriminately on infinite sets. The common technique by which all these paradoxes, and any number of others, can be manufactured, is: (1) start from a well behaved finite result; (2) pass to a limit without specifying how the limit is to be approached; (3) ask a question whose answer depends on how the limit was carried out.

Clearly, the paradoxes of infinite sets have nothing to do with real problems of inference. After all, the number of atoms in our galaxy is only $G < 10^{70}$, a safely finite number, and it is hard to believe that a real problem will ever require a larger set than G! We propose a conjecture: All correct results in probability theory are either combinatorial theorems on finite sets generated by Eqs. (2), or well-behaved limits of such theorems. If this is correct, probability theory may be able to feed for a century on the combinatorial work of Rota (1975).

Let us see some consequences of Eq. (2), interpreted as rules for conducting inference. Take, as above, I = prior information, H = some hypothesis, D = data. Then, since the product rule (2b) is symmetric in A and B, we have

$$p(DH|I) = p(D|I)\,p(H|DI) = p(H|I)\,p(D|HI)\,. \tag{3}$$

If $p(D|I) \neq 0$ (that is, the data set is a possible one), this yields Bayes' theorem:

$$p(H|DI) = p(H|I)\,\frac{p(D|HI)}{p(D|I)}\,, \tag{4}$$

which represents in a very explicit form the process of learning. It shows how the "prior probability" $p(H|I)$ changes to the "posterior probability" $p(H|DI)$ as a result of acquiring new information D. This is exactly the kind of rule we need for inference. Let us note some of its properties.

Suppose we obtain data D_1 today, then additional data D_2 tomorrow. Today's data D_1 will of course be part of tomorrow's prior information, so our inferences about H on the two successive days will take the form

$$p(H|D_1I) = p(H|I)\,\frac{p(D_1|HI)}{p(D_1|I)} \tag{5}$$

$$p(H|D_2D_1I) = p(H|D_1I)\,\frac{p(D_2|HD_1I)}{p(D_2|D_1I)}\,. \tag{6}$$

But substituting Eq. (5) into Eq. (6), and using the product rule (3) in numerator and denominator, reduces Eq. (6) to

$$p(H|D_2D_1I) = p(H|I)\,\frac{p(D_2D_1|HI)}{p(D_2D_1|I)}\,, \tag{6'}$$

which is just of the form (4) of the original Bayes' theorem with $D = D_2D_1$, the total data now available. The result extends by induction to any number of additional data sets D_3, D_4, etc. Thus, when we take into account many different pieces of information, Bayes' theorem updates our state of knowledge at each step, but in such a way that our conclusions always depend only on the total information at hand, and would be the same if we got the information all at once. This is a most welcome consistency property, without which we would be in real trouble. Bayes' theorem has dozens of other nice features that express just the properties that a rational person would demand of a method of inference. Another important one is noted below [Eq. (56)].

But how does orthodoxy view this? Orthodoxy never gets to Eq. (6') because at the start it rejects Bayes' theorem (4) out of hand, as a method of inference, without ever bothering to examine the kind of results it gives. In orthodox ideology, probabilities can be assigned only to "random variables" because a probability, to be respectable, must be also a frequency.

But a hypothesis H is not a "random variable," so $p(H|I)$ and $p(H|DI)$ are held to be meaningless! At this point, orthodox statistics denies itself the use of the single most powerful and useful principle in probability theory, and it is left with no way to take the prior information I into account. This failure of orthodox statistics is just the reason why the new methods to be discussed at this workshop had to be developed.

Further details about the workings of Bayes' theorem are in Section 6, where we shall see that, in the Gull-Daniell image reconstruction problem, both MAXENT and orthodox results are contained in the Bayes solution, in the limits of zero noise and zero prior information respectively.

4. The Maximum-Entropy Formalism

Like any fairly general method, MAXENT can be approached in various ways. It can be based on information theory, on combinatorial theorems, or as a limiting form of Bayes' theorem. That is, all these approaches lead to the same algorithm although conceptually they express three slightly different problems. Arguing over these slight differences may solve the unemployment problems of philosophers for the next 50 years, but we need not dwell on them here.

We introduce "The Maximum-Entropy Formalism" by the information-theory approach. A real variable x can take on the values $\{x_1 \ldots x_n\}$ and we are to assign corresponding probabilities $\{p_1 \ldots p_n\}$ so as to represent our partial information about x. The information theory basis noted that the Shannon entropy

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{7}$$

is a measure of the "amount of uncertainty" in the distribution $\{p_1 \ldots p_n\}$, uniquely determined by certain very elementary consistency and additivity requirements (Shannon, 1948). Intuitively, then, the distribution that most honestly describes what we know, without assuming anything else (that is, is as noncommittal about x as it can be without violating what we know) is the one that maximizes H subject to the constraints imposed by our information.

But this is not a mathematically well-posed problem until we write down the explicit form of the constraints. Some kinds of information are too vague to use in a mathematical theory by any presently known methods, although our intuitive common sense may be able to make some use of them. Various aspects of this are discussed more fully elsewhere (Jaynes, 1968, 1978). With experience one becomes more adept at converting verbal information into mathematical constraints. For present purposes it is enough to give only one special case, which is by far the most useful one so far.

What does it mean to say that a probability distribution $\{p_1 \ldots p_n\}$ "contains" certain information? Presumably, this ought to mean that we can extract that information back out of it. Suppose we had a probability distribution and were asked to make the best estimate $\hat{A}$ of some function $A(x)$.

The exact value of A is not determined by the distribution, so we must introduce some criterion of what we mean by "best."

Since the time of Gauss, the mean square error criterion has been the most popular and easily implemented. Other criteria could be used if there were any advantage in using them; but for the applications we have in mind they would only make the computations longer, while leading to final results practically indistinguishable from the ones given below. If we make the estimate $\hat{A}$, the expected square of the error will be

$$\langle (A - \hat{A})^2 \rangle = \sum_i p_i [A(x_i) - \hat{A}]^2$$

$$= [\langle A^2 \rangle - \langle A \rangle^2] + (\hat{A} - \langle A \rangle)^2 , \tag{8}$$

where the brackets $\langle\ \rangle$ denote averages over the distribution p_i, often called "expectations." The first term, the variance of the distribution of A,

$$\mathrm{var}(A) = \langle A^2 \rangle - \langle A \rangle^2 , \tag{9}$$

is fixed by the distribution p_i, so the mean square error is minimized by choosing as our estimate the average

$$\hat{A} = E(A) = \langle A \rangle = \sum_i p_i\, A(x_i) . \tag{10}$$

Statisticians denote expectations by E(A), while physicists, having already preempted E for energy and electric field, use the bracket notation $\langle A \rangle$. Writing for both, we use both notations.

Conversely, if we are asked to adjust the distribution $\{p_1 \cdots p_n\}$ to incorporate given information about A, we shall understand this as meaning that, by applying the usual prediction rule, (10), we can get that information back out of $\{p_1 \cdots p_n\}$. Thus our mathematical constraints, in the problems considered here, will take the form of fixing expectations of various quantities about which we have some information.

If we wish to incorporate information, not only about the value of A, but also about the accuracy with which A is known, we merely add another constraint, fixing $\langle A^2 \rangle$ as well as $\langle A \rangle$. The general formalism below automatically includes this possibility. But in practice this usually makes no appreciable difference in our conclusions. An exception can arise in the limit $n \to \infty$, where a constraint on $\langle A^2 \rangle$ may be needed to get a convergent solution, $\Sigma p_i = 1$.

More generally, whenever we find we are not getting a normalizable solution, that is how the theory tells us we have not yet specified enough information to justify any definite inferences. As in any situation of insufficient information, we cannot hope to get something for nothing, and the

only remedy is to get more information. Almost always, it turns out that the person using the theory actually did have some more information that he had failed to put into the equations, not realizing that it was essential.

To proceed to "The Maximum-Entropy Formalism," there are m functions $\{A_1(x)\cdots A_m(x)\}$ for which we are given, in the statement of the problem, our "data," that is, a set of numbers $\{A_1'\cdots A_m'\}$ which are the values we want our probability distribution $\{p_1\cdots p_n\}$ to predict. To fit the distribution to our data, we impose the m simultaneous constraints:

$$\sum_{i=1}^{n} p_i A_k(x_i) = A_k', \qquad 1 \leq k \leq m . \tag{11}$$

Much ink has been spilled over the relationship between the numbers A_k' appearing in Eq. (11) and real data. A persistent misconception is that MAXENT requires our data to consist of "mathematical expectations." This does not make sense, and is an egregious case of putting the cart before the horse. If we adjust a mathematical expectation to fit our new state of knowledge after getting some data, it does not follow that the data consisted of expectations. If I adjust my belt to fit my new state of girth after a heavy meal, it does not follow that my meal consisted of belt leather.

In various problems the A_k' could be generated in various ways. For present purposes we note that A_k' is simply a number given to us in the statement of the problem, and our present task is the mathematical one of incorporating the conditions (11) into our probability distribution. For further discussion, see Jaynes (1978, pp. 72-77 and 95-96).

The solution was given by Gibbs (in, of course, different notation):

Define the partition function

$$Z(\lambda_1\cdots\lambda_m) = \sum_{i=1}^{n} \exp[-\lambda_1 A_1(x_i) - \cdots - \lambda_m A_m(x_i)] . \tag{12}$$

Then the MAXENT distribution is

$$p_i = \frac{1}{Z(\lambda_1\cdots\lambda_m)} \exp[-\lambda_1 A_1(x_i) - \cdots - \lambda_m A_m(x_i)] \tag{13}$$

with the λ's determined, as our poet noted, by

$$A_k' = -\frac{\partial}{\partial \lambda_k} \log Z , \qquad 1 \leq k \leq m , \tag{14}$$

a set of m simultaneous equations for the m unknowns $\{\lambda_1\cdots\lambda_m\}$.

The result has also a combinatorial basis, a special case of which was given by Boltzmann (1877). Some process happens N times, and each

time a result is that one of the x_i is chosen. This includes many different scenarios:

(I) BERNOULLI TRIALS. A random experiment is repeated N times, each trial yielding one of the values $\{x_i\}$.

(II) COMMUNICATION. We receive a message of N letters, chosen from the alphabet $\{x_1 \cdots x_n\}$.

(III) KINETIC THEORY. A gas contains N molecules, each in one of the quantum states $\{x_1 \cdots x_n\}$.

(IV) IMAGE RECONSTRUCTION. N elements of luminance are distributed over n pixels, labeled $\{x_1 \cdots x_n\}$, to form an image.

In any of these, the particular result x_i is chosen N_i times; that is, it is generated with frequency $f_i = N_i/N$. Needing a short name for the set $F = \{f_1 \cdots f_n\}$ whatever the scenario, we call it simply "the scene," which seems particularly appropriate in image reconstruction. Any given scene F has a certain multiplicity W(F):

$$W(F) = \frac{N!}{(Nf_1)! \cdots (Nf_n)!}, \tag{15}$$

which is equal to the number of ways in which Nature could have generated it. When N becomes large we have, by Stirling's rule,

$$\frac{1}{N} \log W(F) \rightarrow -\sum_{i=1}^{n} f_i \log f_i = H(f), \tag{16}$$

just the Shannon entropy again, only now depending on "objective" frequencies f_i instead of "subjective" probabilities p_i. For this reason, the combinatorial approach is the one most easily understood by those with orthodox training although the first approach is more general.

Now we do not know the specific sequence of $\{x_i\}$ that was generated; all we know is the resulting averages of m quantities $\{A_1(x) \cdots A_m(x)\}$:

$$\sum_{i=1}^{n} f_i A_k(x_i) = A'_k, \qquad 1 \leq k \leq m. \tag{17}$$

If we are asked which scene we consider most likely, it seems reasonable to favor the one that Nature could have generated in the greatest number of ways consistent with what we know, that is, the one that has maximum multiplicity W(F) subject to the constraints (17). In view of Eq. (16), then, we have formulated the same mathematical problem as in Eqs. (7) to (11), and the same MAXENT algorithm Eqs. (12) to (14) will solve it. We need only write f_i in place of p_i.

For those who have not seen elementary examples of numerical solutions by this method, the extensive analysis of Rudolph Wolf's dice-tossing experiments in Jaynes (1978) is a recommended tutorial introduction that explains many further points of rationale. A still more efficient procedure for testing hypotheses in the light of data by MAXENT appears in Jaynes (1983, Chapter 10). But let us turn now to some of the current nontrivial applications.

5. The No-Noise Spectral Analysis Problem

In the field of spectral analysis, this case is rather special but nevertheless real. Many time series that arise in econometrics and geophysics may be considered essentially uncontaminated by noise, because the variability of the phenomenon being observed greatly exceeds the error of the measurements. Sales on the New York Stock Exchange vary greatly from day to day, yet the value for any one day can be determined exactly. The air temperature in Tulsa may vary by 50 degrees in a few hours, yet its value at any one time can be measured to a fraction of a degree.

We have, then, some discrete time series $\{y_0, y_1, \cdots, y_N\}$ that has a Fourier transform

$$Y_\omega \equiv (N+1)^{-1} \sum_{k=0}^{N} y_k e^{i\omega k} \tag{18}$$

and a power spectrum

$$S_\omega \equiv |Y_\omega|^2 = \sum_{k=-N}^{N} R_k e^{i\omega k} \tag{19}$$

where $R_{-k} = R_k^*$ and

$$R_k \equiv \sum_{j=0}^{N-k} y_j^* y_{k+j}, \qquad 0 \le k \le N, \tag{20}$$

is usually called the "autocovariance" or "autocorrelation" although both terms seem unfortunate and inconsistent with other established usage. The Fourier inversion of Eq. (19) is

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} S_\omega e^{-i\omega k} d\omega = R_k, \quad -N \le k \le N, \tag{21}$$

where $\omega_N = \pi$ is the Nyquist frequency, above which S_ω repeats itself periodically. Note that we have defined these quantities so that the last four equations are exact for finite N, without "end-effects." There is no mathe-

matical advantage, in either precision or simplicity, to be gained by passing to the physically nonexistent limit $N \to \infty$.

The difficulty is that not all the R_k are known. Our data D, although exact, comprise only a subset $D = \{R'_{-m} \cdots R'_m\}$ where $m < N$. The problem we shall consider is how to estimate S_ω from this incomplete information. (Of course, practically everything we say can be carried over mutatis mutandis to the similar problem of estimating Y_ω from a subset of the y_k.)

This no-noise spectrum analysis problem is, then, an example of the standard generalized inverse difficulty; the data cannot distinguish between two spectra S_ω and S'_ω that differ by a solution of the homogeneous equation

$$\int_{-\pi}^{\pi} (S_\omega - S'_\omega)\, e^{-i\omega k}\, d\omega = 0, \qquad |k| \leq m. \tag{22}$$

Therefore, by Eq. (19) the class C of possible spectra compatible with our data consists of all functions of the form

$$\hat{S}_\omega = \sum_{k=-m}^{m} R'_k\, e^{i\omega k} + \sum_{k=m+1}^{m} (\hat{R}_k\, e^{i\omega k} + \hat{R}_{-k}\, e^{-i\omega k}), \tag{23}$$

where the $\hat{R}_k$ $(m < k \leq N)$ may be chosen arbitrarily but for the condition $\hat{S}_\omega \geq 0$ that any power spectrum, by definition, must satisfy. Clearly, the $\hat{R}_k$ represent estimates of the autocovariance, extrapolated beyond the data. Thus the problem of choosing, out of the class C of possible spectra, one "best" estimate of S_ω is equivalent to the problem of extrapolating the autocovariance to all lags $(-N \leq k \leq N)$.

Now, because there is no noise (the R'_k are considered known exactly for lags $|k| \leq m$), there is no sampling distribution to describe it. It appears, then, that sampling theory can provide no basis for choosing one spectral estimate $\hat{S}_\omega$ over another; a pure generalized inverse problem lies entirely outside the domain of sampling theory statistical methods.

Indeed, in the Blackman-Tukey (1958) method of spectrum analysis the problem is never seen as one of extrapolating R_k at all. Blackman and Tukey state unequivocally that, "Surely, no estimate can be made for lags longer than the record." Nevertheless, because of the mathematical connections (19) and (21), any method of estimating S_ω is necessarily also a rule for estimating R_k beyond the record.

A spectral estimate $(\hat{S}_\omega)_{BT}$ found by the Blackman-Tukey method is, in fact, the spectrum of a time series whose autocovariance is zero at all lags beyond the data. As Burg (1975) pointed out, this is tantamount to making an extrapolation of R_k that is almost certainly wrong and that may even stand in violation of the nonnegativity condition $S_\omega \geq 0$, thus making $(\hat{S}_\omega)_{BT}$ outside the class C of logically possible spectra.

Even if this implied extrapolation fails to get us outside the class C, the following difficulty remains. The $\hat{S}_\omega$ that we obtain from Eq. (21) by setting

$\hat{R}_k = 0$, $k > m$, is the spectrum of a time series whose autocovariance is truncated abruptly at $k > m$. It is therefore the convolution of the true spectrum with the "Dirichlet kernel" $D(\omega) = (2\pi)^{-1}\sin[(2m+1)\omega/2]/\sin(\omega/2)$ and will exhibit spurious "side lobe" maxima that are not in the spectrum of the true time series $\{y_k\}$ but are separated from the true spectral lines by odd harmonics of $\omega_{SL} = \pi/(2m+1)$. As Burg realized, these side lobes are only an artifact of the method, caused by the failure to extrapolate R_k in a reasonable way.

However, the BT method proceeded to treat the symptom rather than the disease. To get rid of the abrupt truncation, instead of raising $\hat{R}_k$ for $k > m$, it lowers $\hat{R}_k$ for $k \leq m$ by introducing a "lag window" function W_k $(-m \leq k \leq m)$. It thus replaces Eq. (23) by

$$(\hat{S}_\omega)_{BT} = \sum_{k=-m}^{m} W_k R'_k \, e^{i\omega k}, \tag{24}$$

and indeed, by various choices of window functions, one can reduce the side lobes greatly, at the cost of losing about half the resolution.

But this does violence to Eq. (23), which is the most general estimate that lies in the class C of logically possible spectra. $(\hat{S}_\omega)_{BT}$ is the spectrum of a time series that is known to disagree with our data at every data point k where $W_k \neq 1$. We know, not as a plausible inference but as a logical deduction, that a time series with the spectrum (24) could not have produced our data!

In the noiseless case this criticism by Burg was crystal clear and unanswerable. Then does the use of windows become more defensible if the data are contaminated by noise? Our examination of the theoretically simpler image reconstruction problem below will provide a clue suggesting that allowing for noise must take us not toward window solutions, but away from them, that is, to solutions of still higher entropy.

But the claims of sampling theory are not yet disposed of. One might think that in a pure (noiseless) generalized inverse problem there is no place for sampling theory at all. Yet sampling theory has managed to work itself into the problem anyway, by a remarkable conceptual feat. If the real world has no sampling distribution, then we shall invent one, not by imbedding our data in a set of possible data, but by imbedding the whole real world in a set of possible worlds (just as Everett (1957) did in quantum theory). The one real, finite time series $Y = \{y_0 \ldots y_N\}$ that actually exists is regarded as only a "sample" drawn from some hypothetical ensemble of other infinitely long series—and we are back in business!

Of course, from this standpoint R'_k—which was by definition the exact autocovariance of Y—then appears to be "biased." So it is replaced by the "unbiased estimator"

$$R''_k = \frac{N+1}{N+1-k} R'_k, \qquad 0 \leq k \leq m. \tag{25}$$

However, we stress again that Eqs. (18) to (21) are exact as they stand for finite N, and so we have made a second error of reasoning. But pragmatically, this helps by canceling out some of the effect of the first error. Before falsifying our data R'_k downward by a window function W_k, if we first falsify them upward by using Eq. (25), we end up a little closer to being in the class C of possible spectra.

Here are some of the things that bother me about orthodox reasoning. In the real world no infinitely long time series exist—much less any ensemble of them. What does it mean to say that we are "estimating" something that is only a figment of our imagination? What determines this sampling distribution according to which the real world is now to be selected from this figment? If we are only imagining this ensemble, then it would seem that we are free to imagine it as having any properties we please. Then whose imagination is to take precedence—yours or mine?

If to achieve the "objectivity" of a frequency interpretation of our probabilities we must sacrifice the "objectivity" of dealing with the real world, then it seems to me that we have paid too high a price. True "objectivity" ought to address itself not to estimating figments of our imagination, but rather to representing the state of knowledge about the one real world that actually exists, based on the one data set that actually exists.

Now consider the generalized inverse problem (23) from the standpoint of maximum entropy. The given data

$$\{R'_{-m} \cdots R'_0 \cdots R'_m\}$$

represent, in complex notation where we treat R_k and $R_{-k} = R_k^*$ independently, (2m+1) constraints. (If we revert to the real notation $R_{-k} = R_k$ and treat only $(R_0 \cdots R_m)$ independently, then we have only (m+1) independent conditions.)

The maximum entropy distribution subject to these constraints is a Gibbsian generalized canonical ensemble

$$P(y_0 \cdots y_N) \propto \exp\left[-\sum_{k=-m}^{m} \lambda_k R_k\right]. \tag{26}$$

Writing out the exponent in full, we have

$$\sum \lambda_k R_k = \lambda_0(y_0^2 + y_1^2 + \cdots + y_N^2)$$

$$+ \lambda_1(y_0 y_1 + y_1 y_2 + \cdots + y_{N-1} y_N)$$

$$+ \lambda_2(y_0 y_2 + y_1 y_3 + \cdots + y_{N-2} y_N)$$

$$+ \cdots$$

$$+ \lambda_m(y_0y_m + \cdots + y_{N-m}y_N)$$

$$+ \lambda_{-1}(y_1y_0 + y_2y_1 + \cdots + y_Ny_{N-1})$$

$$+ \cdots$$

$$+ \lambda_{-m}(y_my_0 + \cdots + y_Ny_{N-m}) \,. \tag{27}$$

But this can be rearranged into a matrix product:

$$\sum \lambda_k R_k = \sum_{ij=0}^{N} \Lambda_{ij} y_i y_j \tag{28}$$

where Λ is a Toeplitz matrix:

$$\Lambda_{ij} = \begin{cases} \lambda_{j-i}, & |j-i| \leq m \\ 0, & |j-i| > m \end{cases} \tag{29}$$

with the Lagrange multipliers repeated down bands parallel to the main diagonal.

Thus, without our having assumed any "Gaussian random process," the maximum entropy principle constructs the Gaussian form for us, as the distribution that most honestly represents our data:

$$P(y_0 \cdots y_N) \propto \exp\left[- \sum_{ij} \Lambda_{ij} y_i y_j \right]. \tag{30}$$

The partition function (12) is then proportional to $[\det(\Lambda)]^{-1/2}$, and so

$$\log Z = -\frac{1}{2} \sum_{j=0}^{N} \log \ell_j + \text{constant} \,,$$

where $\{\ell_j\}$ are the eigenvalues of Λ. Determination of the λ's from Eq. (14) then involves solving a number of simultaneous equations. In the limit $N \to \infty$, the eigenvalues $\{\ell_j\}$ are given by known theory of Toeplitz matrices.

The Levinson-Burg numerical algorithm proceeds by a Wiener-Hopf factorization of the polynomial

$$\sum_{k=-m}^{m} \lambda_k z^k = \left| \sum_{k=0}^{m} a_k z^k \right|^2 \tag{31}$$

and evaluating the new coefficients $\{a_k\}$ recursively. The $\{a_k\}$ may be interpreted as the coefficients of a Wiener prediction filter, or as the coefficients of an autoregressive model, if they are chosen to correspond to a minimum delay wavelet. However, from the standpoint of maximum entropy this choice is not mandatory, only the λ's appearing in our final results.

The details of this solution are so well known, and so well covered by others at this workshop, that it would be a needless duplication to repeat them here (they are, however, given from the writer's point of view in Jaynes, 1982). Suffice it to say that maximum entropy uses the form (23), thus guaranteeing that our estimated spectrum $\hat{S}_\omega$ lies in the class C of possible spectra—but we use the optimal extrapolation $\{\hat{R}_{m+1} \cdots \hat{R}_N\}$ of the covariance beyond our data, determined as expectations over the maximum-entropy distribution (30). By a bit of algebraic magic, the final maximum-entropy spectrum estimate turns out unexpectedly simple:

$$(\hat{S}_\omega)_{MAXENT} = \frac{1}{\sum_{k=-m}^{m} \lambda_k e^{-i\omega k}} , \qquad (32)$$

which is the now classic result of Burg (1967).

The virtues of this solution are many. It not only gets us back into the class C of logically possible spectra, but also, instead of forcing us to compromise between side lobes and resolution, it gives us the best of both worlds—eliminating side lobes while giving much greater resolution. The practice of spectrum analysis was revolutionized by this discovery, which made all previous thinking and methods obsolete.

NEW PROBLEMS. Of course, this does not mean that the theory is now finished, and Eq. (32) with the Levinson-Burg evaluation of the $\{\lambda_k\}$ is the optimal solution for every conceivable problem. Some assumptions were made in the derivation of Eq. (32), and although our result is as robust with respect to small departures from those assumptions as were any previous solutions, we are now looking for higher standards of performance than were expected from previous methods. So, instead of being finished, now that our eyes have been opened we can see a mass of new problems in need of solution, some of which could not even be formulated in terms of previous notions.

Two problems now pressing for immediate attention are the variability of Eq. (32) in the presence of noise, pointed out to the writer by David Brillinger in December 1980, and the "line-splitting" phenomenon pointed out to the writer by John Tukey just before this meeting. These can be traced to assumptions made in the derivation of Eq. (32) that are not always satisfied in real problems.

We assumed that our data were noiseless; otherwise the class C of possible spectra would not be sharply defined. With noise, the boundary of C becomes smeared out into a smooth transition region whose thickness repre-

sents the noise level. From the analogous image reconstruction solution to be noted in Section 6, we expect this will lead us to modify the pure MAXENT spectrum (32) in the direction of slightly higher entropy, by an amount of the order of the thickness of the transition region. However, the analytical details for the spectrum analysis case are not yet available.

Also it was assumed implicitly in the algorithm for evaluation of the λ_k that the autocovariance data $\{R_0 \cdots R_m\}$ were from a very long sample of length $N >> m$, since one used the asymptotic distribution of eigenvalues for the limit of an infinitely large Toeplitz matrix. If our data are from a time series $\{y_0 \cdots y_{43}\}$ of only 44 observations, then the exact MAXENT spectrum should use the λ_k's computed from Eqs. (14) and (31) with the exact eigenvalues of the finite Toeplitz matrix with 44 rows and columns.

Again, the analytical details are not yet available, but I am confident, for reasons to be discussed more fully elsewhere, that this small correction of the algorithm will remove the line-splitting difficulty. The point is that when we have only a finite time series $\{y_0 \cdots y_N\}$, the present algorithm gives us only an approximate solution of the real problem; but that happens to be the exact solution for a circular time series that repeats itself forever: $y_0 = y_{N+1}$, $y_1 = y_{N+2}$, etc. A circular time series with the same autocovariances $\{R'_0 \cdots R'_m\}$ would indeed have a spectrum with split lines close together.

These two problems appear to be of high priority but also straightforward in principle, and surely solvable. But the details are not trivial, and some honest labor will be required to find the new algorithms. Fortunately, some of the analytical machinery needed has been provided by Gohberg and Fel'dman (1974) and Bleher (1981).

6. Image Reconstruction

The spectacular recent successes of Frieden, Gull and Daniell, Skilling, and others in the area of image reconstruction have done more than anything else to convert doubters into believers, because here the demonstration of what MAXENT can do stands, literally, before your eyes. Also, at its present stage of development, the problem is simpler theoretically than that of spectrum analysis because correlations in luminance of adjacent pixels are not yet incorporated into the MAXENT equations (whereas in spectrum analysis the whole problem lies in correlations).

This relative simplicity allows us to use the combinatorial approach to MAXENT, Eqs. (15) to (17) above, and to discuss the full Bayes solution. If N equal elements of luminance are distributed over n pixels to form a scene $F = \{f_1 \cdots f_n\}$, $f_i = N_i/N$, the number K of conceivable different scenes that could result is equal to the number of terms in the multinomial expansion of $(f_1 + \cdots + f_n)^N$:

$$K = \frac{(N + n-1)!}{N!(n-1)!}, \tag{33}$$

and when N is large, a given scene of entropy $H(f_1 \cdots f_n)$ has a multiplicity

(number of ways it can be realized) given asymptotically by Eq. (16):

$$W(F) \sim e^{NH(F)} . \tag{34}$$

Our data $D \equiv \{d_1 \ldots d_m\}$ consist of the luminances of m pixels of our blurred image, which we suppose determined by

$$d_k = \sum_{i=1}^{n} A_{ki} N_i + e_k , \qquad 1 \leq k \leq m < n , \tag{35}$$

where $N_i = Nf_i$ is the number of elements of luminance in the ith pixel, A_{ki} is the digitized point spread function of our telescope, and e_k are the traditional "Gaussian noise" terms, which are to have independently the probability distributions

$$p(e_k | \sigma) \propto \sigma^{-1} \exp(-e^2_k/2\sigma^2) , \qquad 1 \leq k \leq m , \tag{36}$$

and σ is the RMS noise level.

Let us proceed directly to the application of Bayes' theorem, (4). Given prior information I and data D, the probability that any specific scene F is the true one is

$$p(F | D,\sigma,I) = p(F | I) \frac{p(D | F\sigma I)}{p(D | \sigma I)} . \tag{37}$$

The denominator, being independent of F, is just a normalizing constant and is not needed in most applications. We have, then, to get the prior probability $p(F | I)$ and the sampling distribution $p(D | F\sigma I)$. The latter is determined from Eqs. (35) and (36): Given that the true scene is $F = \{f_1 \ldots f_n\}$, the probability (density) that we shall obtain the data $D = (d_1 \ldots d_m)$ is just the probability that the noise terms e_k will make up the difference:

$$p(D | F,\sigma,I) = (2\pi\sigma^2)^{-m/2} \exp(-Q/\sigma^2) , \tag{38}$$

where

$$Q(D,F) \equiv \frac{1}{2} \sum_{k=1}^{m} \left[d_k - \sum_{i=1}^{n} A_{ki} f_i \right]^2 \tag{39}$$

is the basic quadratic form of the kind that always gets into problems with Gaussian noise.

Assigning the prior probabilities $p(F | I)$ is not so straightforward because there is no end to the variety of different kinds of prior information that one might have in various problems. Clearly, any cogent prior information about which scenes are a priori likely ought to be taken into account and will increase the reliability of our reconstructed scene. For example,

say

$$I = I_G \equiv \text{'The scene is a distant galaxy'} ;$$

we know in advance, in a general way, what galaxies look like; and a reconstruction depicting a horse or a New Jersey license plate is so improbable that we would not hesitate to assign

$$p(F_H | I_G) = p(F_{NJLP} | I_G) = 0 .$$

Equally clearly, however, it would be a task of unlimited complexity to try to characterize every conceivable nongalactic scene and eliminate all of them from our prior probability assignment, although that would undeniably improve our reconstructions.

Therefore, for the present we are going to do what is feasible, and pretend that we have no prior knowledge at all about what is being looked at, and so we cannot exclude any scene. If N elements of luminance are distributed, one by one, to the n pixels, this can be done in n^N different ways, and we shall look at the consequences of a state of primitive ignorance I_0 that assigns equal probability n^{-N} to each of them.

Indeed, the impressive results achieved thus far correspond to this I_0 although they are only the 'minimal performance' of the theory, and we know that they could be improved with a more elaborate treatment of prior information. We do not know how much further improvement is possible with a feasible amount of further analysis, but the present results prove to be quite good enough for many purposes.

Out of the n^N conceivable ways Nature could have made all possible scenes, any particular one $\{f_1 \ldots f_n\}$ can be realized in W(F) different ways, so our prior probability assignment is

$$p(F | I_0) = n^{-N} W(F) . \tag{40}$$

But since N is large and constant factors are irrelevant, it will suffice to take, from Eq. (34),

$$p(F | I_0) \propto e^{NH(F)} . \tag{41}$$

Then Bayes' theorem, (37), reduces to

$$p(F | D,\sigma,I_0) \propto e^{N[H(F) - wQ(F)]} , \tag{42}$$

where

$$w \equiv \frac{1}{N\sigma^2} . \tag{43}$$

The factor exp(NH) represents our prior information about the multiplicities of different scenes. The factor exp(-NwQ) in its dependence on $\{f_1 \ldots f_n\}$ is the 'likelihood function' that tells what we have learned from the data, making allowance for the noise.

The most probable scene $\{\hat{f}_1 \ldots \hat{f}_n\}$ is then at the peak of this distribution. To locate it we maximize $(H - wQ)$ subject to the constraint $\Sigma f_i = 1$. By Lagrange multipliers, this tells us immediately that, at the peak, the f_i are proportional to

$$\hat{f}_i \propto \exp\left[-w \frac{\partial Q}{\partial f_i}\right], \qquad 1 \leq i \leq n . \tag{44}$$

This relationship, although interesting, is implicit only because all the f_i are still hidden in Q. Still, from Eq. (44) we can understand the transition from full Bayes to pure MAXENT. When w is small (large noise), the data provide only a rather "soft" constraint on the possible f_i, and Eq. (44) allows the solution point to wander up the entropy hill (that is, in the direction of uniform f_i), away from the hyperplane HP whose equation is $Q = 0$. But as the noise decreases and $w \to \infty$, Eq. (44) goes into a rigid "stone wall" constraint, forcing the solution point back to HP. In the zero-noise limit, the term wQ in the Bayes solution thus puts in just the constraint of the pure MAXENT solution.

Conversely, if all scenes had the same multiplicity, $H(F) = \text{constant}$, the most probable scene would revert to the maximum likelihood estimate corresponding to $Q = 0$. But if $m < n$, $Q = 0$ is the equation of the hyperplane HP, not a point, and there is no unique solution, the likelihood is flat at all points of HP and we have a generalized inverse problem of just the kind that orthodox statistics is unable to deal with. Without prior information, there is no criterion for preferring any point of HP satisfying $f_i \geq 0$, $\Sigma f_i = 1$, to any other. The Bayes solution thus contains the MAXENT and orthodox solutions as special cases.

But we have glossed over some conceptual difficulties in getting this result. There are quite a few unspecified quantities (n,m,N,σ) still flapping about loose. How do we decide on their values?

WHAT ABOUT n,m,N? In some cases the circumstances of the problem may determine one or more of these in a way not under our control, but in many problems it is up to us to choose them. That is how we define our "hypothesis space"—the conceptual field on which our game is to be played.

One might question why we bring in all these discrete integers at all. In the real world, Nature seldom divides up her scenes into neat little rectangular cells. The true scene has a continuous structure, and it would seem more realistic to replace Eq. (35) by an integral equation

$$d(x) = \int A(x,y) f(y) d^2y + e(x) \tag{45}$$

to be inverted. The discrete version, (35), that we have adopted is only a crude, inelegant approximation to Eq. (45).

Now if we were going to find an analytical solution, this would be the right way of looking at it, and in principle the general Bayes solution of

Eq. (45) might be given once and for all. Indeed, any specific galaxy, horse, or New Jersey license plate can be portrayed by some analytic function f(y). But an analytic solution that contains enough parameters to include all those possibilities does not seem feasible, and so we are being pragmatists, after useful numbers rather than formal elegance.

It is a practical necessity that our reconstructions be found, not by substituting into some grand and glorious general solution, but in each specific case by numerical processing of the data. Any data set that we can actually record is a collection of a finite number of integers. Any numerical calculation that we can actually perform—whether by hand or by the most powerful computer—is nothing but a finite number of manipulations of a finite number of integers. So if our problem is not solvable in finite, discrete terms, we shall not be able to solve it at all. Introducing discrete integers gives us a solvable problem.

Our choice of (n,m) is of course to be guided by our prior information about the nature of the problem, the quantity and quality of data, and the computing facilities available. For example, regardless of what the theory says, it would be fatal to chose (n,m) so large that our computer memory could not hold all the (f_i, d_k). Likewise, it seems irrational to choose n or m so small that we are obviously throwing away resolution that the data are capable of giving.

It seems, then, that the choice of (n,m) will involve some compromise between performance and computation cost. Larger values increase the potential resolution and squeeze more information out of our data; but beyond a certain point we have extracted essentially all that the data have to tell us, and further increases would only increase the amount of computation without useful return. To make an intelligent choice, we must understand how the resolution depends on our choice. Fortunately, this is not critical, and wide variations—within reason—make little difference in the quality of the reconstruction.

The choice of N is more subtle. As a first orientation, let us get some idea of the rather interesting numbers involved. Gull and Daniell (1978) considered an early example with $n = 128 \times 128 = 16384$ pixels (determined, obviously, by computer considerations). How many primitive elements of luminance N should we think of as strewn over them to make the unknown true scene F? This is something one can argue about for a long time, and I am sure John Skilling will agree with me that more thinking about it is also needed. But for purposes of illustration and to start the discussion, the following seems a reasonable first guess.

The question is important because at issue here is the relative weighting of the prior information (entropy) factor exp(NH) and the likelihood factor $\exp(-Q/\sigma^2)$. Just as the choice of the number n of pixels to use in our reconstructed scene is a pragmatic one (make it large enough that we achieve all the spatial resolution the data are capable of giving, but not much larger), so the choice of the number N of elements of luminance we suppose to be in the scene is a way of stating how fine-grained an intensity resolution the data are capable of giving. (Perhaps we should add, if it were not

already obvious, that the term "luminance" is being used in a loose colloquial sense, not as a technical term of optometrics.)

The little elements of luminance represent basically the smallest increment that we could detect. Conceivably, at very high frequencies and very low intensities they might have something to do with individual photons, but in radio astronomy we are surely very far from that case, and "quantum considerations" are irrelevant. In optical cases they might be identified with individual developed grains in a photograph. If our RMS error σ decreases, then we can detect a smaller increment, and so N increases.

However, looking at it that way may be putting the cart before the horse. Presumably, if our apparatus is reasonably well designed, the "noise" σ is not coming from the apparatus itself but rather is generated by the phenomenon under observation. So perhaps we should state it the other way around: If N increases, then we can detect a smaller relative increment, and so σ decreases.

In any event, the result is that N and σ are linked in some way. Various relationships between them may be appropriate in different problems; we indicate one such connection that appears "crudely reasonable" in some cases.

If we think of our noise σ as generated by variability of the scene to be expected if it were created anew, then if the ith pixel got N_i elements of luminance, we should expect this might vary by about $\pm\sqrt{N_i}$, and so f_i might be uncertain to about

$$\delta f_i \simeq \sqrt{N_i}/N = \sqrt{f_i/N} . \tag{46}$$

The uncertainty σ in our data resulting from this uncertainty in f_i is then

$$\sigma = \Sigma/\sqrt{N} , \tag{47}$$

where Σ depends on details of the smearing matrix A_{ki} and the values of f_i, but is independent of N. But, comparing with Eq. (43), we see that $w = \Sigma^{-2}$. This is the "weight" of the log likelihood factor Q relative to the entropy factor H.

Obviously, many refinements in detail are possible and desirable—in particular, pursuing this line of reasoning will teach us, very quickly, that the errors e_k in Eq. (35) should not be taken as independent and identically distributed, which means that Q in Eq. (39) should be written as a more general quadratic form. But our important conclusion that w is independent of N will remain as long as the errors σ are generated more in the phenomenon than in the measuring instrument.

In the posterior distribution (42), increasing N and decreasing σ^2 so as to keep $w^{-1} = N\sigma^2$ constant has the effect of increasing the sharpness of the peak without affecting its actual location. We do not change our reconstructed scene, but only become more confident about its accuracy.

For most purposes it is w, not N, that is of primary interest. If we know that our reconstructed scene is the best one that can be made on the information used, then whether the peak at $\{\hat{f}_1 \ldots \hat{f}_n\}$ is sharp or broad we shall not be induced to seek another. Some nontrivial analysis may be needed to find the best value of w, but fortunately it appears from the results of Gull and Daniell that this too is not very critical, variations within reason not making a great deal of difference in the reconstructed scene.

Of course, if w is taken unreasonably low, that is tantamount to saying that the data are nearly worthless, and the reconstruction goes to the uninformative uniform gray scene of absolute maximum entropy. If w is taken unreasonably high, we are claiming that every tiny bit of detail in the data is real and not noise, and so the reconstruction starts showing spurious details. Even then, however, the pure MAXENT reconstruction, like the MAXENT spectrum estimate, has at least the merit that it keeps us in the class C of possible scenes and cannot yield a negative f_i for any pixel, as did some previous reconstruction algorithms.

Now we are ready for those interesting numbers. With the value n = 16384 of the Gull-Daniell 1978 work, let us choose arbitrarily N = 10n = 163840. This is probably unfair to Gull and Daniell, since it supposes the data were not very accurate, but even so it gives us big enough numbers to make our point. The number K of conceivable scenes from Eq. (33) is then

$$K = \frac{180332!}{163840!\ 16383!} = 3 \times 10^{23840}. \tag{48}$$

Their multiplicities (15) range from

$$W_{min} = 1 \tag{49}$$

up to

$$W_{max} = \frac{163840!}{(10!)^{16384}} = 4 \times 10^{675703}. \tag{50}$$

Therefore, we could partition the scenes into 67571 categories, those in category c having multiplicity in

$$10^{10c} \leq W \leq 10^{10(c+1)}, \qquad 0 \leq c \leq 67570. \tag{51}$$

Higher W (higher entropy) means a smoother reconstructed scene, but the eye does not distinguish 30000 different degrees of smoothness. Two scenes, as alike as possible except that one is in category (c), the other in (c+2), will be virtually undistinguishable to the eye. But every scene in (c+2) has a greater multiplicity than any scene in (c) by a factor of more than 10^{10}.

So not only do we have variations

$$W_1/W_2 \simeq 10^{10} \tag{52}$$

between scenes, we have chains of thousands of such comparisons, with a factor of 10^{10} at each step.

These numbers should give us an appreciation of the importance of multiplicity factors in inference, and explain why methods that ignore them are at such a disadvantage. The success of MAXENT reconstructions appears almost magical until we realize that, as an elementary combinatorial result, the overwhelming majority of all possible scenes compatible with our data have entropy very close to the maximum. This statement can be refined into an "entropy concentration theorem" (Jaynes, 1982, 1983).

Before closing this discussion, let us note one other aspect of Bayes' theorem. Suppose our problem was different from what was assumed in Eqs. (46) and (47) and we knew N in advance, but this told us nothing about σ. (Perhaps σ is instrumental noise in a new instrument never before used.)

WHAT ABOUT σ? In all the above, we have supposed the noise level σ known. What if it isn't? The answer will give us a good example of the power of Bayesian methods. If σ is unknown, then it too becomes a parameter to which we assign some prior probability $p(\sigma \mid I)$, and Bayes' theorem will give us, instead of Eq. (37), a joint posterior distribution for F and σ:

$$p(F,\sigma \mid D,I_0) = p(F,\sigma \mid I_0)\,\frac{p(D \mid F,\sigma,I_0)}{p(D \mid I_0)}\,. \tag{53}$$

As we have argued very extensively elsewhere (Jaynes 1968, 1978, 1980), "complete ignorance" of any scale parameter such as σ corresponds to the Jeffreys prior $p(\sigma \mid I_0) \propto d\sigma/\sigma$ (strictly, the limit of a sequence of such distributions over finite intervals $(0 < a < \sigma < b < \infty)$, but our present problem has such good convergence that this nit-picking isn't needed). On prior information I_0, the prior probabilities of F and σ are independent: $p(F,\sigma \mid I_0) = p(F \mid I_0)p(\sigma \mid I_0)$; so in place of Eq. (38) we have

$$p(F,\sigma \mid D,I) \propto \sigma^{-(m+2)/2} \exp(NH - Q/\sigma^2)\,. \tag{54}$$

But if we care only about estimating F, then σ is what is called a "nuisance parameter" in statistics. To get rid of it, we integrate it out to get the marginal posterior distribution of various scenes:

$$p(F \mid DI) = \int_0^\infty p(F,\sigma \mid DI)\,d\sigma\,. \tag{55}$$

The integration yields

$$p(F \mid DI) \propto e^{NH}/Q^{m/2}\,. \tag{56}$$

So now, in order to find the most probable scene, instead of maximizing $(NH - \sigma^{-2}Q)$ we should maximize $[NH - (m/2)\log Q]$. This gives, in place of Eq. (44), the implicit relationship

$$f_i \propto \exp\left[-\frac{m}{2N}\frac{\partial}{\partial f_i}\log Q\right]. \tag{57}$$

But this is formally the same as Eq. (44), if we put into Eq. (44)

$$\sigma^2 = \frac{2Q}{m}, \tag{58}$$

which is just the estimate of σ^2 that any rational person would make if he knows the (d_k, f_i) and nothing else.

To make a long story short, then, Bayes' theorem tells us that, if σ is unknown, then: (A) we should choose the scene $\{\hat{f}_1 \ldots \hat{f}_n\}$ as the point of "self-consistency" where the previous connection (44) still holds, with σ^2 replaced by our best estimate of σ^2; (B) the penalty we pay for this ignorance is that the peak of the distribution (56) is not as sharp (no longer exponential in Q) as that of Eq. (42). We still get a definite reconstruction but are not quite so confident of it. This is a good example of what I meant back in Section 3 [Eq. (6)], in saying that Bayes' theorem has other nice features that a rational person would demand of a method of inference.

7. Speculations—Realizable Fantasies

Already in the third issue of Dr. Dobb's Journal of Computer Calisthenics and Orthodontia (March 1976), it seemed time for an editorial with our same title, "Where do we go from here?" Dr. Dobb noted the obvious extensions and more powerful implementations of things already under way, and added: "We will continue the active pursuit of 'realizable fantasies.'" By this was meant projects not yet under way but that appear to be within the bounds of present technology and knowledge, and achievable within the next few years.

Doubtless, each of us has his own favorite list of realizable fantasies: those we would like to do ourselves, and those we would like to persuade others to do. Each of us would like to achieve some new breakthrough that is conceptually difficult but technically easy, and leave to others problems that are technically difficult. Problems that are both conceptually and technically easy are going to be solved automatically.

From a historical perspective, the great achievements of the past always appear to be new advances in conceptual understanding, rather than "mere" technical accomplishments. But we are seeing this through a filter; although technical accomplishment is often a prerequisite for conceptual advance, it tends to be forgotten. Today, everybody knows about the

concept of ellipses, but very few living people know even the formulation, much less the technical details, of Kepler's analysis that led to this concept.

But there is also truth in the opposite emphasis. As we have learned in the theory of irreversible statistical mechanics, problems that seemed horrendously difficult technically may become, if not exactly "easy," at least many orders of magnitude less difficult, as a result of deeper conceptual insight.

There are many problems that may become technically easy as soon as the conceptual cobwebs are swept away. Decades of discussion from the standpoint of sampling theory may have obscured completely the fact that there is no sampling distribution, and the real problem is logically a generalized inverse type where MAXENT, rather than orthodox statistics, is the appropriate method. But the actual implementation of MAXENT may still be very difficult conceptually, because it requires us to define an explicit "hypothesis space" on which our counting of multiplicities is done.

In statistical mechanics we feel we know how to do this: The set of linearly independent "global" quantum states of a system is, according to all present knowledge, the proper field on which our game is to be played. But in forecasting economic time series, it may be perfectly clear that the essence lies not in the "randomness" of our data but in its incompleteness, yet still very unclear just what is the proper hypothesis space of elementary possibilities on which we should define multiplicities W and entropies log W.

There are many problems of this type, in which it seems clear that substantial improvements should be realizable, since these multiplicities must be highly cogent evidence, but present methods do not even recognize their existence. Even in areas such as spectrum analysis and image reconstruction, which are in a sense already "established," there is still much room for creative thinking about what is the appropriate hypothesis space for different kinds of problems. In this respect, I don't think any of our present solutions are optimal except in some very special kinds of problems. Burg's solution is now the classic example, but in time it may be seen as only the first of several MAXENT solutions for various spectral analysis problems.

Let us illustrate the last three paragraphs by the example of turbulence, which has indeed been discussed for decades from the standpoint of sampling theory. Is it a realizable fantasy to find a different conceptual standpoint which makes it clear that the problem is logically a generalized inverse one rather than a sampling theory exercise?

TURBULENCE. In thermodynamics, the first law for closed systems, or a "generalized first law" for open systems (conservation of energy plus the other conservation laws for number of atoms of each type, charge, etc.) determines a class of possible macroscopic states that is permitted by those constraints. Out of all these possibilities allowed by the first law, the one chosen by Nature (second law) is the one that can be realized in the greatest number of ways, that is, that has maximum entropy. This is the principle given by Gibbs (1875-1878), governing heterogeneous equilibrium. For 100 years most of physical chemistry has been based on it.

A similar principle must govern the spectrum of turbulence; a stirred fluid has energy added to it in motions of small wavenumber k, such that $kL \approx 1$, where L is a typical dimension of the system. This is transferred to successively smaller eddies, of greater wavenumber, until it finally degrades into the uniform thermal energy of the highest possible wavenumbers, such that $kd \approx 1$, where d is the molecular separation.

Imagining this degradation process in a momentary steady state, suppose the energy in turbulent eddies in the wavenumber range dk varies as $(1/k^n)$. As far as conservation of mass, energy, and momentum are concerned, many different values of n are allowed. Which one is chosen by Nature? This must be the value that provides the greatest number of channels (sequences of microstates) by which the process can take place. The logarithm of the number of channels W(P) by which a macroscopic process P can take place is a kind of entropy, which we have ventured (Jaynes, 1980) to call the "caliber" of P.

As an analogy, even if individual drivers have no particular preference for one traffic lane over another, still the six-lane road will end up carrying more traffic than the two-lane road to the same place. However, we are concerned here with ratios, not of 3:1, but perhaps of 10^{20}:1.

Among the possible macroscopic paths by which low-k kinetic energy may be degraded into high-k thermal energy, there will be one that opens up overwhelmingly more "traffic lanes" than any other—or indeed, than all others. This is surely the one that Nature will choose. Since the caliber of a path is not determined by the Navier-Stokes equation, but by the multiplicity factors that the N-S equation ignores, we can understand why attempts to base turbulence theory on the N-S equation met with little success, and degenerated into attempts to "guess the right statistical assumptions." Again, the problem is not one of mechanics, but of inference, of the generalized inverse type.

Many years ago, on the basis of some "statistical assumptions" and reasoning that I have been unable to follow, Kolmogorov proposed the value $n = 5/3$, which does seem to have some experimental support. But its theoretical justification and range of validity do not seem at all well understood as yet. This is a problem for the future, on which I think we have a good chance of succeeding.

Now turbulence theory is in so many respects like quantum field theory that a realizable fantasy in one may point to a realizable fantasy in the other. The high-caliber paths in turbulence are undoubtedly "lumpy" in time, small bifurcations suddenly transforming a lump of energy, which would appear as a suddenly generated wavelet rather than a continuous mode-transfer process. This immediately reminds us of quantum theory.

QUANTUM WAVELETS. Having started out as a conventionally trained, if somewhat apostate, quantum theorist, I was of course well indoctrinated in the quantum theorists' view of the world, in which they see themselves at the top of the pecking order in science, and almost disdain to recognize a geophysicist.

Just in the past year I have been shocked by the discovery that geophysicists, uninhibited by the kind of logic taught in quantum theory courses, have answered questions that quantum theorists were afraid to ask. Geophysicists have learned how to deal with wave phenomena in a more complete and sophisticated manner, and are in possession of some fundamental knowledge that quantum physicists need to have, if they could only bring themselves to admit it.

A recent review (Jaynes, 1973) concludes with a very incomplete preliminary sketch, suggesting that today the interesting, relevant physics of blackbody radiation may be contained in the intermode correlations that conventional theory neglects at the outset. To the best of my knowledge, only von Laue (1915) had ever recognized that these correlations might be important. His ideas were not pursued by others because Einstein (1915) argued against them, and, for that reason, the issue remains unresolved to this day.

The ultimate verdict may be that this is the only error Einstein ever made. Apparently, an earlier analysis (Einstein and Hopf, 1910) had convinced him that equilibrium of radiation necessarily leads to independent Gaussian probability distributions for the mode amplitudes. But here we meet again that constant plague of probability theory: What is the Problem? Einstein and Hopf had of course given a correct solution—but to a different problem.

Clearly, in the present problem, every act of emission or absorption must affect not only the field energy but also the intermode correlations. If our probability distribution is independent Gaussian immediately before an emission, it cannot be independent Gaussian immediately afterward.

It is curious that Einstein's theory (1917) of blackbody radiation insisted on, from the standpoint of the atoms, instantaneous transfer of energy and momentum—but blandly ignored, from the standpoint of the field, the intermode correlations that this would necessarily entail. Just at this point, causality was lost, and we have never learned how to restore it.

The probability theory of the early 1900s was not attuned to these realities, because Maxwell was dead and Jeffreys had not yet been heard from. If we define the probability of some condition as the fraction of a very long time in which the condition holds, as Einstein did, then we have sacrificed the very concept of a time-dependent probability. A theory of probability so curtailed cannot be used for inference about situations where we have information referring to specific times. Yet the change in our knowledge about a radiation field when we are told that an emission has occurred at time t must clearly affect our subsequent predictions about it, in a way that depends on t.

Now a correlation in the amplitudes of many modes close together in k-space is what we should describe in ordinary coordinate space (x-space) as a 'wave packet' or 'wavelet.' It is not essentially different from the wavelets envisaged by Huygens. An emitting atom generates a spherical wavelet that propagates outward at the speed of light. Until it has been scattered, reflected, or otherwise perturbed, it retains its integrity and

represents an additional contribution to the amplitudes of many modes, perfectly correlated.

These wavelets are the entities that figure so prominently in the work of H. Wold and have been exploited so well in the work of Enders Robinson and other geophysicists. It is surprising that physicists outside geophysics have made almost no use of the concept of wavelets, and had virtually no awareness of their existence, until the recent interest in solitons, which are wavelets that have become somewhat more indestructible thanks to nonlinearities in the wave equation.

For this reason, conventional physical theory does not use causal mechanisms to account for radiative equilibrium. Indeed, it cannot until it is reformulated in wavelet terms. Emission from an atom does not go into just one mode at a time as one might think from the conventional "Fermi Golden Rule" approximations of quantum theory; it goes into a coherent superposition of many modes simultaneously.

Likewise, absorption of energy by an atom is accomplished by emission of a wavelet of such phase that it cancels out, partially and momentarily, the incident wave in the forward direction (incipient shadow formation). Thus we conjecture that, in both turbulence and blackbody radiation, it is these wavelets that provide the physical mechanism by which the equilibrium energy distribution is brought about and maintained.

This is thrown out as a half-baked idea to ponder: When equilibrium is not maintained by a continuous process, is it possible that the temperature might be related more closely to the average energy of a wavelet than to the average energy of a mode?

After all, when the wavelet gets "hard" and indestructible, we call it a "particle," and then there is no doubt that the temperature relates to its kinetic energy. Then at what degree of hardness does the transition from mode to wavelet take place? Is localization in space a necessary part of that indestructibility? Or does this require only phase persistence for a time long compared to some characteristic interaction time? Indeed, if our concern is with the spectral distribution of the energy, why should its spatial distribution matter? It seems to me that it is the coherent organization of the energy—whether it occupies a cubic micron or a cubic mile—that makes a wavelet a unique "object."

So, to take the final deep plunge into fantasy: Is it possible that wavelets are the missing link, which physics has needed all these years, to resolve the contradictions of the "wave-particle duality"? In quantum theory, two generations of physicists have been taught that waves and particles are so antithetical that when we think about them we must forego using even the notion of an "objectively real" world. That is, present quantum theory cannot, as a matter of principle, answer any question of the form: "What is really happening when . . . ?" It can answer only: "What are the possible results of this experiment and their probabilities?" Yet after being taught that waves and particles form a fundamental, unbridgeable dichotomy, do not wavelets provide a simple, continuous interpolation between them?

References

Blackman, R. B., and J. W. Tukey (1958) The Measurement of Power Spectra (New York: Dover).

Bleher, P. M. (1981) Inversion of Toeplitz matrices, Trans. Moscow Math. Soc., Issue 2.

Boltzmann, L. (1877) Wien. Ber. **76**, 373.

Burg, J. P. (1967) Maximum entropy spectral analysis, Proc. 37th Meeting, Society of Exploration Geophysicists: Reprinted in D. G. Childers, ed. (1978) Modern Spectrum Analysis (New York: Wiley).

Burg, J. P. (1975) Ph.D. thesis, Stanford University.

Einstein, A. (1915) Ann. Phys. **47**, 879-885.

Einstein, A. (1917) Phys. Zeit. **18**, 121-128.

Einstein, A., and L. Hopf (1910) Ann. Phys. **33**, 1096-1104.

Everett, H. (1957) Relative state formulation of quantum mechanics, Rev. Mod. Phys. **29**, 454-462.

Fisher, R. A. (1956) Statistical Methods and Scientific Inference (New York: Hafner).

Gibbs, J. W. (1875-1878) Heterogeneous equilibrium, Conn. Academy of Science (reprinted 1928, New York: Longmans, Green and Co.; 1961, New York: Dover).

Gibbs, J. W. (1902) Statistical Mechanics (reprinted 1928, New York: Longmans, Green and Co.; 1961, New York: Dover).

Gohberg, I. C., and I. A. Fel'dman (1974) Convolution Equations and Projection Methods for Their Solution, Trans. Math. Monographs **41** (Providence, R.I.: Am. Math. Soc.), p. 86.

Gull, S. F., and G. J. Daniell (1978) Nature **272**, 686-690.

Jaynes, E. T. (1957) Information theory and statistical mechanics, Phys. Rev. **106**, 620; **108**, 171.

Jaynes, E. T. (1961) Review of Y. W. Lee, Statistical Theory of Communication, Am. J. Phys. **29**, 276.

Jaynes, E. T. (1967) Foundations of probability theory and statistical mechanics, in M. Bunge, ed., Delaware Seminar in the Foundations of Physics (Berlin: Springer-Verlag).

Jaynes, E. T. (1968) Prior probabilities, IEEE Trans. Syst. Sci. Cybern. **SSC-4**, 227-241. Reprinted in V. M. Rao Tummala and R. C. Henshaw, eds. (1976) Concepts and Applications of Modern Decision Models (Michigan State University Business Studies Series).

Jaynes, E. T. (1973) Survey of the present status of neoclassical radiation theory, in L. Mandel and E. Wolf, eds., Coherence and Quantum Optics (New York: Plenum), pp. 35-81.

Jaynes, E. T. (1976) Confidence intervals vs. Bayesian intervals, in W. L. Harper and C. A. Hooker, eds., Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science (Dordrecht, Holland: D. Reidel).

Harper and C. A. Hooker, eds., Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science (Dordrecht, Holland: D. Reidel).

Jaynes, E. T. (1978) Where do we stand on maximum entropy? in R. D. Levine and M. Tribus, eds., The Maximum Entropy Formalism (Cambridge, Mass.: M.I.T. Press).

Jaynes, E. T. (1980) The minimum entropy production principle, in B. S. Rabinovitch et al., eds., Ann. Rev. Phys. Chem. (Palo Alto, Calif.: Annual Reviews, Inc.), pp. 579-602.

Jaynes, E. T. (1982) On the rationale of maximum entropy methods, Proc. IEEE **70**, 939-952.

Jaynes, E. T. (1983) Papers on Probability, Statistics and Statistical Physics (Dordrecht, Holland: D. Reidel).

Klein, M. J. (1973) The development of Boltzmann's statistical ideas, in E. G. D. Cohen and W. Thirring, eds., The Boltzmann Equation (Berlin: Springer-Verlag), pp. 53-106.

Rota, G.-C. (1975) Finite Operator Calculus (New York: Academic Press).

Shannon, C. E. (1948) Bell Syst. Tech. J. **27**, 379-423, 623-656.

von Laue, M. (1915) Ann. Phys. **47**, 853; **48**, 668.