

THE AXIOMS OF MAXIMUM ENTROPY

John Skilling
Department of Applied Mathematics
and Theoretical Physics
Silver Street
Cambridge CB3 9EW, England

Abstract. Maximum entropy is presented as a universal method of finding a "best" positive distribution constrained by incomplete data. The generalised entropy $\Sigma(f - m - f \log(f/m))$ is the only form which selects acceptable distributions f in particular cases. It holds even if f is not normalised, so that maximum entropy applies directly to physical distributions other than probabilities. Furthermore, maximum entropy should also be used to select "best" parameters if the underlying model m has such freedom.

INTRODUCTION

Many quantities of interest are positive distributions. Typical examples are the pattern of light intensity arriving on a photographic plate, and the power spectrum of some radiative field. Following Jaynes (1984), we shall call such a distribution a "scene", and an estimate of it an "image". Usually a scene is a real function f of a continuous spatial or temporal argument x (which may itself have more than one dimension), requiring an infinite number of bits of information to specify it fully. Our knowledge of it, gleaned ultimately from observation, will only be finite.

Even though our knowledge is incomplete, we still wish to obtain a single image from the many which are consistent with our knowledge. In the first half of this paper, we discuss some guidelines (axioms) for finding a single "best" image, based purely on realistic selection criteria, and not relying on probability or information theory. These axioms lead to the maximum entropy (MaxEnt) method for selecting the best image consistent with our knowledge.

MaxEnt needs a given prior model of the scene. This is useful, because it allows prior insight into the nature of the scene to be incorporated into the formalism. However, the insight may be somewhat vague, and contain unknown parameters, such as the positions and intensities of point stars in an astronomical photograph, or lines in a spectrum. In the second half of this paper, we discuss a related set of guideline axioms for finding the best values of any such

parameters. Again, these rules are based purely upon selection criteria.

Remarkably, the same entropy formula is derived. Thus MaxEnt should be used both to find the best single image and to find the best set of parameters underlying it.

SELECTING AN IMAGE

We aim to provide an image which is the "best" according to an agreed criterion. This involves setting up a ranking procedure which determines which of two images is "better". To avoid circularity, and to ensure that there is always some image which is not "bettered" by any other, we impose the transitivity requirement

$$(f \text{ better than } g) \text{ and } (g \text{ better than } h) \\ \Rightarrow (f \text{ better than } h).$$

Any transitive ranking can be described by real numbers, assigning a number $S(\underline{f})$ to each image f , such that

$$"f \text{ better than } g" \iff S(\underline{f}) > S(\underline{g}) \quad (1)$$

Choosing the "best" image is equivalent to regularising f by maximising $S(\underline{f})$. However, the form of $S(\underline{f})$ remains to be defined.

A fundamental requirement is universal applicability, that S should be independent of the type of data we are given. This assumption is useful because it allows a unified approach to data analysis.

The axioms

Remarkably, a few very simple examples of acceptable reconstructed images suffice to determine the form of S . These examples, considered as axioms, progressively restrict the form of S , until only one form remains (or equivalently a monotonic function of it). Anticipating the result, S is the entropy of f .

For the special case of a probability distribution, it seems that Jaynes (1957a,b) was the first to suppose that consistency arguments alone might suffice to determine the entropy formula in the context of inference. His conjecture was proved by Shore and Johnson (1980) who gave a formal axiomatic derivation. Independently, Tikochinsky, Tishby and Levine (1984) arrived at the same formula from a somewhat more physical viewpoint. Earlier derivations of entropy as an uncertainty or information measure (Shannon 1948, Shannon and Weaver 1949, Kullback 1959, Cox 1961) also treated it as a property of a probability distribution.

based purely upon

is derived. Thus
 st single image and
 ying it.

ch is the "best"
 involves setting up
 ch of two images is
 ensure that there is
 i" by any other, we

than h)
 than h).

d by real numbers,
 such that

$$S(\underline{g}) \quad (1)$$

to regularising f by
 S(f) remains to be

applicability, that S
 a we are given. This
 unified approach to

examples of accept-
 termine the form of
 ions, progressively
 form remains (or
 . Anticipating the

tribution, it seems
 to suppose that
 ce to determine the
 nce. His conjecture
 who gave a formal
 Tikochinsky, Tishby
 me formula from a
 ier derivations of
 n measure (Shannon
 1959, Cox 1961) also
 ility distribution.

However, the theorems are more generally applicable, and indeed the proofs are simpler without the normalisation $\int f(x)dx = 1$ which is imposed on probability distributions.

In the following presentation, the formal statement of each axiom is followed by a justification, then by its consequence, a proof thereof, and a comment. Greek letters denote functions appearing in S except that λ and μ are reserved for Lagrange multipliers such as that appearing in the archetype variational equation

$$\delta(S - \lambda \cdot \text{constraint}) = 0$$

The symbol $f[I,m]$ represents the image f reconstructed by maximising S with respect to constraint information I, over a Lebesgue measure m on x.

Axiom I. Subset independence.

Let I_1 be information pertaining only to $f(x)$ for $x \in D_1$ and similarly let I_2 pertain only to $f(x)$ for $x \in D_2$. Then, if D_1 and D_2 are disjoint,

$$f[I_1,m] \cup f[I_2,m] = f[I_1 \cup I_2, m] \quad (2)$$

where "U" is the union operator.

Justification:

Information about one domain should not affect the reconstruction in a different domain, provided there is no constraint directly linking the domains.

Consequence:

S must be of the form

$$S(\underline{f}) = \int dx m(x) \theta(f(x),m(x),x) \quad (3)$$

where θ is an arbitrary function.

Proof:

Consider first the discrete case. Let D_1 and D_2 be non-intersecting domains with union D. Let there be a linear constraint

$$\sum_{i \in D_1} a_{1i} f_i = b_1, \quad \sum_{i \in D_2} a_{2i} f_i = b_2 \quad (4)$$

on each domain. In D_1 and D_2 respectively, f is separately determined by the variational equations

$$\delta S / \delta f_i = \lambda_1 a_{1i}, \quad \delta S / \delta f_i = \lambda_2 a_{2i} \quad (5)$$

Using both constraints together, f is determined by

$$\partial S / \partial f_i = \begin{cases} \mu_1 a_{1i} & , \quad i \in D_1 \\ \mu_2 a_{2i} & , \quad i \in D_2 \end{cases} \quad (6)$$

Taking two cells j, k both in D_1 , the reconstruction is to be independent of the constraints and values of f in the other domain D_2 . Accordingly, $(\partial S / \partial f_j) / (\partial S / \partial f_k)$ is independent of all f_i for $i \in D_2$. This must hold for arbitrary decomposition of D into D_1 and D_2 . Hence

$$(\partial S / \partial f_j) / (\partial S / \partial f_k) = \alpha_{jk}(f_j, f_k) \quad (7)$$

where α_{jk} is a function which might depend on coordinates j, k but which does not depend on any f_i other than f_j and f_k themselves.

A technical argument now leads from this to the result (3). Consideration of a third cell l yields

$$(\partial S / \partial f_k) / (\partial S / \partial f_l) = \alpha_{kl}(f_k, f_l) \quad (8)$$

$$(\partial S / \partial f_l) / (\partial S / \partial f_j) = \alpha_{lj}(f_l, f_j) \quad (9)$$

Multiplying the latter three equations,

$$1 = \alpha_{jk}(f_j, f_k) \alpha_{kl}(f_k, f_l) \alpha_{lj}(f_l, f_j) \quad (10)$$

Hence

$$0 = \partial^2 (\log \alpha_{jk}) / \partial f_j \partial f_k \quad (11)$$

so that, on using the antisymmetry (7) of $\log \alpha$ in j and k ,

$$\alpha_{jk}(f_j, f_k) = \beta_{jk}(f_j) / \beta_{kj}(f_k) \quad (12)$$

where β_{jk} is an as yet un-determined function. Substituting in (10) and differentiating with respect to f_j yields

$$(\log \beta_{jk}(f_j))' = (\log \beta_{jl}(f_j))' \quad (13)$$

for arbitrary k and l , so that the differential $\beta_{jk}'(f_j)$ does not depend on k . The arbitrary constant which appears on integration can be absorbed in the definition (12), so that the second suffix on β may be dropped. Equation (7) can then be rewritten as

$$(\partial S / \partial f_j) / (\partial S / \partial f_k) = \beta_j(f_j) / \beta_k(f_k) \quad (14)$$

Define

$$R(\underline{f}) = \sum_i \theta_i(f_i) \quad \text{where} \quad \theta_i'(x) = \beta_i(x) \quad (15)$$

Then $\partial R / \partial f_i = \beta_i(f_i)$ for all i , and (14) shows that

$$(6) \quad (\partial S / \partial f_j) / (\partial S / \partial f_k) = (\partial R / \partial f_j) / (\partial R / \partial f_k) \quad (16)$$

This means that the gradients $(\partial / \partial f_i)$ of R and S are parallel. Accordingly, $R(\underline{f})$ and $S(\underline{f})$ produce exactly the same reconstructions from given constraints, because any difference in the gradient magnitudes is absorbed in the Lagrange multiplier(s) of the constraint(s). Hence, without loss of generality, S can be restricted to the form

$$(7) \quad S(\underline{f}) = \sum_i \theta_i(f_i) \quad (17)$$

Passage to the continuum limit requires a Lebesgue measure m to be introduced on the coordinate x , as $\theta_i(f_i)$ is replaced by its continuum equivalent $\theta(f(x), x)$. This completes the proof of (3), in which θ is assigned an explicit additional argument $m(x)$, separate from x , for later convenience.

Comment:

It is not surprising that the axiom is only satisfied by a simple sum over the individual cells i of the scene. The effect of the axiom is precisely to exclude cross-terms between different points.

Axiom II. Coordinate invariance.

Let Γ be a coordinate transformation from x to Γx . Then $f[I, m]$ transforms to

$$(8) \quad \Gamma(f[I, m]) = f[\Gamma I, \Gamma m] \quad (18)$$

Justification:

We expect the same answer when we solve the same problem in two different coordinate systems, in that the reconstructed images in the two systems should be related by the coordinate transformation.

Consequence:

S must be of the form

$$(9) \quad S(\underline{f}) = \int dx m(x) \phi(f(x)/m(x)) \quad (19)$$

Proof:

Write (3) in the form

$$(10) \quad S(\underline{f}) = \int dx m(x) \phi(f(x)/m(x), m(x), x) \quad (20)$$

First, let I be the simple linear constraint $\int dx f(x) = 1$. The variational equation gives

$$(11) \quad \sigma(f(x)/m(x), m(x), x) = \lambda \quad (21)$$

struction is to be
of f in the other
is independent of
rary decomposition

id on coordinates
her than f_j and f_k

to the result (3).

, f_j) (10)

(11)

log α in j and k ,

(12)

tion. Substituting
to f_j yields

(13)

ferential $\beta_{jk}'(f_j)$
tant which appears
inition (12), so
ed. Equation (7)

(14)

= $\beta_i(x)$ (15)

where σ is the derivative of ϕ with respect to its first argument, and λ is the Lagrange multiplier. Suppose there are two points x_1 and x_2 at which m takes the same value, and in the neighbourhood of which m is continuous. Let Γ be the transformation which exchanges equal-volume neighbourhoods D_1 of x_1 and D_2 of x_2 . Then $\Gamma I = I$ and $\Gamma m = m$, hence $\Gamma f = f$. Also $\Gamma \lambda = \lambda$ because (by axiom I) operation of Γ leaves all points other than those in D_1 and D_2 unaffected. Substitution in (21) yields

$$\sigma(f/m, m, x_1) = \sigma(f/m, m, x_2) \quad (22)$$

showing that σ does not depend on its third argument. Integrating, ϕ itself is also independent of its third argument, except for an additive term which does not affect the maximisation over f and may be dropped.

Next, let I be the more general linear constraint

$$\int dx a(x)f(x) = 1 \quad (23)$$

for which the variational equation is

$$\sigma(f(x)/m(x), m(x)) = \lambda a(x) \quad (24)$$

Apply a coordinate transformation $x \rightarrow \Gamma x$ with Jacobian $\Upsilon(x) = \partial(\Gamma x / \partial x)$. This gives

$$dx \rightarrow \Upsilon dx, \quad m \rightarrow \Upsilon^{-1} m, \quad f \rightarrow \Upsilon^{-1} f, \quad a \rightarrow a \quad (25)$$

and the variational equation becomes

$$\sigma(f(x)/m(x), m(x)/\Upsilon(x)) = \mu a(x) \quad (26)$$

where μ is a (possibly different) Lagrange multiplier. Dividing the two forms, we see that

$$\frac{\sigma(f(x)/m(x), m(x)/\Upsilon(x))}{\sigma(f(x)/m(x), m(x))} \text{ is constant in } x. \quad (27)$$

This holds for arbitrary $\Upsilon(x)$, so σ can not depend on its second argument. Integrating, neither does ϕ depend on its second argument, except for a term which does not affect the maximisation over f and can be omitted. This proves the required result (19).

Because S is now constructed purely from invariants $m(x)dx$ and $f(x)/m(x)$, reconstructions obtained from it must clearly satisfy axiom II as well as axiom I. The proof would have been shorter if S itself had been assumed to be invariant. Such an assumption would be plausible, but its truth is a consequence of the prime requirement that the reconstructed

ect to its first
r. Suppose there
the same value,
tinuous. Let Γ be
-volume neighbour-
and $\Gamma_m = m$, hence
operation of Γ
and D_2 unaffected.

(22)

s third argument.
ent of its third
ch does not affect
i.

nstraint

(23)

(24)

with Jacobian $\Upsilon(x)$

, $a \rightarrow a$ (25)

(26)

grange multiplier.

stant in x . (27)

not depend on its
oes ϕ depend on its
does not affect the
. This proves the

n invariants $m(x)dx$
from it must clearly
he proof would have
ed to be invariant.
but its truth is a
t the reconstructed

image should be invariant.

Comment:

It is via this axiom that the additive nature of f is introduced. For example, in incoherent optics, it makes sense to treat the radiative energy flux $\int intensity(x) dx$ in a domain as an invariant quantity under coordinate transformation, whereas it would not make sense to treat the corresponding integral of wave amplitude $\int amplitude(x) dx$ as an invariant. Accordingly, we would identify $f(x)$ with the additive flux density rather than some other function of it such as its square root.

Axiom III. System Independence.

We now restrict the form of $S(\underline{f})$ by requiring a specific reconstruction for a particularly simple problem.

Let $m(x_1, x_2) = 1$ on the unit square $0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1$. Let the constraints I be values of the marginals

$$\int dx_2 f(x_1, x_2) = a_1(x_1) \quad , \quad \int dx_1 f(x_1, x_2) = a_2(x_2) \quad (28)$$

themselves obeying the consistent normalisation condition

$$\int dx_1 a_1(x_1) = \int dx_2 a_2(x_2) = 1 \quad (29)$$

Then we require the reconstructed image to be the direct product

$$f(x_1, x_2) = a_1(x_1) a_2(x_2) \quad (30)$$

Gull, reported in Gull and Skilling (1984) and Livesey and Skilling (1985), has presented a less abstract formulation of this axiom.

Justification:

$f(x_1, x_2)$ represents a distribution of proportions, because clearly it is constrained to satisfy $\iint dx_1 dx_2 f(x_1, x_2) = 1$. If all we know about f are its marginal distributions $a_1(x_1)$ and $a_2(x_2)$, then (in the absence $m = 1$ of any contrary bias) we wish to recover the uncorrelated reconstruction $f = a_1 a_2$. Any other choice of $f(x_1, x_2)$ would imply correlations for which there is evidence neither in the data nor in the measure. Good (1963) showed that this lack of correlation in contingency tables would be a consequence of MaxEnt, but here we reverse the argument and use the lack of correlation in a derivation of MaxEnt.

Consequence:

S must be of the form

$$S(\underline{f}) = - \int dx f(x) (\log(f(x)/m(x)) + c) \quad (31)$$

where c is a constant.

Proof:

With the given constraints, the variational equation

$$\delta \left(S - \int dx_1 \lambda_1(x_1) a_1(x_1) - \int dx_2 \lambda_2(x_2) a_2(x_2) \right) = 0 \quad (32)$$

yields

$$\sigma(f(x_1, x_2)) = \lambda_1(x_1) a_1(x_1) + \lambda_2(x_2) a_2(x_2) \quad (33)$$

in which σ is the derivative of ϕ as before, and where $f(x_1, x_2)$ is given in terms of a_1 and a_2 by the axiom (30). Applying $\delta^2/\delta x_1 \delta x_2$ yields

$$y \sigma''(y) + \sigma'(y) = 0 \quad (34)$$

in which $y=f(x_1, x_2)$ can be chosen to take arbitrary values by suitable choice of constraint functions. Integrating twice,

$$\sigma(y) = A \log y + B \quad (35)$$

where A and B are constants. Integrating again,

$$\phi(y) = A y \log y + (B-A) y \quad (36)$$

plus another constant which does not affect the maximisation of S over f and may be dropped. A should be negative, to ensure that the extremum of S is a maximum ($\delta^2 S < 0$), but is otherwise merely an arbitrary scaling of S . Choosing $A = -1$, we have

$$\phi(y) = -y (\log y + c) \quad (37)$$

($c = \text{constant}$), which immediately gives the required form (31).

Comment:

This is the crucial axiom, which reduces S to the entropic form. The basic point is that when we seek an uncorrelated image from marginal data in two (or more) dimensions, we need to multiply the marginal distributions. On the other hand, the variational equation tells us to add constraints through their Lagrange multipliers. Hence the gradient $\delta S/\delta f$ must be the logarithm,

$$\delta S/\delta f = \log m - \log f \quad (38)$$

which is the only function which converts a product into a sum. Integrating $\log f$ yields the " $f \log f$ " entropic form.

Axiom IV: Scaling.

$$f[\emptyset, m] = m \tag{39}$$

where \emptyset represents the absence of any information.

Justification:

In the absence of any additional information, we wish to recover the initial measure.

Consequence:

The last ambiguity is resolved, and

$$S(\underline{f}) = \int dx (f(x) - f(x) \log(f(x)/m(x))) \tag{40}$$

Proof:

Unconstrained maximisation of (31) over f yields

$$f(x) = m(x) e^{-1-c} \tag{41}$$

It would not actually be inconsistent to have a universal scaling factor e^{-1-c} between initial measure m and reconstruction f , but it would be arbitrary and often inconvenient. The value $c = -1$ avoids the difficulty.

Comment:

This choice may be viewed as a convention defining the units of f to be those of m .

The entropic regularisation function (40), properly written with two arguments as

$$S(\underline{f}, \underline{m}) = \int dx (f(x) - f(x) \log(f(x)/m(x))) \tag{42}$$

is the form to be maximised when selecting an optimal image f . We should note that S does obey all four axioms, so that the axioms are mutually consistent. Before defining S in (42) to be the "entropy of f ", we shall investigate the role of the measure m more closely.

SELECTING A MODEL

The global maximum of S over f , attained in the absence of further constraints, occurs at $f=m$, when the image equals the measure. This suggests a useful interpretation of m . As well as being an abstract Lebesgue measure, m can also be thought of as a prior model for the image. Imposition of further constraints will modify the selected image in such a way that it will always be as "close" (in the sense of maximising S) to the model as possible.

al equation
) $a_2(x_2) = 0$ (32)
 $a_2(x_2)$ (33)
 before, and where
 by the axiom (30).
 (34)
 ce arbitrary values
 ions. Integrating
 (35)
 g again,
 (36)
 ect the maximisation
 ld be negative, to
 um ($\delta^2 S < 0$), but is
 S. Choosing $A = -1$,
 (37)
 the required form
 s S to the entropic
 seek an uncorrelated
 re) dimensions, we
 tions. On the other
 s to add constraints
 Hence the gradient
 (38)
 rts a product into a
 og f " entropic form.

It often happens that a scene contains particular features, such as point sources or spectral lines, which can be described by a fairly small number of parameters. It would be helpful to provide model parameters which were "best" according to an agreed criterion. This involves setting up a ranking of models which itself implies the existence of a functional $H(\underline{m}, \underline{f})$, or $H(\underline{m})$ for short, to be maximised over \underline{m} .

The Axioms

Again, we will use simple properties of selected models to restrict the form of H until only one form remains (or equivalently a monotonic function of it). We assign the symbol J to such freedom, by analogy with possible constraint information I on f . The symbol $m\{J, f\}$ represents the optimal model allowed by J , on the basis of scene f .

Axiom I'. Subset independence.

Let J_1 be freedom allowed to $m(x)$ in $x \in D_1$, and let J_2 be freedom independently allowed to $m(x)$ in $x \in D_2$. Then, if D_1 and D_2 are disjoint,

$$m\{J_1, f\} \cup m\{J_2, f\} = m\{J_1 \cup J_2, f\} \quad (43)$$

Justification:

The model fitted to one domain should not affect the model in a different domain, provided there are no parameters directly linking the domains.

Consequence:

H must be of the form

$$H(\underline{m}) = \int dx f(x) \theta(m(x)/f(x), f(x), x) \quad (44)$$

Proof:

This follows exactly as before (axiom I), with S replaced by H , and f replaced by m , though the function θ may be different.

Axiom II'. Coordinate invariance.

Let Γ be a coordinate transformation from x to Γx . Then $m\{J, f\}$ transforms to

$$\Gamma(m\{J, f\}) = m\{\Gamma J, \Gamma f\} \quad (45)$$

Justification:

We expect the same answer when we solve the same problem in two different coordinate systems, in that the models in the two system should be related by the coordinate transformation.

particular features, which can be parameters. It would which were "best" involves setting up the existence of a t, to be maximised

erties of selected nly one form remains it). We assign the ogy with possible ol $m\{J, f\}$ represents asis of scene f.

$x \in D_1$, and let J_2 be $x \in D_2$. Then, if D_1

$\{, f\}$ (43)

not affect the model are no parameters

$(x), x)$ (44)

, with S replaced by function θ may be

from x to Γx . Then

(45)

the same problem in at the models in the by the coordinate

Consequence:
H must be of the form

$$H(\underline{m}) = \int dx f(x) \phi(m(x)/f(x)) \quad (46)$$

Proof:

This follows exactly as before (axiom II), with S replaced by H, and f replaced by m, though the function ϕ may be different.

Axiom III'. System independence.

We now restrict the form of $H(\underline{m})$ by requiring a specific model for a particularly simple problem. Let $f(x_1, x_2)$, defined on the unit square $0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1$, be normalised $\iint dx_1 dx_2 f(x_1, x_2) = 1$. Let it be modelled by the factorised form

$$m(x_1, x_2) = n_1(x_1) n_2(x_2) \quad (47)$$

with normalisation

$$\int dx_1 n_1(x_1) = \int dx_2 n_2(x_2) = 1 \quad (48)$$

For this, we require

$$n_1(x_1) = \int dx_2 f(x_1, x_2) \quad , \quad n_2(x_2) = \int dx_1 f(x_1, x_2) \quad (49)$$

Justification:

$f(x_1, x_2)$ represents a distribution of proportions, because of its normalisation. In the model, $n_1(x_1)$ represents proportional structure in the x_1 dimension, as does $n_2(x_2)$ in the x_2 dimension, and we wish to recover the correct marginals. The model has no way of displaying correlations between x_1 and x_2 , and we do not wish east-west (x_1) knowledge to influence our reconstruction of overall north-south structure (x_2), neither should x_2 structure influence x_1 .

Consequence:
H must be of the form

$$H(\underline{m}) = \int dx [f(x) \log(m(x)/f(x)) - c m(x)] \quad (50)$$

Proof:

Perturbing the model in (46) yields, in this two-dimensional example

$$\delta H = \iint dx_1 dx_2 \sigma(m(x_1, x_2)/f(x_1, x_2)) \delta m(x_1, x_2) \quad (51)$$

where σ is the differential of ϕ . From the factorised form (47),

$$\delta m(x_1, x_2) = n_1(x_1) \delta n_2(x_2) + \delta n_1(x_1) n_2(x_2) \quad (52)$$

where n_1 and n_2 obey normalisation (48) but are otherwise unrestricted. Substituting this into the variational equation

$$\delta \left(H - \lambda_1 \int dx_1 n_1(x_1) - \lambda_2 \int dx_2 n_2(x_2) \right) = 0 \quad (53)$$

for model parameter perturbations δn_1 and δn_2 gives

$$\begin{aligned} \int dx_2 \sigma \left(n_1(x_1) n_2(x_2) / f(x_1, x_2) \right) n_2(x_2) &= \lambda_1 \\ \int dx_1 \sigma \left(n_1(x_1) n_2(x_2) / f(x_1, x_2) \right) n_1(x_1) &= \lambda_2 \end{aligned} \quad (54)$$

from δn_1 and δn_2 respectively. In (54), $n_1(x_1)$ and $n_2(x_2)$ are to equal the marginals (49), for arbitrary scene f . Perturb f in such a way that the marginals are unchanged. This gives

$$\begin{aligned} \int dx_2 \tau \left(f(x_1, x_2) / n_1(x_1) n_2(x_2) \right) \delta f(x_1, x_2) &= \delta \lambda_1 n_1(x_1) \\ \int dx_1 \tau \left(f(x_1, x_2) / n_1(x_1) n_2(x_2) \right) \delta f(x_1, x_2) &= \delta \lambda_2 n_2(x_2) \end{aligned} \quad (55)$$

where $\tau(y) = d\sigma(y^{-1})/dy$. Select the particular marginal-preserving perturbation proportional to

$$\begin{aligned} \delta f(x_1, x_2) &= \delta(x_1 - a_1) \delta(x_2 - a_2) - \delta(x_1 - b_1) \delta(x_2 - a_2) \\ &\quad - \delta(x_1 - a_1) \delta(x_2 - b_2) + \delta(x_1 - b_1) \delta(x_2 - b_2) \end{aligned} \quad (56)$$

where δ on the right-hand side is the Dirac delta function and a_1, a_2, b_1, b_2 are coordinates between 0 and 1, so that the rectangle of points (x_1, x_2) selected by the four delta functions lies within the given unit square. Values of x_1 other than a_1 and b_1 show that $\delta \lambda_1 = 0$. Putting $x_1 = a_1$ gives

$$\begin{aligned} \tau(f(a_1, a_2) / n_1(a_1) n_2(a_2)) \\ - \tau(f(a_1, b_2) / n_1(a_1) n_2(b_2)) &= 0 \end{aligned} \quad (57)$$

This holds for arbitrary a_1, a_2, b_2 so that the two arguments of τ can each take arbitrarily different values. The only way of satisfying this is to set

$$\tau(y) = A, \quad A = \text{constant} \quad (58)$$

from which

$$\sigma(y) = A y^{-1} + B \quad (59)$$

Integrating again,

$$\phi(y) = A \log y + B y \quad (60)$$

$$n_1(x_1) n_2(x_2) \quad (52)$$

but are otherwise the variational

$$n_2(x_2) = 0 \quad (53)$$

and δn_2 gives

$$n_2(x_2) = \lambda_1 \quad (54)$$

$$n_1(x_1) = \lambda_2$$

, $n_1(x_1)$ and $n_2(x_2)$ arbitrary scene f . Values are unchanged.

$$n_2(x_2) = \delta \lambda_1 n_1(x_1) \quad (55)$$

$$n_2(x_2) = \delta \lambda_2 n_2(x_2)$$

particular marginal-

$$\delta n_1(x_1) \delta(x_2 - a_2) \quad (56)$$

Dirac delta function between 0 and 1, so that defined by the four delta square. Values of $x_1 = 0$. Putting $x_1 = a_1$

$$(57)$$

so that the two different values. Let

$$(58)$$

$$(59)$$

$$(60)$$

plus a constant which does not affect the maximisation over m and may be omitted. A should be positive, to ensure that the extremum of H is a maximum ($\delta^2 H < 0$), but is otherwise merely a scaling factor. Choosing $A = 1$ and setting $B = -c$ gives the quoted form (50).

Axiom IV': Scaling.

$$m\{\emptyset, f\} = f \quad (61)$$

where \emptyset represents the absence of any restriction on the model.

Justification:

In the absence of any restriction, we seek to recover the starting scene.

Consequence:

The last ambiguity is resolved, and

$$H(\underline{m}, \underline{f}) = \int dx [f(x) \log(m(x)/f(x)) - m(x)] \quad (62)$$

Proof:

Unconstrained maximisation of (50) yields

$$m(x) = f(x)/c \quad (63)$$

whereas $m=f$ is required. Thus $c=1$ and (62) is obtained.

Finally, we should note that $H(\underline{m})$ defined in (62) does obey all four axioms, so that the axioms are mutually consistent.

SYNTHESIS

$S(\underline{f}, \underline{m})$ from (42) and $H(\underline{m}, \underline{f})$ from (62) are the same function (apart from additive terms in each which do not affect maximisation of the other). They can be combined into the joint form (denoted by S , because there is now no need for separate symbols):

$$S(\underline{f}, \underline{m}) = \int dx [f(x) - m(x) - f(x) \log(f(x)/m(x))] \quad (64)$$

in the continuous case, and in the discrete case

$$S(\underline{f}, \underline{m}) = \sum_i [f_i - m_i - f_i \log(f_i/m_i)] \quad (65)$$

This can be used to rank and thence to select image-model $(\underline{f}, \underline{m})$ pairs, and is the only function which obeys all the axioms above. We shall call S "entropy" because of its close connection with the classic $-\sum p \log p$ form. The decrease of S from its global maximum of zero quantifies the deviation of f from its model m .

The "maximum entropy method" in data analysis, then, consists of maximising the entropy S , either over an image f subject to given constraints, or over the model m within given degrees of freedom, or both.

CONCLUSIONS

Any universally applicable method of selecting a single positive image ought to give acceptable, sensible results in particular cases. Four such cases, codified as axioms, lead to MaxEnt as the only consistent selection procedure. The MaxEnt method is valid for any type of data, regardless of the normalisation of the image. Of fundamental importance is that the entropy gradient (38) is logarithmic. This neatly ensures that any reconstructed image is both positive and finite, and as close as possible to some prior model. The more complicated formula (64)

$$S(\underline{f}, \underline{m}) = \int dx [f(x) - m(x) - f(x) \log(f(x)/m(x))]$$

for the entropy itself is the integral of this gradient.

Even though MaxEnt has already had considerable practical success in reconstructing various types of positive distribution, theory indicates that the method should be yet more powerful, because the same entropy formula should be used to select optimal parameters in the underlying prior model. Thus MaxEnt should also be used to estimate parameters pertaining to positive distributions, which opens a particularly promising avenue to future research.

In this paper, no attempt has been made to quantify the probabilistic reliability of the MaxEnt estimates. Maximum entropy can stand in its own right as a selection procedure.

lysis, then, consists
an image f subject
al m within given

hod of selecting a
ceptable, sensible
cases, codified as
onsistent selection
or any type of data,
the image. Of
py gradient (38) is
any reconstructed
s close as possible
ed formula (64)

$\log(f(x)/m(x))$]

f this gradient.

siderable practical
of positive distrib-
should be yet more
la should be used to
lying prior model.
estimate parameters
ich opens a partic-
ch.

e to quantify the
estimates. Maximum
selection procedure.

References

- Cox, R.T. (1961). The algebra of probable inference. Johns Hopkins Press, Baltimore, MD.
- Good, I.J. (1963). Maximum entropy for hypothesis formulation, especially for multi-dimensional contingency tables. *Annals.Math.Stat.*, 34, 911-934.
- Gull, S.F. & Skilling, J. (1984). The maximum entropy method. *In* Indirect imaging, ed. J.A. Roberts. Cambridge: Cambridge University Press.
- Jaynes, E.T. (1957a). Information theory and statistical mechanics I. *Phys. Rev.*, 106, 620-630.
- Jaynes, E.T. (1957b). Information theory and statistical mechanics II. *Phys. Rev.*, 108, 171-190.
- Jaynes, E.T. (1984). Monkeys, Kangaroos and N. Presented at fourth maximum entropy workshop, Calgary, ed. J.H. Justice, Dordrecht: Reidel.
- Kullback, S. (1959). Information theory and statistics. New York: Wiley.
- Livesey, A.K. & Skilling, J. (1985). Maximum entropy theory *Acta Cryst.*, A41, 113-122.
- Shannon, C.F. (1948). A mathematical theory of communication. *Bell System Tech. J.*, 27, 379-423 and 623-656.
- Shannon, C.E. & Weaver, W. (1949). The mathematical theory of communication. Urbana, Illinois: University Illinois Press.
- Shore, J.E. & Johnson, R.W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans.Info.Theory*, IT-26, 26-37 and IT-29, 942-943.
- Tikochinsky, Y., Tishby, N.Z. & Levine, R.D. (1984). Consistent inference of probabilities for reproducible experiments. *Phys.Rev.Lett.*, 52, 1357-1360.