On The Rationale of Maximum-Entropy Methods

EDWIN T. JAYNES

Invited Paper

Abstract—We discuss the relations between maximum-entropy (MAXENT) and other methods of spectral analysis such as the Schuster, Blackman-Tukey, maximum-likelihood, Bayesian, and Autoregressive (AR, ARMA, or ARIMA) models, emphasizing that they are not in conflict, but rather are appropriate in different problems. We conclude that:

1) "Orthodox" sampling theory methods are useful in problems where we have a known model (sampling distribution) for the properties of the noise, but no appreciable prior information about the quantities being estimated.

2) MAXENT is optimal in problems where we have prior information about multiplicities, but no noise.

3) The full Bayesian solution includes both of these as special cases and is needed in problems where we have both prior information and noise.

4) AR models are in one sense a special case of MAXENT, but in another sense they are ubiquitous in all spectral analysis problems with discrete time series.

5) Empirical methods such as Blackman-Tukey, which do not invoke even a likelihood function, are useful in the preliminary, exploratory phase of a problem where our knowledge is sufficient to permit intuitive judgments about how to organize a calculation (smoothing, decimation, windows, prewhitening, padding with zeroes, etc.) but insufficient to set up a quantitative model which would do the proper things for us automatically and optimally.

I. Introduction

THIS PAPER concerns what is in one sense a small detail in the context of the vast amount of work done on spectral analysis. But in another sense, we are concerned with the general principles underlying all scientific inference, in which context spectral analysis is only one specialized application. Our aim is to clarify some currently puzzling questions about rationale and method.

There are many different spectral analysis problems, corresponding to different kinds of prior information about the phenomenon being observed, different kinds of data, different kinds of perturbing noise, and different objectives. It is, therefore, quite meaningless to pass judgment on the merits of any proposed method unless one specifies clearly: "In what class of problems is this method intended to be used?"

Most of the current confusion on these questions is, in the writer's opinion, the direct result of failure to define the problem explicitly enough. Today, programming and running a computer is much easier than actually thinking about a problem, so one may program an algorithm appropriate to one kind of problem, and then feed it the data of an entirely different problem. If the result is unsatisfactory, there is an undertandable tendency to blame the algorithm and the method that produced it, rather than the faulty application.

The maximum-entropy (MAXENT) method is particularly vulnerable in this respect, because its rationale is so different from that of "orthodox" statistics that it seems new and mysterious to many (although historically it dates back to Boltzmann, 1877). To compound the confusion, the MAXENT spectral estimate is, for one particular kind of data, identical in analytical form with that of an AR model, as found by Burg [1].

If that were not enough, any MAXENT solution also defines a particular model for which the predictive distribution using the maximum-likelihood estimates of the parameters, is identical with the MAXENT distribution. This is essentially the

Manuscript received March 1, 1982; revised May 21, 1982.
The author is with Arthur Holly Compton Laboratory of Physics, Washington University, St. Louis, MO 63130.

0018-9219/82/0900-0939$00.75

Pitman–Koopman theorem used backwards; given any data, the MAXENT distribution, having exponential form, in effect creates a model for which those data would have been sufficient statistics. This can give one a deeper understanding of the terms "information" and "sufficiency" in statistics, but only after some deep thought. As a result, almost every conceivable opinion about the relation between maximum entropy, maximum likelihood, and autoregression can be found expressed in the current literature.

Therefore, we first point out a class of problems in which MAXENT is demonstrably optimal, in the sense of a simple combinatorial theorem. Secondly, we stress that the analytical form of the MAXENT distribution is determined jointly by the "hypothesis space" representing our prior information about the phenomenon, and by the kind of data we have. To change either will result in a different analytical form of our spectral estimate, that of Burg being only the first discovered.

II. Entropy—Discussion

For many decades it has been recognized, or conjectured, that the notion of entropy defines a kind of measure on the space of probability distributions, such that those of high entropy are in some sense favored over others. The basis for this was stated first in a variety of intuitive forms: that distributions of higher entropy represent more "disorder," that they are "smoother," "more probable," "less predictable," that they "assume less" according to Shannon's interpretation of entropy as an information measure, etc.

While each of these intuitions doubtless expresses an element of truth, none seems explicit enough to lend itself to a "hard," quantitative demonstration of the kind we are accustomed to having in other areas of applied mathematics. Accordingly, many approaching this field are disconcerted by what they sense as a kind of vagueness, the underlying theory lacking solid content.

This has not prevented the useful exploitation of this property of entropy. The MAXENT principle, stated most briefly, is: when we make inferences based on incomplete information, we should draw them from that probability distribution that has the maximum entropy permitted by the information we do have.

Essentially all of the known results of Statistical Mechanics, equilibrium and nonequilibrium, are derivable consequences of this principle. In image reconstruction and spectral analysis, MAXENT takes into account cogent information about multiplicities that orthodox statistics, because of its failure to admit prior probabilities, misses.

But while the pragmatic usefulness of MAXENT is well established in a variety of applications, this leaves an unanswered question in the minds of many. Granted that the distribution of maximum entropy has a favored status, in exactly what sense, and how strongly, are distributions of lower entropy ruled out? Just what are we accomplishing when we maximize entropy?

We shall try to explain this here, not in terms of the most general information theory rationale originally advanced [5], but in terms of an entropy concentration property that is free of all vagueness, at the cost of somewhat restricted application.

Probably most information theorists have considered it obvious that, in some sense, the possible distributions are concentrated strongly near the one of maximum entropy; i.e., that distributions with appreciably lower entropy than the maximum permitted by our data are atypical of those allowed by the data.

Schrödinger [12], noting this concentration property qualitatively, saw it as the reason why, in various problems, the Darwin–Fowler method and the Boltzmann "method of the most probable distribution" lead to the same result in the limit $N \to \infty$, where N is a suitable "size" parameter (in Statistical Mechanics, the number of particles in a system; in Communication Theory the number of symbols in a message; in Statistical Inference, the number of trials of a random experiment, etc.). A general proof of this limiting form is given by Van Campenhout and Cover [14].

But these results, pertaining only to the limiting distribution, leave us in the same unstatisfactory state as did the original limit theorem of Jacob Bernoulli (1713): {as $N \to \infty$, the observable frequency $f = r/N$ of successes converges in probability to p}. This said nothing about how large N must be for a given accuracy. For applications one needed the more explicit deMoivre–Laplace theorem: {Asymptotically, $f \sim N(p, \sigma)$ where $\sigma^2 = N^{-1} p(1 - p)$}.

Similarily, in our present problem it would be desirable to have a quantitative demonstration of this entropy concentration phenomenon for finite N, so that one can see just how the limit is approached. This is so particularly because there are still some who, apparently unaware or unconvinced of the reality of the phenomenon, reject the Principle of Maximum Entropy as a method of inference.

This problem was discussed at the MIT Maximum Entropy Formalism Conference of May 1978, in connection with some alternative solutions that had been proposed for maximum-entropy problems. The result was a lengthy but awkward and unsatisfactory analysis [7] in which real insight into the problem had not yet been achieved. We give here a simpler, more accurate, and more general treatment of entropy concentration.

The general Principle of Maximum Entropy is applicable to any problem of inference with a well-defined hypothesis space and noiseless but incomplete data, whether or not it involves a repetitive situation such as a random experiment. However, we consider below only the special applications where we use entropy as a criterion for estimating frequencies in a random experiment about which incomplete information is available. As will be shown elsewhere, this same concentration theorem establishes the fundamental status of entropy as a criterion for testing hypotheses about systematic effects in experiments where frequency data are available.

III. Entropy Concentration Theorem

A random experiment has n possible results at each trial; thus in N trials there are n^N conceivable outcomes (we use the word "result" for a single trial, while "outcome" refers to the experiment as a whole; thus one outcome consists of an enumeration of N results, including their order). Each outcome yields a set of sample numbers $\{N_i\}$ and frequencies $\{f_i = N_i/N, 1 \leqslant i \leqslant n\}$, with an entropy

$$H(f_1 \cdots f_n) = -\sum_{i=1}^{n} f_i \log f_i. \tag{1}$$

This encompasses many different scenarios, for example:

1) Generalized Loaded Dice: A die with n faces is tossed N times, the ith face turning up N_i times.

2) Statistical Mechanics: A system contains N molecules, N_i of which are in the ith quantum state.

3) Communication: We receive a message of N symbols, chosen from an alphabet of n letters, the ith letter occurring N_i times.

4) Image Reconstruction: N elements of luminance are distributed over n pixels to form a scene, the ith pixel receiving a fraction $f_i = N_i/N$ of the total luminance.

5) Time Series: Nature generates N realizations of a time series $Y \equiv \{y_0, y_1, \cdots, y_T\}$ of which n different sequences $\{Y^{(1)} \cdots Y^{(n)}\}$ are possible. The ith sequence $Y^{(i)} = \{y_0^{(i)} \cdots y_T^{(i)}\}$ is realized N_i times.

Consider the subclass C of all possible outcomes that could be observed in N trials, compatible with m linearly independent constraints ($m < n$) of the form

$$\sum_{i=1}^{n} A_{ji} f_i = d_j, \qquad 1 \leqslant j \leqslant m. \tag{2}$$

The conceptual interpretation is that m different "physical quantities" have been measured, the matrix A_{ji} defines their "nature," and $D = \{d_1 \cdots d_m\}$ is our data set. For example, in image reconstruction A_{ji} might be the digitized point-spread function of our telescope, and D the resulting blurred image, from which we wish to estimate the f_i representing the most plausible true scene. The data D tell us that the actual outcome must have been in class C, but do not determine the frequencies $\{f_i\}$. We examine the combinatorial basis for using—and the consequences of failing to use—the entropy (1) as a criterion for estimating the $\{f_i\}$.

A certain fraction F of the outcomes in class C will yield an entropy in the range

$$H_{\max} - \Delta H \leqslant H(f_1 \cdots f_n) \leqslant H_{\max} \tag{3}$$

where $H_{\max}$ is determined by the well-known algorithm recalled in Appendix I. Their concentration near this upper bound (i.e., the functional relation connecting F and ΔH) is given by the

Concentration Theorem: Asymptotically, $2N\Delta H$ is distributed over class C as chi-squared with $k = n - m - 1$ degrees of freedom, independently of the nature of the constraints. That is, denoting the critical chi-squared for k degrees of freedom at the $100P$ percent significance level by $\chi_k^2(P)$, ΔH is given in terms of the upper tail area $(1 - F)$ by

$$2N\Delta H = \chi_k^2(1 - F). \tag{4}$$

The proof is relegated to Appendix II, since it consists of little more than repeating *mutatis mutandis* Karl Pearson's original derivation of the chi-squared distribution, taking note of the reduction of dimensionality due to constraints. Note that the theorem is combinatorial, expressing only a counting of the *possibilities*; it does not become a statement of *probabilities* unless one assigns equal probability to each outcome in class C.

IV. Example—Loaded Dice

Consider the case $n = 6$, $N = 1000$. Can we estimate the six frequencies on the basis of no information except that there are 6 faces and it was tossed 1000 times? On orthodox statistical theory the problem is hopelessly ill-posed and we have no basis for making any estimate at all.

Yet intuition is strong, even when a rational justification for it is not apparent. Almost everybody, when faced with this problem, will suggest hesitatingly the uniform distribution $f_i = \frac{1}{6}$. Pressed for the reason, he will probably say something like: "Well, I didn't see any reason to think any face was more likely than any other."

It would appear, from such a reply, that he is invoking that infamous "Principle of Insufficient Reason" and we know with what withering scorn that is regarded by orthodox statistical theory. Our unfortunate guesser would get a stern lecturing—but he would get an even sterner one if he gave any other answer. Even the most devout Orthodoxian still has a little inner voice calling for that uniform distribution.

Now let us consider this problem from the standpoint of the entropy concentration theorem. With no constraints other than normalization $\Sigma f_i = 1$, the entropy reaches its maximum value $H_{\max} = \log_e 6 = 1.79176$ just for that uniform distribution. Applying the concentration theorem, we have $6 - 1 = 5$ degrees of freedom. Entering the chi-squared tables at the conventional 5-percent significance level, we find $\chi_5^2(0.05) = 11.07$. Thus 95 percent of all *possible* outcomes have entropy in the range $2N\Delta H = 11.07$, or

$$1.786 \leqslant H \leqslant 1.792. \tag{5}$$

Likewise, $\chi_5^2(0.005) = 16.75$, and so 99.5 percent of all possible outcomes have entropy in the interval

$$1.783 \leqslant H \leqslant 1.792. \tag{6}$$

It is, therefore, pretty clear which estimate we shall wish to make. Without invoking either empirical evidence, or any probability model of Bernoulli trials we know, as an elementary combinatorial theorem, that the vast majority of all possible outcomes have frequencies close to uniform.

Once aware of this, therefore, unless we had additional evidence of systematic influences that are keeping the frequencies away from uniformity—*and* indicating in what specific way they depart from uniformity—*it* would seem highly irrational to make any other estimate than the uniform one. The entropy concentration theorem provides a quantitative justification for that intuitive predilection that we all feel for the uniform distribution.

In fact, this rationale was well understood by Jacob Bernoulli and Laplace, although they did not use the logarithmic form that we now call "entropy." They calculated multiplicities, such as

$$W = \frac{N!}{N_1! N_2! \cdots N_n!} \tag{7}$$

but today we prefer to invoke the Stirling approximation to derive

$$\lim_{N \to \infty} N^{-1} \log W = -\sum (N_i/N) \log (N_i/N) \tag{8}$$

the Shannon entropy form. Intuitively, then, distributions of higher entropy have higher multiplicity—i.e., they can be realized by Nature in more ways—and that provides a clear justification for thinking that they are more likely to be observed. As Max Planck put it, Nature will appear to have a "strong preference" for situations of higher entropy. When N becomes very large, like Avogadro's number, this relative preference $(W_2/W_1) \sim \exp[N(H_2 - H_1)]$ becomes so overwhelming that exceptions to it are never seen; and we call it the Second Law of Thermodynamics.

But the multiplicity of a parameter θ is not a frequency in any random experiment; and so with the rise of the "orthodox" view which sought to restrict the meaning of the word "probability" to frequencies, this principle of reasoning was lost to statistics. As long as one considered only problems where the multiplicities of the parameters estimated did not vary greatly, this did little harm. But today, in image reconstruction and spectrum analysis, the multiplicities of different

scenes or different spectra constitute highly cogent information that one needs for any rational predictions. As we shall see below, it is just the failure to take this information into account that leads to such anomalies as spurious sidelobes.

New Evidence: Now suppose we do acquire evidence for some systematic influence causing the distribution to depart from uniformity. We learn that in the 1000 tosses the average number of spots up was not 3.5, as we would have predicted from the uniform distribution, but

$$\sum_{i=1}^{6} i f_i = 4.5 \tag{9}$$

which is a special case of (2). Given this constraint and nothing else (i.e., not making use of any additional information that you or I might get from inspection of the die or from past experience with dice in general), what estimates should we now make of the frequencies $\{f_i\}$ with which the various faces appeared? This is a kind of caricature of a class of real problems that arises constantly in physical applications.

The distribution $\{f_i\}$ which has maximum entropy subject to the constraint (9) is found by the method of Appendix I with $n = 6$, $m = 1$, $A_{ji} = i$. The numerical results, derived in more detail before [7] are: $\{f_i = e^{-\lambda i}(\Sigma e^{-\lambda i})^{-1},\ 1 \leqslant i \leqslant 6\}$ with $\lambda = -0.37105$, or

$$\{f_1 \cdots f_6\} = \{0.0543, 0.0788, 0.1142, 0.1654, 0.2398, 0.3475\}. \tag{10}$$

This distribution has entropy

$$H_{\max} = 1.61358 \tag{11}$$

far below the range (6), indicating that the new constraint is very strong, confining us to an extremely small subclass of all the 6^N outcomes conceivable *a priori*.

Again applying the concentration theorem, we have $6 - 1 - 1 = 4$ degrees of freedom; the chi-squared tables tell us that 95 percent of all *possible* outcomes allowed by the constraint (9) have entropy in a range of width $\Delta H = (2N)^{-1}\ \chi_4^2(0.05) = 0.00474$; or, to sufficient accuracy,

$$1.609 \leqslant H \leqslant 1.614. \tag{12}$$

Thus on the "null hypothesis" which supposes that no further systematic influence is operative in the experiment other than the one taken into account (i.e., which assigns equal probability to all outcomes in the new class C), there is less than a 5-percent chance that the frequency distribution has entropy outside the interval (12).

A remarkable feature is that the "95-percent concentration range"

$$H_{\max} - \frac{4.74}{N} \leqslant H \leqslant H_{\max} \tag{13}$$

is valid asymptotically for any experiment with four degrees of freedom, although the value of $H_{\max}$ may vary widely with other details.

More interesting numerical results are found at more extreme significance levels. Thus in an experiment with 1000 trials and four degrees of freedom, 99.99 percent of all outcomes allowed by the constraints have entropy in a range of width $\Delta H = (2N)^{-1}\ \chi_4^2(0.0001) = 0.012$. In the above example this is

$$1.602 \leqslant H \leqslant 1.614 \tag{14}$$

and only one in 10^8 of the possible outcomes has entropy outside

$$1.592 \leqslant H \leqslant 1.614. \tag{15}$$

Thus given incomplete information, the distribution of maximum entropy is not only the one that can be realized in the greatest number of ways; in fact, for large N the overwhelming majority of all *possible* distributions compatible with our information have entropy very close to the maximum.

Note that the width of this region of concentration goes down like N^{-1}; and not like $N^{-1/2}$ as one might have guessed. Thus in 20 000 tosses agreeing with (9), 95 percent of the possible outcomes have entropy in the tiny interval $\Delta H = 0.00024$, and only one in 10^8 has $H < H_{\max} - 0.001$. As $N \to \infty$, any frequency distribution other than the one of maximum entropy thus becomes highly atypical of those allowed by the constraints. This is the asymptotic optimality property.

In view of this result, we can now appreciate the prophetic wisdom in the remark of Burg [1] that ". . . a reasonable goal is to find a single function, $P(f)$, which will be representative of the class of all possible spectra." This is precisely what he did accomplish; and indeed in a deeper sense than he may have realized at the time.

Even more interesting numbers are readily found. Rowlinson [11] rejected the principle of maximum entropy for this problem, and proposed as an alternative solution in place of (10) the binomial distribution

$$f_i' = \binom{5}{i-1} p^{i-1}(1-p)^{6-i}, \qquad i \leqslant i \leqslant 6 \tag{16}$$

which also satisfies the constraint (9) if $p = 0.7$. But the distribution (16) has entropy $H' = 1.4136 = H_{\max} - 0.200$, far below the limit (15). We now have $2N\Delta H = 400 = \chi_4^2(1 - F)$; or from (A12)

$$1 - F \simeq 2.94 \times 10^{-84}. \tag{17}$$

This indicates that in 1000 tosses, less than one in 10^{83} of the outcomes compatible with the constraint (9) have entropy as low as H'.

But the concentration theorem is valid only asymptotically, because of the approximation (A8) made in its derivation; and even for $N = 1000$ we might distrust its numerical accuracy that far out in the tail of the distribution. However, we can check the magnitude of (17) by direct counting.

The number of ways W in which a specific set of sample numbers $\{N_1 \cdots N_6\}$ can be realized is given by the multinomial coefficient (A5). The asymptotic formula (A7) for the ratio W/W' (which is free from any errors that might result from the aforementioned approximation) says that, for every way in which the binomial distribution (16) can be realized, there are about $\exp(N\Delta H) = \exp(200)$, or more than 10^{86} ways, in which the maximum-entropy distribution (10) can be realized (about 10^{62} ways for every microsecond in the age of the universe). While this result does not take into account the volume element factors $(r^{k-1}\,dr)$ of the full concentration theorem, it does indicate that (17) did not mislead us.

Even if we come down to $N = 50$, we find the following. The sample numbers which agree most closely with (10), (16) while summing to $\Sigma N_k = 50$ are $N_k = \{3, 4, 6, 8, 12, 17\}$ and $\{N_k'\} = \{0, 1, 7, 16, 18, 8\}$, respectively. With such small numbers, we no longer need asymptotic formulas; for every way in which Rowlinson's binomial distribution $\{N_k'\}$ can be

realized, there are exactly $W/W' = (7!\ 16!\ 18!)/(3!\ 4!\ 6!\ 12!\ 17!) = 38\ 220$ ways in which the maximum-entropy distribution $\{N_k\}$ can be realized.

Such numbers give a rather clear answer to our query, "Just what are we accomplishing when we maximize entropy?" If noiseless data do not fully determine a distribution $\{f_i\}$, it is prudent to adopt for purposes of inference that distribution which has maximum entropy subject to the data we do have. It is prudent, not for any vague, mystical reason; but for the very clear and pragmatic reason that the MAXENT predictions are the most reliable ones that can be made on the given information. It is a combinatorial theorem that to choose any other estimate would amount to ignoring the vast majority of all the possibilities allowed by the data, and concentrating our attention on a small and unrepresentative subclass of them.

Such a small subclass could exhibit almost any kind of wild anomaly, that would not be seen in practice in millions of repetitions of the experiment. In the case $N = 1000$ and four degrees of freedom, the person who chooses any distribution with entropy 0.2 units less than $H_{\max}$ is picking out some subclass containing less than one in 10^{80} of the outcomes allowed by his data. In these tiny subclasses, 10^{80} different kinds of anomalies might be hidden; the particular one he would predict would be determined not by his data, but by which particular subclass he happened to get. Yet in all probability not one of these anomalies would ever be seen in many lifetimes of observation.

Entropy maximization is a kind of insurance policy that protects us against predicting spurious details (such as sidelobes) for which there is no evidence in the data.

V. Time Series

In the preceding example each toss of the die could be considered a separate "trial" because the available information took the form of constraints on quantities that involved only sums $\Sigma\, g_i$ of properties g_i of individual trials; and not mutual properties h_{ij} of different trials. Entropy maximization then led to independent probabilities for different trials, indicating that in the majority of all possible outcomes allowed by the constraints, successive trials are uncorrelated.

This result is, again, hardly surprising intuitively although perhaps not quite so obvious; MAXENT tells us that if our data give no evidence for correlations of different trials, then we should not assume any such correlations to exist. To do so would lower the entropy of the joint distribution for N tosses—which would, again, amount to choosing arbitrarily a very small subclass of all outcomes allowed by the data.

However, in analysis of a discrete time series $\{y_0 \cdots y_T\}$ it is typical that we have some information about mutual properties $h(y_i, y_j)$ at different times in addition to properties of the form $g(y_i)$. In general, systematic effects that tend to make successive values y_i correlated will then manifest themselves, and the maximum-entropy distribution taking into account this more detailed information will exhibit these correlations. In this case, the entire times series $\{y_0 \cdots y_T\}$ must be considered a single "trial" and the combinatorial arguments refer to a collection of many different realizations of it.

This corresponds, in the original statistical mechanics applications, to the transition from the Boltzmann "molecular" point of view to the Gibbs "global" point of view. The combinatorial arguments are then at a more abstract level, but the MAXENT algorithm still applies, leading to distributions that are analogs of the Gibbs "canonical ensemble" and "grand canonical ensemble."

A real-time series is of finite length T, and the individual y_i can be recorded only to finite accuracy $\pm\epsilon$ over a finite range $|y_i| < M$; thus the number n of different realizations is finite, of the order of $(M/\epsilon)^T$. Indeed, any real problem of inference is concerned with finite sets; it is hard to believe, for example, that we will ever need to consider a larger set than $G!$, where $G \approx 10^{80}$ is the number of atoms in the known universe. Thus the paradoxes of infinite set theory are never relevant in real problems, and in case of doubt we can always retreat to the safety of a finite set, where such pathology as nonconglomerability cannot occur.

Nevertheless, infinite limits of finite sets are usually well-behaved, and if we are concerned with a huge dense set it may be a convenient mathematical approximation to consider the limit of a continuous set, where analytical methods can be used.

At this point we need a compact notation, that keeps the number of different symbols to a minimum. We denote various physical quantities by $\{A, B, \ldots\}$ and adopt the conventions:

$A \equiv$ the true value, known or unknown;
$A' \equiv$ a numerical value of A given to us in the statement of a problem; i.e., the "data";
$\hat{A} \equiv$ any estimate of A that we make, usually the mean value over a MAXENT distribution.

The MAXENT problem for our time series then takes the form: find the probability density $p(y_0 \cdots y_T)$ which has maximum entropy

$$H = -\int dy_0 \cdots \int dy_T\, p(y_0 \cdots y_T) \log p(y_0 \cdots y_T) \quad (18)$$

subject to constraints that represent all our information about the time series. If these constraints take the form of mean values of m different quantities A_k, our data set $D = \{A_1' \cdots A_m'\}$ imposes the constraints

$$A_k' = \int dy_0 \cdots \int dy_T\, p(y_0 \cdots y_T)\, A_k(y_0 \cdots y_T),$$

$$1 \leqslant k \leqslant m \quad (19)$$

and this is, but for notation, the problem that was formulated and solved by Gibbs. It is a continuous analog of (1) and (2).

Constraints of the form (19) appear general enough (with a little ingenuity in defining the functions A_k) to deal with almost any real problem yet thought of. The algorithm proceeds as before; define the partition function analogous to (A1)

$$Z(\lambda_1 \cdots \lambda_m) = \int dy\, e^{-\lambda \cdot A} \quad (20)$$

and the maximum-entropy distribution is, analogous to (A4),

$$p(y) = Z^{-1} e^{-\lambda \cdot A} \quad (21)$$

where we have passed to the compact notation $dy \equiv dy_0 \cdots dy_T$

$$\lambda \cdot A \equiv \sum_{k=1}^{m} \lambda_k A_k(y_0 \cdots y_T) \quad (22)$$

and the Lagrange multipliers λ_k are still determined by (A3).

The maximum entropy attained is a function of the data (compare (A2))

$$H_{\max} = S(A'_1 \cdots A'_m) = \log Z + \lambda \cdot A' \tag{23}$$

and if this function were known, the Lagrange multipliers would be given by $\lambda_k = \partial S/\partial A'_k$. That is, $\log Z(\lambda)$ and $S(A')$ are equivalent representations, each containing full information about the distribution, and differing by the Legendre transformation (23).

VI. Example: The Burg Problem

Our information consists of measured values R'_k of the autocovariance

$$R_k(y_0 \cdots y_T) = \frac{1}{T+1} \sum_{j=0}^{T-k} y_j^* y_{j+k}, \qquad 0 \leq k \leq m \tag{24}$$

for $m+1$ lags, where $m < T$. Put, for formal reasons, $R_{-k} = R_k^*$, although these quantities are real in most applications. If we have no other information, then the probability density that has maximum entropy while yielding the correct autocovariance will contain the Lagrange multipliers $\{\lambda_{-m} \cdots \lambda_0 \cdots \lambda_m\}$.

The quantities $A_k(y_0 \cdots y_T)$ in the general formalism may be defined with any coefficients we please, and the scalar product $\lambda \cdot A$ and final conclusion will, of course, be independent of our choice. The choice

$$A_k = \frac{T+1}{2} R_k, \qquad -m \leq k \leq m \tag{25}$$

will be convenient, making the Lagrange multipliers λ_k independent of T.

The maximum-entropy distribution is then

$$p(y_0 \cdots y_T) = Z^{-1} \exp\left[-\Sigma \lambda_k A_k(y_0 \cdots y_T)\right] \tag{26}$$

which is a Gibbsian generalized canonical ensemble. But in this case, the exponent is a quadratic form in the y_i; from (24) and (25) we have

$$p(y_0 \cdots y_T) \propto \exp\left[-\tfrac{1}{2}(y^\dagger \Lambda y)\right] \tag{27}$$

where y is the column vector $(y_0 \cdots y_T)$, $y^\dagger$ its Hermitian conjugate row vector $(y_0^* \cdots y_T^*)$, and Λ is the matrix with $(T+1)$ rows and columns

$$\Lambda_{ij} = \begin{cases} \lambda_{j-i}, & |j-i| \leq m \\ 0, & |j-i| > m \end{cases} \tag{28}$$

in which the Lagrange multipliers are assembled in the Toeplitz form.

With this kind of information, the MAXENT distribution is, therefore, multivariate Gaussian. Note that in this derivation, which differs from that of Burg, we did not assume any "Gaussian random process"; the MAXENT principle constructed the Gaussian form for us, as the distribution that could be realized by Nature in the Greatest Number of Ways (GNW) while agreeing with our autocovariance data. Henceforth, for brevity, we call this the GNW criterion. The status of Gaussian distributions in this field calls for further comment to be given elsewhere.

The partition function (20) is given by

$$\log Z = -\frac{1}{2} \sum_{j=0}^{T} \log g_j + (\text{const}) \tag{29}$$

where g_j are the eigenvalues of Λ. If we define the polynomial

$$g(z) \equiv \sum_{k=-m}^{m} \lambda_k z^k \tag{30}$$

then from Toeplitz theory, for $T \gg m$ the eigenvalues go into

$$g_j = g(z_j), \qquad 0 \leq j \leq T \tag{31}$$

where z_j are the roots of $z^{T+1} = 1$; i.e.,

$$z_j = \exp\left[2\pi i j/(T+1)\right]. \tag{32}$$

In fact, for a "circular" time series, $y_{T+1} = y_0$, Λ is a circulant matrix and (31) is exact for finite T, a fact that will prove essential in understanding the "line-splitting" phenomenon.

As $T \to \infty$, then, from (29) and (31), $\log Z$ goes asymptotically into an integral over the unit circle in the z-plane

$$\frac{2}{T+1} \log Z(\lambda_k) \to -\frac{1}{2\pi} \int_0^{2\pi} \log g(e^{i\theta})\, d\theta. \tag{33}$$

The general MAXENT formalism then determines the λ_k from $A'_k = -\partial \log Z/\partial \lambda_k$, or

$$R''_k = \frac{1}{2\pi} \int_0^{2\pi} \frac{e^{ik\theta}\, d\theta}{g(e^{i\theta})}, \qquad -m \leq k \leq m. \tag{34}$$

At this point we note an interesting property of the MAXENT formalism. Suppose we wish to extrapolate the autocovariance R_k beyond the data, for $k > m$. The estimate R_k which minimizes the expected square of the error is the expectation $R_k = E(R_k)$ taken over the distribution (26). But these Gaussian integrals are elementary and lead us to the same analytical expression (34).

Equation (34) is therefore valid for all k, but it has a different meaning in different ranges. When k is in the "information-gathering" region ($-m \leq k \leq m$), it represents the constraints $R''_k = R'_k$ determining the Lagrange multipliers λ_k. When $|k| > m$ it represents the predicted MAXENT extrapolation of the covariance function: $R''_k = \hat{R}_k$.

But a further generalization is then obvious: if we have data on any information set I; i.e., we are given R'_k for $k \in I$, then define

$$g(z) \equiv \sum_{k \in I} \lambda_k z^k \tag{35}$$

whereupon (34) will be valid for all k, with the meaning

$$R''_k = \begin{cases} \text{data } R'_k, & k \in I \\ \text{prediction } \hat{R}_k, & \text{otherwise.} \end{cases} \tag{36}$$

There is a Lagrange multiplier for each item of data we have, and the same formula (34) gives the optimal interpolation of missing data as well as the optimal extrapolation beyond the data—"optimal" by the GNW criterion.

We have noted before [7, eq. D54] that this convenient double meaning of our constraint equations holds quite generally for perturbation expansions of MAXENT solutions. Full MAXENT is a highly nonlinear algorithm; however, in first order of perturbation about a "reference" solution P_0, changes in predictions become linear functions of changes in the data (including introduction of new kinds of data not in P_0). One consequence of this double meaning is that the theory of the Wiener prediction filter is contained in the linear approximation to MAXENT (extrapolating the function itself instead of its autocovariance).

When space-time variations are considered, this result is extended to include the standard Callen-Green-Kubo theory of

response functions and transport coefficients (diffusion, electrical and thermal conductivity, etc.) of Irreversible Statistical Mechanics. All of the static transport theory summarized in the treatise of Zubarev [15] is a consequence of this phenomenon, in the case where constraints are confined to conserved quantities. Had Zubarev used more general kinds of input data, this same double meaning would have generated the algorithm for more general predictions, such as ultrasonic dispersion and attenuation [7], [8].

We interject these remarks to stress that the double meaning found in (34) is a very general and important property of the MAXENT method. The same principle that Burg used to extrapolate the autocovariance beyond the data, also generates virtually all of presently known Irreversible Statistical Mechanics.

Once the λ_k are determined from (34), the MAXENT prediction for any property of the time series follows from the distribution (29). If we wish only to predict the power spectrum $P(f)$, the result is trivial, for this is given by autocovariance

$$P(f) = \sum_{k=-\infty}^{\infty} R_k \exp(+i2\pi fk), \quad |f| \leq \tfrac{1}{2}. \tag{37}$$

But then (34) is just the inversion of this Fourier series, and we have, by inspection, $P(f) = 1/g$; the predicted spectrum is

$$\hat{P}(f) = \frac{1}{\sum_{k\in I} \lambda_k e^{-i2\pi fk}}, \quad |f| \leq \tfrac{1}{2}. \tag{38}$$

This formula, first derived by Burg [1], is one of the most beautiful analytical results in statistics. Note some of its properties; by the MAXENT principle it is the "smoothest" (by the Burg $\int \log P(f)\, df$ criterion) spectrum consistent with the data; it is therefore "fail-safe" in the sense that the MAXENT predictor (38) cannot show any details for which there is no evidence in the data.

It is interesting to note the manner in which (38) strives constantly for all the uniformity that the data will allow. Consider two problems:

Problem (I): We have data $D = \{R_0' R_1' \cdots R_9'\}$. Then in general, (38) has nine poles inside the unit circle of the z-plane, and it is potentially capable of representing nine sharp peaks in $P(f)$ in the interval $(-\frac{1}{2} < f \leq \frac{1}{2})$.

Problem (II): Same except that the particular datum R_5' is missing, and so must be estimated as $\hat{R}_5$ by interpolation from the other data using (34).

Now if the estimate $\hat{R}_5$ in problem (II) is the same as the datum R_5' in problem (I), then knowing R_5' does not give us any information that we did not have already in the other data. The Lagrange multiplier λ_5 is then zero in both problems, and it contributes nothing to the structure of the spectrum (38).

As any datum becomes "more nearly irrelevant" in the context of the other data, its Lagrange multiplier tends to zero, contributing less and less to the structure of the spectrum. The amount of detail in the MAXENT spectrum is determined, not by the number of data points we have, but by the "effective number of logically independent pieces of information" contained in them.

This is, again, a general property of the MAXENT formalism [6]; it is never necessary, when setting up a MAXENT problem, to ascertain whether the different pieces of information used are independent. Any redundant information will drop out automatically—for in any variational problem, adding a redundant constraint cannot change the solution.

Consider now *Problem (III):* Same as (I) except that R_8', R_9' are missing and must be estimated as $\hat{R}_8$, $\hat{R}_9$ by extrapolation from $\{R_0' \cdots R_7'\}$. Again, if the data R_8', R_9' are redundant in the sense that they agree with the estimates $\hat{R}_8$, $\hat{R}_9$, that we would have made without them, the Lagrange multipliers λ_8, λ_9 will be zero—but now there can be at most seven poles within the unit circle. The number of poles in (38) is equal to the maximum lag for which we have *relevant* data; redundant data do not count.

The Lagrange multipliers λ_k in the MAXENT formalism have therefore a deep meaning: λ_k is the "potential" of the datum R_k', that measures how important a constraint it represents. Redundant data are at zero potential and are therefore "invisible" in the MAXENT distribution and the predictions that come from them. A highly relevant datum R_j' is one without which our predictions would be very different; then its λ_j is large and its presence greatly lowers the entropy $S(A')$ of (23).

In (38) the spurious sidelobes of any method that extrapolates R_k to zero beyond the data are eliminated. Any distribution that gave $R_k = 0$ beyond the data would require many additional constraints beyond those in (27), and would have an entropy far below that of (27). Therefore, by the entropy concentration theorem, to use it would amount to ignoring the vast majority of all possible spectra consistent with our data, in favor of a negligibly small subclass of them. The particular subclasses picked out by the various B–T lag windows happen to be ones in which sidelobes of various amplitudes are present, separated from the true spectral lines by odd multiples of $[2(T+1)]^{-1}$. The millions of other subclasses of comparable size (each of which is just as plausible in the light of the data) would exhibit millions of other spurious features—ramps, plateaus, sidelobes with every conceivable spacing law, etc.

Unless further systematic constraints are operating, of which we were not told in the statement of the problem, the great majority of the true spectra would be close to MAXENT prediction, because the great majority of all possible spectra have that property. Conversely, if the MAXENT prediction turns out to be significantly wrong, then we have statistically significant evidence for the existence of a systematic effect pushing us into one of those subclasses; and a clue to its nature.

Pursuing this idea leads us into the significance tests mentioned in Section II above. Given frequency data $\{f_1 \cdots f_n\}$ and any hypothesis about which systematic effects are present, calculate its maximum entropy (23) conditional on our data, and compare to the entropy (1) of the data. A large discrepancy is evidence against the hypothesis, since from the entropy concentration theorem, if the hypothesis were true the data should have shown a higher entropy. Although the entropy test uses almost the same philosophy and even the same chi-squared tables as the original test of Karl Pearson, there is no need to compute all the squares of the residuals and their weighted sum for each hypothesis. That information is contained already in the maximized entropy $S(A_k')$ in (23), whose factors must in any event be calculated in fitting the hypothesis to the data (i.e., in achieving $E(A_k) = A_k'$). To test any number of hypotheses in the light of our data, we need only compare their entropies $S(A_k')$ with the entropy of the data.

In dice tossing, any imperfection in the die represents such a systematic influence tending to make the frequencies nonuniform. An analysis of the famous dice data of Rudolph Wolf by the entropy test, revealing two strong imperfections and barely significant evidence for a third very weak one, is given in [9].

Also eliminated in (38) is the possibility of a negative esti-

mate; if the data are the sampled autocovariance of a possible time series, (38) cannot become negative. Yet (38) can also exhibit sharp lines with arbitrarily high resolution when the data do contain evidence for them.

Finally, (38) even yields the correct "Lorentzian" analytical form of the resonance curve of a physical damped oscillator. At first glance, this seems almost magical—more than we had any right to expect from a mere statistical analysis that did not go into the physics of the damping process.

But the very beauty of this result—coupled with the convenient Levinson–Burg numerical algorithm to evaluate the λ_k's—may tempt one to use it in problems different from the one for which it was derived. We stress again; a method that is optimal in one class of problems can be dangerously misleading in another. At this point, we enter into the confusion of the current literature.

VII. When Does It Apply?

In the derivation just given, we have been led to the solution (38) for one very specific problem, namely one in which

a) the data consist of the exact values of

$$R'_k = \frac{1}{T+1}\sum_{t=0}^{T-k} y_t y_{t+k}, \qquad 0 \leqslant k \leqslant m.$$

b) $T \gg m$, because (38) is only the asymptotic form of (29), (A3).

If condition a) is not met, then our state of knowledge is different from that portrayed by (27), (29), and the whole problem must be reconsidered from the beginning. If condition a) is met but b) is not, then (38) may not be (in general will not be) the correct MAXENT spectral estimate that comes from (29), and the problem must be reconsidered from that point.

There is a need for some creative computer programming here. We indicate briefly the exact calculation that should, in principle, be done when condition a) is met but b) is not. Then the MAXENT distribution will always have the Toeplitz form (27), (28); but the λ_k should be determined from (A3) which now reads, using (20), (25)

$$R'_k = \frac{1}{T+1}\frac{\partial}{\partial \lambda_k}\log\det(\Lambda), \qquad -m \leqslant k \leqslant m. \tag{39}$$

This is, in complex notation $R_k = R^*_{-k}$, a set of $(2m+1)$ simultaneous equations for $(2m+1)$ unknowns; or if we revert to real notation $R_k = R_{-k}$, (39) becomes $(m+1)$ independent equations for $(m+1)$ unknowns.

Now we need to extrapolate R_k beyond the data, to get $\{\hat{R}_{m+1} \cdots \hat{R}_T\}$. But the same double meaning holds also for the constraint equations (39), if we use the trick explained above. The Toeplitz matrix (28) has nonzero potentials λ_k only up to a maximum lag of $|k| \leqslant m$ because we were given data only that far.

Now imagine *Problem (IV):* we are given as data all the R'_k for $(-T \leqslant k \leqslant T)$; but $\{R'_{m+1} \cdots R'_T\}$ happened to be redundant, equal to the MAXENT extrapolation $\{\hat{R}_{m+1} \cdots \hat{R}_T\}$. Then all the R_k are present in the general MAXENT distribution (26) and all rows and columns of Λ are occupied by λ's. But $\{R_{m+1} \cdots R_T\}$ are invisible because they are at zero potential: $\lambda_{m+1} = \cdots = \lambda_T = 0$. Looking at the problem this way, it is clear that (39) holds for all k in $(-T \leqslant k \leqslant T)$, with the derivatives evaluated at $\lambda_k = 0$ for $(m+1 \leqslant k \leqslant T)$.

The exact predicted MAXENT spectrum is then

$$P(f) = \frac{1}{T+1}\sum_{k=-T}^{T} e^{+i2\pi fk}\frac{\partial}{\partial \lambda_k}\log\det(\Lambda) \tag{40}$$

and the essence of the computer program is to find the determinant of an arbitrary finite-dimensional Toeplitz matrix.

One might think that all this could be done analytically; but to the best of the writer's knowledge, this has not yet been accomplished. For analytical properties of Toeplitz matrices we seem to have little beyond the theorems of Szegö, which concern asymptotic properties as $T \to \infty$; just the case that (40) seeks to avoid.

There is, as already noted, one case where the analytical solution is easily carried beyond (40), leading to an explanation of the "line-splitting" phenomenon. It is described in the following section.

A general-purpose computer program capable of carrying out the calculations (39), (40) would have many useful applications, not only in spectrum analysis but in statistical mechanics and various problems at experimental physics and engineering. Note that if the one-dimensional array $\{y_j\}$ could be extended to two dimensions $\{y_{jk}\}$ the algorithm would be general enough to include a new form of image reconstruction, and also vector-valued time series with every conceivable kind of intra- and inter-vector correlations.

However, the above solutions apply to only a small part of the real problems confronting us. The state of knowledge presupposed is rather unusual; we know a few exact values of the autocovariance R_k, but no values of the actual times series y_t. One can think of cases (some kind of automatic processing of the raw time series before data become available to us, as in measurement of optical coherence by interferometry) where this is indeed the situation; but in most cases of geophysical or economic time series, our raw data are the actual values $\{y_0 \cdots y_m\}$ of a short sequence, with or without appreciable noise. The autocovariance is not then something given to us in the statement of the problem, but rather R_k must itself be inferred from the raw data.

These cases call for a different analytical treatment than that given above. It would be a kind of *ad hoc* patchwork to make approximate estimates of R_k from the data; and then use them in the MAXENT solution as if they were exact (indeed, how is one to choose the "information set" I on which to make these estimates?).

The MAXENT estimator (38) stands ready to give us arbitrarily high resolution if the data contain evidence for very sharp structure in $P(f)$. But it is able to do this only because we have assured it that the values of R'_k in (36) are exact; as the zeroes of $g(z)$ approach the unit circle, their exact positions become more and more critical, and the data R'_k need to be more and more accurate if the result is to be trusted. In short, (38) is a precision, high-performance machine; but like other such machines it is specialized and can deliver that high performance only when fed very-high-quality fuel.

When our information is different from that presupposed above, then in some cases a near-optimal solution might still be found by patching up the above equations, but we really need to reconsider the entire formulation of the problem from the beginning. Some examples of MAXENT solutions appropriate to other kinds of information will be given elsewhere.

Basically, the problem is open-ended because there is no end to the variety of different kinds of information we might have in real problems. However, we expect the field to develop as did the theory of analog filters; while there is no end to the

variety of different kinds of filters that one might want, in practice a study of three or four well-analyzed standard solutions (the idealized rectangular filter, the Chebyshev filter, etc.) provides the design engineer with the understanding he needs to deal adequately with virtually all real problems. In this sense, the Burg solution may, in time, be seen as the first of a small collection of basic MAXENT solutions that cover the field adequately for practical purposes.

VIII. Line Splitting

Consider a circular time series, for which $y_0 = y_{T+1}, y_1 = y_{T+2}$, etc. All y_k are now defined modulo $(T+1)$, and our data consist of $\{R'_0 \cdots R'_m\}$ as in (24), but with the upper limits T instead of $T-k$. However, the number $(m+1)$ of possible independent pieces of data is now reduced by the circularity of the process; there is no longer any distinction between a correlation over a lag less than half way around the circle and more than half way around. More precisely, because $y_j = y_{j+T+1}$, we have necessarily not only $R_k = R^*_{-k}$, but also

$$R_k = R^*_{T+1-k}, \qquad 1 \leqslant k \leqslant m \tag{41}$$

and so the maximum possible m is $T/2$ or $(T+1)/2$, whichever is an integer.

The form (27) of the MAXENT distribution still holds, but Λ is now a circulant matrix, the "circularity" bringing in new elements in the northeast and southwest corners, $\{\lambda_T = \lambda_{-1}, \lambda_{T-1} = \lambda_{-2}$, etc.$\}$. Its eigenvalues are given by (31), (32).

Equation (38) determining $(\lambda_0 \cdots \lambda_m)$ then reduces to

$$R'_k = \frac{1}{T+1} \sum_{j=0}^{T} \frac{z_j^k}{g(z_j)}, \qquad -m \leqslant k \leqslant m \tag{42}$$

a discrete version of (34). But because of the circularity we have: a) the power spectrum need be defined only at $(T+1)$ discrete frequencies

$$f_j = \frac{j}{T+1}, \qquad 0 \leqslant j[\text{mod}\,(T+1)] \leqslant T \tag{43}$$

and b) only $(T+1)$ consecutive values of R_k are different, so

$$P(f) = \sum_{k=0}^{T} R_k e^{-2\pi i k f} \tag{44}$$

and any other $(T+1)$ consecutive values would yield the same sum. Then, because of orthogonality of $z_j^k = \exp[2\pi ijk/(T+1)]$ on the unit circle, using (42) in (44) yields simply

$$P(f_j) = \frac{1}{g(z_j)}, \qquad 0 \leqslant j \leqslant T \tag{45}$$

which is identical with (38).

This identity of the asymptotic solution to the linear time series problem and the exact solution for the circular time series problem, reveals to us the cause of the "line-splitting" phenomenon. If our computer is programmed to evaluate the Burg solution (38) which presupposes that the data $\{R'_0 \cdots R'_m\}$ are from an arbitrarily long time series, but we feed it data obtained from only a finite run $\{y_0 \cdots y_T\}$, then the computer can return to us only an approximate solution to the Burg problem—but the approximate solution to that problem happens also to be the exact solution to a different problem, that of the circular time series.

Therefore, if the autocovariance data R'_k are obtained from a finite sample $\{y_0 \cdots y_T\}$ and there is a sinusoidal component in y_t that does not go through an integer number of cycles in the sampled interval $(0 \leqslant t \leqslant T)$, then (38) is the optimal estimate of the spectrum of a process that is circular in time but has a phase jump at regular intervals $(T+1)$. Of course, a sine wave whose phase jumps suddenly by 10 degrees every 50 cycles does not have a single line spectrum; it is more complicated with two strong lines close to the "nominal" frequency and the computer program, not knowing any better, will dutifully report those two lines back to us.

From this hint, we can simulate the line splitting very easily. The phenomenon is simplest mathematically in the continuous case. The function $y(t) = \cos(\nu|t| + \phi)$ has a nominal frequency ν, but with a phase jump of 2ϕ at $t = 0$. If it persists for an interval T, its spectrum is proportional to

$$|Y(\omega)|^2 = \left| \int_{-T/2}^{T/2} y(t) \cos \omega t \, dt \right|^2 \tag{46}$$

and in the vicinity of $\omega = \nu$ this varies like

$$\left| \frac{\sin(\alpha + \phi) - \sin\phi}{\alpha} \right|^2 \tag{47}$$

where $\alpha \equiv (\nu - \omega)T/2$. Without the phase jump this would be the conventional Dirichlet $\sin\alpha/\alpha$ function, but the phase jump gives a skewed spectrum with two main peaks. In the case $\phi = \pi/4$, the spectrum peaks occur at frequencies $\omega = \nu + 6.8/T$, $\omega = \nu - 2.4/T$. However, they have different heights, in such a way that the center of gravity of the amplitude function $|Y(\omega)|$ is still at $\omega = \nu$. Computer plots of the function (47) for various values of ϕ reproduce quite nicely the various shapes of line-split spectra that have been reported previously [2], [10].

A person who tries to use the solution (38) without understanding in what problem it is appropriate, might easily find himself in just the situation noted in the Introduction; rejecting a method that is giving the correct, optimal solution to a problem on the grounds that it is not the solution to a different problem.

The difficulties we face in trying to define "What is the Problem?" are not confined to spectral analysis. As we have noted recently [7], for 200 years applications of probability theory have been plagued by the seeming impossibility of communicating to another person exactly what problem is being solved. Dating back at least to Laplace, almost every writer on probability theory has had the experience of giving the correct solution to a problem, only to have it rejected because it is not the solution to some different problem. And indeed, an old joke among mathematicians runs, "I have found a beautiful solution, but I have not yet found the problem."

We predict, rather confidently, that the line-splitting difficulty will go away if workers will reprogram their computers to get the λ's from (39).

Like any other mathematical machine, MAXENT has available to it only the information that we have put into it. The solution (38) does not estimate the spectrum of some hypothetical infinitely long time series that exists only in somebody's imagination; it has no way of knowing what you or I may be imagining. It is estimating the spectrum of the specific real time series $\{y_0 \cdots y_T\}$ that generated the data $\{R'_0 \cdots R'_m\}$.

If we believe that this is only the beginning of an arbitrarily long and stationary time series $Y^* = \{y_0 \cdots y_N\}$, $N \gg T$ and we want to predict the spectrum of Y^*, that additional information must be put into our probability model. This also leads

to a solvable problem; but it is a different problem than the one solved by (38). However, a kind of probability model that may be appropriate is suggested by the MAXENT solution.

IX. Relation to Autoregressive Models

The MAXENT spectrum (38) is of the same analytical form as that resulting from an autoregressive (AR) model of order m, driven by white noise. Because of this, some have dismissed the whole MAXENT principle as nothing but AR, thereby missing some points that we think important for present understanding, and crucial for future theoretical progress.

Note first that the mathematics of AR models is ubiquitous in this field; a power spectrum determines a covariance function, which in turn determines a Wiener prediction filter, whose coefficients can always be interpreted as the coefficients in an AR model. Thus whenever we study the power spectra of discrete time series, it is inevitable that we shall produce mathematical relations that could have been interpreted in terms of an AR model.

But this very ubiquitousness shows that merely invoking an AR model is not a method for solving a spectrum estimation problem, but only an alternative way of formulating the problem. Instead of asking "What is the power spectrum?" it is mathematically equivalent to ask, "What are the AR coefficients?" Whatever is stated in one language, can be stated as well in the other.

So when van den Bos [13] announced his discovery that the MAXENT estimate (38) is equivalent to estimating AR coefficients, he might have announced far more; not only MAXENT, but any spectrum estimation method whatsoever, can with equal justice be interpreted as solving an AR problem.

Curiously, the MAXENT principle has been criticized several times on the grounds that it fails to determine the AR order. Equation (38) seems to point to just the opposite conclusion; indeed, it is, to the best of the writer's knowledge, the only theoretical relation that *does* determine a definite AR order for us. It tells us that given data D, the optimal spectrum estimate corresponds to an AR model whose order is the maximum lag for which we have relevant data.

van den Bos expressed concern that the number of poles generated by MAXENT could be too small and thought it might be "more adequate" to use other methods for fitting AR or ARMA models to the data, "since these approaches provide in addition tests for the order of the model." These and other such comments in the literature show that the MAXENT method is in need of some fundamental clarification.

It appears to us that the MAXENT order is eminently reasonable; surely, whatever method we use, if we have data $\{R'_k\}$ only up to a maximum lag of 6, these data (even if noiseless, as we have supposed thus far) can provide no evidence for any AR coefficients beyond lag 6.

It is the essence of the above entropy concentration theorem that, to assume the existence of correlations for which there is no evidence in the data, is tantamount to ignoring arbitrarily the great majority of the possible time series consistent with our data, and concentrating on some small and unrepresentative subclass of them. Any method which did so would, far from being "more adequate" than MAXENT, be subject to criticism on just those grounds.

As noted above, all sorts of anomalies can be present in estimates that imply nonzero AR coefficients beyond the data, the most familiar example being the spurious sidelobes of the Blackman–Tukey method.

If our data are noisy, we have even less grounds for assuming AR structure beyond the data; indeed we then have grounds for doubting some of the structure that the data do indicate.

But van den Bos also expressed the opposite fear that MAXENT might give us too many poles "that were not actually present in the process" and thus generate spurious detail. Closer examination of the MAXENT method would have allayed such fears. As a general argument, the variational principle that generates it ensures that the MAXENT estimator (38), properly applied, cannot show spurious detail. But it is also important to understand the mechanism by which (38) accomplishes this.

It does not appear to us meaningful to say that a real world time series is autoregressive of order 6 but not of order 7, as if this were an "objective physical property" of the time series. However, it may be that the coefficients beyond lag 6 are so small that they make a negligible contribution to the predictability of the series.

If we have data to lag 15 which give evidence for this, then as discussed above the data $\{R'_7 \cdots R'_{15}\}$ will be nearly redundant and the Lagrange multipliers $\{\lambda_7 \cdots \lambda_{15}\}$ will be negligibly small. If λ_{15} is not strictly zero there will still be 15 poles inside the unit circle; but nine of them will be collected very near the origin where they do not contribute to the structure of the spectrum estimate. It is by this mechanism that MAXENT protects us against spurious details.

Therefore, if by the "order of the model" we mean the number of poles which are far enough from the origin to make an appreciable contribution to the predictability of the time series, then MAXENT provides a "test" for the order of the model, that is optimal for noiseless data: take data until they start becoming redundant, and note at what lag this occurs.

However, as stressed above, the MAXENT principle is more general than AR. Had we been given, in (25), different data than the covariances R_k, the canonical distribution (26) would have had a different analytical form than the almost-AR sampling distribution (27). If the data referred to specific times, the MAXENT distribution would represent a nonstationary process.

X. Current Ad Hockeries

An often-expressed goal is to estimate the spectrum of a process $\{y_0 \cdots y_T\}$ that is known to be stationary, although we have data from only a small part of it (and here we leave it for the reader to imagine how, in the real world, one could ever acquire the knowledge that a process, although unobserved, is nevertheless stationary). Then we face the problem of how to put this additional information (that the process is stationary over a longer interval than the data) into our probability model. One way of doing this, which if not optimal is at least a popular and useful *ad hoc* device, is simply to re-interpret the MAXENT distribution (27) as a sampling distribution with unknown parameters λ_k to be estimated from whatever data we have.

Now if our data consist of the values $\{y_0 \cdots y_m\}$, then as discussed above they can give no evidence about any AR coefficients beyond lag m, and so it seems reasonable to estimate at most only $\{\lambda_0 \cdots \lambda_m\}$, setting the remaining λ's equal to zero as in the Toeplitz matrix (28). If the λ's were known exactly, then in the limit $T \to \infty$ this knowledge would overwhelm any finite amount of data, and our spectrum estimate would still be given by (38); this is not obvious, but requires some derivation to show.

At this point, another popular *ad hockery* is to estimate the power spectrum by (38), in which the λ's are replaced by "estimates" of the λ's.

Note the subtle logical distinction here: in MAXENT, the λ's had no previous existence, but were Lagrange multipliers created out of the data by the process of entropy maximization; they were not estimated, but *defined*, by the constraint equation (A3). The resulting MAXENT distribution is not a sampling distribution, but a predictive distribution. Now we are reinterpreting the mathematics; (27) is regarded as a sampling distribution with parameters λ_k that were assumed to exist before acquiring the data. Presumably, therefore, they are considered to have some "objective" physical meaning; so they are now to be estimated rather than defined.

If we accept this parametric interpretation, then the *ad hockery* just mentioned will lead us to a spectral estimate that would be optimal if the λ's were known to be exact values obtained from data on the entire time series. When we have only approximate estimates of the λ's, the results have an additional uncertainty that is not easy to estimate because the positions of the poles depend on the data in a complicated way.

Clearly, if some poles approach the rim of the unit circle, putting sharply detailed structure into the estimated spectrum, the problem of "choosing the AR order" becomes a serious one. But it is serious, not as some have stated, for the MAXENT principle itself, but for this *ad hockery* with which the MAXENT principle is now being mutilated. If we estimate too few AR coefficients, the result lacks detail; if we estimate too many it will show spurious detail. So still another *ad hockery*, such as the FPE criterion, must be invoked; and we have departed rather far from a neat, theoretically justifiable method.

According to the principles of probability theory, however, this uncertainty ought to be allowed for in a quite different way; one should calculate the joint posterior distribution of the λ's and average the spectrum estimate (38) over this distribution. This in effect "hedges our bets" by smearing over the likely range of positions of the poles.

The analytical theory of this full Bayesian solution has not, to the best of the writer's knowledge, been developed, although it ought to be. Without this theory, we cannot judge how close these *ad hoc* methods are to optimal, even if we believe the model (27). Furthermore, there is not just one analytical theory to be developed, but several different ones corresponding to different kinds of data, different kinds of physical phenomena, and different objectives. However, some features can be anticipated from other Bayesian solutions as noted below.

XI. Bayesian Image Reconstruction

To sum up what has been found thus far, the combinatorial basis for MAXENT is nothing but an application of the principle that was stated clearly in the *Ars Conjectandi* of Jacob Bernoulli (1713) as his definition of probability, and used repeatedly by Laplace; it reached its present mathematical form in the work of Boltzmann (1877). With the rise of the frequency definition of probability (Venn, von Mises, Fisher) this principle was lost to statistics, with the consequence that the presently taught "orthodox" methods are able to deal with only a part of the real problems of inference. But the part they ignore (prior information) is crucial for many current statistical problems, including irreversible thermodynamics, image reconstruction, and spectral analysis.

MAXENT utilizes our prior information about the multiplicity factors (A5) of different distributions. Distributions of higher entropy are more likely, because Nature can generate them in more ways; and MAXENT is simply taking that fact into account.

Indeed, we need only the crudest form of Bernoulli's principle. Suppose we are trying to decide between two hypotheses A and B, but our data D are equally consistent with either: $p(D|A) = p(D|B)$. Then orthodox statistical theory has no criterion for choosing between them. Yet if our prior information tells us that for every way in which A could be true, there are a million ways in which B could be true [i.e., their multiplicities satisfy $W(B)/W(A) = 10^6$], then the "art of conjecture" as Bernoulli calls it, will surely lead us to choose rather confidently. In the last analysis, this is all that MAXENT amounts to, and the entropy concentration theorem is just a quantitative refinement of the reasoning. In current examples of MAXENT image reconstruction, the multiplicity ratios of two possible scenes are generally much larger than 10^6, and there is no question which one we should prefer.

However, MAXENT fails to take noise into account, a factor that orthodox methods do deal with usefully and sometimes even optimally. So they represent, in a sense, opposite reasoning formats; orthodox methods apply in cases where we have a sampling distribution (i.e., noise) of known properties, but no prior information about multiplicities; while MAXENT is for cases where we have known multiplicities but no noise.

Neither method is appropriate in problems where we have both noise and prior information, although there have been unceasing attempts to use them in such problems, with the predictably unsatisfactory results. Many important real problems, in geophysics, astronomy, and economics, are of this type; and only a full Bayesian solution is adequate to deal with them.

In the analogous but theoretically simpler problem of image reconstruction, a start on the full Bayesian solution has been made by Gull and Daniell [4]. They find, as one would expect, that taking the uncertainty due to noise into account leads one to modify the MAXENT solution, moving to a point higher up on the "entropy hill" corresponding to a smoother scene. In other words, the predicted scene has a higher entropy than would be possible if the data were noiseless, imposing a "hard" constraint on the possible solutions.

The nature of the solution can be seen by the following "hand-waving" argument; a similar result must hold also in spectrum analysis.

To incorporate noise, one represents the data $D = \{d_1 \cdots d_m\}$ by a modification of (2) above

$$d_j = \sum_{i=1}^{n} A_{ji} f_i + e_j, \qquad 1 \leqslant j \leqslant m \tag{48}$$

where e_j are the traditional noise terms. If $e_j \sim N(0, \sigma_j)$ independently, define the quadratic form

$$Q(f_1 \cdots f_n) \equiv \frac{1}{2} \sum_{j=1}^{m} \sigma_j^{-2} \left(\sum_{i=1}^{n} A_{ji} f_i - d_j \right)^2 . \tag{49}$$

Then given the prior information I and data D, any scene $(f_1 \cdots f_n)$ of entropy $H(f_1 \cdots f_n)$ has a posterior probability proportional to

$$p(\text{scene}|D, I) \propto \exp\,[NH(\text{scene}) - Q(\text{scene})] \tag{50}$$

in which both the prior probability (multiplicity factor) $W \sim \exp(NH)$ of (A5) and the likelihood $\exp(-Q)$ are present. In

the limit of zero noise, the mode of this distribution goes into the MAXENT solution (A4), and if all scenes have the same entropy the mode goes into the maximum-likelihood estimate; thus both the MAXENT and orthodox solutions are contained in the full Bayesian solution as special cases.

The maximization is unique and has a simple geometrical meaning. The conceivable scenes are constrained to a convex set $\{S\colon f_i \geq 0,\ \Sigma f_i = 1,\ 1 \leq i \leq n\}$, whose vertices are the points $f_i = 1$. By well-known properties of entropy, on S the subset $\{S_h\colon H(f_1 \cdots f_n) \geq h\}$ is strictly convex. Also $Q(f_1 \cdots f_n)$ vanishes on the hyperplane (HP) of dimension $k = n - m$, whose equation is $Af = d$; and $Q(f_1 \cdots f_n)$ is essentially the square of the distance from HP. The set $\{S_q\colon Q(f_1 \cdots f_n) \geq q\}$ is convex, its boundary forming, for $k > 1$, an elliptic hypercylinder whose "axis" is HP. As h and q vary, S_h and S_q each form a nested sequence of convex sets.

Maximization of $NH - Q$ will then lead us to a solution point of tangency of a set S_h and a set S_q. As the mean-square noise increases from zero to infinity, the solution point moves from the pure MAXENT solution to the uniform grey scene $f_i = n^{-1}$ of absolute maximum entropy $\log n$.

Higher entropy means a smoother reconstructed scene; intuitively, some of the detail in the data, which we would have to interpret as real if we knew there was no noise, is reinterpreted as more probably due to noise and is ignored. As the noise increases, our reconstructions become smoother and smoother. The Bayesian algorithm thus returns the smoothest scene compatible with the information fed into it. It is still "fail-safe" in the sense that the reconstruction cannot show any detail for which there is no evidence in the data.

But this geometrical picture shows that the Bayesian solution can be reinterpreted as a pure MAXENT solution for different constraints. We need only Lagrange's seemingly trivial observation that in a variational problem, imposing a new constraint does not change the solution if the old solution already satisfied that constraint; it only reduces the class of variations being considered.

The Bayesian solution point $P = (\hat{f}_1 \cdots \hat{f}_n)$ maximizes $NH - Q$ with respect to arbitrary variations $\{\delta f_i\}$ on S. Suppose that at P we find the value $Q(\hat{f}_1 \cdots \hat{f}_n) = Q_0$. Then the solution point *P a fortiori* maximizes $NH - Q$ with respect to the smaller class of variations that hold Q fixed at Q_0. Thus it maximizes H subject to the constraint $Q = Q_0$. This property, the *modus operandi* of Lagrange multipliers, is used here in the opposite direction to show that the unconditional Bayesian maximum may be interpreted also as a constrained maximum.

Therefore it appears that, if one had a pure MAXENT computer program already running, it could be used also to generate full Bayes solution by feeding it "preadjusted" data which amounts to fixing Q at a nonzero value to allow for noise. Indeed, this is just what Gull and Daniell [4] did in a beautiful early example of MAXENT in radio astronomy.

XII. Periodograms and Lag Windows

Historically, spectrum estimation started in the last century with the Schuster periodogram. Nobody seemed to like it, and the proposal of Blackman and Tukey (1958) to use lag windows was an easily implemented, empirical approach to correct its shortcomings.

Yet in defense of the periodogram one could note that it is, by definition, the exact power spectrum of the one real-world time series that actually exists before us, in the form of our data. Why then should we be dissatisfied with it? Different answers to this imply different "corrections" to the periodogram.

Probably the most common reason for dissatisfaction, although not always articulated, is that we do not believe all the fine detail in the periodogram. But this must mean that we do not want the spectrum of our data; we want the spectrum of something else. Until we specify exactly what that "something else" is, we are in the standard quandary of "What is the problem?" and there can be no analytical theory of optimal estimation, only empirical guesswork.

It appears to the writer that one may want to depart from the periodogram in a variety of different directions, for a variety of different reasons, but these have not been defined clearly enough to provide any definite criterion of optimality by which various algorithms could be judged. Nevertheless, we can perceive two broadly different philosophies about the undefined problem:

a) We believe that a repetition of the measurements would yield a different set of data $\{x'_1 \cdots x'_N\}$ and a periodogram P' in which the fine details would be entirely different. We want to remove details that differ erratically from one data set to another, and to retain only features that are common to all data sets. In other words, we think the data are noisy, and we want to reduce the variability of our estimates from one data set to another. Of course, if we have only one data set, it cannot tell us which features are common to all data sets. But it is clear that some kind of smoothing is needed; and introduction of a lag window is a computationally simple way of accomplishing this.

b) We view our data as incomplete, rather than noisy. The fine details in the periodogram therefore signify, not variability from noise, but artifacts caused by our Fourier-transforming only a short run of the real-time series. Again, the remedy is some kind of smoothing—but its purpose is to eliminate these artifacts by taking account of prior information we have about the possible spectra that could have generated our data. Introducing a lag window is the last thing we should wish to do; instead of further compressing our already too short run of data, we need to do the opposite, and find the most reasonable extrapolation of the time series beyond our data.

It is disconcerting that each of these two philosophies, which advise us to do opposite things, could be construed as a realistic description of the same actual physical situation!

In judging whether lag windows are appropriate we cannot escape the crucial role of prior information. If we knew in advance that our spectrum has only two sharp lines and we wanted to estimate their positions, we could learn to ignore sidelobes and would want whatever presentation gave the highest resolution. A lag window would only decrease the accuracy of our estimates. But in trying to interpret a spectrum about which nothing is known in advance, spurious features such as sidelobes are so intolerable that in the past one was willing to sacrifice half the resolution in order to keep them down. But the lag windows that did this were not a final solution, only a temporary expedient giving symptomatic relief without going after the real cause of the disease.

The major advance of Burg [1] was to see clearly that the cause of the disease lay in the unwarranted and almost surely wrong extrapolation of R_k to zero beyond the data. The MAXENT extrapolation not only got rid of the sidelobes, for noiseless data one now had much higher resolution.

In the noiseless case, Burg's criticism of windows was clear and unanswerable. It seems now generally conceded that lag windows are not appropriate in that case; and we have conceded, as noted above, that the presence of noise makes pure MAXENT inappropriate. But does the presence of noise make lag windows any more appropriate? Is the cogency of Burg's argument affected by noise?

Surely, whether noise is or is not present, all our instincts must tell us how unreasonable it is to suppose that in a real time series the autocovariance drops abruptly to zero just beyond the point at which we made our last measurement. Note, for example, what would have to happen in the coefficients λ_k of (27) to bring this about. Suppose we have data $\{R_0' \cdots R_m'\}$ from which MAXENT predicts $\hat{R}_{m+1} > 0$. Suddenly, a negative correlation λ_{m+1} would have to appear reaching across all the span of our data, of just the right magnitude to cancel out $\hat{R}_{m+1}$. Then more new coefficients $\lambda_{m+2}, \lambda_{m+3}, \ldots$ would have to put in an appearance, each precisely determined by the previous ones, so as to keep $R_{m+2} = R_{m+3} = \cdots = 0$. Every new λ would further lower the entropy of (27) taking us down into smaller and smaller subclasses of the spectra allowed by the data.

On the other hand, from the analogous image reconstruction solution (50) it seems clear that allowing for noise must take us to distributions (27) of higher entropy, and therefore even smoother extrapolations of R_k. Starting from the MAXENT solution, which is optimal in the absence of noise, making proper allowance for noise should take us not toward lag window solutions, but away from them.

This writer has not been able to envisage any situation in which there would be a theoretical justification for using lag windows in spectrum estimation, although it is clear that they are often useful as a temporary expedient—good enough for the purpose at hand, and easy to implement. However, the ambiguities of "What is the problem?" are still very great. Advocates of windows may be able to point out to us a class of well-defined problems—perhaps with some particular kind of prior information about the spectrum—in which window methods have a demonstrable optimality property.

XIII. Conclusion

The methods of spectral analysis now in use, having conquered the (resolution versus sidelobes) problem, are routinely extracting information from data in a way that would not have been possible before the major breakthrough accomplished by Burg [1]. However, present methods are still not quite optimal, always involving some *ad hoc* patchwork as noted above.

In particular, the problems of "choosing the AR order" and of "getting a variance estimate" are still dealt with by a variety of *ad hoc* devices, and decisions between them are usually based on comparing computer simulations. Yet one is convinced that a fully developed analytical theory of spectrum analysis would provide unequivocal answers to such questions, out of the principles of probability theory, with no need for any *ad hockery*.

Thus having digested the advance of 1967, practice has again outrun theory, and the analytical theory of spectrum estimation needs to be developed much further before we can know how close present methods are to the best that could ever be hoped for. In this paper we have not attempted to present such a theory, but have tried to achieve the preliminary conceptual understanding without which further theoretical development could not proceed.

Appendix I

To summarize the MAXENT algorithm, define the partition function

$$Z(\lambda_1 \cdots \lambda_m) \equiv \sum_{i=1}^{n} \exp\left(-\sum_{j=1}^{m} \lambda_j A_{ji}\right). \tag{A1}$$

Then

$$H_{\max} = \log Z + \sum_{j=1}^{m} \lambda_j d_j \tag{A2}$$

in which the Lagrange multipliers $\{\lambda_j\}$ are found from

$$\frac{\partial}{\partial \lambda_j} \log Z + d_j = 0, \qquad 1 \leqslant j \leqslant m \tag{A3}$$

a set of m simultaneous equations for m unknowns. The frequency distribution which has this maximum entropy is then

$$f_i = Z^{-1} \exp\left(-\sum_j \lambda_j A_{ji}\right), \qquad 1 \leqslant i \leqslant n. \tag{A4}$$

Other distributions $\{f_i'\}$ allowed by the constraints (2) will have various entropies less than $H_{\max}$.

Appendix II

In N trials of a random experiment, the ith result occurs $N_i = Nf_i$ times, $1 < i < n$. Out of the n^N conceivable outcomes, the number which yield a particular set of frequencies $\{f_i\}$ is the multiplicity factor

$$W(f_1 \cdots f_n) \equiv \frac{N!}{(Nf_1)! \cdots (Nf_n)!} \tag{A5}$$

and as $N \to \infty$ we have by the Stirling approximation

$$N^{-1} \log W \to H(f_1 \cdots f_n) \tag{A6}$$

the entropy function (1). Given two sets of frequencies $\{f_i\}$ and $\{f_i'\}$, the ratio (number of ways f_i can be realized)/(number of ways f_i' can be realized) is asymptotically

$$\frac{W}{W'} \sim A e^{N(H-H')}. \tag{A7}$$

The conceivable frequencies $\{f_1 \cdots f_n\}$ may be regarded as Cartesian coordinates of a point P in an n-dimensional space, restricted to $\{S: 0 \leqslant f_i, \Sigma f_i = 1\}$, the $(n-1)$-dimensional convex set noted above. On S, the entropy (1) varies continuously, taking on all values in $(0 \leqslant H(P) \leqslant \log n)$ as P moves from a vertex to the center.

But now we obtain information that imposes the m linearly independent constraints (2), which define an $(n-m)$-dimensional hyperplane M. P is now confined to the intersection $S' = M \cap S$, a closed set comprising a bounded portion of the hyperplane M, of dimensionality $k = n - m - 1$.

On S', the entropy attains a maximum $H_{\max} \leqslant \log n$ at a unique point of S'. For the set $\{S_x: P \in S, H(P) \geqslant x\}$ is strictly convex; entropy maximization with constraints linear in $\{f_i\}$ thus amounts to finding the value of $x = H_{\max}$ for which S' is a supporting tangent plane to S_x. In S' we may define new coordinates $\{x_1 \cdots x_k\}$ as appropriate linear functions of $\{f_1 \cdots f_n\}$ such that the new origin is at the maximum-entropy point, and there is a distance $r = (\Sigma x_i^2)^{1/2}$ such that near the origin a power series expansion yields

$$H(P) = H_{\max} - ar^2 + \cdots, \qquad a > 0. \tag{A8}$$

We then have a volume element in S' proportional to $r^{k-1}\, dr$. The domain of all possible frequency distributions $\{f_1 \cdots f_n\}$ which satisfy the constraints and whose entropy is in the range (3) is a k-sphere of radius R, given by $aR^2 = \Delta H$.

In N trials, this sphere contains a fraction F of all possible outcomes in class C. From (A7), (A8) this is given asymptoti-

cally by

$$F \sim I(R)/I(\infty) \tag{A9}$$

where

$$I(r) \equiv \int_0^R e^{-Nar^2} r^{k-1}\, dr. \tag{A10}$$

But, setting $NaR^2 = N\Delta H = (1/2)\chi^2$, this is just the cumulative chi-squared distribution with k degrees of freedom, in conventional notation the relation between ΔH and F is given by (4).

In our applications we are generally concerned with numerical values for large $N\Delta H$, beyond the range of tables. The chi-squared distribution $F(N\Delta H)$ may be expressed analytically as

$$F(x) = \frac{1}{s!}\int_0^x t^s e^{-t}\, dt \tag{A11}$$

where $s = (k/2) - 1$. For large $x = N\Delta H$, this yields the asymptotic expansion

$$1 - F(x) \sim (s!)^{-1} x^s e^{-x}[1 + sx^{-1} + s(s-1)x^{-2} + \cdots]. \tag{A12}$$

When s is an integer (k even) (A12) terminates and is exact. Most of the numerical results cited in the text have been obtained from (A12).

Acknowledgment

I have profited from several discussions with J. W. Tukey, who pointed out the line-splitting phenomenon. The entropy concentration theorem was first presented at the Nineteenth NBER-NSF Seminar on Bayesian Inference in Econometrics, Montreal, Que., Canada, October 1979. The participants are thanked for several useful comments.

References

[1] J. P. Burg, "Maximun entropy spectral analysis," in *Proc. 37th Meet. Soc. Exploration Geophysicists*, 1967; Stanford Thesis, 1975.
[2] P. F. Fougere, *J. Geophys. Res.*, vol. 82, pp. 1051, 1054, 1977.
[3] S. F. Gull and G. J. Daniell, "Image reconstruction from incomplete and noisy data," *Nature*, vol. 272, p. 686, 1978.
[4] ——, "The maximum entropy algorithm applied to image enhancement," *IEEE*, vol. 5, p. 170, 1980.
*[5] E. T. Jaynes, *Phys. Rev.*, vol. 106, p. 620; also vol. 108, p. 171, 1957.
*[6] ——, "Prior probabilities," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-4, pp. 227–241, Sept. 1968. Reprinted in *Concepts and Applications of Modern Decision Models*, V. M. Rao Tummala and R. C. Henshaw, Eds. (Michigan State Univ. Business Studies Series), 1976.
*[7] ——, "Where do we stand on maximum entropy?" in *The Maximum Entropy Formalism*, R. D. Levine and M. Tribus, Eds. Cambridge, MA: MIT Press, 1978, pp. 15–118.
*[8] ——,"The minimum entropy production principle," in *Annual Review of Physical Chemistry*, vol. 31 (Annual Reviews, Inc., Palo Alto, CA, 1980, pp. 579–601.
[9] ——, *Papers on Probability, Statistics, and Statistical Physics*, a reprint collection, Dordrecht, The Netherlands: D. Reidel, 1982.
[10] S. Kay and S. L. Marple, Jr., in *Rec. IEEE ICASSP*, pp. 151–154, 1979.
[11] J. S. Rowlinson, *Nature*, vol. 225, pp. 1196–1200, 1970.
[12] E. Schrödinger, *Statistical Thermodynamics.* Cambridge, England: Cambridge Univ. Press, 1948.
[13] A. van den Bos, *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 493–494, 1971.
[14] J. Van Campenhout and T. M. Cover, "Maximum entropy and conditional probability," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 483–489, 1981.
[15] D. N. Zubarev, *Nonequilibrium Statistical Thermodynamics.* New York: Plenum, 1974.

*These papers are reprinted in [9].