SIAM-AMS PROCEEDINGS
Volume 14
1984

PRIOR INFORMATION AND AMBIGUITY IN INVERSE PROBLEMS

E. T. Jaynes

ABSTRACT. Mathematically ill-posed problems, asking us to invert a singular or nearly singular operator, appear constantly in applications. Many attempts have been made to deal with such problems by inventing ad hoc algorithms which imitate the direct mathematical inversion that one would like to carry out.

But if we view these as problems of inference rather than inversion there is a formal Decision Theory that, by taking into account prior information and value judgments about the purpose of the inversion, can often guide us without ad hockery to an algorithm that is unique, numerically stable, and demonstrably optimal by some rather basic criteria of rational choice.

The method is illustrated by two simple examples, inversion of an integral equation in which the instability is resolved by prior information about the class of possible solutions; and inversion of a singular matrix (image reconstruction) in which the ambiguity is resolved by entropy factors; i.e., prior information about multiplicity.

1. ILL-POSED PROBLEMS. The terms "Well-Posed" and "Ill-Posed" are commonly attributed to Hadamard.[1] However, in the Nineteenth Century Bertrand[2] applied the epithet "mal posée" to his famous paradox in probability theory, in which one asks for the distribution of lengths of chords drawn at random on a circle. He evidently meant the term in the sense of "underdetermined".

In applications, underdetermined problems are the rule rather than the exception. In physics, engineering, or statistics it usually requires creative imagination (inventing models, or prior information such as initial conditions that we do not actually possess) to convert a real problem into one which is "well-posed" in the sense that the statement of the problem gives just enough information to determine one unique solution. In biology, econometrics, geophysical exploration, medical diagnosis, and synthesis of electrical filters or optical systems, a truly well-posed problem is virtually unknown.

We must therefore, of necessity, learn to reason somehow in logically indeterminate situations. G. Polya[3] termed this "plausible reasoning" and showed that even a pure mathematician uses it constantly. Polya's plausible reasoning remained qualitative, although he noted a loose correspondence with the relations of probability theory.

0080-5084/84 $1.00 + $.25 per page

The quantitative use of probability theory for this purpose is what we shall call inference. As we use it, this term includes "Bayesian statistics", but is more general in that Bayes' theorem is only one of the useful principles of inference. Others presently known include group invariance, maximum entropy, and coding theory.

Examples of underdetermined problems are spectrum analysis, image reconstruction, determining the shape of a target from its radar reflections, determining crystallographic or macromolecular structure from X-ray scattering patterns; and as we shall see, the Statistical Mechanics of J. Willard Gibbs for predicting thermodynamic properties--in all cases from incomplete and/or approximate data.

A problem might be ill-posed, in the more general sense of "not well-posed", in other ways. Usually, an overdetermined problem would be considered not merely ill-posed, but wrongly posed, calling for reformulation rather than inference. But a problem may also be formally well-posed, but nevertheless pragmatically without a unique solution because of practical difficulties such as instability, that make it impossible to use the solution with real data.

Examples of such "morally ill-posed" problems are analytic continuation from numerical data, some Fredholm integral equations, extrapolating a solution of the diffusion equation backward in time, determining subsurface structure from surface gravimetric or seismic data, and the mechanics of billiard balls.

An unstable problem may be much like an underdetermined one, in that the formal solution must be supplemented by additional means, such as a preliminary smoothing or other "massaging" of the data in a particular way, or putting in a preliminary bias favoring some possibilities over others. The rationale for these is not always clear to the uninitiated, since it often arises out of prior knowledge of the subject-matter that is too extensive to be repeated in the statement of the problem, and can only be presumed "understood by the expert".

Furthermore, in both underdetermined and unstable problems we may require not only inference making use of expert prior information, but also value judgments indicating what we want the solution to accomplish, in order to arrive at useful results.

The above remarks sound very much like an introduction to Statistical Decision Theory, which shows us how to take prior information and value judgments into account, in a way that is proved optimal by some very fundamental and pretty nearly inescapable criteria of rational behavior. Indeed, the problem of inverting some singular or nearly singular operator A (i.e., given $y = Ax$, estimate x) would seem to call out for decision theory, just as clearly as the problem of driving a nail calls out for a hammer. This makes it curious that so little use has been made of this theory--or even probability theory--in dealing with ill-posed problems.

2. AD HOC ALGORITHMS. Tikhonov and Arsenin[4] explore a variety of unstable inverse problems, but do not see them as problems of inference or decision at all. Instead they invent various ad hoc algorithms that serve as a substitute for inversion; their "regularized" solutions force continuity in a neighborhood of a known exact solution, but in a way that does not necessarily make use of any prior information, or even any properties of the operator A--and for which it is therefore hard to give any convincing rationale.

When they take the nature of A into account to the extent of minimizing a mean square error metric, they have in effect rediscovered the Wiener[5] filter algorithm. But that too can become unstable in just the problems of greatest interest. In forecasting a time series with the Wiener prediction filter, for example, as we approach the limit where the Paley-Wiener criterion ceases to be satisfied, correlations persist for longer and longer times and formally the time series becomes more and more predictable. Actually, the prediction algorithm approaches analytic continuation and becomes less and less stable.

The limit is reached just for the physicist's favorite random function, Planck black-body radiation with power spectrum $I(\omega) \propto \omega^3/(e^{b\omega}-1)$. The Paley-Wiener integral $\int \log I(\omega)/(1+\omega^2)d\omega$ then diverges, and the conventional Wiener theory thus tells us that the time series (say, the x-component of electric field) is perfectly predictable from its past.

Of course, nobody familiar with the realities would believe this for an instant. The Wiener prediction algorithm here reduces to analytic continuation, not only impossible from numerical data, but even physically wrong for reasons apparent to physicists but neglected in the Wiener theory.

This reminds us that in many inverse problems, as the solution approaches instability, not only do our conclusions become highly sensitive to small changes in the data; they become equally sensitive to the exact physical assumptions underlying the theory itself. When the numerical algorithm becomes shaky, the whole foundation of the theory may also become shaky and a direct mathematical inversion, even if achieved, could be more misleading than useful.

Thus in a variety of real problems a different kind of philosophy and rationale is needed. Unique deductively obtained results (i.e., direct mathematical inversion) being impossible, we must set our sights on some other goal, perhaps more modest but attainable.

3. THE ROLE OF PROBABILITY THEORY. It appears to us that the reason for this neglect of inference/decision theory methods in dealing with ill-posed problems lies in the attitude toward probability theory itself that is instilled by most current pedagogy. As currently taught, probability theory does not seem

applicable unless a problem has some evident "element of randomness". Even then, the conventional "frequentist" interpretation places strong restrictions on the allowable forms of its application.

This was stressed by L. J. Savage[5] who noted that on the conventional view probabilities may be assigned only to "random variables" and not to hypotheses or parameters; and evidence as to the magnitude of a probability is to be obtained "by observation of some repetitions of the event, and from no other source whatsoever." But this means that (a) we are prohibited from using probability theory for inference about the very things about which we are most interested; and (b) we are prohibited from making use of any prior knowledge we might have--however cogent--if it does not happen to consist of frequency data.

As Savage and others have noted, such a program is almost never workable in problems of the real world and nobody can really adhere to it; yet it continues to be taught in most statistics courses.

However, we view probability theory, as did Good,[6] Savage[5], Jeffreys,[7] de Finetti[11] and many others, as basically a set of normative rules for conducting inference, with no necessary connection with the notion of "random variables". As we have expounded elsewhere,[9] the usual equations of probability theory are uniquely determined, as rules for inference, by some very elementary desiderata of consistency that make no reference to random experiments.

This broader view (actually, the original view of Jacob Bernoulli) is not in conflict with the conventional interpretation of probability as frequency in a random experiment; rather, the latter is included as a special case of probabilistic inference, for certain kinds of propositions or prior information (indeed, just the kind that the "frequentist" would want before using probability theory at all). But unlike the frequentist we consider it legitimate to assign probabilities to any clearly stated proposition; in the special case that it happens to be a proposition about frequencies, then the usual connections between probability and frequency are found to appear automatically, as a consequence of our theory. Bernoulli's "weak law of large numbers" was only the first of many such connections; another important one is contained in the famous de Finetti[8] exchangeability theorem.

We hope to show here that this reinterpretation of probability theory can convert an ill-posed problem of deductive reasoning into a well-posed problem of inference. Indeed, on the viewpoint advocated here, Bertrand's original "ill-posed" problem proves to be well-posed after all, with a unique solution that was conjectured by Borel and has been verified experimentally.[10]

In a sense, the following considerations might be called pre-mathematical rather than mathematical. At least in the applications we have studied, once

a definition algorithm has been decided upon, realizing it explicitly tends to be straightforward. The difficulties that still hold up progress involve rather the preliminary rationale by which one decides which specific algorithm we should seek. That is the problem we address here.

4. A SIMPLE UNSTABLE PROBLEM. A proposal of Wolf and Mehta[12] to measure fluctuations in light intensity from data on fluctuations in the counting rate of the photoelectrons ejected by the light, gives us an excellent example of a problem of this type. It is formally well-posed and mathematically quite simple; but nevertheless exhibits almost all of the practical difficulties noted in the various more complicated inversion problems described at this Symposium. So let us see how to deal with it by inference rather than inversion.

The variable λ is the "light intensity" in units such that given λ, the conditional probability of observing n ejected photoelectrons in some nominal time interval, say a microsecond, is the Poisson distribution:

$$p(n|\lambda) = \exp(-\lambda)\lambda^n/n! \quad , \tag{1}$$

(More explicitly, $\lambda = qE/h\nu$, where ν = light frequency, E = light energy incident on the photocathode during that microsecond, and q = quantum efficiency).

But λ fluctuates from one microsecond to another according to some probability distribution $P(\lambda)$, and so the probability distribution for observed photoelectrons is a mixture of Poisson distributions:

$$p(n) = \int_0^\infty p(n|\lambda)P(\lambda)d\lambda \quad . \tag{2}$$

Wolf and Mehta note that this integral equation can be inverted, yielding the formal solution

$$P(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} ds\ e^{-is\lambda} \sum_{n=0}^{\infty} (1+is)^n\ p(n) \tag{3}$$

and it seems at first glance natural to conclude, with them, that by use of (3) we can determine $P(\lambda)$ from experimental measurements of the distribution of observed counts n.

But anyone who tries to do this using for p(n) the observed counting distribution will discover three difficulties with the formal solution:

(A) With any finite amount of data there will be some $N = n_{max}$, the maximum number of photoelectrons observed in any microsecond. The sum in (3) is then a polynomial of degree N, and we obtain the startling conclusion that $P(\lambda)$ is a sum of derivatives of delta-functions:

$$P(\lambda) = \sum_{n=0}^{N} a_n \delta^{(n)}(\lambda) \quad . \tag{4}$$

Evidently, in order for this procedure to yield any acceptable distribution at all (i.e., a $P(\lambda)$ that is non-negative, normalized, and zero for $\lambda < 0$), the raw data must be extrapolated to $n \to \infty$ in a way that makes the sum in (3) a rather special, well-behaved analytic function of s. But the resulting $P(\lambda)$ will depend, not only on our data, but on how this extrapolation is carried out, and the statement of the problem does not seem to provide any criterion for preferring one extrapolation over another. This is an example where massaging of the data, guided by some of that expert prior information, is necessary.

(B) Even if we had an infinite amount of data, thus avoiding the extrapolation problem, we would not in general get an acceptable $P(\lambda)$, because the algorithm (3) has an unacceptable instability. A small change in the data, for example, $p_2 \to p_2 + 0.001$, $p_3 \to p_3 - 0.001$, would make an unbounded change in the right-hand side of (3). Common sense tells us that our conclusions ought to depend continuously on our data.

(C) Finally, we need to distinguish between the theoretically assigned probability p_n and the empirically measured frequency f_n. From the standpoint of frequentist probability theory, the observed f_n is equal to the "true" p_n plus some "random error" e_n. We would prefer to verbalize this a little differently, but the pragmatic result is the same: common sense tells us that, with incomplete data of finite accuracy, $P(\lambda)$ cannot be recovered with deductive certainty and perfect accuracy. We can make only rather crude estimates, and the accuracy of the estimate surely must depend on the amount of data we have. Yet the proposed solution (3) by inversion makes no reference at all to the amount of data or the accuracy of the result!

These same difficulties infect most of the inverse problems discussed at this Symposium and by Tikhonov and Arsenin. That these difficulties exist is, of course, well recognized; that their resolution by inference rather than inversion also exists, is our main message.

5. INFERENCE. It is clearly asking too much to expect that a finite amount of noisy data can be determine a full continuous function $P(\lambda)$ without further restrictions; and so we seek solutions only within some prescribed class of conceivable functions $P(\lambda|\theta)$, the different functions of the class being characterized by a parameter θ, which may be multidimensional. We then ask of the data only that they provide us with some "best" estimate of θ, and a statement about the reliability of the estimate. The particular class of functions considered would be chosen on the basis of some of that "expert prior information" about what kind of distributions are likely, given what one knows about the source of the light.

Fortunately, in the present problem little expertise is needed to see that, for most light sources the electric field $E(x,t)$ is a sum of an enormous number

of small and nearly independent contributions (emission from individual atoms), and so by the Central Limit Theorem we expect a Gaussian distribution for E, therefore an exponential distribution for λ (which is essentially a space-time average of E^2) if the observation time is short compared to the coherence time.

An estimate of θ should, of course, be based on all the available evidence, and not just the evidence of one experiment or measurement. The same problem was faced by Laplace in the 18'th Century, as he sought to combine astronomical data from various sources into a single "best" estimate of some parameter, such as the mass of Saturn, appearing in the equations of celestial mechanics. He showed that probability theory can often tell us, uniquely, how evidence from different sources is to be combined into a final estimate, and what accuracy we are entitled to claim for that estimate.

In Laplace's problems, the final algorithm usually turned out to be weighted least squares. Unfortunately, then as now there was a tendency to confuse the algorithm with the method. In the nineteenth Century, losing sight of Laplace's rationale, "Least Squares" came to be considered a principle in its own right, to be applied indiscriminately in all problems whether or not it could be justified by the principles of probability theory. To avoid repeating past mistakes, it is important that in each new problem we re-examine the probabilistic basis for our algorithm.

In our present problem, the experiment consists of K repetitions of the measurement of n; denote the data obtained by $D \equiv \{n_1, n_2, \ldots, n_K\}$, and any additional prior information (which might be the result of previous experiments or a theoretical analysis) by I and let T stand for the statement: "θ is in the interval $(\theta, \theta+d\theta)$". The evidence I contains some information about θ, described by a probability $p(T|I)$. From the product rule of probability theory: $p(T,D|I) = p(T|D,I)p(D|I) = p(D|T,I)p(T|I)$ we have, if $p(D|I) > 0$ (i.e., the data set is a possible one),

$$p(T|D,I) = p(T|I)\,\frac{p(D|T,I)}{p(D|I)} \quad . \tag{5}$$

Then, indicating a probability density $f(\theta|X)$ of θ conditional on any information X according to $p(T|X) = f(\theta|X)d\theta$, the final probability distribution for θ, given the prior information and the data, will have the form

$$f(\theta|D,I) = A\, f(\theta|I)\, p(D|\theta,I) \tag{6}$$

where A is a normalizing constant, independent of θ. The evidence contained in the experimental data D thus resides entirely in the θ dependence of the factor $p(D|\theta,X)$; all other details of the data are irrelevant for the estimation of θ. Usually, I will be relevant to the probability of obtaining the data D only through its relevance to the value of θ, in which case it is superfluous in

this factor: $p(D|\theta,I) = p(D|\theta)$.

For example, suppose we seek the unknown distribution $P(\lambda)$ in the aforementioned class of exponential densities $P(\lambda|\theta) = \theta^{-1}\exp(-\lambda/\theta)$ corresponding to Gaussian distributions for the field components. θ is then the average light intensity; given θ, the probability of obtaining exactly n counts in any one measurement is

$$p(n|\theta) = \int_0^\infty p(n|\lambda)P(\lambda|\theta)d\lambda = \theta^n/(1+\theta)^{n+1} \tag{7}$$

and if successive measurements are statistically independent (i.e., separated by a time long compared to the correlation time of the light), the probability of obtaining the entire run of data D is a product of K such factors:

$$p(D|\theta) = \frac{\theta^N}{(1+\theta)^{N+K}} \tag{8}$$

where $N = \Sigma_i\, n_i$ is the total number of counts observed.

If the additional evidence I yields a probability density $f(\theta|I)$ which varies little in the range of θ where (8) is appreciably large, then I is very uninformative compared to the evidence of the experiment (i.e., the experiment is well-designed); and from (6) the final probability density $f(\theta|D,I)$ will be, for all practical purposes, proportional to (8) in its dependence on θ.

However, a slight formal refinement may be achieved by considering the prior probability density $f(\theta|I)$ a little more carefully; and if we have very little data the difference might be noticeable. Suppose we wish to express "complete prior ignorance" of the value of θ; what function $f(\theta|I)$ does this? Stated in this way, the question has been rightly rejected in the past as ill-posed; the phrase "complete ignorance" is too vague to define any specific mathematical problem. But in fact we are not completely ignorant; if we know the distribution $P(\lambda|\theta)$ we can hardly be ignorant of the fact that θ is a scale parameter.

Presumably, by ignorance of the absolute scale of the problem one ought to mean a state of knowledge that is not changed by a small change in that absolute scale; just as ignorance of one's location is a state of knowledge that is not changed by a small change in that location. One may, therefore, view "ignorance" as an invariance property; the probability density that is invariant under the group of scale changes ($\theta \to \theta' = a\theta$) satisfies the functional equation $f(\theta) = af(a\theta)$; i.e., it is $f(\theta|I) = (1/\theta)$.

Indeed, this prior was advocated long ago, on partly intuitive grounds, by Jeffreys.[7] The group invariance argument[13] is at least a strong heuristic principle taking a step toward a more rigorous derivation, but it still depends

on intuition to the extent that the user must choose the group. We have now taken another step in proving, via the integral equations of marginalization theory,[14] that in a problem with two parameters (θ,α) with θ a scale parameter, the Jeffreys prior $(1/\theta)$ is uniquely determined as the only prior that is "completely uninformative about α" without further qualifications.

Because of this convergence of quite different lines of argument, and the good pragmatic success we have had in using it for many years, we shall advocate using the Jeffreys prior here if we wish to express a completely open prior opinion, that leaves the entire decision to the evidence of the data (thus achieving R. A. Fisher's goal of "letting the data speak for themselves").

The fact that the Jeffreys prior is improper (i.e., not normalizable) could be dealt with, if needed, by approaching it as the limit of a sequence of proper priors, as we have shown elsewhere.[14] However, in the present problem the integrals converge so well that this is not necessary; our result is the same.

If we do have cogent prior information and relatively little data, then adopting a different prior distribution which expresses that prior information may improve the reliability of our estimates. This has been found particularly in recent work on forecasting economic time series,[15] where incorporating prior information about regression coefficients can make a quite noticeable improvement in the forecasts; and in the problem of seasonal adjustment,[16] where prior information about the smoothness of the seasonal component can make a major change in our estimate of the irregular component.

With the Jeffreys prior, (6) and (8) yield the posterior density

$$f(\theta|D,I) = \frac{\Gamma(N+K)}{\Gamma(N)\Gamma(K)} \quad \frac{\theta^{N-1}}{(1+\theta)^{N+K}} \quad , \qquad 0 < \theta < \infty \tag{9}$$

and it is usually sufficient to express our conclusions in the form of a few moments or percentiles of this distribution. To find the percentiles, note that (9) is a Beta distribution in the variable $x \equiv \theta/(1+\theta)$, so that the identity of the incomplete Beta function and the incomplete Binomial sum[17] gives the cumulative distribution

$$p(\theta<\alpha|D,I) = \sum_{r=0}^{K-1} \frac{\Gamma(N+r)}{\Gamma(N)\, r!} \quad \frac{\alpha^N}{(1+\alpha)^{N+r}} \tag{10}$$

in a form for computer evaluation (the "Snedecor F-tables" of the statistician could be used also, for the particular percentiles tabulated).

The moments of the distribution (9) are found to be

$$E(\theta^m|D,I) = \frac{\Gamma(N+m)\Gamma(K-m)}{\Gamma(N)\Gamma(K)} \quad , \qquad -N < m < K \quad . \tag{11}$$

The estimate of θ which minimizes the expected square of the error (our "value judgment" for this case) is then $t \equiv E(\theta|D,I) = N/(K-1)$, and the variance of (9) is

$$\sigma^2 = t(t+1)/(K-2) \tag{12}$$

The (mean) $\pm$ (standard deviation)

$$(\theta)_{est} = t \pm \sigma \tag{13}$$

is then a reasonable statement of our "best" estimate and its accuracy. For example, if we wish to determine θ to $\pm 1\%$ accuracy for a light source that gives a counting rate of about t counts/microsecond, we shall require a number of observations $K > 10^4\ (1 + t^{-1})$, a result which is hardly surprising to common sense, but of which the attempted inversion (3) gives no hint. As $(N,K) \to \infty$, (9) goes into a normal distribution: $\theta \sim N(t,\sigma)$.

This example shows how the inference (9) can answer, successfully, a more modest question than the direct inversion (3) tried to answer, unsuccessfully. Our conclusions are numerically stable, and the way in which the accuracy of those conclusions depends on the amount of data is now exhibited explicitly in (12).

Note, however, that in answering a more modest question, inference is not giving us any less information than inversion; for when a reliable inversion is possible, the likelihood factor $p(D|\theta,I)$ in (6) develops a single sharp peak and inference will reduce to inversion. An unstable inversion, with the superficial appearance of giving more information, is actually giving false/unreliable information without warning us of that fact. Inference gives us those conclusions that are actually justified by the prior information and data, and it tells us, by the probable error σ, how reliable our estimates are. It may also take into account prior information that inversion ignores.

We note how the appearance of an unstable geophysical inverse problem might be changed by a similar approach. Suppose we wish to infer some subsurface property Q(z) (density, conductivity, elastic constants) from surface data D (gravimetric, electromagnetic, seismic). The data depend on Q(z) through some relation expressing physical law (potential theory, electromagnetic or acoustical wave equations, etc.); abstractly,

$$D = AQ(z) + N \tag{14}$$

where A is some operator, presumed known and N is whatever "noise"--unavoidable, uncontrollable, and unknown--places the ultimate limit on the accuracy of D.

The trouble is that the effect of Q(z) on D falls off, often exponentially fast, with z. Any attempt to reconstruct Q(z) by direct inversion of (14)

must then become increasingly unstable and unreliable with depth. An inversion algorithm may be at least usable--although with unknown accuracy--up to a certain depth, beyond which it fails entirely.

If the problem were treated by inference the final result would be, not a single value for each depth z, but a family of probability densities parameterized by z:

$$F_z(Q) \equiv f(Q|z,D,I) \propto f(Q|z,I)p(D|Q,z,I) \tag{15}$$

such that $F_z(Q)\,dQ$ is the probability that, at depth z, Q lies in (Q,Q+dQ). As we go to greater depths and the estimates become, necessarily, less accurate, $F_z(Q)$ would indicate this by becoming broader. At depths where the data can give no information, $F_z(Q)$ would reduce to the prior density $f(Q|z,I)$. At depth z, our estimate will have the form

$$(Q)_{est} = t(z) \pm \sigma(z) \tag{16}$$

in which t(z) is our "best" estimate, that inversion had tried to give, and $\sigma(z)$ indicates how reliable that estimate is. As z increases, $\sigma(z)$ would increase from values small enough to make the estimate t(z) useful, to values so large that the data have told us nothing beyond whatever prior information we had. But the algorithm is stable, the increase is smooth and continuous, and there is no point at which the method suddenly fails.

It appears to us that conclusions stated in the format (15), (16) would indicate, more usefully and more honestly, what the data actually have to tell us about the question being asked.

6. GENERALIZED INVERSE PROBLEMS. In another class of problems arising constantly in applications, the trouble is not merely that the inversion is unstable; it is in principle impossible because the operator A is singular. Here too there have been unceasing efforts to resolve the ambiguity by inventing ad hoc algorithms that imitate inversion, but take no note of the principles of inference.

For example, given values of a function

$$y(t) = \int d\omega Y(\omega) e^{i\omega t} \tag{17}$$

over only a part of its support, estimate its fourier transform $Y(\omega)$. The most obvious algorithm (take the transform of the data) gives us the fourier transform $Y_T(\omega)$ of a function $y_T(t)$ that is truncated to zero outside the measured region. In almost all real cases this would be arbitrary and unrealistic, and in some cases it would be unacceptable because it contradicts

our prior information. Thus if $y(t)$ is the autocorrelation function of a time series, $Y(\omega)$ is its power spectrum, by definition non-negative. But as Burg[17] has emphasized, $Y_T(\omega)$ is not in general non-negative.

It is clear that there is a fundamental ambiguity here, since the data cannot distinguish between two estimates $Y_1(\omega)$, $Y_2(\omega)$ whose fourier transforms $y_1(t)$, $y_2(t)$ differ only outside the measured region. For a choice between them, one must appeal to prior information and/or value judgments.

To formulate problems of this type in a general, abstract way, there is an unknown "state of Nature" x, which for brevity and with a view to image reconstruction, we shall call "the scene". It might be a number, a vector, or a function. Intuitively (and even this step may require some of that creative imagination) we think of x as belonging to some set $X = \{x_1 \ldots x_n\}$ of possible scenes.

We would like to know the true scene x, but our information is incomplete. Instead, we know only the "blurred scene"

$$y = Ax \tag{18}$$

where A is an operator, supposed known but noninvertible. That is, we cannot recover x in the manner $x = A^{-1}y$ because the data y cannot distinguish between two scenes x, x' that satisfy the "homogeneous equation" $Ax - Ax' = 0$ (a true homogeneous equation if A is linear). The best we can do is to make an inference, in which we choose some estimate of x from our data:

$$\hat{x} = Ry \tag{19}$$

where R is a "resolvent" operator to be chosen. The conceptually difficult pre-mathematical problem is: by what criterion do we choose R?

It appears that deductive logic is able to give us only one restriction on R. Given the data (18), we know at least that x must lie in the class C (subset of X) of scenes x_i that satisfy $y = Ax_i$. Thus for all possible x we should have $y = Ax = A\hat{x} = ARy = ARAx$, or

$$ARA = A \tag{20}$$

and so R, in order not to conflict with deductive reasoning, must be a generalized inverse operator. Stated differently, as seen through the "distorting window" A, the estimated scene $\hat{x}$ should be indistinguishable from the true scene x.

A problem with this simple logical structure, in which A is considered known exactly, and the data are noiseless, will be called a pure generalized inverse problem. To have the data contaminated with noise or A unknown makes the problem "impure" in our terminology.

Even in the seemingly straightforward pure problem, there have been conceptual difficulties so serious that the logically necessary condition (20) is not always recognized; the literature of spectrum analysis, image reconstruction, and quality control contains various proposed algorithms that violate it.

Then what prior information and value judgments are available to guide us to one specific choice within C? No general answer can be given once and for all; basically, this must be pondered separately for each new problem, and the following suggestions surely will not be applicable in all cases. Yet there is a large class of problems in which a single "new" principle and a rather well-developed formalism resolve this ambiguity, in a way that proves to have demonstrable optimality properties and pragmatic success. With better understanding of its rationale[18] this class is growing, and its ultimate limits are not yet in sight.

In many problems, it develops that we actually have some highly cogent prior information of which most of us are hardly consciously aware, because--as our quotation from L. J. Savage shows--conventional probability theory adjures us to ignore it. But this adjuration is just the reason why "orthodox" statistics is incapable of dealing with pure generalized inverse problems.

Although, as we have shown elsewhere,[9,18] the following rationale applies without essential change to many different kinds of problems, it will suffice here to consider a finite, discrete, linear version which amounts to inverting a singular matrix. Our general "scene" x is then represented by a set of "true but unknown" numbers $\{x_1 \ldots x_n\}$ which we wish to estimate, the general "data" y by a smaller set $\{y_1 \ldots y_m\}$ of observations, $m < n$, the general operator A by a known $(m \times n)$ matrix:

$$y_j = \sum_{i=1}^{n} A_{ji}\, x_i \quad , \quad 1 \leq j \leq m \quad . \tag{21}$$

The resolvent operator R might, conceivably, be an $(n \times m)$ matrix; but this is not required. Indeed, if we restrict R to be linear we shall hardly get past the Wiener filter type of algorithm. We advocate below a highly nonlinear R, whose performance could not be matched by any linear operation.

At present, then, there are an infinite number of different operators R, linear and nonlinear, which all satisfy the necessary condition (20), and therefore yield estimates $\{\hat{x}_1 \ldots \hat{x}_n\}$ in the class C of possible scenes.

If we have no prior information about the phenomenon being observed, which would make some scenes in C inherently more likely than others, then it appears to us that the ambiguity is fundamentally irremediable and there can be no justification for any algorithm that picks out only one scene. In that case, the only honest "solution" to the problem would seem to be:

specify the entire class C, which contains (n-m) arbitrary parameters.

Suppose, however, that we know (or wish to adopt as a working hypothesis) that Nature is generating the scene x by N repetitions of some process (call it a "random experiment" if you like) which can, at each trial, produce any one of n results $\{r_1 \ldots r_n\}$. In image reconstruction we might think of the scene as produced by distributing N little "elements of luminance" over the n pixels of the scene, such that the i'th pixel receives a total of N_i elements, and taking x_i as the fraction $x_i = N_i/N$ of total luminance in the i'th pixel. It is much like tossing N pennies onto a floor whose square tiles are numbered 1 to n, and noting how many land on the ith tile.

But this picture of the mechanism constitutes relevant prior information; there are a priori n^N different conceivable things that could happen in this sequence of tosses, and of these a given scene x could be realized in a number of ways given by the multiplicity factor

$$W(\text{scene}) = \frac{N!}{(Nx_1)! \ldots (Nx_n)!} \quad . \tag{22}$$

For large N the Stirling approximation gives asymptotically

$$\frac{1}{N} \log W(\text{scene}) \sim -\sum_i x_i \log x_i = H(\text{scene}) \quad , \tag{23}$$

the Shannon entropy of that scene. For all practical purposes, then, we may take the multiplicity of a scene as

$$W(\text{scene}) = e^{NH(\text{scene})} \quad . \tag{24}$$

With this observation, the ambiguity of our inversion problem is resolved. Scenes of higher entropy are inherently more likely because they have higher multiplicity; i.e., they can be realized by Nature in more ways. The scene which has maximum entropy subject to the constraints (21) is the one with the greatest multiplicity of all those in the class C of scenes permitted by our data; and so unless we have further prior information not yet brought to bear on the problem, it would seem irrational to choose any estimate other than the scene of maximum entropy.

There remain the questions of uniqueness and sharpness of this result. For most purposes uniqueness is disposed of by noting that the set $\{S_h : H > h\}$ of scenes with entropy greater than h is, by well-known properties of entropy, strictly convex. A sufficient (stronger than necessary) condition for uniqueness of the maximum-entropy point is then that the set C picked out by our constraints be convex, as is evidently the case for the constraints (21).

Our solution point is then a point of tangency of the set C with one of the sets S_h.

The sharpness of the result is indicated by the Entropy Concentration Theorem[18] presented recently in some detail; suffice it to say here that if we count the scenes by their multiplicities, then not only is the scene of maximum entropy favored over all others, for large N the overwhelming majority of all possible scenes have entropy very close to the maximum. For example, asymptotically, 99% of all scenes in class C have entropy in the range

$$H_{max} - \Delta H \leq H(\text{scene}) \leq H_{max} \tag{25}$$

where $H = \chi^2_k(0.01)/2N$, and $\chi^2_k(q)$ is the critical Chi-squared statistic at the 100q%significance level, for $k = n - m - 1$ degrees of freedom.

The analytical solution of this maximization problem is well known; define the partition function

$$Z(\lambda_1 \ldots \lambda_m) \equiv \sum_i \exp(-\lambda_1 A_{1i} \ldots - \lambda_m A_{mi}) \quad . \tag{26}$$

Then the maximum-entropy scene is given by

$$\hat{x}_i = Z^{-1} \exp(-\lambda_1 A_{1i} - \ldots - \lambda_m A_{mi}) \tag{27}$$

in which the λ's (Lagrange multipliers in the constrained maximization) are chosen to fit the data (21).

To give further details here would duplicate what is in the presentations of J. Shore and J. Skilling at this Symposium. Another recent application to crystallographic inversion is given by Wilkins, et al.[19] We close with the observation that (26) is nothing but a generalized Gibbsian canonical distribution, which has been the basis of Statistical Mechanics for some 60 years. From the inference point of view, therefore, Statistical Mechanics was, historically, the first example of a pure generalized inverse problem in which the ambiguity was resolved by entropy maximization. There was no necessary connection with thermodynamics; but unfortunately, the generality of Gibbs' method was concealed by attempts to put frequentist interpretations on it, and it is only in very recent years that we have realized how much we still had to learn from Gibbs. The overwhelming "preference of Nature" for scenes of high entropy indicated by the Entropy Concentration Theorem is just what we have been calling the "Second Law of Thermodynamics" for a Century. The maximum-entropy methods expounded at this Symposium are only new applications of the Second Law, generalized beyond its original domain.

BIBLIOGRAPHY

1. J. Hadamard, Lectures on Cauchy's Problem, Yale University Press, New Haven, 1923.

2. J. Bertrand, Calcul des probabilités, Gauthier-Villars, Paris, 1889.

3. G. Polya, Mathematics and Plausible Reasoning, 2 Vols., Princeton University Press, Princeton, 1954.

4. A. N. Tikhonov & V. Y. Arsenin, Solutions of Ill-Posed Problems, V. H. Winston & Sons, Washington, D.C., 1977

5. L. J. Savage, The Foundations of Statistics, J. Wiley & Sons, Inc., New York, 1954.

6. I. J. Good, Probability and the Weighing of Evidence, C. Griffin & Co. Ltd., London, 1950.

7. H. Jeffreys, Theory of Probability, Oxford University Press, 1939.

8. B. de Finetti, "Prevision: its Logical Laws, its Subjective Sources", Translated from the French in H. E. Kyburg & H. I. Smokler, Studies in Subjective Probability, 2nd Edition, J. Wiley & Sons, Inc., New York, 1981.

9. E. T. Jaynes, "Where do we Stand on Maximum Entropy?", in R. D. Levine & M. Tribus, Eds., The Maximum Entropy Formalism, MIT Press, Cambridge, MA, 1978.

10. E. T. Jaynes, "The Well-Posed Problem", Foundations of Physics 3, (1973), 477-492.

11. B. de Finetti, The Theory of Probability, 2 vols., J. Wiley & Sons, Inc., New York, 1974.

12. E. Wolf & C. L. Mehta, "Determination of Statistical Properties of Light from Photoelectric Measurements", Phys. Rev. Lett. 13 (1964), 705-707.

13. E. T. Jaynes, "Prior Probabilities", IEEE Trans. Syst. Sci. Cybern, SSC-4 (1968), 227-241.

14. E. T. Jaynes, "Marginalization and Prior Probabilities", in A. Zellner, Ed., Bayesian Analysis in Econometrics and Statistics, North-Holland Publishing Co., Amsterdam, 1980, pp. 43-78.

15. R. B. Litterman, "A Bayesian Procedure for Forecasting with Vector Autoregression", Ph.D. Thesis, University of Minnesota, 1979.

16. E. T. Jaynes, "Highly Informative Priors", Proceedings of the Second International Meeting on Bayesian Statistics, Valencia, University of Valencia Press (in press).

17. J. P. Burg, "Maximum Entropy Spectrum Analysis", Ph.D. Thesis, Stanford University, 1975.

18. E. T. Jaynes, "On the Rationale of Maximum Entropy Methods", Proc. IEEE, 70 (1982), 939-982.

19. S. W. Wilkins, J. N. Varghese, and M. S. Lehmann, "Statistical Geometry: A Self-Consistent Approach to the Crystallographic Inversion Problem Based on Information Theory", Acta Cryst. A39 (1983), 47-60.