MAXENT, MATHEMATICS, AND INFORMATION THEORY

I. CSISZÁR
Mathematical Institute of the Hungarian Academy of Sciences
H-1364 Budapest
P.O. Box 127
Hungary[‡§]

Abstract. This is a mathematically oriented survey about the method of maximum entropy or minimum I-divergence, with a critical treatment of its various justifications and relation to Bayesian statistics. Information theoretic ideas are given substantial attention, including "information geometry". The axiomatic approach is considered as the best justification of maxent, as well as of alternate methods of minimizing some Bregman distance or f-divergence other than I-divergence. The possible interpretation of such alternate methods within the original maxent paradigm is also considered.

Key words: Bregman distance, entropy, f-divergence, information divergence, information geometry, large deviations, maxent axioms, maxent in mean, prior distribution.

1. The scope of MAXENT

Maxent is a method to infer a function $p(x)$ defined on a given set $\mathcal{X}$ when the available information specifies only a feasible set F of such functions. Originally coming from physics (Boltzmann, Gibbs), maxent has been promoted to a general method of inference primarily by Kullback and Jaynes, [1],[2].

Typically but not always, the feasible set is determined by linear constraints on p, i.e.,

$$F = \{p : \int a_i(x)p(x)\lambda(dx) = b_i,\ i = 1, \ldots, k\}, \tag{1}$$

for some given functions $a_i(x)$ and constants $b_i, i = 1, \ldots, k$. Here and in the sequel, λ denotes a given σ-finite measure on (a σ-algebra of subsets of) $\mathcal{X}$. The reader may assume with little loss of generality that $\mathcal{X}$ is finite or countably infinite and

[‡]This work was supported by the Hungarian National Foundation for Scientific Research, Grant T016386.

[§]Email: csiszar@math-inst.hu

K. M. Hanson and R. N. Silver (eds.), Maximum Entropy and Bayesian Methods, 35–50.

λ is the counting measure (then integrals become sums) or $\mathcal{X}$ is a subset of a finite dimensional space and λ is the Lebesgue measure.

The function p to be inferred is often a probability density or mass function. Then we will speak of Problem (i) or (ii) according as the only available information is $p \in F$, or also a *default model* $q(x)$ is given, such that $p = q$ would be inferred were $q \in F$. We will speak of Problem (iii) when a non-negative $p(x)$ not necessarily a probability density (such as a power spectrum or an intensity function) has to be inferred, knowing only a feasible set F and a default model $q(x)$.

The maxent solution to Problem (i) is that $p^* \in F$ whose entropy

$$H(p) = -\int p(x) \log p(x) \lambda(dx) \tag{2}$$

is maximum. For Problems (ii) and (iii), the maxent solution is that $p^* \in F$ whose information divergence (I-divergence) from q, i.e.,

$$D(p \parallel q) = \int [p(x) \log \frac{p(x)}{q(x)} - p(x) + q(x)] \lambda(dx) \tag{3}$$

is minimum. If $\lambda(\mathcal{X}) < \infty$, and $p_0(x) = \text{constant} = 1/\lambda(\mathcal{X})$ is the uniform distribution on $\mathcal{X}$, then $H(p_0) = \log \lambda(\mathcal{X})$ and

$$H(p) = H(p_o) - D(p \parallel p_0) \tag{4}$$

for every probability density p. Hence the maxent solution to Problem (i) is the same as that to Problem (ii) with default model $q = p_0$.

I-divergence has a variety of other names (and notations) such as Kullback-Leibler information number, relative entropy, information gain, etc. It is a (non-symmetric) measure of distance of p from q, i.e., $D(p \parallel q) \geq 0$, with equality iff $p = q$ (interpreted as $p(x) = q(x)$ for λ-almost all x). The funtion p^* minimizing $D(p \parallel q)$ subject to $p \in F$ is called the I-projection of p onto F.

It should be emphasized that a maxent solution as defined above does not always exist. On the other hand, if a solution exists it is unique, providing the feasible set F is convex. For more details, cf. Section 5.

2. Information – theoretic heuristic

The amount of information provided by the outcome x of a drawing from a finite or countably infinite $\mathcal{X}$, governed by a probability mass function p, is defined to be $-\log_2 p(x)$ bits. One motivation for this definition is Shannon's

Theorem 2.1: Any uniquely decodable code that assigns a binary codeword of length $l(x)$ to each $x \in \mathcal{X}$, has average length

$$\Sigma p(x) l(x) \geq -\Sigma p(x) \log_2 p(x). \tag{5}$$

The lower bound can be approached closer than 1 bit by a code with $l(x) = \lceil -\log_2 p(x) \rceil$.

Often, $-\log_2 p(x)$ is referred to as the ideal codelength of x for p; the right hand side of (5) is the ideal average codelength. Notice that the average length (under p) of an ideal code for some $q \neq p$, with $l(x) = -\log_2 q(x)$, exceeds the ideal average codelength for p by $\Sigma p(x) \log_2(p(x)/q(x))$.

In information theory, the entropy of a probability mass function p and the I-divergence of p from q are usually defined as $-\Sigma p(x) \log_2 p(x)$ and $\Sigma p(x) \log_2 \frac{p(x)}{q(x)}$. The difference from our definitions (2) and (3), with λ=counting measure, is only a constant factor, i.e., the choice of unit. Thus $H(p)$ measures the average amount of information provided by the outcome of a random drawing governed by p , which may also be interpreted as a measure of uncertainty about that outcome before observing it, or of the amount of randomness represented by p. $D(p \parallel q)$ is an information theoretic distance of p from q. It measures how less informed is about the outcome of a random drawing one who believes this drawing is governed by q than one who knows the true p. Alternatively, it is a measure of information gained when learning that the true distribution is p rather than q.

If $\mathcal{X}$ is a subset of R^k and λ is the Lebesgue measure, the entropy (2) can not be directly interpreted as a measure of average amount of information or degree of randomness; e.g., it may be negative. Still, if the outcome of a drawing governed by a density p is observed with a prescribed (large) precision, say each of its k components is rounded to N fractional binary digits, the probability of any particular observed outcome x will be $p(x)2^{-kN}$ with good approximation, supposing p is smooth. Thus the amount of information provided by this observation will be $-\log_2 p(x) + kN$ bits, and the average amount of information will be $H(p)$ + constant. Thus a density with larger entropy still represents larger randomness. The intuitive interpretation of I-divergence can be similarly extended to the continuous case, indeed, its extension is more straightforward than that of entropy.

For Problem (i), maxent may be regarded as an extension of the *principle of insufficient reason*, i.e., that in case of complete ignorance $p(x)$=constant should be assumed. Of course, this makes sense only when $\lambda(\mathcal{X}) < \infty$, and in that case the uniform distribution on $\mathcal{X}$ represents maximum randomness (formally, (4) shows that the uniform distribution has maximum entropy). If instead of complete ignorance we know that $p \in F$, but nothing else, maxent still suggests to adopt the most random feasible p. The information theoretic meaning of I-divergence provides a reasonable heuristic background of maxent also for Problem (ii). The same, however, can not be said about Problem (iii).

3. Large deviations and maxent

Maxent is closely connected with the subject of large deviations in probability theory, which, in turn, is closely related to information theory. The simple large deviation results stated below are key ingredients of information theory, cf. Csiszár and Körner [3]. However, they have been known in statistical physics much earlier, dating back to Boltzmann. A proof can be given by straightforward calculation using Stirling's formula; an even simpler proof appears in [3], p.30.

Theorem 3.1: Given a finite set $\mathcal{X}$ of size $|\mathcal{X}|$, let $N_n(\hat{p})$ denote the number of n-tuples $(x_1, \ldots, x_n) \in \mathcal{X}^n$ with a given empirical distribution $\hat{p}$, where

$$\hat{p}(x) = \frac{1}{n} \text{ (number of indices i with } x_i = x). \tag{6}$$

Then

$$N_n(\hat{p}) = \exp[nH(\hat{p}) - r_n(\hat{p})], \tag{7}$$

with

$$0 \leq r_n(\hat{p}) \leq |\mathcal{X}| \log n. \tag{8}$$

Corollary: If $x_1, \ldots, x_n$ are drawn independently from $\mathcal{X}$, governed by q, the empirical distribution will be $\hat{p}$ with probability

$$\exp[-nD(\hat{p} \| q) - r_n(\hat{p})]. \tag{9}$$

Suppose now that $\hat{p}$ is known to belong to a closed, convex feasible set F, and let $p^* = \text{argmax}_{p \in F} H(p)$. Then, providing F contains empirical distributions arbitrarily close to p^* if n is sufficiently large, the Theorem implies that among the n-tuples $(x_1, \ldots, x_n)$ with empirical distribution $\hat{p} \in F$, all but an exponentially small fraction will be in an arbitrarily small neighborhood of p^*, if n is large. Similarly, the Corollary implies that if $x_1, \ldots, x_n$ are drawn independently from $\mathcal{X}$, governed by q, the conditional probability on the condition $\hat{p} \in F$ that $\hat{p}$ will be in an arbitrarily small neighborhood of $p^* = \text{argmin}_{p \in F} D(p \| q)$, is exponentially close to 1, if n is large. These facts represent very strong arguments for maxent inference, providing the probability mass function to be inferred is an empirical distribution. This covers many applications of maxent in physics.

Theorem 3.1 and its corollary also imply that if all $(x_1, \ldots, x_n) \in \mathcal{X}^n$ are a priori equally likely, or if $x_1, \ldots, x_n$ are independently drawn governed by q, the conditional distribution of x_1 (or of x_k for any fixed k) on the condition that $\hat{p} \in F$, will be arbitrarily close to the maxent solution p^* as above. More general results of this kind, for general rather than finite $\mathcal{X}$, were proved by Campenhout and Cover [4] and Csiszár [5]. They may be used to argue that maxent is "right" more generally than for inferring an empirical distribution, [6]. It should be added, however, that such arguments do not seem to cover the majority of applications of maxent, in particular, they do not apply to Problem (iii).

4. Maxent and Bayesian statistics

Given a family $\{p_\vartheta(x), \vartheta \in \Theta\}$ of probability densities on $\mathcal{X}$, where $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$ is a parameter vector taking values in a given set $\Theta \subset R^k$, for Bayesian inference about ϑ one needs a prior distribution on Θ. If the family is reparameterized as $\{p_\varphi(x), \varphi \in \Phi\}$, where φ is a one-to-one function of ϑ, one may as well want a prior on Φ. A natural invariance requirement is that the latter be the transform of the prior on Θ under the mapping $\vartheta \to \varphi$.

It has been suggested that the prior should be selected by maxent. In the absence of constraints on feasible priors, this amounts to selecting the uniform

distribution on Θ. Unfortunately, in addition to other shortcomings, this method of prior selection does not meet the invariance requirement.

Another method, motivated by information theory, might be to select that prior π under which the average amount of information obtained about ϑ from observing x, known as mutual information, is maximum. By one of several equivalent definitions, this mutual information equals the expectation $E_\pi D(p_\vartheta \parallel \bar{p})$, where $\bar{p} = E_\pi p_\vartheta$ is the marginal density of x. The maximum mutual information is mathematically equivalent to channel capacity in information theory.

The latter method does satisfy invariance but, unfortunately, it often yields a prior concentrated on a finite subset of Θ (always if $\mathcal{X}$ is a finite set). As an improvement, one might take that prior under which a sequence of observations $x_1, \ldots, x_n$, conditionally independent given ϑ, provides maximum information about ϑ in the limit $n \to \infty$. As first established by Bernardo [7] and later rigorously proved by Clarke and Barron [8] (under suitable regularily conditions), this leads to the so-called Jeffreys prior, with density proportional to the square-root of the determinant of the Fisher information matrix $J(\vartheta) = [J_{ij}(\vartheta)]$,

$$J_{ij}(\vartheta) = \int \frac{\partial \log p_\vartheta(x)}{\partial \vartheta_i} \; \frac{\partial \log p_\vartheta(x)}{\partial \vartheta_j} p_\vartheta(x) \lambda(dx). \tag{10}$$

It may be reassuring to have recovered a prior that had been derived earlier by other considerations, even though this prior is not beyond criticism, cf. Bernardo and Smith [9].

Let us briefly discuss a rather debatable connection of maxent to Bayesian statistics. In many applications, the feasible set is supposed to be defined by linear constraints, cf. (1), but the known values of the constants are subject to errors. In other words, for the unknown p and the known values b_i, the errors

$$e_i = b_i - \int a_i(x) p(x) \lambda(dx) \quad , \quad i = 1, \ldots, k \tag{11}$$

are small but not necessarily 0. One way to infer p in this case in the spirit of maxent is to subtract from $H(p)$ or $-D(p \parallel q)$ a regularization term penalizing large values of the errors, and maximize the so obtained functional of p, without constraints. Typically, the regularization term is chosen as $\alpha \Sigma_{i=1}^k e_i^2$, where e_i is defined by (11) and $\alpha > 0$ is a regularization parameter.

This very reasonable technique is sometimes given a Bayesian interpretation which, in the opinion of this author, is questionable even for those who fully accept Bayesian philosophy. Namely, maximizing $H(p) - \alpha \Sigma e_i^2$ or $-D(p \parallel q) - \alpha \Sigma e_i^2$ is interpreted as maximizing

$$\log (\text{prior density of } p) + \log (\text{joint density of } b_i, \ldots, b_k \text{ given } p) \tag{12}$$

which amounts to interpreting the regularized maxent solution p^* as a MAP (maximum a posteriori probability) estimate of p. Supposing that the e_i's are independent Gaussian random errors with common variance σ^2, the regularization term $-\alpha \Sigma e_i^2$ gives the second term in (12) multiplied by $2\alpha\sigma^2$. To make $H(p)$

or $-D(p \parallel q)$ give the first term in (12) multiplied by the same factor, an *entropic prior* proportional to $\exp(\alpha^1 H(p))$ or $\exp(-\alpha^1 D(p \parallel q))$ is assigned to p, where $\alpha^1 = (2\alpha\sigma^2)^{-1}$. An immediate objection is that in the general case what is meant by a prior density of p ? What is the underlying measure in the space of functions on $\mathcal{X}$? If $\mathcal{X}$ is finite and p represents the empirical distribution of a sequence $x_1, \ldots, x_n$, a prior density for p proportional to $\exp(\alpha^1 H(p))$ or $\exp(-\alpha^1 D(p \parallel q))$ is a reasonable approximation, on account of (7), (9), specifically with $\alpha^1 = n$. In this case, the MAP interpretation of the regularized maxent solution p^* is justified, but only when the regularization parameter α specifically equals $(2n\sigma^2)^{-1}$.

5. Mathematical properties. Information geometry

Recall Problems (i), (ii), (iii) stated in Section 1, and their maxent solution defined there. Here we discuss some mathematical results about this solution, including the problem of its existence, when $\mathcal{X}$ and the given measure λ on $\mathcal{X}$ are arbitrary. Suppose first that F is defined by linear constraints, cf.(1). Given the functions $a_i(x)$ in (1) and the default model q, let Θ denote the set of those vectors $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$ for which

$$Z(\vartheta) = \int q(x) \exp \sum_{i=1}^{k} \vartheta_i a_i(x) \lambda(dx) \tag{13}$$

is finite; for Problem (i), $q(x) = 1$ is set. For Problem (i) or (ii), consider the exponential family of densities

$$p_\vartheta(x) = Z(\vartheta)^{-1} q(x) \exp \sum_{i=1}^{k} \vartheta_i a_i(x), \quad \vartheta \in \Theta. \tag{14}$$

For Problem (iii), $p_\vartheta(x)$ is defined similarly but without the factor $Z(\vartheta)^{-1}$.

Theorem 5.1 If $p_\vartheta \in F$ for some $\vartheta \in \Theta$ then this p_ϑ is the maxent solution p^*. If $p_\vartheta \notin F$ for all $\vartheta \in \Theta$ then the maxent solution does not exist, providing

$$\{x : p(x) > 0\} = \{x : q(x) > 0\} \text{ for some } p \in F. \tag{15}$$

Next, for any convex feasible set F write

$$H(F) = \sup_{p \in F} H(p) \tag{16}$$

for Problem (i), and

$$D(F \parallel q) = \inf_{p \in F} D(p \parallel q) \tag{17}$$

for Problems (ii), (iii). The following theorem and corollary comprise results of Csiszár [10], [5] and Topsoe [11]; Problem (iii) was not considered there, but the proofs easily extend to that case.

Theorem 5.2. For either of Problems (i), (ii), (iii), providing $H(F)$ resp. $D(F \parallel q)$ is finite, there exists a unique function p_0, not necessarily in F, such that for each $p \in F$

$$H(p) \leq H(F) - D(p \parallel p_0) \tag{18}$$

respectively

$$D(p \parallel q) \geq D(F \parallel q) + D(p \parallel p_0). \tag{19}$$

If F is of form (1) and condition (15) holds then $p_0 = p_\vartheta$ for some $\vartheta \in \Theta$.

Corollary. (a) A maxent solution p^* exists iff $p_0 \in F$, in which case $p^* = p_0$, thus p^* is unique. (b) For any sequence of functions $p_n \in F$ with $H(p_n) \to H(F)$ resp. $D(p_n \parallel q) \to D(F \parallel q)$, we have $D(p_n \parallel p_0) \to 0$. (c) A sufficient condition for $p_0 \in F$, i.e., for the existence of the maxent solution, is the closedness of F in $L_1(\lambda)$. If F is of form (1) and (15) holds, the weaker condition that Θ is an open subset of R^k is also sufficient.

An example of non-existense of the maxent solution is Problem (i) for the set of densities on the real line

$$F = \{p : \int xp(x)dx = 0, \int x^2p(x)dx = 1, \int x^3p(x)dx = 1\}.$$

The integral (13) with $q(x) = 1$, $a_1(x) = x$, $a_2(x) = x^2$, $a_3(x) = x^3$ can be finite only if $\vartheta_3 = 0$, then the densities (14) (with $q(x) = 1$) are Gaussian. As no Gaussian density belongs to F, the non-existence follows from Theorem 5.1. In this example, the p_0 of Theorem 5.2 will be the Gaussian density with mean 0 and variance 1.

It follows from Theorem 5.1 that if the maxent solution exists for Problem (ii) or (iii) with F defined by linear constraints, then eq. (19) holds with equality. In other words, if the I-projection p^* of q onto F exists then

$$D(p \parallel q) = D(p \parallel p^*) + D(p^* \parallel q) \quad \text{for each} \quad p \in F. \tag{20}$$

Actually, this holds even without the hypothesis (15).

The identity (20) is an analogue of the Pythagorean Theorem of Euclidean geometry, I-divergence regarded as an analogue of squared Euclidean distance. Indeed, if p^* were the Euclidean projection of a point q onto a plane F and D would denote squared Euclidean distance, (20) would be just the Pythagorean Theorem for the right-angled triangle pqp^*. Notice that if F is an arbitrary convex set, and the I-projection p^* of q onto F exists, (20) always holds with $\geq$ rather than $=$, cf. (19); this, too, is analogous to a result in Euclidean geometry with D=squared distance.

One simple but important consequence of the Pythagorean identity (20) is that if $F_1 \subset F$, the same $p = p_1^*$ attains both $\min_{p \in F_1} D(p \parallel q)$ and $\min_{p \in F_1} D(p \parallel p^*)$, providing either minimum is attained. Thus I-projection has the

Transitivity property: The I-projection onto $F_1 \subset F$ of the I-projection of q onto F equals the I-projection of q onto F_1, if F is defined by linear constraints.

For arbitrary F_1 and F_2 , both defined by linear constraints, let p_1^* be the I-projection of q onto F_1 and p_2^* the I-projection of p_1^* onto F_2. In general, p_2^* will differ from the I-projection of q onto F_2. However, taking the I-projection p_3^* of p_2^* onto F_1, then the I-projection p_4^* of p_3^* onto F_2, etc., we obtain a sequence $\{p_n^*\}$ that converges to the I-projection of q onto $F = F_1 \cap F_2$, again in analogy with Euclidean geometry. The geometric view turned out very useful in proving the convergence of this and related iterative algorithms, [10], [12], [13] (to be precise, a convergence proof of the I-projection iteration is available for the case when $\mathcal{X}$ is finite).

Simplest to determine I-projection is when the functions $a_1(x), \ldots, a_k(x)$ that define F by (1) are indicator functions of disjoint sets $A_1, \ldots, A_k$ with $\cup A_i = \mathcal{X}$, i.e., the constraints on p are that $\int_{A_i} p(x)\lambda(dx) = b_i, i = 1, \ldots, k$. This is because of the obvious

Scaling property: The I-projection of any q onto F as above is given by $p^*(x) = \lambda_i q(x)$ if $x \in A_i$, where $\lambda_i = b_i / \int_{A_i} q(x)\lambda(dx)$.

If the $a_i(x)$ are indicator functions of non-disjoint sets A_i, such as in the problem of inferring a bivariate probability mass function $p(x, y)$ when only its marginals $p_1(x)$ and $p_2(y)$ and a default model $q(x, y)$ are known, one may apply the above method of iterative I-projections. This amounts to *iterative scaling*, an algorithm that had been used in various fields well before Kullback pointed out that it converged to the maxent solution.

Let us mention one more "geometric" result, not directly related to maxent. It says that channel capacity may be regarded as "information radius", and plays a key role in the theory of universal data compression, [14].

Consider an arbitrary family of densities on $\mathcal{X}$, represented as $\{p_\vartheta(x), \vartheta \in \Theta\}$ as in Section 4, but now Θ may be any set, not necessarily finite dimensional. Let C denote the supremum of the mutual information $E_\pi D(p_\vartheta \parallel \bar{p})$ (where $\bar{p}(x) = E_\pi p_\vartheta(x)$) with respect to the choice of the prior π. Further, for any density q on $\mathcal{X}$, let

$$r(q) = \sup_{\vartheta \in \Theta} D(p_\vartheta \parallel q) \tag{21}$$

denote the "radius" of the smallest "I-divergence ball with center q" that contains the given family.

Theorem 5.3: C is finite iff $r(q)$ is finite for some q. In that case

$$C = \min_q r(q) \tag{22}$$

where the minimum is attained for a unique q^*. Moreoever, the maximum mutual information is attained for a prior π iff the corresponding marginal of x equals q^*, i.e., $E_\pi p_\vartheta(x) = q^*(x)$.

6. Other entropy functionals

In the literature, methods similar to maxent but based on other "entropy functionals" have also been suggested. In spectrum analysis, "maximum entropy" usually means maximizing $\int \log p(x)dx$, the Burg entropy of the power spectrum p or, if a non-constant default model q is given, minimizing $\int(\log \frac{q(x)}{p(x)} + \frac{p(x)}{q(x)} - 1)dx$, the Itakura-Saito distance of p from q.

The author is indebted to John Burg [15] for having pointed out that he does not consider $\int \log p(x)dx$ as an entropy functional on its own right. His idea in developing maximum entropy spectral analysis was to maximize the (Shannon) entropy rate of a stochastic process under spectral constraints, and he was using Kolmogorov's formula for the entropy rate of a Gaussian process in terms of its power spectrum.

Without entering philosophy, let us formally define ([16], [17], [18]) the f-entropy of a function $p(x) \geq 0$ as

$$H_f(p) = -\int f(p(x))\lambda(dx), \tag{23}$$

where $f(t)$ is any strictly convex differentiable function on the positive reals, and a corresponding measure of distance of p from q as

$$B_f(p,q) = \int [f(p(x)) - f(q(x)) - f'(q(x))(p(x) - q(x))]\lambda(dx). \tag{24}$$

As strict convexity implies $f(s) > f(t) + f'(t)(s-t)$ for any positive numbers $s \neq t$, $B_f(p,q)$ is a distance in the same sense as I-divergence is, cf. Section 1. It is called Bregman distance, [19].

As examples, take $t\log t$, $-\log t$, or t^2 for $f(t)$. Then $H_f(p)$ will be the Shannon entropy, Burg entropy, or the negative square integral of p. $B_f(p,q)$ will be I-divergence, Itakura-Saito distance, and squared $L_2(\lambda)$-distance, respectively.

Theorem 5.2 remains valid if $f(t) = t\log t$ is replaced by another f, except that the $L_1(\lambda)$-closedness of F will be sufficient for $p_o \in F$, and hence for the existence of $p^* \in F$ attaining $\max_{p\in F} H_f(p)$ or $\min_{p\in F} B_f(p,q)$, only if f has the following property:

$$\inf_{t\geq 1}(f'(Kt) - f'(t)) > 0 \quad \text{for some} \quad K > 1, \tag{25}$$

[20]. Notice that $f(t) = t\log t$ does have this property but $f(t) = -\log t$ does not. Indeed, the maximum of Burg entropy need not be attained even if $\mathcal{X} = [0,1]$ and F is defined by linear constraints with continuous functions $a_i(x)$, [18].

The Pythagorean identity (20) with D replaced by B_f, for p^* minimizing $B_f(p,q)$ subject to $p \in F$, where F is of form (1), will always hold if

$$\lim_{t\to 0} f'(t) = -\infty. \tag{26}$$

Then, of course, the transitivity property stated in Section 5 for I-projections, will also hold for "B_f−projections".

Notice that if f does not satisfy (26) then it can be extended to a (strictly convex, differentiable) function defined on the whole real line. With such as extension, $B_f(P, Q)$ can be defined for P and Q taking negative values, as well. Now, in (1) it has been tacitly assumed that the feasible functions P are non-negative, but if not needed for $B_f(P, Q)$ to be defined, this assumption might be dropped. Then the analogue of (20) will hold unconditionally, but with P^* that may take negative values, as well. This highlights why least squares (minimization of L_2 distance), a preferred method of inference for real-valued functions, is not recommended for inferring functions constrained to be non-negative.

Another family of distances for non-negative functions, with properties somewhat similar to I-divergence, are the f-divergences

$$D_f(P \parallel Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) \lambda(dx), \tag{27}$$

with f as above, and additionally satisfying $f(1) = f'(1) = 0$. Originally, these distances were defined for probability densities, without the additional conditions on f ([21][22]).

As examples, take $t \log t - t + 1$, $-\log t + t + 1$ or $(\sqrt{t} - 1)^2$ for $f(t)$. Then $D_f(P \parallel Q)$ will be the I-divergence $D(P \parallel Q)$, the reversed I-divergence $D(Q \parallel P)$, and the squared Hellinger distance $\int (\sqrt{p} - \sqrt{q})^2 \lambda(dx)$, respectively.

Results similar to those in Theorem 5.2 can be proved also for f-divergences in the role of I-divergence, [20]. We only mention the simple fact that projections defined by D_f share the scaling property of I-projections, stated in Section 5.

7. Axiomatic approach

Whereas maxent is no doubt an attractive method, it is hard to come up with mathematical results that distinguish it unequivocally. For certain applications, mostly for those in statistical physics, the large deviation arguments in Section 3 provide a very strong justification. Apparently, no similarly convincing arguments are available for the majority of applications, in particular for those to Problem (iii).

One difficulty is that the problem of inferring a function from insufficient information, stated in Section 1 in a deliberately vague form, does not easily admit a mathematical formulation, with well defined optimality criteria. Under such circumstances, an axiomatic approach appears most promising: Start from certain properties that a "good" method of inference is supposed to have, and investigate whether the postulated properties uniquely characterize a distinguished method, or else what alternatives come into account. Such an approach was first put forward by Shore and Johnson [23]. The currently available best axiomatic results are those of Csiszár [24]. Below some of these are briefly reviewed, restricted to Problem (iii), for simplicity.

In this Section we assume that $\mathcal{X}$ is a finite set of size $|\mathcal{X}| \geq 3$, and restrict attention to strictly positive functions on $\mathcal{X}$.

By an inference method for Problem (iii), we mean any rule that assigns to every feasible set F defined by linear constraints, and any default model q, an inferred function $p^* \in F$, denoted by $p^*(F, q)$. It should be emphasized that it is *not* postulated that p^* is the minimizer of some measure of distance of p from q. This does follow from the regularity and locality axioms below, but it is to be proved, which represents a major step of the approach.

Regularity axiom: If $F_1 \subset F$ and $p^*(F, q) \in F_1$ then $p^*(F_1, q) = p^*(F, q)$.

Some technical assumptions additionally postulated in [24] as part of this axiom are omitted here.

Locality axiom: If the constraints defining F can be partitioned into two sets, the first resp. second involving functions $a_i(x)$ that vanish for $x \notin \mathcal{X}_1$ resp. for $x \in \mathcal{X}_1$, where $\mathcal{X}_1$ is some subset of $\mathcal{X}$, then the values of $p^*(F, q)$ for $x \in \mathcal{X}_1$ depend only on the first set of constraints and on the values of q for $x \in \mathcal{X}_1$.

The regularity axiom formalizes the intuitive idea that if the inference based on some knowledge happens to be consistent also with some additional knowledge then the new knowledge provides no reason to change that inference. The locality axiom means that if the available knowledge consists of pieces pertaining to disjoint subsets of $\mathcal{X}$ then on each of these subsets, the inference must be based on the knowledge pertaining to that subset.

We recall the transitivity and scaling properties of I-projection, cf. Section 5. These might be used as axioms, requiring them to hold for our $p^* = p^*(F, q)$ in the role of I-projection. Instead of the scaling properly, we will consider a weaker version of it.

Weak scaling axiom: If $F = \{p : p(x_1) + p(x_2) = t\}$ for some x_1, x_2 in $\mathcal{X}$ and $t > 0$, then $p^* = p^*(F, q)$ satisfies $p^*(x_i) = \lambda q(x_i), i = 1, 2,\ \lambda = t(q(x_1) + q(x_2))^{-1}$.

An even weaker postulate is the

Semisymmetry axiom: For F as above, if $q(x_1) = q(x_2)$ then $p^*(x_1) = p^*(x_2)$.

Theorem 7.1: The regularity, locality, transitivity and weak scaling axioms are satisfied iff $p^*(F, q)$ is the I-projection of q onto F. If weak scaling is replaced by semisymmetry, the possible inference rules will be those where $p^*(F, q)$ minimizes the Bregman distance $B_f(p, q) = \Sigma(f(p(x)) - f(q(x)) - f'(q(x))(p(x) - q(x)))$ for some f satisfying (26). If transitivity is dropped but weak scaling is retained, $p^*(F, q)$ will be the minimizer of the f-divergence $D_f(p \parallel q) = \Sigma q(x) f(\frac{p(x)}{q(x)})$, for some f satisfying (26).

The maxent solution (i.e., I-projection) can be uniquely characterized also by postulating, in addition to regularity and locality, a consistency property for inferring a bivariate function from its marginals, similar to the "system independence" axiom of [23] (this, however, appears less compelling for Problem (iii) than for Problem (ii) treated in [23]).

Another natural postulate might be the *scale-invariance* of inference, i.e., that if $b_1, \ldots, b_k$ in the definition (1) of F and the default model q are multiplied by a constant, the effect on $p^*(F, q)$ be just multiplication by the same constant. This

is automatically fulfilled by f-divergence minimization. On the other hand, we have

Theorem 7.2 Minimizing a Bregman distance gives scale invariant inference iff $B_f(p,q)$ corresponds to one of the following functions:

$$f_1(t) = t\log t, f_0(t) = -\log t, \quad f_\alpha(t) = -\frac{1}{\alpha}t^\alpha, \ 0 < \alpha < 1 \text{ or } \alpha < 0.$$

Among the alternatives to maxent, the most promising might be the minimization of a Bregman distance as in Theorem 7.2. Recall that $f_0(t)$ gives the Itakura-Saito distance; of course, $f_1(t)$ gives I-divergence. Practical success with $f_\alpha(t)$, with a small positive α, has been reported by Jones and Trutzer [17].

The axiomatic approach may help to recognize those situations when maxent should not be used. E.g., the regularity axiom is inappropriate when such a function should be selected from the feasible set F that "best represents" the functions in F. Then a more adequate method than maxent might be to select that $q \in F$ which minimizes the information radius $r(q) = \sup_{p\in F} D(p \parallel q)$, cf. (21) (barycenter method, Perez [25], suitable for F as in (1) only if $\mathcal{X}$ is finite.)

The locality axiom is inappropriate if the inferred function p^* is required to be smooth in some sense. This is more typical in the continuous case, but a discrete version of smoothness may also be a natural requirement. One way to deal with this problem is to modify maxent by a regularization term that penalizes departures from the desired smoothness, similarly to regularization for constraints subject to errors, cf. Section 4. Of course, it may well be that both kinds of regularization are needed.

8. Maximum entropy in the mean (MEM)

Space permits to give but a flavor of MEM, an extension of maxent developed by French researchers, [26], [27], [28], [29]. Though MEM is not restricted to non-negative functions, we keep this restriction here. As a more serious simplification, we restrict attention to functions on a finite set of size m. This preempts significant features of the theory but still affords remarkable conclusions, some apparently new. For convenience, the functions considered for MEM inference will be regarded as m-vectors (with non-negative components), $u = (u_1, \ldots, u_m) \in R_+^m$.

The MEM idea is to visualize the vector to be inferred as the expectation of some probability distribution on R_+^m. Given a reference measure λ on R_+ (typically the Lebesgue measure, or the counting measure on a countable subset of R_+), associate with each $u \in R_+^m$ the set of all probability densities with respect to λ^m whose expectation is equal to u:

$$E(u) = \{p : \int x_i p(x_1, \ldots, x_m)\lambda(dx_1)\ldots\lambda(dx_m) = u_i, \ i = 1, \ldots, m\}. \tag{28}$$

To infer u knowing only its membership to a feasible set $F \subset R_+^m$, MEM suggests to take the feasible set of densities $\tilde{F} = \cup_{u\in F} E(u)$, determine the maxent density $p^* \in \tilde{F}$, and declare its expectation u^* to be the solution.

One easily sees from Theorem 5.1 that if λ=Lebesgue measure then

$$\max_{p \in E(u)} H(p) = \sum_{i=1}^{m} (\log u_i + 1), \tag{29}$$

attained for $p(x_1, \ldots, x_m) = \Pi_{i=1}^{m} u_i^{-1} \exp(x_i/u_i)$. The MEM solution u^* is attained by maximizing (29) subject to $u \in F$, which is the same as maximizing Burg entropy.

Suppose next that a default model $v = (v_1, \ldots, v_m)$ with positive components is also given, and let $q(x_1, \ldots, x_m)$ be a corresponding default model for densities, again with respect to λ^m. Then one can proceed as before, taking for p^* the I-projection of q onto $\tilde{F}$. As any density in $E(v)$ might be chosen as q, the MEM idea does not lead to a unique inference method, even when λ is fixed. Rather, any assignment $v \mapsto q_v \in E(v)$ gives rise to a different method. We will consider assignments of product form

$$q_v(x_1, \ldots, x_m) = \prod_{i=1}^{m} q_{v_i}(x_i) \tag{30}$$

where $\{q_t(x),\ t > 0\}$ is a given family of one-dimensional densities, with expectations

$$\int x q_t(x) \lambda(dx) = t. \tag{31}$$

For any choice of q, if the I-projection p_u^* of q onto $E(u)$ exists, the MEM solution u^* will be the minimizer of $D(p_u^* \parallel q)$ subject to $u \in F$. For $q = q_v$ as in (30), Theorem 5.1 implies that if

$$p_{\vartheta_1, \ldots, \vartheta_m}(x_1, \ldots, x_m) = \prod_{i=1}^{m} (Z_{v_i}(\vartheta_i))^{-1} q_{v_i}(x_i) \exp(\vartheta_i x_i) \tag{32}$$

belongs to $E(u)$ then $p_u^* = p_{\vartheta_1, \ldots, \vartheta_m}$. Here

$$Z_t(\vartheta) = \int q_t(x) \exp(\vartheta x) \lambda(dx) \tag{33}$$

is the moment generating function of the distribution with density $q_t(x)$. The condition $p_{\vartheta_1, \ldots, \vartheta_m} \in E(u)$ is equivalent to

$$\frac{\partial}{\partial \vartheta} \log Z_{v_i}(\vartheta)|_{\vartheta = \vartheta_i} = u_i,\ i = 1, \ldots, m. \tag{34}$$

Thus, assuming that the equations (34) have solutions $\vartheta_1, \ldots, \vartheta_m$, we obtain that

$$D(p_u^* \parallel q) = \sum_{i=1}^{m} (\vartheta_i u_i - \log Z_{v_i}(\vartheta_i)). \tag{35}$$

Two types of choice of the family $\{q_t(x), t > 0\}$ deserve special attention.

(i) Let $q_1(x)$ be given, with expectation 1, and let

$$q_t(x) = (Z_1(\vartheta_t))^{-1} q_1(x) \exp(\vartheta_t x), \quad t > 0 \tag{36}$$

where $Z_1(\vartheta)$ is defined by (33) and ϑ_t is determined to make (31) hold, i.e.,

$$\frac{d}{d\vartheta} \log Z_1(\vartheta)|_{\vartheta=\vartheta_t} = t. \tag{37}$$

It is assumed that (37) has a solution ϑ_t for every $t > 0$, this also ensures that the equations (34) have solutions $\vartheta_1, \ldots, \vartheta_m$. Let

$$f(t) = t\vartheta_t - \log Z_1(\vartheta_t) \tag{38}$$

be the Legendre transform or convex conjugate of the convex function $\log Z_1(\vartheta)$. It is a convex function of $t > 0$ with $f'(t) = \vartheta_t$. A trite calculation shows that in case of (36) the right hand side of (35) equals the Bregman distance $B_f(u, v)$. Thus MEM with the assignment $v \mapsto q_v$ defined by (30), (36) leads to Bregman distance minimization.

(ii) Given an infinitely divisible distribution on R_+ with expectation 1, let $Z_1(\vartheta)$ be its moment generating function. Then for each $t > 0, (Z_1(\vartheta))^t$ is the moment generating function of a distribution with expection t. Let $\{q_t(x), t > 0\}$ be the family of densities of these distributions (with a suitable choice of λ). Since now $Z_t(\vartheta) = (Z_1(\vartheta))^t$, we obtain that the right hand side of (35) equals the f-divergence of u from v, with f as in (38). Thus MEM with the assignment $v \to q_v$ defined by (30) with the present $\{q_t(x), t > 0\}$ leads to f-divergence minimization.

Examples. 1. Let λ be the counting measure on the natural numbers, and $q_1(x) = (ex!)^{-1}$. Then $Z_1(\vartheta) = \exp(e^\vartheta - 1), \vartheta_t = \log t, f(t) = t \log t - t + 1$. Now the family $\{q_t(x), t > 0\}$ consists of the Poisson distributions in both cases (i) and (ii) above, and MEM leads to regular maxent.

2. Let λ be the Lebesgue measure on R_+, and $q_1(x) = e^{-x}$. Then $Z_1(\vartheta) = (1 - \vartheta)^{-1} (\vartheta < 1)$, $\vartheta_t = 1 - \frac{1}{t}$, $f(t) = t - 1 - \log t$. Now the family $\{q_t(x),\ t > 0\}$ consists of the exponential densities $t^{-1} \exp(x/t)$ in case (i), and of the Γ densities $\gamma_t(x) = x^{t-1} e^{-x}/\Gamma(t)$ in case (ii); thus with these choices in (30), MEM leads to minimizing Itakura-Saito distance resp. reversed I-divergence, cf. Section 6.

3. Let λ be the Lebesgue measure on R_+ plus a unit mass at 0, and consider a compound Poisson distribution with λ-density $q_1(x)$ given by

$$q_1(0) = e^{-\tau},\ q_1(x) = \sum_{k=1}^{\infty} \tau^k e^{-\tau} \gamma_{k/\tau}(x)/k! \ \ (x > 0). \tag{39}$$

Here $\gamma_t(x)$ denotes Γ-density as above, and $\tau > 0$ is a parameter. Then $Z_1(\vartheta) = \exp\{\tau[(1-\vartheta)^{1/\tau} - 1]\}\ (\vartheta < 1)$, $\vartheta_t = 1 - t^{-\tau/1+\tau}$, $f(t) = t - (1+\tau)t^{1/1+\tau} + \tau$. Thus in case (i), MEM leads to minimizing a Bregman distance as in Theorem 7.2, with $0 < \alpha < 1$ ($\alpha = (1 + \tau)^{-1}$), and among the f-divergences whose minimization we are lead to in case (ii) is the Hellinger distance ($\tau = 1$), cf. Section 6.

It is remarkable that via MEM, several alternatives to maxent can be arrived at within the maxent paradigm. Still, the implications of this remain debatable, except for those cases when the MEM idea can be justified by a physical model.

References

1. S. Kullback, *Information Theory and Statistics*, John Wiley and Sons, "New York, 1959.
2. E.T. Jaynes (R.D. Rosenkrantz ed.), *Papers on Probability, Statistics and Statistical Physics, Reidel, Dordrecht, 1983.*
3. *I. Csiszár and J. K{orner, Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.
4. J.M. Van Campenhout and T. Cover, "Maximum entropy and conditional probability," *IEEE Trans. Inform. Theory*, **27**, 483-489, 1981.
5. I. Csiszár, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Probability*, **12**, 768-793, 1984.
6. I. Csiszár, "An extended maximum entropy principle and a Bayesian justification (with discussion)," *Bayesian Statistics 2*, J.M. Bernardo et al., pp. 83-89, North-Holland, Amsterdam, 1985.
7. J.M. Bernardo, "Reference posterior for Bayesian inference (with discussion)," *J. Roy. Statist. Soc. B*, **41**, 113-147, 1979.
8. B. Clarke and A.R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Statist. Planning and Inference*, **41**, pp. 37-60, 1994.
9. J.M. Bernardo and A.F.M. Smith, "Bayesian Theory," John Wiley and Sons, New York, 1994.
10. I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probability*, **3**, pp. 146-158, 1975.
11. F. Topsoe, "Information theoretical optimization techniques", *Kybernetika*, **15**, pp. 7-17, 1979.
12. I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statist. Decisions*, **Suppl. 1**, pp. 205-237, 1984.
13. I. Csiszár, " A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling," *Ann. Statist.*, **17**, pp. 1409-1413, 1989.
14. L.D. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, **26**, pp. 166-174, 1980.
15. J. Burg, "Personal Communication," 1995.
16. C. R. Rao and T. K. Nayak, "Cross entropy, dissimilarity measures, and characterization of quadratic entropy", *IEEE Trans. Inform. Theory*, **31**, pp. 589-593, 1985.
17. L. Jones and V. Trutzer, "Computationally feasible high-resolution minimum-distance procedures which extend the maximum-entropy method," *Inverse Problems*, **5**, pp. 749-766, 1989.
18. J. M. Borwein and A. S. Lewis, "Partially-finite programming in L_1 and the existence of maximum entropy estimates," *SIAM J. Optimization*, **3**, pp. 248-267, 1993.
19. L.M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR Comput. Math. and Math. Phys.*, **7**, pp. 200-217, 1967.
20. I. Csiszár, "Generalized projections for non-negative functions", *Acta Math. Hungar.*, **68**, pp. 161-185, 1995.
21. I. Csiszár, "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizit{at von Markoffschen Ketten," *Publ. Math. Inst. Hungar. Acad. Sci*, **8**, pp. 85-108, 1963.
22. S.M. Ali and S.D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Roy. Statist. Soc. Ser. B*, **28**, pp. 131-142, 1966.
23. J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*, **26**, pp. 26-37, 1980.
24. I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Ann. Statist.*, **19**, pp. 2032-2066, 1991.
25. A. Perez, "Barycenter" of a set of probability measures and its application in statistical decision, *Compstat Lectures*, pp. 154-159, Physica, Heidelberg, 1984.
26. J. Navaza, "The use of non-local constraints in maximum-entropy electron density reconstruction," *Acta. Crystallographica*, **A42**, pp. 212-223, 1986.
27. D. Dacunha-Castelle and F. Gamboa, "Maximum d'entropie et problème des moments,"

Ann. Inst. H. Poincarè, **4**, pp. 567-596, 1990.

28. F. Gamboa and G. Gassiat, "Bayesian methods and maximum entropy for ill posed inverse problems," *Ann. Statist.*, Submitted, 1994.
29. J.F. Bercher G. LeBesnerais and G. Demoment, "The Maximum Entropy on the Mean, Method, Noise and Sensitivity", in *Proc. 14th Int. Workshop Maximum Entropy and Bayesian Methods*, S. Sibisi and J. Skilling, Kluwer Academic, 1995.