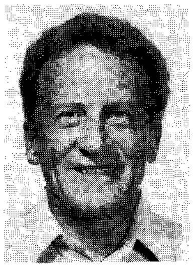
Maximum entropy and the generalized inverse problem

Entropie maximale et problème inverse généralisé



Walter T. GRANDY Jr.

Department of Physics and Astronomy, University of Wyoming, LARAMIE, Wyoming 82070, USA

For many years Professor Grandy has been engaged in theoretical research in the fields of statistical mechanics and electrodynamics, and in the latter area is author of the monograph "Introduction to Electrodynamics and Radiation". A particular interest in the maximum-entropy based foundations of statistical mechanics has led to recent interest in generalized inverse problems. Currently he is involved in a study of alternative theories of elementary particles, and in some problems of biophysics.

He frequently is a visiting professor at the University of Tubingen in Germany and the University of Sao Paulo in Brasil.

SUMMARY

Real-World inverse problems generally consist of two major but equally important parts. In all but the simplest cases incomplete and noisy data are the rule rather than the exception, and here it is emphasized strongly that careful analysis of such data must take place prior to employment of powerful mathematical techniques of inversion. It is argued that the quintessential element in solving most nontrivial inverse problems is one of inference, not mathematical deduction, though the latter remains an important component.

KEY WORDS

Maximum entropy, inverse problems.

RÉSUMÉ

Les problèmes inverses du monde physique consistent en deux parties d'importance égale. A l'exception des cas les plus simples, les données incomplètes et bruyantes sont la règle, et dans cet exposé on souligne vigoureusement que l'analyse attentive de ces données doit avoir la préséance sur l'emploi de techniques mathématiques puissantes d'inversion. Il est soutenu que l'élément clé pour résoudre la plupart des problèmes inverses importants est celui d'inférence et non pas la déduction mathématique, bien que celle-ci en reste une partie constituante importante.

MOTS CLÉS

Entropie maximale, problèmes inverse.

CONTENTS

1. Introduction
2. Inverting Dice Data
3. Generalization
4. Summary
Appendix
References

1. Introduction

In abstract form the quintessence of the inverse problem is conveyed by the expression:

$$(1) \qquad u = \mathrm{KU}\,(+e).$$

The direct problem is, given a state U and operator K, determine u. Often one must cope with errors, or noise, which we denote here by e, and this usually complicates matters considerably. Much deeper is the inverse question: given u and a specific K, what is the true state U? If K should also be a functional of U the problem becomes arbitrarily nonlinear, as it does if we wish to associate K with the human brain, say. In order to focus on the essential features of inverse problems we shall discuss here only the pure (noiseless) linear inverse problem, for which we set $e=0$, although a few remarks about noise will be made later.

A more explicit example of equation (1) is provided by the linear Fredholm integral equation of the first kind:

$$(2) \qquad u(x) = \int_a^b \mathrm{K}(x, y)\,\mathrm{U}(y)\,dy,$$

or in a more familiar discrete form:

$$(3) \qquad u_i = \sum_j \mathrm{K}_{ij}\,\mathrm{U}_j.$$

It will be useful in the sequel to focus on this last form as a specific example, because it also encompasses the well-known problem of matrix inversion in n dimensions ($i=1, 2, \ldots, n, j=1, 2, \ldots, n$), and that is a well-defined mathematical problem. If K^{-1} exists the problem is solved. But if K is singular, or is not square, then what? For example, if $j=1, \ldots, m \ll n$, how do we proceed? Of course, this is just what happens in some image-reconstruction problems.

Thesis: The most common scenario of inversion presents not so much a mathematical problem as one of inference, for only in the simplest of situations is there a unique solution, if any exist at all.

As an illustration of this position we note that most physicians make their livings by solving inverse problems. But often it is found that numerous diseases fit the same set of symptoms, thereby complicating the problem significantly. Sometimes, however, a knowledge of the patient's medical history can reduce the possibilities to only a few.

2. Inverting Dice Data

Consider now a much simpler example: throw an ordinary die a *large* number of times n, and count the number of spots up. For an honest die we would expect the mean number up to be close to 3.5, but suppose our observation yields instead:

$$(4) \qquad 4.5 = \sum_{i=1}^{6} i f_i,$$

where the frequencies are defined as n_i/n. If asked to provide an estimate of the set of frequencies which would most likely yield this number, it seems a very difficult problem. There does not seem to be enough information available, for it is possible that many sets of f_i can be found that will fit the single datum of equation (4). Which set is to be considered correct?

Although this is clearly a problem of incomplete information, it is just as clear that there is a good deal more information available in the statement of the problem. Because each throw of the die, or trial, is independent of all the others, it follows that of the 6^n possible outcomes the number yielding a particular set of frequencies is just the multinomial coefficient:

$$(5) \qquad \mathrm{W} = \frac{n!}{(nf_1)!\ldots(nf_6)!}.$$

This is a multiplicity factor, and we need only find the set of frequencies maximizing it in order to find the set that can be realized in the greatest number of ways. It is an equivalent procedure to maximize log W, so that for very large n we can employ Stirling's formula and obtain:

$$(6) \qquad \mathrm{H} = n^{-1} \log \mathrm{W} = -\sum_i f_i \log f_i.$$

That is, in asking for the set of frequencies that can be realized in the greatest number of ways, we have simply reformulated the problem of maximum entropy. But, rather than an appeal to a particular theorem, we have only an application of common sense! The relation to maximum entropy in the present context provides a frequency correspondence, but the same analysis can be carried out in terms of probabilities, and we demonstrate this in the Appendix.

What we have illustrated with the dice problem is that for pure inverse problems of the type (3), in which the states can be interpreted as frequencies and n is very large, we now have the optimium solution. Indeed, the solution was spelled out long ago by Boltzmann and Gibbs. Given data of the form:

$$(7)\qquad u_i=\sum_{j=1}^{M} K_{ij}\, f_j, \qquad 1 \leqq i \leqq m < M,$$

the set of f_i which can be realized in the overwhelmingly greatest number of ways is that which maximizes the entropy (6) subject to the constraints (7). The frequencies are:

$$(8)\qquad f_j = Z^{-1} \exp\,[-\sum_i \lambda_i K_{ij}],$$

where:

$$Z=\sum_j \exp\,[-\sum_i \lambda_i K_{ij}],$$

is called the *partition function*. The Lagrange multipliers λ_j are found by substituting the f_j of (8) into the equations of constraint (7), which yields a set of coupled differential equations. In addition to the above scenario of a loaded die, the class of problems solved in an optimal way by the above procedure includes statistical mechanics, time-series analysis, and image reconstruction. In the latter case the indices refer to a square array of pixels and K_{ij} is a point-spread function. Note, also, that the maximum-entropy construction provides a completely justifiable means for interpolation and extrapolation of missing data. As an aside, one finds that the set of frequencies solving the dice problem is [1]:

$$(10)\quad (f_1, \ldots, f_6) = (0.054,\ 0.079, 0.114, 0.166, 0.240, 0.348).$$

3. Generalization

Let us now return to the general situation described by equation (1). In most scientific problems of this kind the set of equations determined by $u=KU$ is very much underdetermined, K is singular, and there is no unique solution. The problem is simply not well enough defined, and without further information there can be no definite solution. (Note that the ambiguities at issue here are not merely those of space attenuation and lack of phase information in scattering, say. More immediate and severe is the real-world situation in which there is a genuine paucity of data.) Put another way, the specified data can only determine the class C of possible solutions, but which single solution is to be chosen within that class? Mathematically, incomplete information forces one to conclude that the best we can do is obtain an estimate:

$$(11)\qquad U^* = R\, u,$$

where the resolvent R is to be determined. At the very least R must be chosen so that U* lies in the class C of those possible states which *could* have produced the data u. The mathematical statement of this requirement is that KRK=K, meaning that R is a generalized inverse. In the case of matrices, at least, R is then guaranteed to exist by means of a theorem of Penrose [2].

Therefore, just as with the case of multiplicity factors in the dice problem, one always needs some sort of prior information in order to contract the range of choices C and indeed make an optimal choice. Of course, it may happen that no further information is forthcoming, in which case no one choice within C is any better than any other. But this is an extraordinarily rare occurrence in serious problems.

At this point it is useful to expand our general example (3) to include linear noise:

$$(12)\qquad u_i=\sum_j K_{ij}\, U_j + e_i,$$

in which the data points are distorted away from sharp values in an essentially random way. That is, the class C of possible solutions is enlarged and its boundaries have become somewhat uncertain, or fuzzy. One still needs prior information in order to make any progress, but now the problem has actually changed in a qualitative way. The element of randomness now appears jointly with that of incomplete information, and one must invoke a broader variational principle so as to arrive at an optimal solution. At this time there does not seem to be complete agreement on the specific statement of this larger principle, although definite progress has been made [3]. Further details can be found elsewhere [4], and compared with other approaches [5].

4. Summary

If it is possible to carry out a direct mathematical inversion and obtain a unique result, then by all means one should do so. In most problems of intellectual interest, however, this is not possible, owing to incomplete information, and the problem becomes one of inference. We must then locate all relevant prior information to use as constraints in an optimizing principle for making a choice from the class C. Usually there is a great deal of such prior information available if one will simply look for it in the statement of the problem. For a large class of problems, as we have seen, the principle of maximum entropy leads to a unique solution in the sense of minimum squared error, and it can also be shown that the overwhelming majority of states compatible with the data (3) have entropy very close to the maximum [3].

The presence of noise complicates the problem, of course, precluding a sharp specification of the class C. Because this situation contains elements of both incomplete information and randomness (error), one perceives the need for a complete Bayesian solution. To the author's knowledge, this has yet to be constructed.

Appendix

It is not always possible to interpret data in a frequency context, in which case one must ultimately refer to a probability interpretation. As a first step toward developing such an approach, let us consider the possibility that two variables x and y may be necessary for a complete description of some problem. Suppose, for example, that separate measurements of these two variables are obtained and we represent these numbers as expectation values:

$$\text{(A.1}\,a) \qquad \langle x \rangle = \sum_{i=1}^{I} P(x_i)\, x_i,$$

$$\text{(A.1}\,b) \qquad \langle y \rangle = \sum_{j}^{J} P(y_j)\, y_j,$$

where it is presumed that x and y are capable of taking on values in descrete and exhaustive sets of mutually exclusive alternatives. In this event the total system of interest can be desbribed in part by a joint probability distribution $P(x_i, y_j)$, with normalization:

$$\text{(A.2)} \qquad \sum_{i,\,j} P(x_i, y_j) = \sum_i P(x_i) = \sum_j P(y_j) = 1.$$

For independent variables the distribution factors into the product $P(x_i)\,P(y_j)$, where the single-variable probability distributions are defined as:

$$\text{(A.3}\,a) \qquad P(x_i) = \sum_j P(x_i, y_j),$$

$$\text{(A.3}\,b) \qquad P(y_j) = \sum_i P(x_i, y_j).$$

Up to a positive constant, here chosen as unity, the joint entropy is written:

$$\text{(A.4)} \quad H(x, y) = -\sum_{i,\,j} P(x_i, y_j) \log P(x_i, y_j),$$

and the individual entropies are:

$$\text{(A.5}\,a) \quad H(x) = -\sum_{i,\,j} P(x_i, y_j) \log P(x_i) = -\sum_i P(x_i) \log P(x_i),$$

$$\text{(A.5}\,b) \quad H(y) = -\sum_{i,\,j} P(x_i, y_j) \log P(y_j) = -\sum_j P(y_j) \log P(y_j).$$

Note that $H(x, y) \leqq H(x) + H(y)$, with equality if and only if x and y are independent.

Thus, if x and y constitute independent events, the principle of maximum entropy (PME) with constraints (A.1) yields:

$$\text{(A.6}\,a) \quad P(x_i, y_j) = Z_{12}^{-1} \exp[-\lambda_1 x_i - \lambda_2 y_j] = P(x_i)\,P(y_j),$$

where:

$$\text{(A.6}\,b) \quad Z_{12}(\lambda_1, \lambda_2) = \sum_{i,\,j} \exp[-\lambda_1 x_i - \lambda_2 y_j] = Z_1(\lambda_1)\, Z_2(\lambda_2).$$

It may happen, however, that the given data actually constitute a single datum:

$$\text{(A.7)} \qquad w = A_{11} \langle x \rangle + A_{12} \langle y \rangle,$$

such that:

$$\text{(A.8}\,a) \quad \langle x \rangle = \sum_{i,\,j} P(x_i, y_j)\, x_i = \sum_i P(x_i)\, x_i,$$

$$\text{(A.8}\,b) \quad \langle y \rangle = \sum_{i,\,j} P(x_i, y_j)\, y_j = \sum_j P(y_j)\, y_j.$$

Although equations (A.8) appear equivalent to equation (A.1), this is somewhat deceiving. The latter are *two* pieces of data, whereas here the single datum is given by equation (A.7). There is now only one Lagrange multiplier and maximization of the entropy (A.4) subject to the constraint (A.7) yields:

$$\text{(A.9}\,a) \quad P(x_i, y_j) = Z_{12}^{-1}(\lambda) \exp[-\lambda(A_{11} x_i + A_{12} y_j)],$$

$$\text{(A.9}\,b) \quad Z_{12}(\lambda) = \sum_{i,\,j} \exp[-\lambda(A_{11} x_i + A_{12} y_j)],$$

and λ is determined by:

$$\text{(A.9}\,c) \qquad w = -\frac{\partial}{\partial \lambda} \log Z_{12}(\lambda).$$

It is still true, of course, that $P(x_i, y_j)$ and $Z_{12}(\lambda)$ factor in this case, and H is additive. But we must notice that only the datum is correlated in the two variables, and *not* the variables themselves. A physical realization of this scenario will be presented below.

If there are two pieces of data:

$$\text{(A.10)} \qquad \begin{cases} w_1 = A_{11} \langle x \rangle + A_{12} \langle y \rangle, \\ w_2 = A_{21} \langle x \rangle + A_{22} \langle y \rangle, \end{cases}$$

the same procedure as above will yield immediately the expressions:

$$\text{(A.11}\,a) \quad P(x_i, y_j) = Z_{12}^{-1}(\lambda_1, \lambda_2) \times \exp[-\lambda_1(A_{11} x_i + A_{12} y_j) - \lambda_2(A_{21} x_i + A_{22} y_j)],$$

$$\text{(A.11}\,b) \quad Z_{12}(\lambda_1, \lambda_2) = \sum_{i,\,j} \exp[-\lambda_1(A_{11} x_i + A_{12} y_j - \lambda_2(A_{21} x_i + A_{22} y_j)],$$

$$\text{(A.11}\,c) \qquad w_k = -\frac{\partial}{\partial \lambda_k} \log Z_{12}(\lambda_1, \lambda_2), \qquad k = 1, 2.$$

An obvious generalization is suggested now, in which we have M independent variables x_m that take on discrete values $x_m(i)$, such that $1 \leqq i \leqq \mathrm{I}_{\mathrm{m}}$ and $1 \leqq m \leqq \mathrm{M}$. A set of data is specified by:

$$(\mathrm{A}.12)\quad w_k = \sum_{m=1}^{\mathrm{M}} \mathrm{A}_{km} \langle x_m \rangle, \qquad 1 \leqq k \leqq \mathrm{K} \leqq \mathrm{I}_m,$$

for all m. If we adopt the vector notation $x=(x_1, x_2, \ldots, x_m)$, then the distribution of maximum entropy is:

(A.13 *a*)

$$\mathrm{P}(x) = \mathrm{Z}^{-1} \exp\left[\sum_{k=1}^{\mathrm{K}} \lambda_k \sum_{m=1} \mathrm{A}_{km}\, x_m\right],$$

(A.13 *b*)

$$\mathrm{Z}(\lambda_1, \ldots, \lambda_k) = \sum_{i,\, j} \exp\left[-\sum_{k=1}^{\mathrm{K}} \lambda_k \sum_{m=1}^{\mathrm{M}} \mathrm{A}_{km}\, x_m\right],$$

with:

$$(\mathrm{A}.13\,c)\quad \mathrm{w}_k = -\frac{\partial}{\partial \lambda_k} \log \mathrm{Z}, \qquad 1 \leqq k \leqq < \mathrm{I}_m.$$

Equations (A.13) represent the general solution to a broad class of noiseless, or pure inverse problems. The best estimate that npw can be made for a state of the system, by the criterion of minimum squared error, is just the expectation value:

$$(\mathrm{A}.14)\qquad \langle x_m \rangle = \sum_{i,\, j,\, \ldots} \mathrm{P}(x)\, x_m.$$

But from equation (A.14) we notice immediately the factorization:

$$(\mathrm{A}.15)\quad \mathrm{Z}(\lambda_1, \ldots, \lambda_{\mathrm{k}}) =$$
$$\sum_i \exp\Big[-\sum_k \lambda_k \mathrm{A}_{\mathrm{k}1}\, x_1(i)\Big] \ldots \sum_j \exp\Big[-\sum_k \lambda_k \mathrm{A}_{k\mathrm{M}}\, x_{\mathrm{M}}(j)\Big],$$

and similarly for P(x). Hence, equation (A.14) can be written more explicitly as a single sum:

(A.16 *a*)

$$\langle x_m \rangle = \mathrm{Z}_m^{-1} \sum_i \exp\Big[-\sum_k \lambda_{\mathrm{k}} \mathrm{A}_{km}\, x_m(i)\Big]\, x_m(i),$$

where:

$$(A.16\,b)\quad \mathrm{Z}_m = \sum_j \exp\Big[-\sum_{\mathrm{k}} \lambda_k \mathrm{A}_{km}\, x_m(j)\Big].$$

That is, dependence on all other variables has dropped out and one need only deal with single-variable distributions in any calculation.

This last result might have been expected, owing to the fact that the data contain no information about possible correlations among the variables. But this can also be deceiving, because one needs all the variables in order for the Lagrange multipliers to reproduce the data (A.12). Moreover, equation (A.14) does not really describe a state of the entire system. Rather, one will eventually be interested in quantities of the type $f(x_1, \ldots, x_{\mathrm{M}})$.

A particular advantage of the preceding formulation is that there is no need for $\langle x_m \rangle$ in equation (A.12) to be positive, for it is an expectation value and neither a frequency nor a probability. Further advantages can be adduced by considering the specific example of image reconstruction. Suppose the experiment seeks to measure the actual energy or intensity in a pixel, and our instruments record precisely that. We attempt to measure the energy $x_m(i)$ received in the m-th pixel, and this energy itself can vary over a range of values determined by bandlimits and other factors. The data then take the form (A.12) in terms of expected energies in each pixel, so that A_{km} is indeed a point-spread function. Reconstruction of the image is now a pure probability problem and the expectation values (A.16 *a*) provide the solution. Note that one need not define either n_i or N in this scenario, although it is still necessary to determine the Lagrange multipliers from equation (A.13 *c*).

Manuscrit reçu le 9 *septembre* 1984.

REFERENCES

[1] E. T. Jaynes, in *The Maximum Entropy Formalism*, R. D. Levine and M. Tribus eds., MIT Press, Cambridge, MA, 1978.

[2] R. Penrose, *Proc. Camb. Phil. Soc.*, 51, 1955, p. 406.

[3] E. T. Jaynes, *Proc. IEEE*, 70, 1982, p. 939.

[4] C. R. Smith and W. T. Grandy Jr., *Maximum-Entropy and Bayesian Methods in Inverse Problems*, eds. Reidel, Dordrecht, 1985.

[5] H. P. Baltes ed., *Inverse Scattering in optics*, ed., Springer-Verlag, Berlin, 1980.