Invariance-Based Justification of the Maximum Entropy Method and of Generalized Maximum Entropy Methods in Data Processing

Hung T. Nguyen[1], Olga M. Kosheleva[2], and Vladik Kreinovich[3]

[1]Department of Mathematical Sciences, New Mexico State University
Las Cruces, NM 88003, USA, hunguyen@nmsu.edu
[2]Department of Electrical and Computer Engineering and
[3]Department of Computer Science, University of Texas at El Paso
El Paso, TX 79968, USA, olga@ece.utep.edu, vladik@cs.utep.edu

Abstract

Maximum entropy method and its generalizations are very useful in data processing. In this paper, we show that these methods naturally follow from reasonable invariance requirements.

1 Introduction to the Problem

Traditional statistical data processing methods assume that we know the probability distribution of the corresponding errors. This distribution can be given, e.g., in terms of a (multi-D) probability density function $\rho(\vec{x})$. In many real-life situations, we do not know the exact type of the distribution; at best, we have partial information about this distribution, which is consistent with many different distributions. If we use different distributions, we may get drastically different data processing results. In such situations, in order to apply appropriate statistical techniques, it is desirable to select, from the set of all distributions which are consistent with our knowledge, a distribution which is in some reasonable sense "the best".

There exist many successful semi-heuristic methods for selecting such a distribution. One of the most well known is the the *maximum entropy* method (see, e.g., [8]), according to which we select, from all the probability distributions which are consistent with the observations, the distribution for which the entropy is the largest possible:

$$-\int \rho(\vec{x}) \cdot \log(\rho(\vec{x}))\, \mathrm{d}\vec{x} \to \max . \tag{1}$$

In the discrete case, when a probability distribution is characterized by the probabilities $p_1, \ldots, p_n$ of different alternatives, the maximum entropy method means selecting a distribution for which

$$-\sum_{i=1}^{n} p_i \cdot \log(p_i) \to \max . \tag{2}$$

In some real-life situations, it turns out that better results can be obtained if, instead of maximizing entropy, we maximize entropy-like functions called *generalized entropies*. The first such function $\sum p_i^p$ for some real number p, was introduced by Rényi in 1960 [24] (see also [25, 26]); in [5], another function $\sum \log(p_i)$ was introduced by Burg [5] (see also [6]). Maximization of these functions leads to *generalized maximum* *entropy methods*, in which we choose the distribution $\rho(\vec{x})$ for which

$$\int \log(\rho(\vec{x}))\,\mathrm{d}\vec{x} \to \max \tag{3}$$

or

$$\int (\rho(\vec{x}))^p\,\mathrm{d}\vec{x} \to \max \tag{4}$$

for some real number p.

The generalized maximum entropy methods have been successfully used in geophysics [5, 6], in coding theory [1, 2, 3], in speech processing [13], in signal restoration [4], in computerized tomography [7, 14, 23], in radar imaging and planetary radar imaging [11, 12, 20], and in many other application areas.

Since these methods often work well, the natural question is: why? In [27], p. 229, Gian-Carlo Rota expresses the opinion of many mathematicians in the following words:

> *The maximum entropy principle is one of the hot potatoes of our day. It has not yet split the world of statistics as has Bayes' law, but no one has yet succeeded in finding a justification for it. Maybe we should make it one of the axioms of statistics.*

In several application areas, there is an application-specific justification of these techniques; see, e.g., [1, 2, 3, 5, 6, 18, 16, 17].

In [15, 22], we used fuzzy logic techniques to show that these methods are, in some reasonable sense, the best in describing the natural common-sense ideas.

In this paper, we provide a new justification of the maximum entropy method and of generalized maximum entropy methods. a justification which is both general (i.e., application-independent) and objective (i.e., does not depend on the expert's opinions and heuristics). Namely, we prove that, if we take into consideration natural symmetries, then the natural description of our preferences leads exactly to maximum entropy and generalized maximum entropy techniques.

2 Towards Mathematical Formulation

How to describe preferences? There exists a general formalism. In order to describe what is the *best*, we must describe what is *better*, i.e., we must describe *preferences*. The necessity to describe preferences, i.e., to describe the *utility* of different alternatives for different people, is extremely important in decision making, including decision making under conflict (also known under a somewhat misleading name of *game theory*). To describe these preferences (utilities), a special *utility theory* has been developed; see, e.g., [9, 19, 21, 28].

The mathematical formalism of utility theory comes from the observation that sometimes, when a person faces several alternatives $A_1, \ldots, A_n$, instead of choosing one of these alternatives, this person may choose a *probabilistic* combination of them, i.e., A_1 with probability p_1, A_2 with a probability p_2, etc. For example, if two alternatives are of equal value to a person, that person will probably choose the first one with probability 0.5 and the second one with the same probability 0.5. Such probabilistic combinations are called (somewhat misleadingly) *lotteries*. In view of this realistic possibility, it is desirable to consider the preference relation not only for the original alternatives, but also for arbitrary lotteries combining these alternatives. Each original alternative A_i can be viewed as a *degenerate* lottery, in which this alternative A_i appears with probability 1, and every other alternative $A_j \neq A_i$ appear with probability 0.

The main theorem of utility theory states that if we have an ordering relation $L_1 \succ L_2$ between such lotteries (with the meaning "L_1 is preferable to L_2"), and if this relation satisfies natural consistency conditions such as transitivity, etc., then there exists a function u from the set $\mathcal{L}$ of all possible lotteries into the set R of real numbers for which:

- $L_1 \succ L_2$ if and only if $u(L_1) > u(L_2)$, and
- for every lottery L, in which each alternative A_i appears with probability p_i, we have

$$u(L) = p_1 \cdot u(A_1) + \ldots + p_n \cdot u(A_n).$$

This function u is called a *utility function*. Each consistent preference relation can thus be described by its utility function.

General formalism (cont-d): how unique is the utility function? The correspondence between preference relations and utility functions is not 1-1: different utility functions may correspond to the same preference. For example, if the preference relation is consistent with the utility function $u(L)$, then, as one can easily see, it is also consistent with the function $\widetilde{u}(L) = a \cdot u(L) + b$, in which a is an arbitrary positive real number, and b is an arbitrary real number.

It is known that such linear transformations form the only non-uniqueness of utility function: namely, if two utility functions $u_1(L)$ and $u_2(L)$ describe the same preference relation, then $u_2(L) = a \cdot u_1(L) + b$ for some real numbers $a > 0$ and b.

Application to our problem. In general, preference relations can be described by utility functions. Therefore, to describe preferences between distributions, we need a utility function that is defined on the set of all possible distributions.

In particular, if we consider realistic distributions ρ that concentrate on finitely many points $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)}$, and are thus described by the probabilities $p_1, \ldots, p_n$ of these points, then we need a utility function that depends on n parameters p_i: $u(\rho) = u(p_1, \ldots, p_n)$.

What function should we choose?

Localness property. An important feature of many data processing problems is their *localness*: different parts of the probability distribution are pretty much "independent" on each other. For example, the relative quality of two possible reconstructions of a "tail" of this distribution (i.e., a part that contains large values $\vec{x}$), does not seem to depend on the remaining part of the distribution. In mathematical terms, if

- a distribution p_i is preferable to the distribution q_i that differs from p_i only in points $\vec{x}$ from some set P, and
- distributions $\widetilde{p}_i$ and $\widetilde{q}_i$ coincide with each other for $\vec{x}^{(i)} \notin P$ and with, correspondingly, p_i and q_i for $\vec{x}^{(i)} \in P$,

then $\widetilde{p} \succ \widetilde{q}$.

This "localness" ("independence") is a frequent feature in practical problems, and utility theory has developed a precise description of utility functions that satisfy this property. Namely, it has been shown that when alternatives are characterized by n parameters $z_1, \ldots, z_n$, then the localness of the preference is equivalent to the utility function $u(z_1, \ldots, z_n)$ being of one of the two types [10]:

- *additive* $u(z_1, \ldots, z_n) = u_1(z_1) + \ldots + u_n(z_n)$ for some functions $u_i(z_i)$; or
- *multiplicative* $u(z_1, \ldots, z_n) = u_1(z_1) \cdot \ldots \cdot u_n(z_n)$ for some functions $u_i(z_i)$.

In utility theory, the values $u_i(z_i)$ are called *marginal utilities*.

For distributions, parameters $z_1, \ldots, z_n$ are the probabilities $p_1, \ldots, p_n$, and the resulting forms of the utility function are $u(\rho) = \sum u_i(p_i)$ and $u(\rho) = \prod u_i(p_i)$.

The quality (utility) of a distribution should not change under a permutation, so all the functions $u_i(p)$ must be the same function; hence, $u(\rho) = \sum U(p_i)$ or $u(\rho) = \prod U(p_i)$.

Continuous case. A general distribution can be described as a limit of its finite-point approximations. The denser the points (i.e., the smaller the distances h_x and h_y between the neighboring points), the closer the discrete distribution to the continuous one. Therefore, as a utility $u(\rho)$ of a function ρ, we can take the *limit* of the utilities of its discrete representation as $h_x \to 0$ and $h_y \to 0$. How can we describe such a limit?

This limit is easy to describe for the case when utility is a *sum* of marginal utilities: in this case, the sums are, in effect, integral sums, and therefore, as the pixels get denser, the sums tend to the integral $u(\rho) = \int U(\rho(\vec{x}))\, \mathrm{d}\vec{x}$.

For the case when utility is a *product* of marginal utilities, the limit can be obtained indirectly: indeed, since utility is a product of marginal utilities, its *logarithm* is the *sum* of logarithms of marginal utilities: $v(\rho) = \log(u(\rho)) = \sum V(p_i)$, where $V = \log(U)$. For these logarithms, we also get integral sums and therefore, a reasonable limit expression: $v(\rho) = \int V(\rho(\vec{x}))\, \mathrm{d}\vec{x}$, and $u(\rho) = \exp(v(\rho))$.

Comments. Before we go into further details, let us make two comments.

- At first glance, the multiplicative case seems to lead to a more complicated formula than the additive one. However, our goal is to find an distribution $\rho(\vec{x})$ for which $u(\rho) \to \max$. Since logarithm is monotonic, the condition $u(\rho) \to \max$ is equivalent to $v(\rho) = \log(u(\rho)) \to \max$. Therefore, in multiplicative case, we get the same problem $\int V(\rho(\vec{x}))\,\mathrm{d}\vec{x} \to \max$ as in the additive case.

- Since our goal is optimization, we would like to restrict ourselves to *smooth* (differentiable) functions U and V, because for smooth functions, optimization is as easy as computing the derivatives and equating them to 0. Moreover, since many useful optimization techniques use the second derivatives as well, we will require that these functions are *twice* differentiable.

  Fortunately, we can impose this restriction without losing generality, because, as it is well known, every continuous function can be, with an arbitrary accuracy, approximated by twice differentiable functions (even by polynomials). Since we are dealing with not 100% accurate data anyway, there is no reason to represent the expert's preferences absolutely precisely. Therefore, even if the actual expert preferences are described by a non-smooth function, we can, within an arbitrary accuracy, still approximate it by a smooth function.

Scale invariance. The relative quality of two distributions should not change if we simply change the measuring units for the components $x_1, \ldots, x_d$ of the underlying vector $\vec{x}$. If we change each unit to a λ times smaller one, then the numerical value of the corresponding variable x_i will change to $x_i = \lambda \cdot x_i$. Change of units replaces the original probability distribution $\rho(\vec{x})$ by a new distribution $\widetilde{\rho}(\vec{x}) = \lambda^d \cdot \rho(\lambda \cdot \vec{x})$ (where d is the dimension of the underlying space).

It is therefore natural to require that if a distribution ρ_1 is better than the distribution ρ_2, then the re-scaled distribution $\widetilde{\rho}_1$ should still be better than the re-scaling $\widetilde{\rho}_2$ of the distribution ρ_2.

How can we express the invariance of the preference relation in terms of utility function? The fact that the preference relation does not change does not necessarily mean that the utility function is necessarily invariant, because, as we have mentioned, different utility functions can correspond to the same preference relation. What it does mean is that the utility function corresponding to re-scaled distributions must describe the same preference relation as the original utility function. We know that two utility functions describe the same preference relation if and only if they can obtained from each other by a linear transformation $u \to a \cdot u + b$; so, we arrive at the requirement that for every re-scaling λ there exist numbers $a(\lambda)$ and $b(\lambda)$ for which, for all distributions,

$$\int U(\lambda^d \cdot \rho(\lambda \cdot \vec{x}))\,\mathrm{d}\vec{x} = a(\lambda) \cdot \int U(\rho(\vec{x}))\,\mathrm{d}\vec{x} + b(\lambda).$$

Of course, we have to consider only probability distributions, i.e., only functions $\rho(\vec{x}) \geq 0$ for which $\int \rho(\vec{x})\,\mathrm{d}\vec{x} = 1$.

As a result, we arrive at the following definitions:

3 Definitions and the Main Result

Definition 1.

- *By an additive utility function, we mean an expression of the type* $u(\rho) = \int U(\rho(\vec{x}))\,\mathrm{d}\vec{x}$, *where* $\vec{x} \in R^d$, *and* $U(\rho)$ *is a twice differentiable function.*

- *By an multiplicative utility function, we mean an expression of the type* $u(\rho) = \exp(v(\rho))$, *where* $v(\rho) = \int V(\rho(\vec{x}))\,\mathrm{d}\vec{x}$ *and* V *is a twice differentiable function.*

- *By an distribution utility function* $u(\rho)$, *we mean either an additive utility function, or a multiplicative utility function.*

Definition 2. *We say that an distribution utility function* $u(\rho)$ *is scale-invariant if there exists functions* $a(\lambda)$ *and* $b(\lambda)$ *such that for every distribution* $\rho(\vec{x})$, *and for every real number* $\lambda > 0$, *we have*

$$u(\lambda^d \cdot \rho(\lambda \cdot \vec{x})) = a(\lambda) \cdot u(\rho(\vec{x})) + b(\lambda). \tag{5}$$

Definition 3. *We say that an distribution utility function $u(\rho)$ is equivalent to the functional $F(\rho)$ if for every two distributions ρ_1 and ρ_2, $u(\rho_1) > u(\rho_2)$ if and only if $F(\rho_1) > F(\rho_2)$.*

Comment. Thus, if a functional $F(\rho)$ is equivalent to the distribution utility function $u(\rho)$, then, for every case in which we have have to select between the distributions, the selection $u(\rho) \to \max$ is equivalent to the selection $F(\rho) \to \max$.

Theorem. *If an distribution utility function is scale-invariant, then it is equivalent to one of the functionals*

$$F(\rho) = \pm \int \rho(\vec{x}) \cdot \log(\rho(\vec{x}))\, \mathrm{d}\vec{x},\ F(\rho) = \pm \int \log(\rho(\vec{x}))\, \mathrm{d}\vec{x},\ \text{or } F(\rho) = \pm \int (\rho(\vec{x}))^p\, \mathrm{d}\vec{x}.$$

Comment. As a particular case of these functionals, we get the corresponding discrete formulas. Thus, natural invariance requirements justify the use of maximum entropy method (1)–(2) and of generalized maximum entropy methods (3)–(4).

4 Proof

We want to transform the scale-invariance requirement into a differential equation. For that, we will have to use the differentiability assumptions. Namely, let us take an distribution $\rho(\vec{x})$ that is positive in a certain area $\mathcal{A}$. Then, for every smooth function $\delta\rho(\vec{x})$ which is equal to 0 outside this area $\mathcal{A}$ and for which

$$\int \delta\rho(\vec{x})\, \mathrm{d}\vec{x} = 0, \tag{6}$$

we can consider a 1-parametric family of distributions $\rho_\varepsilon(\vec{x}) = \rho(\vec{x}) + \varepsilon \cdot \delta\rho(\vec{x})$ with a real parameter ε, and then use the above equality (5) for the distributions from this family.

Due to our definition of utility in terms of one of the functions U or V, and due to the assumption that functions U and V are differentiable, we can conclude that the expressions $u(\rho_\varepsilon)$, and $u(\lambda^d \cdot \rho_\varepsilon(\lambda \cdot \vec{x}))$ are differentiable with respect to ε, and their derivatives at the point $\varepsilon = 0$ can be explicitly computed.

From (5), it follows that if the derivative of $u(\rho)$ is equal to 0, then the derivative of $u(\lambda^d \cdot \rho(\lambda \cdot \vec{x}))$ must also be equal to 0.

For *additive* utility functions, the derivative of

$$u(\rho_\varepsilon) = \int U(\rho + \varepsilon \cdot \delta\rho)\, \mathrm{d}\vec{x}$$

is equal to

$$\int U'(\rho(\vec{x})) \cdot \delta\rho(\vec{x})\, \mathrm{d}\vec{x},$$

where by U', we denoted the derivative of the function $U(\rho)$.

To compute the derivative of

$$u(\lambda^d \cdot \rho(\lambda \cdot \vec{x})) = \int U(\lambda^d \cdot \rho(\lambda \cdot \vec{x}))\, \mathrm{d}\vec{x},$$

we introduce auxiliary variables $\vec{y} = \lambda \cdot \vec{x}$; then we get

$$u(\lambda^d \cdot \rho(\lambda \cdot \vec{x})) = \lambda^{-d} \cdot \int U(\lambda^d \cdot \rho(\vec{y}))\, \mathrm{d}\vec{y}.$$

Now, the desired derivative is equal to

$$\int U'(\mu \cdot \rho(\vec{y})) \cdot \delta\rho(\vec{y})\, \mathrm{d}\vec{y},$$

where we denoted $\mu = \lambda^d$. Thus, the above property means that if (6) holds and

$$\int U'(\rho(\vec{x})) \cdot \delta\rho(\vec{x})\, \mathrm{d}\vec{x} = 0, \tag{7}$$

then

$$\int U'(\mu \cdot \rho(\vec{x})) \cdot \delta\rho(\vec{x})\, \mathrm{d}\vec{x} = 0. \tag{8}$$

For *multiplicative* utility functions, we get a similar property, but with V instead of U.

The above if-then property is formulated in (not very intuitive) analytical terms, but it can be reformulated in more intuitive geometric terms if we take into consideration that all our functions are smooth and located on $\mathcal{A}$, and therefore, belong to $L^2(\mathcal{A})$. In terms of L^2, the conditions (6) and (7) mean that the function $\delta\rho$ is orthogonal to a constant function 1 (appropriately bounded outside $\mathcal{A}$, to make it an element of L^2), and to $U'(\rho(\vec{x}))$, and the conclusion means that $\delta\rho$ is orthogonal to $U'(\mu \cdot \rho(\vec{x}))$. In these terms, the above property says that every element of L^2 that is orthogonal to 1 and to $U'(\rho)$ is also orthogonal to $U'(\mu \cdot \rho)$ (it actually says so not about *every* element of L^2, but about every *smooth* element of L^2, but since smooth elements are everywhere dense in L^2, we can easily extend this property to all possible functions from L^2).

In geometric terms, it is easy to prove that if a vector v is orthogonal to every vector x that is orthogonal to two given vectors v_1 and v_2, then v belongs to the linear space generated by v_1 and v_2: indeed, otherwise, we could take a projection $\pi(v)$ of v on the orthogonal complement to that linear space; this projection is orthogonal to both v_i, but not to v.

Thus, for every $\mu > 0$, the function $U'(\mu \cdot \rho)$ is a linear combination of the functions 1 and $U'(\rho)$, i.e., $U'(\mu \cdot \rho(\vec{x})) = \alpha(\mu) + \beta(\mu) \cdot U'(\rho(\vec{x}))$ for some values $\alpha(\mu)$ and $\beta(\mu)$. This is true for all points $\vec{x}$, and therefore, this equality must be true for all possible values of ρ. Hence, the function $u'(\rho)$ must satisfy the following functional equation: for every $\mu > 0$ and for every ρ, we have

$$U'(\mu \cdot \rho) = \alpha(\mu) + \beta(\mu) \cdot U'(\rho). \tag{9}$$

We would like to use differentiability to solve the functional equation (9). The function U' is differentiable, so we need to prove the differentiability of the functions α and β. Let us do it.

Indeed, if we consider the equation (9) for two different values ρ_1 and ρ_2, and subtract the resulting equations, we conclude that $U'(\mu \cdot \rho_1) - U'(\mu \cdot \rho_2) = \beta(\mu) \cdot (U'(\rho_1) - U'(\rho_2))$, and, therefore,

$$\beta(\mu) = \frac{U'(\mu \cdot \rho_1) - U'(\mu \cdot \rho_2)}{U'(\rho_1) - U'(\rho_2)}.$$

Since the function U is twice differentiable, the right-hand side of this equality is differentiable, and so, $\beta(\mu)$ is a differentiable function.

Now, from the equation (9), we conclude that $\alpha(\mu) = U'(\mu \cdot \rho) - \beta(\mu) \cdot U'(\rho)$. Since all the terms in the right-hand side of this equality are differentiable, the function $\alpha(\mu)$ is differentiable as well.

Now, we are ready to deduce the differential equation from the functional equation (9).

Since all three functions $U'(\rho)$ (we will denote it by $W(\rho)$), $\alpha(\mu)$, and $\beta(\mu)$, are differentiable, we can differentiate both sides of the equation (9) with respect to μ and substitute $\mu = 1$. As a result, we get the following differential equation: $W'(\rho) \cdot \rho = A + B \cdot W$, where we denoted $A = \alpha'(1)$ and $B = \beta'(1)$. Hence,

$$\frac{dW}{d\rho} \cdot \rho = A + B \cdot W. \tag{10}$$

To solve the equation (10), let us first simplify it. To simplify this equation, let us separate the variables W and ρ by multiplying both sides by $d\rho/(\rho \cdot (A + B \cdot W))$; then, the equation takes the form

$$\frac{dW}{A + B \cdot W} = \frac{d\rho}{\rho}. \tag{11}$$

This equation is easy to integrate; the resulting solution is slightly different for $B = 0$ and $B \neq 0$.

If $B = 0$, then integrating both parts of (11), we get $A^{-1} \cdot W = \ln(\rho) + C_1$ ($C_1, C_2, \ldots$ will denote constants). Hence, $U'(\rho) = W = A \cdot \ln(\rho) + C_2$, and integrating again, we get $U(\rho) = A \cdot \rho \cdot \log(\rho) + C_2 \cdot \rho + C_3$ for some constants C_i.

If $C_3 \neq 0$, then the expression for $U(\rho)$ would include an infinite integral; therefore, $C_3 = 0$, and $U(\rho) = A \cdot \rho \cdot \log(\rho) + C_2 \cdot \rho$. Hence,

$$u(\rho) = A \cdot \int \rho \cdot \log(\rho)\, \mathrm{d}\vec{x} + C_2 \cdot \int \rho\, \mathrm{d}\vec{x}.$$

Since for every distribution ρ, we have $\int \rho \, \mathrm{d}\vec{x} = 1$, we have

$$u(\rho) = A \cdot \int \rho \cdot \log(\rho) \, \mathrm{d}\vec{x} + C_2,$$

and therefore, the condition $u(\rho_1) > u(\rho_2)$ is equivalent to the condition $F(\rho_1) > F(\rho_2)$ for an entropy functional $F(\rho) = \pm \int \rho \cdot \log(\rho) \, \mathrm{d}\vec{x}$ (the sign is equal to the sign of the constant A).

If $B \neq 0$, then

$$\frac{dW}{A + B \cdot W} = \frac{d(W + A/B)}{A \cdot (W + A/B)},$$

and therefore, after integrating both parts of the equation (11), we get

$$A^{-1} \cdot \ln(W + A/B) = \ln(\rho) + C_1;$$

hence $\ln(W + A/B) = A \cdot \ln(\rho) + C_2$, and so, after exponentiating, we get $W + A/B = C_3 \cdot \rho^A$. Thence, $W = U' = C_3 \cdot \rho^A + C_4$.

- If $A \neq -1$, we get $U = C_5 \cdot \rho^{A+1} + C_4 \cdot \rho + C_6$. Similarly to the case $B = 0$, we can now conclude that $C_6 = 0$, and that the corresponding utility function is equivalent to a functional $F(\rho) = \pm \int \rho^p \, \mathrm{d}\vec{x}$ (for $p = A + 1$).

- If $A = -1$, we similarly get $U(\rho) = C_5 \cdot \ln(\rho) + C_4 \cdot \rho + C_6$, in which case the utility function $u(\rho)$ is equivalent to a functional $\pm \int \log(\rho) \, \mathrm{d}\vec{x}$.

In both cases $B = 0$ and $B \neq 0$, the utility function is equivalent to one of the three functionals $F(\rho)$ from the formulation of the theorem. The theorem is thus proven.

Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209, by NSF grants No. DUE-9750858 and CDA-9522207, by United Space Alliance, grant No. NAS 9-20000 (PWO C0C67713A6), by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant number F49620-95-1-0518, and by the National Security Agency under Grant No. MDA904-98-1-0561. This work was supported in part by NSF grant CDA-9522207.

References

[1] J. Aczél, "Determination of all additive quasimetric mean codeword lengths", *Z. Wahrsch. Verw. Gebeite*, 1974, Vol. 29, pp. 351–360.

[2] J. Aczél and Z. Dároczy, *On measures of information and their characteristics*, Academic Press, New York, 1979.

[3] J. Aczél and J. Dhombres, *Functional equations in several variables*, Cambridge University Press, Cambridge, 1989.

[4] C. Auyeung and R. M. Mersereau, "A dual approach to signal restoration", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Glasgow, May 1989, pp. 1326–1329.

[5] J.P. Burg, "Maximum entropy spectral analysis", *Proc. of 37th Meeting of the Society of Exploration Geophysicists*, Oklahoma City, 1967.

[6] J.P. Burg, *Maximum entropy spectral analysis*, Ph.D. Dissertation, Department of Geophysics, Stanford University, 1975.

[7] N.J. Dusaussoy and I. E. Abdou, "The extended MENT algorithm: a maximum entropy type algorithm using prior knowledge for computerized tomography", *IEEE Transactions on Signal Processing*, 1991, Vol. 39, No. 5, pp. 1164–1180.

[8] G.J. Erickson, J.T. Rychert, and C.R. Smith (eds.), *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht, 1998.

[9] P.C. Fishburn, *Utility Theory for Decision Making*, John Wiley & Sons Inc., New York, 1969.

[10] P.C. Fishburn, *Nonlinear preference and utility theory*, The John Hopkins Press, Baltimore, MD, 1988.

[11] B.C. Flores, A. Ugarte, and V. Kreinovich, "Choice of an entropy-like function for range-Doppler processing", In: *Proceedings of the SPIE/International Society for Optical Engineering*, 1993, Vol. 1960, *Automatic Object Recognition III*, pp. 47–56.

[12] B.C. Flores, V. Kreinovich, and R. Vasquez, "Signal design for radar imaging and radar astronomy: genetic optimization", In: D. Dasgupta and Z. Michalewicz, eds., *Evolutionary Algorithms in Engineering Applications*, Springer-Verlag, Berlin-Heidelberg, 1997, pp. 406–423.

[13] R.W. Johnson and J.E. Shore. Which is the better entropy expression for speech processing: $-s\log(s)$ or $\log(s)$? *IEEE Trans on Acoustics, Speech and Signal Processing*, February 1984.

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

[15] O. Kosheleva, "Symmetry-group justification of maximum entropy method and generalized maximum entropy methods in image processing", In: G.J. Erickson, J.T. Rychert, and C.R. Smith (eds.), *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht, 1998, pp. 101–113.

[16] O.M. Kosheleva and V. Kreinovich, "A letter on maximum entropy method," *Nature*, 1979, Vol. 281, No. 5733 (Oct. 25), pp. 708–709

[17] O.M. Kosheleva and V. Kreinovich, *Utility functions that describe invariant preferences*, Technical Report, Center for New Information Technology "Informatika", Leningrad, 1989 (in Russian).

[18] O. Kosheleva, V. Kreinovich, B. Bouchon-Meunier, and R. Mesiar, "Operations with Fuzzy Numbers Explain Heuristic Methods in Image Processing", *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'98)*, Paris, France, July 6–10, 1998, pp. 265–272.

[19] D.R. Luce and H. Raiffa, *Games and Decisions, Introduction and critical survey*, John Wiley & Sons, Inc., New York, 1957.

[20] J.L. Mora, B.C. Flores, and V. Kreinovich, "Suboptimum binary phase code search using a genetic algorithm", In: S.D. Udpa and H.C. Han, eds., *Advanced Microwave and Millimeter-Wave Detectors.* Proceedings of the SPIE/International Society for Optical Engineering, Vol. 2275, San Diego, CA, 1994, pp. 168–176.

[21] R.B. Myerson, *Game theory. Analysis of conflict*, Harvard University Press, Cambridge, MA, 1991.

[22] H.T. Nguyen, V. Kreinovich, and B. Bouchon-Meunier, "Soft Computing Explains Heuristic Numerical Methods in Data Processing and in Logic Programming", preliminary version appeared in *Working Notes of the AAAI Symposium on Frontiers in Soft Computing and Decision Systems*, Boston, MA, November 8–10, 1997, pp. 40–45; final version is in L. Medsker (ed.), *Frontiers in Soft Computing and Decision Systems*, AAAI Press (Publication No. FS-97-04), 1997, pp. 30–35.

[23] M.L. Reis and N.C. Roberty, "Maximum entropy algorithms for image reconstruction from projections", *Inverse Problems*, 1992, pp. 623–644.

[24] A. Rényi, “On measures of entropy and information”, *Proc. 4th Berkeley Symposium on Math. Statistics and Probability, Berkeley, 1960*, Univ. of Calif. Press, Berkeley, CA, 1961, Vol. 1, pp. 547–561.

[25] A. Rényi, “On the foundations of information theory”, *Rev. Inst. Intern. Statist.*, 1965, Vol. 33, pp. 1–14.

[26] A. Rényi, *Probability theory*, North Holland, Amsterdam, London, New York, 1970.

[27] G.-C. Rota, *Indiscrete thoughts*, Birkhäuser, Boston, MA, 1997.

[28] P. Suppes, D.M. Krantz, R.D. Luce, and A. Tversky, *Foundations of measurement. Vol. II. Geometrical, threshold, and probabilistic representations*, Academic Press, San Diego, CA, 1989.