General Entropy Criteria for Inverse Problems, with Applications to Data Compression, Pattern Classification, and Cluster Analysis

LEE K. JONES, MEMBER, IEEE, AND CHARLES L. BYRNE, MEMBER, IEEE

Abstract—Minimum distance approaches are considered for the reconstruction of a real function from finitely many linear functional values. An optimal class of distances satisfying an orthogonality condition analogous to that enjoyed by linear projections in Hilbert space is derived. These optimal distances are related to measures of distance between probability distributions recently introduced by Rao and Nayak and possess the geometric properties of cross entropy useful in speech and image compression, pattern classification, and cluster analysis. Several examples from spectrum estimation and image processing are discussed.

I. INTRODUCTION

WE SHALL be concerned with the reconstruction of an unknown real function $R(x)$ from finitely many linear functional (inner product) values; the reconstruction from samples of the Fourier transform of R is a special case. The function $R(x)$ is assumed to be defined for x within some domain Ω of real n-dimensional Euclidean space R^n and to have range contained in a specified open interval Θ. Given that the finite data typically are inadequate to determine $R(x)$ uniquely, prior information plays an important role in such reconstructions; we shall consider the inclusion of a prior estimate $P(x)$ with range in Θ. Minimum distance reconstruction then directs us to accept as our estimate of $R(x)$ that $Q(x) = \tilde{Q}(x)$ consistent with the data for which the directed distance to $P(x)$, given by $D(Q, P)$, is minimum. Our distances are of the form $D(Q, P) = \int f(Q(x), P(x))w(x)\,dx$, with $w(x)$ positive (and with all integrals taken over Ω, unless otherwise stated). The wide class of reconstruction procedures these provide is the subject of this paper.

For particular choices of the functions $f(y, z)$ and $w(x)$, we obtain distances already considered in the literature: the Itakura–Saito distortion measure [1] is one such, leading to Burg entropy [2] when the prior $P(x)$ is constant; for the class of probability densities the cross entropy, or Kullback–Leibler discrimination information is another. The minimization of cross entropy (MCE) approach has been studied extensively by Shore and Johnson, who derive this method from axioms of consistent inference [3] and use it for speech processing and spectrum estimation [4].

Although the maximum entropy method (MEM) of Burg and the MCE approach are based on probabilistic arguments, they find application in the reconstruction of functions that are not essentially probabilistic in nature, such as energy distributions in space, X-ray attenuation functions, and optical images. The properties of the reconstruction, in such cases, are not easily determined from the axioms of probabilistic inference upon which the methods are based. The basic problem is one of approximating an unknown function and the properties of the estimate are best understood when the procedure involves the minimization of a distance.

The MCE and the method of minimizing the Itakura–Saito distance employ (nonsymmetric) measures of distance between functions and obey a principle of directed orthogonality similar to that exhibited by orthogonal linear projections in Hilbert space [5], [6]. In [6], these methods were characterized, among distances of the smooth Ali–Silvey–Csiszár class and the class of regular ratio distortion measures, respectively, as the only methods for which directed orthogonality ever holds. In this article we consider more general classes of distances and investigate those distances for which directed orthogonality holds in all cases.

Our results provide, for a fixed compact domain Ω and a fixed set of linear functionals, necessary and sufficient conditions on the distance D for directed orthogonality to hold in all cases. The class of such distances is easily described and includes the ordinary functional versions of the generalized "cross entropies" introduced in [7]. We show that these distances share other useful geometric properties of cross entropy, making them ideal for pattern recognition, data compression, and cluster analysis.

In Sections II–IV we shall assume that $w(x) = 1$ and all functions P, Q, R considered are positive (so Θ is the open positive half-line) but shall not impose the constraint that the integral be unity: we extend cross entropy to this more general class in an appropriate way. In Section V we

Manuscript received April 19, 1988; revised April 19, 1989. This work was supported in part by the Office of Naval Research under Contract N00014-87-K-0394.

The authors are with the Department of Mathematics, University of Lowell, Lowell, MA 01854.

IEEE Log Number 8932718.

0018-9448/90/0100-0023$01.00

extend these results to the general case of positive $w(x)$ and open interval Θ.

II. Directed Orthogonality, Prior Information and Regular Distortion Measures

The problem is to reconstruct the positive (measurable) function $R(x)$, defined on the measurable set Ω of positive measure, subject to the data constraints

$$r_k = \int R(x) g_k(x)\, dx, \qquad k = 0, 1, \cdots, K \tag{1}$$

where the integral is over Ω, the constraint set $\{g_k(x)\}$ consists of known measurable, locally bounded and linearly independent functions on Ω and equality of functions is taken to be almost everywhere (in practice "piecewise continuous" can replace "measurable" and "rectangle" can replace "measurable set"). In most applications $R(x)$ itself is either a probability density or, at least, a distribution of a quantity (energy, e.g.) the totality of which is known or can be estimated, so we assume that g_0 is constantly unity over Ω, so that $r_0 = \int R(x)\, dx$. Call a measurable function on Ω *admissible* if it is positive, integrable and bounded above and away from zero locally on Ω; let $\boldsymbol{Q}$ be the collection of all admissible functions satisfying (1), and assume that R is a member of $\boldsymbol{Q}$.

We may view $R(x)$ as the unknown input to a signal processing system that records $\{r_k\}$, the set of known features or measurements related to $R(x)$. For the given set of data, we seek, as output of the system, an admissible data-consistent reconstruction, $\tilde{Q}(x)$ in $\boldsymbol{Q}$, that is "optimal" in some appropriate sense. The distance from R to $\tilde{Q}$ will be measured by a distortion (also called input–output discrepancy measure or directed distance from input to output) $D(R, \tilde{Q})$. Our general approach will be to posit a prior admissible estimate of $R(x)$, call it $P(x)$, and then to select as our posterior estimate that $Q = \tilde{Q}$ in $\boldsymbol{Q}$ for which $D(Q, P)$ is minimum. Choosing the distortion D is the subject of this paper.

If we regard the reconstruction process as occurring repeatedly, for various input R and recorded data $\{r_k\}$, we obtain for each set of data an optimal solution. The collection of all these optimal solutions then belongs to a parametrized family or manifold, obtained using the first variation, which we label $\boldsymbol{T}$; the set $\boldsymbol{T}$ which we construct explicitly below, then depends on D, the prior P, and the g_k, but not on the particular values of the r_k. Given a particular set of data $\{r_k\}$, our optimal solution $\tilde{Q}$ is that member of $\boldsymbol{T}$ consistent with the given data; each data set $\{r_k\}$ should determine uniquely a set of parameter values and hence a unique member of $\boldsymbol{T}$.

For arbitrary distortion D the member of $\boldsymbol{T}$ consistent with the data need not be the member of $\boldsymbol{T}$ closest from R; D will be said to obey the directed orthogonality principle if this pathology never occurs:

Definition: Let $\tilde{Q}$ and $\boldsymbol{T}$ be as before. Then D obeys the *directed orthogonality principle* with respect to admissible $\tilde{Q}$ if $D(R, \tilde{Q}) = \min D(R, T)$, the minimum taken over all members T of $\boldsymbol{T}$, and this holds for all admissible R consistent with the data $\{r_k\}$.

Remark: We show in Appendix II that, for a given Ω and constraint set, a distortion of the form $D(Q, P) = \int f(Q(x), P(x))\, dx$ obeys the principle of directed orthogonality in all cases for which an admissible solution exists if and only if, for those same cases, D satisfies the triangle equality: $D(R, P) = D(R, \tilde{Q}) + D(\tilde{Q}, P)$.

If D is the mean square discrepancy, $2D(R, T) = \int (R - T)^2\, dx$, and $\boldsymbol{T}$ is affine (the translation of a linear subspace), then directed orthogonality reduces to ordinary orthogonality in Hilbert space; directed orthogonality implies that the difference $R - \tilde{Q}$ is orthogonal to $\boldsymbol{T}$. Viewing $\boldsymbol{T}$ as the set of all possible reconstructions our method produces, as the data vary, directed orthogonality requires that the data-consistent member of $\boldsymbol{T}$ be the member of $\boldsymbol{T}$ closest from R, for each R. The mapping from input to output is a projection.

Others have also endorsed the principle of directed orthogonality. In speech analysis it is known as the "correlation matching property." There the speech patterns consist of power spectral densities of a Gaussian speech process and the data are certain autocorrelation values. It is advantageous to know that the modeled speech pattern closest to the true one is that consistent with the observed autocorrelation values (see [8], [9]). In the context of probability density reconstruction, the principle of directed orthogonality is called "expected value matching" [9] and is applied to pattern recognition and cluster analysis.

We now turn to the construction of the set $\boldsymbol{T}$ and to a discussion of the possibilities for the distortion measure D. For admissible Q and P let

$$D(Q, P) = \int f(Q(x), P(x))\, dx \tag{2}$$

be the directed distance from Q to P. With P our prior estimate of R, we select as our posterior estimate that $Q = \tilde{Q}$ in $\boldsymbol{Q}$ minimizing $D(Q, P)$. The function P is our best estimate of R prior to observing the data. After we observe the data we move away from P only to the extent to which the data force us.

We assume that the following partial derivatives of $f(y, z)$ exist and are continuous in the first quadrant: $f_y, f_z, f_{yy}, f_{zyy}$. If $Q = \tilde{Q}$ is our optimal admissible solution, then it must satisfy the Euler–Lagrange equations

$$f_y(Q(x), P(x)) = \sum_{k=0}^{K} t_k g_k(x) \tag{3}$$

with $Q = \tilde{Q}$ and some choice of the real constants $t_0 = \tilde{t}_0, \cdots, t_k = \tilde{t}_K$ (for a proof assuming only the continuity of f_y in the arbitrary measurable case see [6]). For every set of data $\{r_k\}$ for which there is an admissible solution $\tilde{Q}$ there will be a choice of parameters for which (3) holds with $Q = \tilde{Q}$. The set of admissible solutions so generated will be a subset of $\boldsymbol{T}$, where $\boldsymbol{T}$ is defined as the set of all admissible solutions to (3). For the case of the mean-square

discrepancy, (3) becomes

$$Q(x) = P(x) + \sum_{k=0}^{K} t_k g_k(x). \qquad (3')$$

making T an affine set; since $R - \tilde{Q}$ has zero inner product with each g_k, directed orthogonality holds.

We now consider additional conditions for $f(y, z)$ so that the distortion D will have the reasonable properties of a directed distance. For $D(Q, P) \geq D(P, P)$ for all Q and P we impose the following condition.

Condition 1: $f_y(y, y) = 0$, for all $y > 0$.

Then the Euler equation becomes $f_y(P(x), P(x)) = 0$ for all admissible priors.

To have $D(P, P) = 0$, we impose another condition.

Condition 2: $f(y, y) = 0$, for all $y > 0$.

Then it follows from Conditions 1 and 2 that $f_z(y, y) = 0$ for all $y > 0$.

To make $D(Q, P)$ strictly convex in Q, so that if an admissible solution $\tilde{Q}$ exists it is unique, we impose Condition 3.

Condition 3: $f_{yy}(y, z) > 0$, for all $y, z > 0$.

In the next section we obtain necessary and sufficient conditions for directed orthogonality; at that time we present a fourth condition on f that guarantees the admissibility of any data-consistent solution of (3).

We rewrite the statement of the principle of directed orthogonality as follows:

$$\arg\min_{T \text{ in } \boldsymbol{T}} D(R, T) = \tilde{Q} = \arg\min_{Q \text{ in } \boldsymbol{Q}} D(Q, P) \qquad (4)$$

where arg min means that value of the variable for which the (unique) minimum is attained. Under our assumption that D is convex in Q, the left side of (4) is equivalent to the Lagrange dual [10] of the convex programming problem on the right side of (4); under our assumptions, directed orthogonality can be viewed as "self duality." The dual problem is locally unconstrained for general convex primals. This has been exploited previously for cross entropy [11] and for Itakura–Saito distortion measures with uniform prior [12].

III. Necessary and Sufficient Conditions for Directed Orthogonality, with Examples of Projective Distortion

The following lemma (the proof of which is given in Appendix I) is basic to our characterization of distortions that satisfy the principle of directed orthogonality in all cases. We say that a distortion D of the form (2) is a *projective distortion* if, for some compact domain Ω of positive measure, some set of constraint functions g_k and all choices of $\{r_k\}$ and admissible P, directed orthogonality holds whenever an admissible solution exists. It will follow from our characterization of projective distortion that directed orthogonality then holds for all other (possibly noncompact) domains Ω and all sets of constraint functions.

Fundamental Lemma: If D is a projective distortion then $f_{zyy}(y, z) = 0$ for every (y, z) in the interior of the first quadrant.

Theorem 1: A distortion is projective if and only if the associated integrand has the form

$$f(y, z) = J(y) - J(z) + (z - y)\, j(z) \qquad (5)$$

where $j(u) = dJ/du$ and dj/du is continuous and positive.

Proof: First, suppose that the distortion is projective, so that, by the lemma, f_{zyy} is identically zero inside the first quadrant. Then f_z is a linear function of y, with coefficients that are functions of z. Therefore, f has the form $f(y, z) = -yj(z) + L(z) + M(y)$. From Condition 1 we have that $dM/dy = j(y)$ and from Condition 2 we have that $L(y) = yj(y) - J(y)$; (5) is satisfied. The rest follows from Condition 3 and the continuity of f_z.

For the remainder of the proof, note that the Euler–Lagrange equation (3) becomes

$$j(Q(x)) = j(P(x)) + \sum_{k=0}^{K} t_k g_k(x). \qquad (3'')$$

Now let $R(x)$ be the true function to be reconstructed, $\{r_k\}$ the data, and $\tilde{Q}(x)$ an admissible solution. For T in $\boldsymbol{T}$ we have

$$\begin{aligned} D(R, T) &= \int_\Omega \big[J(R(x)) - J(T(x)) \\ &\qquad + (T(x) - R(x))\, j(T(x)) \big]\, dx \\ &= \int_\Omega \big[J(\tilde{Q}(x)) - J(T(x)) + (T(x) - \tilde{Q}(x)) \\ &\qquad \times j(T(x)) + J(R(x)) - J(\tilde{Q}(x)) \\ &\qquad + (\tilde{Q}(x) - R(x)) \big(j(P(x)) + \textstyle\sum t_k g_k(x) \big) \big]\, dx \\ &= D(\tilde{Q}, T) + \int_\Omega \big[J(R(x)) - J(\tilde{Q}(x)) \\ &\qquad + (\tilde{Q}(x) - R(x)) \big(j(P(x)) + \textstyle\sum \tilde{t}_k g_k(x) \big) \big]\, dx \\ &= D(R, \tilde{Q}) + D(\tilde{Q}, T), \end{aligned} \qquad (5')$$

which is uniquely minimized at $T = \tilde{Q}$. Note that the positivity of (5) for $y \neq z$ follows from the mean value theorem and the monotonicity of j. Q.E.D.

As an immediate consequence, we have the first part of the following proposition, which is a well-known property of cross entropy [13].

Proposition 1: Any projective distortion D obeys the triangle equality $D(R, P) = D(R, \tilde{Q}) + D(\tilde{Q}, P)$. Conversely, if, for fixed Ω and constraint set, the triangle equality holds whenever an admissible solution exists, then the distortion is projective. (*Note*: if we extend the theory to include distortions that are not ordinary functionals, such as $\sinh(\int f(Q(x), P(x))\, dx)$, D can still be projective, but the triangle equality will not hold.)

Proof: For the first half, rewrite (5'), with $T = P$ and $t_k = 0$. The converse statement is proved in Appendix II. Q.E.D.

Writing the projective distortion as

$$D(Q, P) = \int [J(Q(x)) - J(P(x)) + (P(x) - Q(x)\, j(P(x)))]\, dx,$$

we see that $D(Q, P) = \mathcal{J}(P) - \mathcal{J}(Q) + \delta\mathcal{J}(P; Q - P)$, where $\delta\mathcal{J}(x; h)$ is the Gateaux differential of the abstract concave entropy functional $\mathcal{J}(Q) = -\int J(Q(x))\,dx$. For the case of probability densities these distances are the ordinary functional versions (with uniform weighting) of those considered in [7]. In the case of constant prior $\tilde{Q}$ is the abstract maximum entropy solution, i.e., that admissible function $Q = \tilde{Q}$ maximizing $\mathcal{J}(Q)$ subject to the data constraints. Solving our original problem is equivalent to solving (3'') subject to the data constraints.

For a data-consistent solution of (3'') to be admissible, it is sufficient that $f(y, z)$ satisfy a fourth condition.

Condition 4: The range of $j(y)$, for $y > 0$, is the whole real line.

Then the solution of (3'') is written as

$$\tilde{Q}(x) = j^{-1}\left(j(P(x)) + \sum_{k=0}^{K} t_k g_k(x) \right)$$

and the local boundedness and positivity follow from the local boundedness of $P(x)$ and the $g_k(x)$.

The choice $j(u) = u$ gives the mean square distortion, which frequently leads to nonpositive solutions [14]. We now give other examples of projective distortions, some of which lead to well-known techniques for solving inverse problems.

Example 1: Let $j(u) = \ln(u)$. Condition 4 is satisfied, and we have

$$D(Q, P) = \int [Q(x) \ln(Q(x)/P(x)) + P(x) - Q(x)]\, dx,$$

which we call the Kullback distortion, since it is just the cross entropy or Kullback–Leibler discrimination information for the case of probability densities. That directed orthogonality holds for cross entropy is well known [5], [9]. The set $\boldsymbol{T}$ consists of all admissible functions of the form

$$T(x) = P(x) \exp\left(\sum_{k=0}^{K} t_k g_k(x) \right) \tag{6}$$

and clearly any data consistent function of the form (6) is admissible. Further sufficient conditions for the existence of data-consistent solutions of the form (6) can be found in [13]. If $P(x)$ is constant, then the minimum Kullback distortion method reduces to finding the data consistent function $Q = \tilde{Q}$ with maximum Shannon differential entropy, $-\int Q \ln Q$; solutions in this case are the familiar "maximum Shannon entropy" solutions well known in image processing [15]–[18]. For such applications the g_k can be integrated point spread functions corresponding to the kth pixel, r_k the image intensity in the kth pixel, and r_0 total image intensity. The solution of the dual problem for such cases, which is globally unconstrained, has been studied in [11].

Example 2: Let $j(u) = -1/u$. The distance now becomes the Itakura–Saito distortion measure $D(Q, P) = I(Q, P) = \int [Q/P - \ln(Q/P) - 1]$. The solutions of interest have the form

$$Q(x) = \left(1/P(x) - \sum_{k=0}^{K} t_k g_k(x) \right)^{-1}. \tag{7}$$

Condition 4 fails, and the positivity of (7) is not guaranteed. For constant prior we have the maximum entropy method of Burg [2]. For the case of a discrete stationary random process and one-dimensional power spectrum with autocorrelation values known out to time lag N, Burg's method gives a positive data-consistent estimate $Q = \tilde{Q}$ of the spectrum, which maximizes the entropy $\int \log Q$. That entropy is actually maximized follows from the convexity of $I(Q, P)$. For higher dimensional spectral estimation iterative solutions of the locally unconstrained dual problem (with constant prior) are used [12]. Here the dual problem is equivalent to minimizing

$$\int \ln\left[\left(-\sum t_k g_k(x) \right)^{-1} \right] dx + \sum t_k r_k \tag{7'}$$

where the g_k are multiple cosines and the t_k are subject only to the condition $-\sum t_k g_k(x) > 0$; the constant $j(P)$ is absorbed into the constant term $t_0 g_0$. At each step of the algorithm correlation coefficients are evaluated by numerical quadrature.

Example 3: Let $j(u) = -1 - (1-u) - (1-u)^2 - \cdots - (1-u)^M$, which for large M is approximately $-1/u$. Woods [19] approximates the logarithmic argument in (7') by $-j(-\sum t_k g_k)$. He then uses a dual algorithm with a positivity constraint, which involves the gradient of (7') and requires, at each step, the evaluation of correlation coefficients for a spectrum of the form $-j(-\sum t_k g_k)$. These correlation coefficients are easily computed at each step, via the fast Fourier transform algorithm for the case of trigonometric polynomials. Woods, without stating it, has introduced computational simplicity by using the current $j(u)$ to approximate the j function in Example 2. Again there is no guarantee of solutions. While there is some debate over which entropy functionals to use, we feel that approximations for computational advantage have their place; most likely the best j function will depend on the problem.

Example 4: Let $j(u) = u^{1/m}$, for $m = 2, 3, \cdots$. Then (3) becomes

$$Q(x) = \left[P(x)^{1/m} + \sum t_k g_k(x) \right]^m \tag{8}$$

and

$$D(Q,P)=\int\left[(m/m+1)Q^{1+1/m}-QP^{1/m}\right.$$
$$\left.+(1/m+1)P^{1+1/m}\right]dx. \quad (9)$$

For m even, Q has the advantage of being positive in all cases. The computational advantages are even greater than in Example 3 because $j^{-1}(u)$ is a monomial. Positivity of the bracketed term in (8) is required for the dual problem. Subject to this, we have the simple and computationally attractive problem

$$\text{minimize} -\sum t_k r_k+\int(1/m+1)\left(P^{1/m}+\sum t_k g_k\right)^{m+1}dx. \quad (10)$$

We compare reconstructions of the object shown in Fig. 1 (which is also discussed in [20]). We use $g_k=\cos(kx)$, $\Omega=[0,\pi]$, $K=20$ and $P(x)=\pi$. Fig. 2 shows the MEM reconstruction (i.e., $j(v)=-1/v$). Fig. 3 shows the mean-square reconstruction ($j(v)=v$). Finally, Fig. 4 gives the reconstruction for $j(v)=v^{1/9}$. This was computed using a simple convex programming algorithm with a data consistency tolerance of ± 0.02. It is clear from the figures that the new method (with $j(v)=v^{1/9}$) gives better support and scale estimates that the mean square case and does not suffer from the spurious peaks of the MEM; we have some indication, then, that the principle of directed orthogonality may lead to superior new reconstruction methods. It is shown in [20] that resolution superior to that of MEM can feasibly be achieved by iterating the distance of Example 4. Our examples have not been symmetric distortions; in fact, this is not surprising, considering Theorem 2.

Theorem 2: If a projective distortion is symmetric in Q and P, then it is a multiple of the mean-square distortion.

Proof: For any compact domain Ω we can choose Q and P to be constant functions, so the symmetry of

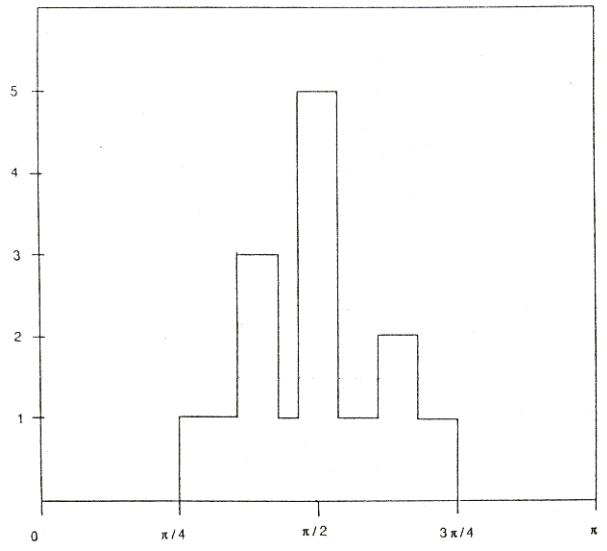

Fig. 1. Object to be reconstructed.

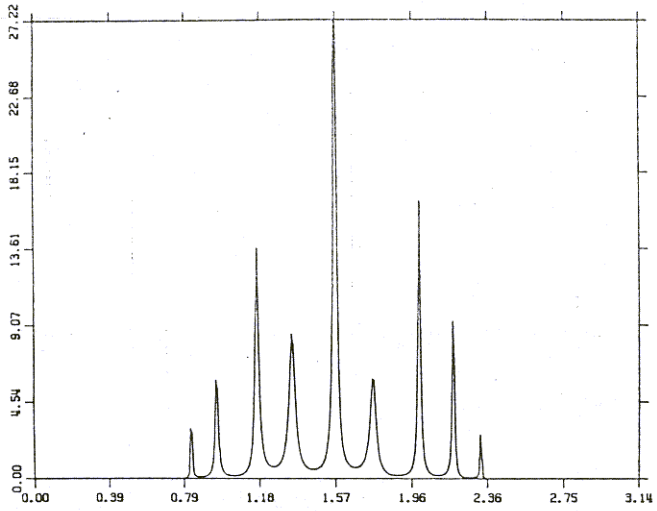

Fig. 2. Maximum entropy reconstruction.

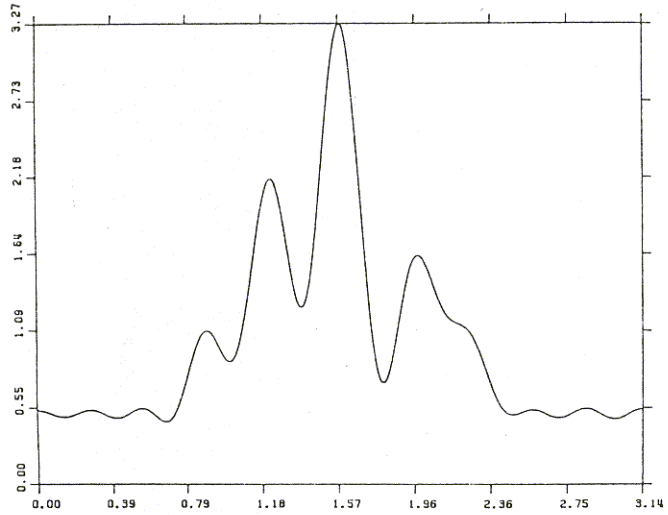

Fig. 3. Mean square reconstruction.

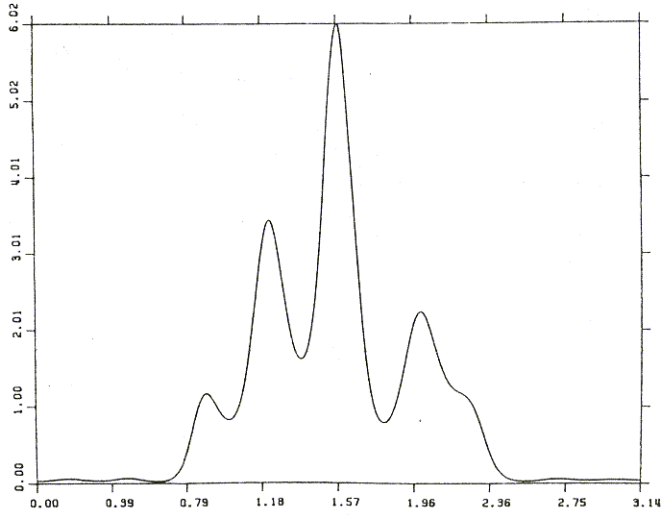

Fig. 4. Proposed reconstruction.

$D(Q, P)$ implies that $f(y, z) = f(z, y)$ for all y and z in the interior of the first quadrant. From (5) it follows that $J(y) - J(z) = (y - z)(j(y) + j(z))/2$. Differentiating with respect to y yields, at a fixed $y = y_0$, $j(y_0) = (j(y_0) + j(z))/2 + (y_0 - z)j'(y_0)/2$, which means that $j(z)$ is linear in z. The assertion then follows. Q.E.D.

IV. Geometric Consequences of Directed Orthogonality, with Applications to Data Compression, Pattern Recognition, and Cluster Analysis

Cross entropy obeys a principle of iterated information [9], as a consequence of directed orthogonality; consequently, this principle holds for general projective distortions. Let D be any projective distortion and let P be a fixed admissible prior function. Values $r_k, k = 0, \cdots, K$, are given and later revised to $\tilde{r}_k$. Let $Q = \tilde{Q}_1$ be the admissible function consistent with the first set of data that minimizes $D(Q, P)$, and let $Q = \tilde{Q}_2$ be the admissible function consistent with the revised set of values that minimizes $D(Q, P)$. Let $Q = \tilde{Q}_3$ be the admissible function consistent with the revised set of values that minimizes $D(Q, \tilde{Q}_1)$.

Theorem 3: With notation as described, $\tilde{Q}_3 = \tilde{Q}_2$. For any Q consistent with the revised set of values and for the parameters $\tilde{t}_0, \cdots, \tilde{t}_K$ for which the equation $j(\tilde{Q}_1) = j(P) + \sum \tilde{t}_k g_k(x)$ holds, we have the equality

$$D(Q, P) = D(Q, \tilde{Q}_1) + D(\tilde{Q}_1, P) + \sum \tilde{t}_k (\tilde{r}_k - r_k). \quad (11)$$

Proof: With Q as above we have

$$D(Q, P) = \int \big[J(Q) - J(\tilde{Q}_1) + J(\tilde{Q}_1) - J(P) + (\tilde{Q}_1 - Q)\big(j(P) + \sum \tilde{t}_k g_k\big) + \big(Q - \tilde{Q}_1\big) \sum \tilde{t}_k g_k + \big(P - \tilde{Q}_1\big) j(P) \big] \, dx$$

which simplifies to (11). That $\tilde{Q}_3 = \tilde{Q}_2$ follows from (11) by noticing that minimizing $D(Q, P)$, subject to the $\{\tilde{r}_k\}$ data constraints, is equivalent to minimizing $D(Q, \tilde{Q}_1)$ subject to the $\{\tilde{r}_k\}$ data constraints. Q.E.D.

We now turn to applications that have been discussed previously for cross entropy [9]: a) pattern classification, compression, and quantization; b) cluster analysis.

A. Pattern Classification, Compression, and Quantization

We consider now three applications of minimum distance methods for which the property of directed orthogonality is particularly useful. A lookup table approach to reconstruction can be an efficient and effective way to glean information from observed data. In this approach one has a "training set" of system inputs $R_m(x)$, $m = 1, \cdots, M$, with associated linear functional values $r_{mk} = \int R_m(x) g_k(x) \, dx$, $k = 0, \cdots, K$. Let us assume that we have a fixed common prior estimate for the R_m, call it $P(x)$, and that $Q = \tilde{Q}_m(x)$ is that function consistent with the data $r_{mk}, k = 0, 1, \cdots, K$, minimizing $D(Q, P)$. We are given information $r_k = \int R(x) g_k(x)$, $k = 0, \cdots, K$, about an unknown $R(x)$, from which we calculate the function $Q = \tilde{Q}$ consistent with these data and minimizing $D(Q, P)$, and we may wish to do the following.

1) Find that $m = \hat{m}$ minimizing $D(R, \tilde{Q}_m)$, so that we may associate with the unknown R other attributes known of the $R_{\hat{m}}$ or $\tilde{Q}_{\hat{m}}$. This is the *nearest neighbor rule* of pattern recognition using the (typically nonsymmetric) distortion D. There is substantial research on optimal metrics for this rule [21] and some evidence of practical advantage to the use of the (nonmetric) cross entropy [22].

2) Approximate $R(x)$ by the member of the family $\{\tilde{Q}_m\}$ minimizing the distortion $D(R, \tilde{Q}_m)$. This is a *quantized estimate* of R.

3) Compress the data $\{r_k\}$ (and hence R or $\tilde{Q}$) by transmitting only the value $m = \hat{m}$ of the $\tilde{Q}_m$ closest to R and using $\tilde{Q}_{\hat{m}}$ later as the decompressed estimate of R.

Each of these tasks is facilitated by the fact that the following quantized triangle equality holds: $D(R, \tilde{Q}_m) = D(R, \tilde{Q}) + D(\tilde{Q}, \tilde{Q}_m)$ for $m = 1, \cdots, M$; this follows by setting $T = \tilde{Q}_m$ in (5′) of the proof of Theorem 1, or by direct application of Theorem 3 and Proposition 1. It follows that minimizing $D(R, \tilde{Q}_m)$ is equivalent to minimizing

$$D(\tilde{Q}, \tilde{Q}_m) = \int \big[J(\tilde{Q}) - J(\tilde{Q}_m) + (\tilde{Q}_m - \tilde{Q})\big(j(P) + \sum t_{mk} g_k\big) \big] \, dx,$$

which is equivalent to minimizing $\int [\tilde{Q}_m(x) j(P(x)) - J(\tilde{Q}_m(x))] \, dx + \sum t_{mk}(r_{mk} - r_k)$, where the t_{mk} are the constants for which $j(\tilde{Q}_m) = j(P) + \sum t_{mk} g_k$. The terms $\int [\tilde{Q}_m(x) j(P(x)) - J(Q_m(x))] \, dx$ are data independent and can be calculated in advance. The result is that the optimal rule for finding the index $\hat{m}$ of the best approximation is linear in the data $\{r_k\}$.

B. Cluster Analysis

In the above applications and others as well, the prototypes $\tilde{Q}_m$, $\{r_{mk}\}$ have to be chosen from a much larger class of possibilities by some procedure that represents a cluster of functions by a single function. Suppose we have a large class $\{Q_n\}$, $n = 1, \cdots, N$, of solutions to (3), for various data sets $\{r_{nk}\}$. We want to determine a best center Q_0 for this cluster of functions and, simultaneously, a best set of data $\{r_{0k}\}$ to represent the various $\{r_{nk}\}$; that is, we seek $Q = Q_0$ so as to minimize

$$N^{-1} \sum_{n=1}^{N} D(Q_n, Q)$$

and its corresponding linear functional values, $r_{0k} = \int Q_0(x) g_k(x) \, dx$ where Q_0 has the form (3).

Theorem 4: For a projective distortion D we have $r_{0k} = N^{-1} \sum_{n=1}^{N} r_{nk} = \bar{r}_k$ and $Q_0 = \overline{Q}$, the function consistent with the values $\bar{r}_k$ that minimizes $D(Q, P)$, as a function of Q.

Proof: For any admissible Q satisfying (3) we have

$$\begin{aligned} N^{-1}\sum D(Q_n, Q) &= N^{-1}\sum \int \big[J(Q_n) - J(\bar{Q}) + J(\bar{Q}) - J(Q) \\ &\quad + (\bar{Q} - Q_n + Q - \bar{Q})\big(j(P) + \textstyle\sum t_k g_k\big)\big]\, dx \\ &= N^{-1}\sum \int \big[J(Q_n) - J(\bar{Q}) + (\bar{Q} - Q_n) j(P)\big]\, dx \\ &\quad + D(\bar{Q}, Q) \\ &= D(\bar{Q}, Q) + N^{-1}\sum \int \big[J(Q_n) - J(P) + J(P) \\ &\quad - J(\bar{Q}) + (P - Q_n + \bar{Q} - P) j(P)\big]\, dx \\ &= D(\bar{Q}, Q) - D(\bar{Q}, P) + N^{-1}\sum D(Q_n, P), \end{aligned}$$

which is minimized uniquely in Q for $Q = \bar{Q}$. It follows that $Q_0 = \bar{Q}$. Q.E.D.

V. Extension to Arbitrary Θ and $w(x)$

The foregoing results extend straightforwardly when a positive $w(x)$ is used and the functions R, Q, P are constrained to have range in a fixed open interval Θ: the definition of *admissibility* must be altered, by requiring that P, Q, and R take values locally in a compact subset of Θ. The function $f(y, z)$ is now defined on $\Theta \times \Theta$, and $j(v)$ is defined for v in Θ. In all proofs, the integration of f is with respect to $w(x)\, dx$, instead of an ordinary Lebesgue measure. The factor $1/w(x)$ appears on the right side of (3) and in the Appendices. The projective distortions now take the form

$$D(Q, P) = \int \big[J(Q(x)) - J(P(x)) + (P(x) - Q(x))\, j(P(x))\big] w(x)\, dx.$$

The only symmetric projective D is now the weighted mean square discrepancy. Finally, we note that the Itakura–Saito and Kullback distortions may not be used with arbitrary Θ but that other projective distortions may be introduced for those cases: for instance, use $j(v) = v^{1/m}$ for $m = 3, 5, 7, \cdots$ and $\Theta = (-\infty, \infty)$.

VI. Summary

We have considered the problem of reconstructing a real function $R(x)$ from finitely many linear functional values. In particular, we have concentrated on the use of distortion measures of the form $D(Q, P) = \int f(Q(x), P(x)) w(x)\, dx$, where $P(x)$ is a prior estimate of $R(x)$ and the posterior estimate is taken to be that Q consistent with the data minimizing the above distance to $P(x)$. For positive functions a commonly used measure of distance is cross entropy, obtained from $f(y, z) = y \ln(y/z) + (z - y)$ and weight $w(x) = 1$. The approximation by minimization of cross entropy is known to obey a directed orthogonality principle analogous to that exhibited by orthogonal projection in Hilbert space. In this paper we characterize those functions $f(y, z)$ for which this directed orthogonality always holds and show that, as a result of directed orthogonality, a number of the useful approximation-theoretic properties of cross entropy carry over to the more general class.

By using directed orthogonality as the fundamental property of our general class of distances, we obtain methods for information-theoretic inference from limited data based on approximation-theoretic rather than probabilistic assumptions.

We relate the methods derived here to others discussed in the spectrum estimation and optics literature and suggest avenues for further investigation. An example is presented showing that some of the new distances may be superior to the commonly used ones for the generation of reconstruction algorithms. Directed orthogonality is shown to lead to measures of discrepancy that are useful in pattern classification, data compression, and cluster analysis.

Appendix 1
Proof of the Fundamental Lemma

A) Differentiating formally with respect to t_k in (3), we have, for T in $\boldsymbol{T}$,

$$\partial T(x)/\partial t_k = g_k(x)/f_{yy}(T(x), P(x)). \tag{A1}$$

Let $\tilde{Q}$ be an admissible solution for some admissible R and P. Fix x in the compact domain Ω, and consider (A1) as a partial differential equation for $T(x; t)$, viewed as a function of $t = (t_0, \cdots, t_K)$ near $t = \tilde{t} = (\tilde{t}_0, \cdots, \tilde{t}_K)$, the parameters for $\tilde{Q}$. We rewrite (A1) as

$$\partial T/\partial t_k = g_k/h(T) \tag{A2}$$

where $h(y) = h(y; x) = f_{yy}(y, P(x))$. Let $G(y) = G(y; x)$ be the (increasing) antiderivative of the function $h(y)$ having the "constant of integration" such that

$$G(\tilde{Q}(x)) = \sum_{k=0}^{K} \tilde{t}_k g_k(x). \tag{A3}$$

Repeat this procedure for all x in Ω. Using the positivity and continuity of f_{yy} and the local boundedness of g_k and P, we find that

$$T(x; t) = G^{-1}\left(\sum_{k=0}^{K} t_k g_k(x)\right)$$

solves the partial differential equation (A1), equals $\tilde{Q}(x)$ for $t = \tilde{t}$, and in a neighborhood of $\tilde{t}$ is bounded above and away from zero, uniformly in x. Hence $T(x; t)$ is admissible in a neighborhood of $\tilde{t}$.

b) By the hypothesis of the lemma, directed orthogonality holds for any admissible R consistent with $\{r_k\}$, provided $\tilde{Q}$ is an admissible solution. Hence, by the compactness of Ω, the discussion in a), and the bounded convergence theorem (so we may differentiate inside the integral),

$$\begin{aligned} &\partial D(R, T(x; t))/\partial t_0\big|_{t=\tilde{t}} \\ &\quad = \int \big[f_z(R(x), \tilde{Q}(x))/f_{yy}(\tilde{Q}(x), P(x))\big]\, dx = 0. \end{aligned} \tag{A4}$$

c) If the assertion of the lemma were false, then there would be (y_0, z_0) in the interior of the first quadrant with $f_{zyy}(y_0, z_0) \neq 0$.

By continuity $f_{zyy}(y, z_0)$ is of one sign in a neighborhood N of y_0. Take $P(x)$ to be identically z_0 and $r_k = \int P(x) g_k(x)$. Then $Q = P$ is admissible and (A4) becomes

$$\int f_z(R(x), z_0)\, dx = 0 \tag{A5}$$

for any admissible R consistent with $\{r_k\}$. Let x_0 be a point of Lebesgue density one in Ω and consider

$$R(x) = P(x) + (y_0 - z_0) \cdot \left(\exp\left(-m(x - x_0)^T (x - x_0) \right) + \sum_{k=0}^{K} \beta_k g_k(x) \right). \tag{A6}$$

We may assume without loss of generality that the g_k have been orthogonalized in Ω. Hence for each m the β_k may be chosen so that the $R(x)$ is consistent with the data $\{r_k\}$ and the β_k converge to 0 as m increases. So, for sufficiently large m, $R(x)$ is admissible and $R(x_0)$ is in the neighborhood N. Because x_0 is a point of density one, we could also ensure, for large enough m, the existence of a closed subset S of Ω of positive measure such that $R(x)$ is in N for x in S.

d) Use $R(x) + \epsilon\tau(x)$ in (A5) as the unknown input function, where $\tau(x)$ is not identically zero, has support contained in S, and is orthogonal to all the g_k. For small enough $\epsilon > 0$, the function $R + \epsilon\tau$ is admissible. Differentiating (A5) twice with respect to ϵ at $\epsilon = 0$ yields

$$\int f_{zyy}(R(x), z_0)(\tau(x))^2\, dx = 0,$$

which is impossible because $f_{zyy}(y, z_0)$ is of one sign in N.

Appendix II
Equivalence of Directed Orthogonality and the Triangle Equality

We assume that for some fixed domain Ω and constraint set $\{g_k\}$, we have the triangle equality $D(R, P) = D(R, \tilde{Q}) + D(\tilde{Q}, P)$ whenever the problem has an admissible solution; that is,

$$\int f(R(x), P(x))\, dx - \int f(R(x), \tilde{Q}(x))\, dx = \int f(\tilde{Q}(x), P(x))\, dx$$

is independent of R and depends only on the data. For any $\bar{w}$ and $\bar{z}$ positive constants, take $P(x) = \bar{z}$ and $r_k = \bar{w} \int g_k\, dx$, $k = 0, \cdots, K$. Then $\tilde{Q}(x) = \bar{w}$ is the unique admissible solution: for any admissible $Q(x) \neq \tilde{Q}(x)$ we have

$$D(\tilde{Q}, P) = m(\Omega) f(\bar{w}, \bar{z}) < m(\Omega) \int f(Q(x), \bar{z})\, m(\Omega)^{-1}\, dx = D(Q, P)$$

by the strict convexity of f in the first variable and Jensen's inequality. Now it follows that for data-consistent R, $\int [f(R(x), \bar{z}) - f(R(x), \bar{w})]\, dx$ is a function only of $\bar{w}$ and $\bar{z}$. Applying the same techniques used in parts c) and d) of Appendix I, we can show that $f_{yy}(y, \bar{z}) = f_{yy}(y, \bar{w})$ for all y. It follows that $f_{yy}(y, z)$ is independent of z, so that f_{yyz} is identically zero. Proceeding as before, we show that $f(y, z)$ has the form (5).

References

[1] F. Itakura and S. Saito "Analysis synthesis telephony based on the maximum likelihood method," in *Proc. 6th. Int. Conf. Acoustics*, Tokyo, Japan, 1968, pp. C-17–C20.

[2] J. P. Burg, "Maximum entropy spectral analysis," in *Proc. 37th Meeting Soc. of Exploration Geophysicists*, Oklahoma City, OK, Oct. 1967.

[3] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*, vol. IT-26, no. 1, pp. 26–37, Jan. 1980.

[4] J. Shore, "Minimum cross-entropy spectral analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, no. 2, pp. 230–237, Apr. 1981.

[5] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[6] L. K. Jones, "Approximation theoretic derivation of logarithmic entropy principles for inverse problems and unique extension of the maximum entropy method to incorporate prior knowledge," *SIAM J. Appl. Math.*, vol. 49, pp. 650–661, Apr. 1989.

[7] C. R. Rao and T. K. Nayak, "Cross entropy, dissimilarity measures, and characterizations of quadratic entropy," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 5, Sept. 1985.

[8] R. Gray, A. Buzo, A. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 367–376, Aug. 1980.

[9] J. Shore and R. Gray, "Minimum cross-entropy pattern classification and cluster analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-4, no. 1, pp. 11–17, Jan. 1982.

[10] D. G. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.

[11] A. Charnes, W. W. Cooper, and L. Seiford, "Extremal principles and optimization qualities for Khinchin-Kullback-Leibler estimation," *Math. Operationsforsch, Statist.*, vol. 9, pp. 21–29, 1978.

[12] J. McClellan, "Multidimensional spectral estimation," *Proc. IEEE*, vol. 70, pp. 1029–1039, Sept. 1982.

[13] I. Csiszár, "*I*-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, pp. 146–168.

[14] R. Goodrich and A. Steinhardt, "L_2 spectral estimation," *SIAM J. Appl. Math.*, vol. 40, no. 2, pp. 417–426, June 1986.

[15] B. R. Frieden, "Restoring with maximum likelihood and maximum entropy," *J. Opt. Soc. Amer.*, vol. 62, pp. 511–518, Apr. 1972.

[16] R. Gordon and G. T. Herman, "Reconstruction of pictures from their projections," *Quart. Bull. Center for Theor. Biol.*, vol. 4, pp. 71–151, 1971.

[17] S. F. Gull and G. J. Daniell, "Image reconstruction from incomplete and noisy data," *Nature*, vol. 272, pp. 666–670, Apr. 1978.

[18] J. Skilling, "Maximum entropy and image processing—Algorithms and applications," in *Proc. 1st Maximum Entropy Workshop*, Univ. Wyoming, 1981.

[19] J. W. Woods, "Two-dimensional Markov spectral estimation," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 5, Sept. 1976.

[20] L. K. Jones and V. Trutzer, "Computationally feasible high resolution minimum distance procedures which extend the maximum entropy method," *Inverse Problems*, vol. 5, pp. 749–766, 1989.

[21] K. Fukunaga and T. E. Flick, "An optimal global nearest neighbor metric," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 314–318, May 1984.

[22] W. Gersch *et al.*, "Automatic classification of electroencephalograms," *Science*, vol. 205, pp. 193–195, July 13, 1979.