Reliable Computing 2 (1) (1996), pp. 63–79

Maximum entropy and interval computations

(September notes on summer impressions)

VLADIK KREINOVICH

Максимум энтропии и интервальные вычисления

(Сентябрьские заметки о летних впечатлениях)

В. КРЕЙНОВИЧ

Conference. From July 30 to August 4, 1995, I attended the Fifteenth International Workshop on Maximum Entropy Methods in Santa Fe, New Mexico, USA. This is a major international annual meeting on maximum entropy methods; the 1995 workshop attracted about 150 people from all over the world; the next one will be in South Africa.

At first glance, maximum entropy and intervals are not related. The most widely used word at this workshop was "probability", and very few speakers talked about intervals. At first glance, it seems that we and they are miles apart. So why write about them in this journal?

Because these fields *are* closely related, and since this relation is not yet a common knowledge, I would like it to be known to the interval community. Yes, *terms* that we use are different, and mathematical methods are different, but these different methods are used to solve the same application problems. In solving these problems, interval and maximum entropy methods not only do not compete, they *complement* each other. In this report, I will try to explain how.

1. The origin of maximum entropy methods: in some cases, interval computations are not sufficient

Case study: indirect measurements. To explain where maximum entropy methods come from, let us consider a typical problem for which interval computations are useful: estimating errors of indirect measurements. Suppose that we are interested in the value of a physical quantity y, and it is either difficult, or even impossible, to measure y directly. For example, in medical tomography, we may be interested in the tissue density y at a certain point inside the brain. To determine the values of the quantities that we cannot measure *directly*, we measure them *indirectly*, i.e.:

- we measure other, easier-to-measure quantities $x_1, \ldots, x_n$ that are related with y, and from which y can be reconstruct by means of some known algorithm f, as $y = f(x_1, \ldots, x_n)$; and then,

- we compute an estimate $\tilde{y}$ for y from the results $\tilde{x}_1, \ldots, \tilde{x}_n$ of measuring x_i, as $\tilde{y} = f(\tilde{x}_1, \ldots, \tilde{x}_n)$.

In the above medical example, we scan the brain in different directions with, say, ultrasound, measure the signals x_i that have passed through the brain, and find y by solving the corresponding integral equation (in this example, f is a numerical algorithm for solving this equation).

Since measurements are never absolutely accurate, the measured values $\tilde{x}_i$, generally speaking, differ from the actual values x_i. Additional uncertainty is caused by the fact that the algorithm f often gives only an *approximate* solution to the desired equation. However, even if f is exact (i.e., if the actual values $x_1, \ldots, x_n, y$ satisfy the exact equality $y = f(x_1, \ldots, x_n)$), the errors $\Delta x_i = \tilde{x}_i - x_u$ lead to the error $\Delta y = \tilde{y} - y = f(\tilde{x}_1, \ldots, \tilde{x}_n) - f(x_1, \ldots, x_n)$ in the result of data processing. So, the actual value y may differ from the result $\tilde{y}$ of indirect measurement. We want to know the possible values of y.

Interval computations lead to a guaranteed estimate for the problem. Usually, we know the bounds Δ_i that bound measurement errors Δx_i; as a result, we know that $x_i \in [x_i^-, x_i^+] = [\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]$. From these intervals, we must compute the interval of possible values of $y = f(x_1, \ldots, x_n)$. This is a typical problem in which interval computations are useful: these computations lead to a interval $[y^-, y^+]$ that is *guaranteed* to contain all possible values of y.

In some (ill-posed) problems the guaranteed interval bound is so wide that the results of the measurement become meaningless. If the guaranteed interval $[y^-, y^+]$ is narrow enough, the problem is solved. But what if it is so wide as to be of no use? For example, what if in the above medical situation, when we wanted to find out whether a patient has a brain tumor, we get an interval that contains both the density values corresponding to the tumor and the density values corresponding to the healthy brain? In this case, we are as unclear about the brain as before the measurement, and therefore, the entire procedure was a waste of time and money.

If the guaranteed error bound (that *always* contains the error) is too wide, it is desirable (if possible) to produce a smaller bound that *almost always* contains the error. When the guaranteed error bound leads to too wide intervals, it is desirable to narrow down this interval. In some cases, when the interval is wide because of an overestimation (a phenomenon well known in interval computations), we can simply select a better method (e.g., use a centered form), and get a narrower estimate as a result. In some other cases, however, the exact guaranteed interval is still way too wide. These are typical situations in so-called *ill-posed problems* (of which medical tomography is an example), where a small deviation in x_i can lead to a huge change in $y = f(x_1, \ldots, x_n)$. What can we do then?

The situation is not so hopeless as it may seem. In real engineering problems, we never have a 100% guarantee of success: an unexpectedly powerful earthquake can destroy even a very strong construction; a volcano that has been dormant for thousand years can start erupting and thus vibrating the neighborhood, etc. In mathematical terms, it means that our intervals for x_i may turn out to be incorrect. Since there is always a possibility that our models are not precise (i.e., that with some small probability, the actual values x_i lie outside the intervals $[\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]$), it makes sense not to require that the computations with these models *always* lead to guaranteed results, but to require instead that these computations lead to an interval that contains y *almost always* (i.e., *with a probability close to 1*).

"Almost always" is easy to interpret if we know probabilities of different errors. We formulated this idea informally (we used such informal terms as "almost", "close to", etc). This idea can be easily formulated if we know the probabilities of different values of x_i. In this case, we can calculate the probability of different values of $y = f(x_1, \ldots, x_n)$, and select the narrowest interval that contains y with a probability $\geq 1 - p_0$ for some *a priori* chosen value $p_0 \ll 1$.

In many real-life situations, we do not know probabilities. How do we then interpret "almost always"? Simplest case: finite number of alternatives. In many real-life situations, however, we do not know these probabilities: we either only know the intervals for x_i, or we may have some *partial* information about the probabilities. Can we interpret "almost always" in this case?

Such an interpretation is given by *maximum entropy* (*Maxent*) methods. During the workshop, the foundations of Maxent were presented in a tutorial by Imre Csiszár (Hungary). I will try to reformulate the main ideas of his description for the interval community (for technical details, see, e.g., [3]).

2. The idea of Maxent

Let us first consider the simplest case when:

- we have no information about the probabilities, and we only know the *set* A of possible states of an object; and
- this set A is *finite*.

Let us denote the number of elements in a set A by n, and i-th element by a_i; in these denotations, we will have $A = \{a_1, \ldots, a_n\}$.

We want to formalize the notion of "almost all" cases. If there is only *one* object with this uncertainty, the notion of "almost all" cases makes no sense. Therefore, we must consider the situation in which there are *several* objects described by the same set A. Let us denote the number of such objects by N. Then, each of these N objects is in one of the states $a \in A$. To describe the set of all N objects, we thus need the N-dimensional vector $(a^{(1)}, \ldots, a^{(N)})$ with $a^{(k)} \in A$. We will call such a vector a *world view*. Since each object can be in one of n states, there are N^n possible world views.

Now, a natural idea is to consider all world views equally probable, and thus consider a property to be true in "almost all" cases if this property is true for almost all world views.

Let us describe this idea in mathematical terms. Our problem started with the fact that we do not know the probability $p(a)$ for different values $a \in A$. So, let us see what we can say about these probabilities now. Since we only have finitely many (N) objects, we cannot really talk about the probability, we can only talk about the *frequencies* $f(a)$ of different values $a \in A$; then, if there is a probability distribution, in the limit $N \to \infty$, these frequencies will turn into probabilities.

Let us fix a set of frequencies $f = \big(f(a_1), \ldots, f(a_n)\big)$ $\left(\sum f(a_i) = 1\right)$, and count the number $N(f)$ of world views with these very frequencies. Frequency $f(a_i)$ means that a_i occurs $N \cdot f(a_i)$ times. Therefore, we have a well-known combinatorial problem: we have N objects and n boxes, and we are interested in knowing the number of ways in which we can distribute these objects between these boxes so that i-th box gets $N_i = N \cdot f(a_i)$ objects. Let us recall the solution:

- First, we must select N_1 objects out of N that go into the first box. The number of such assignments is known to be equal to

$$\frac{N!}{N_1! \cdot (N-N_1)!}.$$

- For each of these assignments, we must select N_2 objects that go into the second box out of the remaining $N - N_1$ objects. This number of such assignments is equal to

$$\frac{(N-N_1)!}{N_2! \cdot (N-N_1-N_2)!}.$$

  The total number of assignments to the two boxes can be computed as a product of the two numbers:

$$\frac{N!}{N_1! \cdot (N-N_1)!} \cdot \frac{(N-N_1)!}{N_2! \cdot (N-N_1-N_2)!} = \frac{N!}{N_1! \cdot N_2! \cdot (N-N_1-N_2)!}.$$

- Similarly, if we have assigned elements to boxes 1 and 2, we must assign N_3 of the remaining $N - N_1 - N_2$ elements to the third box, which leads to

$$\frac{N!}{N_1! \cdot N_2! \cdot (N-N_1!-N_2)!} \cdot \frac{(N-N_1-N_2)!}{N_3! \cdot (N-N_1-N_2-N_3)!}$$
$$= \frac{N!}{N_1! \cdot N_2! \cdot N_3! \cdot (N-N_1-N_2-N_3)!}$$

  assignments, etc.

Finally, for all boxes, this number is equal to

$$N(f) = \frac{N!}{N_1! \times \cdots \times N_n!} = \frac{N!}{\big(N \cdot f(a_1)\big)! \times \cdots \times \big(N \cdot f(a_n)\big)!}.$$

For large N, we can apply Stirling's formula $N! \approx (N/e)^N$. The product is easier to handle after we go to logarithms; as a result, we get $\ln(N!) \approx N(\ln N - 1)$, and

$$\ln\big(N(f)\big) \approx N \ln N - N - \sum_{i=1}^{n} N \cdot f(a_i)\Big[\ln\big(N \cdot f(a_i)\big) - 1\Big].$$

Since $\ln\big(Nf(a_i)\big) = \ln N + \ln\big(f(a_i)\big)$, and $\sum f(a_i) = 1$, we get $\ln\big(N(f)\big) \approx N \cdot S(f)$, where we denoted

$$S(f) = -\sum_{i=1}^{n} f(a_i) \ln\big(f(a_i)\big)$$

this expression $S(f)$ is called an *entropy* of the frequency distribution f. Therefore, $N(f)/N \approx \exp\big(NS(f)\big)$. Hence, if we have two distributions f and f' for which $S(f) > S(f')$, then for large N, we get $N(f)/N(f') \to \infty$. So, for large N, almost all world views correspond to the frequencies for which $S(f)$ is either equal to the maximum possible value $S(f)$, or close to $S(f)$. Our arguments are not exactly proving it, because we are using approximate formulas that are only asymptotically correct, but one can show (see, e.g., [3]) that indeed, for every

$\delta > 0$ and $\varepsilon > 0$, there exists an N_0 such that for all $N \geq N_0$, the fraction of world views for which the frequencies are *not* δ–close to the frequencies with $S(f) \rightarrow \max$, is $< \varepsilon$.

In other words, "almost all" world views correspond to the frequency distribution $f(a_i)$ that has the largest possible entropy $S(f)$. In the limit $N \rightarrow \infty$, frequencies tend to probabilities, so, we can conclude that *for almost all world views, the probabilities correspond to the maximum possible entropy.*

We justified this conclusion in the case when we have no information about the probabilities. It turns out that the same conclusion can be made if we have *some* information about probabilities: namely, if this information restricts us to a certain set $\mathcal{P}$ of possible probability distributions on the set A, we can conclude that for almost all world views (i.e., less formally, in almost all cases) the actual probability distribution p is the one for which $S(p) \rightarrow \max$ under the condition that $p \in \mathcal{P}$.

General case: possibly infinite number of alternatives. In real-life situations, the set of possible states is infinite: e.g., usually, we know that the value of a certain parameter x belongs to an interval $[x^-, x^+]$; in this case, every real number from this interval describes a possible state; hence, there are infinitely many states.

To handle this situation, we can take into consideration that in reality, we cannot know x precisely. In the most distant and technologically advanced future, we will still be able to measure x only with some accuracy $\delta > 0$. If in two measurements with this accuracy, we get two different values $\tilde{x}_1$ and $\tilde{x}_2$ for which $|\tilde{x}_1 - \tilde{x}_2| < 2\delta$, then, it could be that the corresponding *actual* values x_1 and x_2 coincide (and are equal to $(1/2) \cdot (\tilde{x}_1 + \tilde{x}_2)$), and the difference between x_i and $\tilde{x}_i$ is caused by the measurement inaccuracy. Therefore, we will have *finitely many* distinguishable outcomes, that correspond to the values $x^-, x^- + 2\delta, x^- + 4\delta, \ldots, x^+$.

If we apply the above-justified maximum interval idea to these outputs, we can conclude that the probabilities $p\Big([x^- + 2k\delta, x^- + 2(k+1)\delta]\Big)$ of x being in each of these interval satisfy the condition $S(p) \rightarrow \max$. In the limit $\delta \rightarrow 0$, the sum in $S(p)$ tends to an integral $S(\rho) = -\int \rho(x) \ln \Big(\rho(x)\Big)\, dx$, where by ρ, we denoted the *probability density.* Therefore, we can also conclude that *in almost all cases, the actual probability density is the one for which the entropy $S(\rho)$ is the largest possible.* (As in the finite case, we only gave a heuristic justification, but this justification can be made rigorous.)

The use of this *maximum entropy* (*Maxent*) distribution is called *Maximum entropy method*, or *Maxent*, for short.

A brief history of maximum entropy methods. The methodology of maximum entropy originated from statistical physics (this is where the notion of an *entropy* came from). The first person to notice that these methods can be used in general data processing was E. T. Jaynes [8]. Different areas of physics still form the main area of Maxent applications. Suffice it to say that proceedings of the annual Maxent workshops are published by Kluwer as a part of the "Fundamental Theories of Physics" book series (for a reasonably current survey, see, e.g., [6]).

3. Maxent really works

For indirect measurements, Maxent really helps to make the resulting interval narrower. For a measuring instrument for which we only know the interval $[-\Delta, \Delta]$ of possible values of error, Maxent requires that we consider the probability density $\rho : [-\Delta, \Delta] \rightarrow R^+$ for which $S(\rho) = -\int \rho(x) \ln \Big(\rho(x)\Big)\, dx \rightarrow \max$ under the normalizing condition $\int \rho(x)\, dx = 1$. This problem is easily solvable: it leads to the *uniform distribution* with the density $\rho(x) = 1/(2\Delta)$. For several measurements with interval uncertainty, similar formulation leads to the conclusion that each error is uniformly distributed, and the corresponding distributions are statistically independent. The independent uniform distributions are therefore recommended for usage (in the case described above) by the guidelines of the Western European Calibration Cooperation (WECC) [24] and the Comité International des Poids et Mesures (CIPM) ([20, 21]; see also [1]).

Let us show that the use of this distribution really decreases the interval, on the example of a simple function $y = x_1 + \cdots + x_n$. Let us assume that each value x_i is measured with an accuracy Δ. If, as a result of n measurements, we get n values $\tilde{x}_i$, this means that the actual values of each x_i belong to the interval $[\tilde{x}_i - \Delta, \tilde{x}_i + \Delta]$. Let us describe the results of applying interval computations and Maxent to this problem.

- *Interval computations.* The resulting interval of possible values of y is $[\tilde{x} - n\Delta, \tilde{x} + n\Delta]$, where we denoted $\tilde{x} = \tilde{x}_1 + \cdots + \tilde{x}_n$. The width of this interval is $2n\Delta$.
- *Maxent.* If we use-Maxent, then the resulting error $\Delta y = \tilde{y} - y$ is the sum of n independent random variables $\Delta x_i = \tilde{x}_i - x_i$ with 0 average, each of which is uniformly distributed on the interval $[-\Delta, \Delta]$. For large n, due to the central limit theorem (see, e.g., [24]), the resulting distribution of Δy will be close to Gaussian, with the mathematical expectation equal to the sum of averages of component distribution (i.e., to 0), and the second central moment (square σ^2 of the standard deviation σ) equal to the sum of the corresponding moments. Hence, $\sigma^2 = Cn\Delta^2$ for some constant C, and hence, $\sigma = \sqrt{n}\sqrt{C}\Delta$. For a Gaussian distribution, almost all values are located within $k\sigma$ from the average (for $k = 2$, we get $\approx$ 90%; for $k = 3$, we get $\approx$ 99.9%, etc). Substituting the above expressions for the average and for σ, we can conclude that *almost all* values of y belong to the interval $[-k\sqrt{C}\sqrt{n}\Delta, k\sqrt{C}\sqrt{n}\Delta]$.

The width of this interval grows as $\sqrt{n}$ as opposed to n for interval arithmetic, so, for large n, *the resulting "Maxent" interval is indeed drastically smaller.*

Maxent is not a magic bullet. It is important to emphasize that *Maxent is not a magic bullet*: all we can say about the resulting narrower interval is that in "almost all" cases (in some reasonable sense) y belongs to this interval. We *cannot* claim that y *always* belongs to this interval, because we can have $\Delta x_i = \Delta$ for all i; then $\Delta y = n\Delta$. So, the guaranteed interval is still wide. However, for practical applications, it is nice to know, in addition to the wider *guaranteed* interval, a narrower interval that contains y in almost all cases. This is what Maxent does.

4. Applications of Maxent

According to the above description, Maxent methods make sense if the guaranteed intervals are too wide to be meaningful, and if it is very difficult or even impossible to get any additional information. If:

- an object of our study is reasonably *simple* in the sense that it can be characterized by the values of a few physical quantities, and if
- we can easily measure these quantities,

then it is reasonable to measure these quantities and thus get the desired information.

But this is often not possible, so, Maxent is needed. In this section, we will describe all such situations, and illustrate them by applications presented at the workshop.

Many objects are *complex*; so many independent factors influence the behavior of an object that it is practically impossible to measure enough of them. Such complex objects include:

- systems described by *statistical physics*; this is the area where Maxent originated, and from which it spread to other fields [8]; during the workshop:
  - an application to *hydrodynamics* was presented by D. Montgomery (USA);
  - applications to the description of multi-electron atoms, magnetic substances, etc, were presented by H. Akhlaghpur, M. Jarrell, H. Panf, and R. N. Silver (USA);
  - a typical real-life example of a complex system is *weather*: even with the nowadays ability to measure millions of values per second, it is still practically impossible to get guaranteed weather predictions; applications of Maxent to weather prediction were presented by M. Berliner (USA);
- *biological systems*, including living organisms and ecosystems; during the workshop:
  - general applications to the analysis of biological data were presented by I. Tchoumatchenko and J. G. Ganascia (France); there were also two specific applications:
  - to the analysis of *cellular structure*, by N. Rivier, B. Dubertet (France) and G. Schliecker (Germany); and
  - to *predicting fish density*, by S. Lizamore, M. Vignaux, and G. A. Vignaux (New Zealand);
- *financial and economic systems* (world markets, stock prices, etc, are influenced by too many factors to be easily predictable); during the workshop:
  - applications to *predicting prices on financial markets* were presented by G. J. Daniell (UK), and by R. J. Hawkins and M. Rubinstein (USA);
- *experts*, when we try to make a computer simulate the way they make decisions (i.e., to design an *expert system*); expert's reasoning is definitely a very complicated object, difficult to describe; the idea of applying Maxent to expert systems was first proposed by Cheeseman [2]; see also [11, 16–19]; during the workshop:
  - applications to expert systems and intelligent control were presented by V. Kreinovich, H. T. Nguyen, and E. A. Walker (USA);
  - applications to *speech recognition* by R. Laboissiere (France); and
  - applications to *natural language processing* by J. D. Lafferty and B. Suhm (USA).

Another case when Maxent is useful is when the object is simple (in the sense that it can be described by a few parameters), but due to some fundamental reasons, there is no way that we can measure all of them.

- One such case is *quantum mechanics*, where due to Heisenberg's inequality, measuring one of the characteristics (e.g., location) changes the systems so drastically that the information about the values of other characteristics is lost. During the workshop, the applications to quantum mechanics were described by S. Youssef (USA).

In some cases, the system is simple, and it is in principle possible to measure the values of its parameters, but at the current technological level, we do not know how to measure them: the signal is too small, and the noise is too high. Examples include:

- *astrophysics*, where signal are weak, either
  - in *passive observation*, because they come from very distant sources (see, e.g., [10]), or
  - in *active*, radar astronomy, because over interplanetary distances, the reflected radar signal becomes extremely weak (see, e.g., [5, 15]);

  during the workshop, applications to imaging in astrophysics were presented by D. D. Dixon, W. N. Johnson, J. D. Kurfess, R. C. Peutter, R. K. Pina, W. R. Purcell, O. T. Tumer, W. A. Wheaton, and A. D. Zych (USA);
- *high-energy physics*, where experiments are extremely costly;
  - applications of Maxent to *high-energy physics* were presented by G. S. Cunningham and K. M. Hanson (USA), and also
  - to *nuclear fusion* by V. Dose, A. Garrett, W. von der Linden (Germany).

Finally, for some simple systems it is possible to measure all their characteristics, but we do not want to do it, because these measurements may endanger or even ruin the system itself; such cases include:

- *medical applications*, e.g., in *tomography*, where we want to learn about the state of the internal organs without performing a surgery; such applications were presented by M. P. Anderson, R. Bajcsy, D. G. Brown, J. C. Gee, G. Gindi, K. M. Hanson, D. R. Haynor, K. J. Meyers, A. Rangarajan, and R. F. Wagner (USA);
- *engineering*, where we want to learn about the inside of a system without destroying it; such methods are called *non-destructive testing*; corresponding applications of Maxent were presented by G. Le Besnerais, S. Gautier, B. Lavayssieère, and A. Mohammad-Djafari (France);

In addition to talks that described specific applications, several talks describe the applications of Maxent to general data processing techniques:

- to *signal processing* (G. L. Bretthorst, A. Ramaswami, USA);
- to *dynamic signal processing* (L. Borland, USA);
- to general *inverse problems* (A. Mohammad-Djafari, France; V. Dose, R. Fischer, W. von der Linden, Germany);
- to *image processing* (P. Boulanger, M. Rioux, Canada; J. Skilling, S. Sibisi, UK; J. Besag, USA);
  - to *pattern recognition* (R. Snapp, USA); and
  - to *improving computational algorithms* (specifically, so-called genetic algorithms; A. Prügel-Bennett, Denmark; M. Rattray, J. L. Shapiro, UK).

We only enumerated the *basic* applications. Many interesting applications were also presented in the posters, and mentioned in major talks. For details, one can see the abstracts [12, 13], and the forthcoming Proceedings [14].

5. Maxent and intervals

5.1. Intervals return

When we use Maxent, we replace intervals (or any other information) with a probability distribution (namely, the one with the maximum entropy among all distributions that are consistent with the given information). If, in addition to the interval, we know some probabilities, we simply add these probabilities to the set of conditions that restrict the desired probability distribution. At first glance, as soon as we say "Maxent", intervals disappear and statistics takes the stage. Well, intervals do not completely disappear.

We will now describe two problems that combine Maxent and intervals. These problems stem from the fact that the a priori known probabilities that we have just mentioned are often determined from the experiments, and therefore, instead of their precise values, we only know *intervals* of their possible values.

5.2. The first problem: interval-valued probabilities

In the first problem, we *almost* know probabilities, i.e., we have n alternatives $a_1, \ldots, a_n$, and we know the approximate values of each of the probabilities p_i. In mathematical terms, for each i, we know the *interval* $[p_i^-, p_i^+]$ such that the (unknown) probabilities p_i belongs to this interval. The question is: what is the corresponding Maxent distribution? To describe this distribution, let us first formulate the problem in precise mathematical terms:

Definition 1. *Let a finite set $A = \{a_1, \ldots, a_n\}$ be given.*

- *By an probability distribution p on the set A, we mean a tuple $(p_1, \ldots, p_n)$, where $\sum p_i = 1$.*
- *By an interval probability distribution $\mathbf{p}$ on the set A, we mean a tuple $(\mathbf{p}_1, \ldots, \mathbf{p}_n)$, where $\mathbf{p}_i = [p_i^-, p_i^+] \subseteq [0,1]$ is an interval.*
- *We say that a probability distribution p is consistent with an interval probability distribution $\mathbf{p}$ is $p_i \in \mathbf{p}_i$ for all i.*
- *We say that an interval probability distribution $\mathbf{p}$ is consistent if it is consistent with some probability distribution p.*
- *By a Maxent distribution corresponding to a consistent interval probability distribution $\mathbf{p}$, we mean a vector $p = (p_1, \ldots, p_n)$ for which $S(p) = -\sum p_i \log p_i \to \max$ among all distributions p consistent with $\mathbf{p}$.*

To compute the Maxent distribution, we will need the following auxiliary results:

Proposition 1. *An interval probability distribution $\mathbf{p}$ is consistent iff $\sum p_i^- \le 1 \le \sum p_i^+$.*

Proposition 2. *If p is a Maxent distribution corresponding to $\mathbf{p}$, then there exists a number $p_0 \in [0,1]$ such that for all $i = 1, \ldots, n$:*

- *If $p_i^+ \le p_0$, then $p_i = p_i^+$.*
- *If $p_0 \le p_i^-$, then $p_i = p_i^-$.*
- *If $p_i^- \le p_0 \le p_i^+$, then $p_i = p_0$.*

Using this result, we can formulate the following theorem:

Theorem 1. *There exists a quadratic-time algorithm that, given an interval probability distribution* $\mathbf{p}$, *returns the Maxent distribution corresponding to* $\mathbf{p}$.

The proofs of Propositions and of Theorem 1 (including the description of the corresponding algorithm and the proof of correctness of this algorithm) are given in the Appendix. This algorithm can be parallelized:

Theorem 2. *There exists an algorithm that, given an interval probability distribution* $\mathbf{p}$, *returns the Maxent distribution corresponding to* $\mathbf{p}$ *in time* $O(\log n)$ *on* $O(n^2)$ *processors.*

5.3. The second problem: interval-valued probabilities in expert systems

A typical *expert system* consists of n logically independent statements $E_1, \ldots, E_n$, with their ptobabilities p_i (or, intervals of probabilities). Let us give exact definitions:

Definition 2. *Let* $E_1, \ldots, E_n$ *be statements (in a certain language).*

- *By a world, we mean an expression of the type* $E_1^{\varepsilon_1} \& \cdots \& E_n^{\varepsilon_n}$, *where* $\varepsilon_i \in \{+, -\}$, $E^+ = E$, *and* E^- *means* $\neg E$. *Worlds will be denoted by* $W, W', \ldots$ *and the set of all the worlds will be denoted by* $\mathcal{W}$. *If a world* W *contains* E_i^+, *we say that* E_i *is true in* W, *and denote it by* $W \vdash E_i$.
- *We say that the statements* $E_1, \ldots, E_n$ *are logically independent if all* 2^n *worlds* W *are consistent.*
- *By a probability distribution, we mean a probability distribution* $\{p(W)\}$ *on the set* $\mathcal{W}$ *of all worlds, i.e.. a set of non-negative values* $p(W)$ *for which* $\sum p(W) = 1$.
- *For a given probability distribution, by a probability* $p(E_i)$ *of* E_i *we mean the value*

$$p(E_i) = \sum_{W : W \vdash E_i} p(W).$$

Definition 3.

- *By a knowledge base, we mean a finite set of pairs* (E_i, p_i), *where* E_i *is a statement, and* $p_i \in [0, 1]$.
- *We say that a probability distribution* $\{p(W)\}$ *is consistent with the knowledge base if for all* i, $p(E_i) = p_i$.
- *By a Maxent distribution corresponding to a given knowledge base, we mean a probability distribution* $\{p(W)\}$ *for which the entropy is the largest*

$$S\big(\{p(W)\}\big) = -\sum p(W) \log p(W) \to \max$$

among all probability distributions consistent with the given knowledge base.

For logically independent knowledge bases, it is well known how to describe the Maxent distribution:

Proposition 3. *If the statements from the knowledge base are logically independent, then the Maxent distribution has the form $p(W) = p_1^{\varepsilon_1} \times \cdots \times p_n^{\varepsilon_n}$, where we denoted $p^+ = p$ and $p^- = 1 - p$. For this distribution,*

$$S\Big(\{p(W)\}\Big) = -\sum_{i=1}^{n} \Big(p_i \log p_i + (1 - p_i)\log(1 - p_i)\Big).$$

In real life, we only know *intervals* $\mathbf{p}_i$ of possible values of the probability. In this case, we arrive at the following definition:

Definition 4.

- *By an interval-valued knowledge base, we mean a finite set of pairs $(E_i, \mathbf{p}_i)$, where E_i is a statement, and $\mathbf{p}_i \subseteq [0,1]$ is an interval.*
- *We say that a probability distribution $\{p(W)\}$ is consistent with the knowledge base if for all i, $p(E_i) \in \mathbf{p}_i$.*
- *By a Maxent distribution corresponding to a given knowledge base, we mean a probability distribution $\{p(W)\}$ for which the entropy is the largest*

  $$S\Big(\{p(W)\}\Big) = -\sum p(W) \log p(W) \rightarrow \max$$

  among all probability distributions consistent with the given knowledge base.

The description of Maxent distribution is given by the following theorem:

Theorem 3. *For every interval-valued knowledge base $(E_i, \mathbf{p}_i)_i$ with logically independent statements E_i, the corresponding Maxent distribution is the one that corresponds to the knowledge base (E_i, p_i), where for $i = 1, \ldots, n$:*

- *If $p_i^+ \leq 0.5$, then $p_i = p_i^+$.*
- *If $0.5 \leq p_i^-$, then $p_i = p_i^-$.*
- *If $p_i^- \leq 0.5 \leq p_i^+$, then $p_i = 0.5$.*

Comment. According to this theorem, if we are unsure about probabilities, i.e., if instead of the actual probabilities p_i, we know their δ−approximations $\tilde{p}_i$ (for some small $\delta > 0$), then it is best:

- to *overestimate* small probabilities (i.e., to assume that $p_i = \tilde{p}_i + \delta$) if $\tilde{p}_i$ is small (i.e., if $\tilde{p}_i \leq 0.5 - \delta$); and
- to *underestimate* small probabilities (i.e., to assume that $p_i = \tilde{p}_i - \delta$) if $\tilde{p}_i$ is large (i.e., if $\tilde{p}_i \geq 0.5 + \delta$).

It is interesting to mention that this is exactly what we human do if we are not 100% sure about the probabilities. This phenomenon was experimentally discovered in [22]. Thus, *Maxent provides a fundamental theoretical explanation for this phenomenon.*

Back to the workshop: We also had some fun

The city of Santa Fe. The city of Santa Fe was founded by the Spaniards more than 300 years ago; its name literally means *Saint Faith.* For several centuries, it was a quiet provincial town. There was little construction here, and, as a result, the city still has many old and beautiful buildings that make it a major tourist attraction. Not only tourists come here: artists, sculptors, musicians, craftsmen flock to the city, to be inspired by its ancient beauty; as results, there are lots of unusually dressed artists everywhere.

St.John's College. The conference itself was held in St.John's College, an interesting educational institution where everything is taught from the original sources:

- students of *mathematics* start with Euclid in Greek;
- students of *physics* start with Newton in Latin, and
- students of *interval computations* start with Moore in English.

About Moore, it is a joke (some day they may start to learn it, however), but Greek and Latin they do study, and many students even study Russian to read the original Russian-language sources, Dostoevsky for one (that's where I borrowed the subtitle; times are faster now, so I cannot wait until winter to describe my travel impressions as he did).

Sponsors. The workshop was sponsored by *Santa Fe Institute* and by *Los Alamos National Laboratory*, so we had three banquets for three days in a row: at the College and at each of the sponsoring organizations.

- *Santa Fe Institute* is a world-renown multi-disciplinary center where physicists, computer scientists, mathematicians, philosophers, and others get together to study complexity. The small charming Institute is overcrowded with famous people (for example, Murray Gell-Mann, who discovered quarks, is working there).
- *Los Alamos* is the place where the first atomic bombs were made. In the local museum, we could see how the physicists who made it lived, and take photos of each other in front of the exact copies of the first atomic bombs. My family and I were probably the participants of the conference who got interested in one of the most unusual Santa Fe attractions called *Spymaster tour*: to see the bars where the Soviet spies would try to get the atomic secrets from the scientists, and the places where the atomic secrets were actually passed to Stalin's agents.

A less exotic attraction that we visited is located in a few miles from Los Alamos: an ancient Indian city made of numerous caves carved in the mountain; this city was never conquered or destroyed: about a thousand years ago, it was suddenly and mysteriously abandoned.

Stairway to Heaven. In the middle of downtown Santa Fe stands an old Loretto chapel, with a stairway inside that has no railings, no support, just goes straight from Earth to Heaven. No one knows how this engineering wonder became possible. The legend says that a saint descended from Heaven and built this stairway. Who knows, maybe he used interval computations to guarantee its 300 years of stability? Or maybe, since this stairway looks so light, he used maximum entropy methods to guarantee that it is *almost always* stable?

Appendix: proofs

Proof of Proposition 1 is straightforward.

Proof of Proposition 2.

1°. First, let us show that it is impossible to have $p_k < p_l$, $p_k < p_k^+$, and $p_l > p_l^-$ for some k and l.

Indeed, in this case, by changing p_k and p_l to $p_k' = p_k + \Delta$ and $p_l' = p_l - \Delta$ for some small $\Delta > 0$ and leaving all other values p_i unchanged, we get a new vector p_i' for which:

- $\sum p_i' = \sum p_i = 1$;
- $p_i' \in [p_i^-, p_i^+]$, if Δ is small enough (to be more precise, if $\Delta \le \min(p_k^+ - p_k, p_l - p_l^-)$); and
- since $S(p) = -\sum p_i \ln p_i$ is a smooth function, with

$$\frac{\partial S}{\partial p_i} = -\ln p_i - 1$$

we have

$$S(p') = S(p) + \left(\frac{\partial S}{\partial p_k} - \frac{\partial S}{\partial p_l}\right)\Delta + o(\Delta) = S(p) + (\ln p_l - \ln p_k)\Delta + o(\Delta)$$

since $p_l > p_k$, and ln is an increasing function, we have $S(p') > S(p)$ for small Δ.

This conclusion contradicts to our assumption that p is a Maxent distribution. So, the case described in the formulation of this point is indeed impossible.

2°. From 1°, in particular, we can conclude that two different values $p_k \neq p_l$ cannot be both inner points of the corresponding intervals $\mathbf{p}_k$ and $\mathbf{p}_l$. So, there is no more than one such point.

If such a point exists, let us denote it by p_0.

3°. If p_0 exists, i.e., if $p_0 = p_l \in (p_l^-, p_l^+)$ for some l, then, we can prove the conclusion of the theorem by considering three possible cases:

- If $p_i^+ \le p_0 = p_l$, then, from the fact that $p_l^- < p_l$, and from 1°, it follows that we cannot have $p_i < p_i^+$. Since $p_i \in [p_i^-, p_i^+]$, we must have $p_i \le p_i^+$, and therefore, $p_i = p_i^+$.
- Similarly, if $p_i^- \ge p_0 = p_l$, then, due to $p_l < p_l^+$ and 1°, we cannot have $p_i > p_i^-$, so, $p_i = p_i^-$.
- Finally, if $p_i^- \le p_0 = p_l \le p_i^+$, then due to 1°, we cannot have $p_i < p_l$ and we cannot have $p_l < p_i$, so the only remaining possibility is $p_i = p_l = p_0$.

4°. If no such p_0 exists, i.e., if for every i, either $p_i = p_i^-$, or $p_i = p_i^+$, then from 1°, we can conclude that all the values p_i with $p_i = p_i^-$ are smaller than each of the values p_j for which $p_j = p_j^+$, so, as p_0, we can take any separating point. □

Proof of Theorem 1. The corresponding algorithm is as follows:

1°. First, we order $2n + 2$ values 0, 1, p_i^-, $1 \le i \le n$, and p_i^+, $1 \le i \le n$, into a sequence $u_1 \le u_2 \le \cdots \le u_{2n+2}$.

Ordering requires $O(n \ln n)$ computational steps (see, e.g., [4]).

2°. For each element u_k from this sequence, let us denote by u'_k the first next element that is different from u_k (it is not necessarily u_{k+1}, because some of the bounds $p_i^{\pm}$ may coincide). Thus, the entire interval $[0. 1]$ is divided into several sub-intervals that intersect only in their boundary points. The critical value p_0 (whose existence is described in Theorem 1) must belong to one of these subintervals. Let us show that as soon as we know the subinterval to which to which p_0 belongs, we can uniquely determine p_i. Then, we will compute entropy for the vectors p that correspond to different subintervals, and choose the vector with the largest entropy as the desired Maxent distribution.

So, let us fix a subinterval $[u_k, u'_k]$, and describe the probabilities that correspond to the case $p_0 \in [u_k, u'_k]$:

- For all i for which $p_i^+ \leq u_k$, we take $p_i = p_i^+$.
- For all i for which $p_i^- \geq u'_k$, we take $p_i = p_i^-$.

Then, we have two possible cases:

- If all $i = 1, \ldots, n$ are covered by one of the two formulas, then we have defined p_i for all i. Then, we check whether $\sum p_i = 1$, and, if yes, compute $S(p)$. We will denote the computed $S(p)$ by $S^{(k)}$.
- Let us now consider the case when not all i are covered by these formulas. In this case, since $p_i^+ \not\leq u_k$, we have $p_i^+ > u_k$. But since p_i^+ is one of the values u_j, it is therefore $\geq$ that the next value to u_k, i.e., that u'_k. From $p_i^+ \geq u'_k$ and $u'_k \geq p_0$, we conclude that $p_0 \leq p_i^+$. Similarly, from the fact that i is not covered by one of the above two cases, we can conclude that $p_i^- \leq p_0$. Hence, $p_i^- \leq p_0 \leq p_i^+$, so, according to Theorem 1, for all uncovered i, p_i have the same value p_0. This value p_0 can be thus determined from the condition $\sum p_i = 1$, as
$$p_0 = \frac{\sum\{p_i \,|\, i \text{ is not covered}\}}{\#\{i \,|\, i \text{ is not covered}\}}.$$
If the resulting p_0 is outside the interval $[u_k, u'_k]$, this case is impossible. If p_0 is in this interval, then we compute $S(p)$ for the resulting probability distribution p. We will also denote the resulting $S(p)$ by $S^{(k)}$.

3°. Now, we can take the largest of the values $S^{(k)}$, and, as a Maxent distribution, take the probability distribution that correspond to this largest $S^{(k)}$.

4°. There are $2n + 2$ subintervals, and computations that correspond to each of them take $\leq Cn$ computation steps. So, totally, we need $\leq Cn(2n + 2) = O(n^2)$ steps. Together with sorting $\big(O(n \ln n)\big)$, we still need quadratic time. □

Example. Let $n = 3$, $\mathbf{p}_1 = [0, 1/3]$, $\mathbf{p}_2 = [1/3, 2/3]$, and $\mathbf{p}_3 = [2/3, 1]$. In this case, out of $2n + 2 = 8$ values, only 4 are different: 0, 1/3, 2/3, and 1. So, we must consider three subintervals $[u_k, u'_k]$: $[0, 1/3]$, $[1/3, 2/3]$, and $[2/3, 1]$.

1) For the first subinterval $[u_k, u'_k] = [0, 1/3]$, we have $u'_k \leq p_2^- = 1/3$ and $u'_k \leq p_3^- = 2/3$; therefore, according to the described algorithm, we take $p_2 = p_2^- = 1/3$ and $p_3 = 2/3$. The only uncovered value is $i = 1$, so we define $p_1 = 1 - p_2 - p_3 = 1 - (1/3) - (2/3) = 0$. This value belongs to the interval $[u_k, u'_k] = [0, 1/3]$, therefore, we can compute the corresponding value of $S(p)$.

2) For the second subinterval $[u_k, u'_k] = [1/3, 2/3]$, we have $p_1^+ = 1/3 \leq u_k$ and $u'_k \leq p_3^- = 2/3$; therefore, according to the described algorithm, we take $p_1 = p_1^+ = 1/3$ and $p_3 = 2/3$.

The only uncovered value is $i = 2$, so we define $p_2 = 1 - p_1 - p_3 = 1 - (1/3) - (2/3) = 0$. This value, however, does not belongs to the interval $[u_k, u'_k] = [1/3, 2/3]$, therefore, we discard this subinterval.

3) For the third subinterval $[u_k, u'_k] = [2/3, 1]$, we have $p_1^+ = 1/3 \leq u_k$ and $p_2^+ = 2/3 \leq u_k$; therefore, according to the described algorithm, we take $p_1 = p_1^+ = 1/3$ and $p_2 = 2/3$. The only uncovered value is $i = 3$, so we define $p_3 = 1 - p_1 - p_2 = 1 - (1/3) - (2/3) = 0$. This value, however, does not belong to the interval $[u_k, u'_k] = [2/3, 1]$, therefore, we discard this subinterval.

So, we have only one value $S^{(k)}$, and hence, as a Maxent distribution, we take the values p_i that correspond to this $S^{(k)}$, i.e., $p_1 = 0$, $p_2 = 1/3$, and $p_3 = 2/3$.

Proof of Theorem 2. First, we need to sort $2n + 2$ values. Sorting can be done in $O(\log n)$ time (see, e.g., [7], Section 4.3). Then, we can take $2n + 1$ groups of n processors (totally, $O(n^2)$), and target group $\#k$ for computing the probability distribution that corresponds to $[u_k, u'_k]$.

For each k, i-th processor from each group checks whether this i is covered by one of the two conditions from the proof of Theorem 1, and if it does, computes the value p_i. This is done in finitely many ($O(1)$) steps. If not all i are thus covered, we compute p_0 by computing two sums; this is done in $O(\log n)$ steps ([7], Section 1.3).

Computing $S(p)$ consists of computing the values $-p_i \ln p_i$ for all i ($O(1)$ steps on each of n processors), and adding up the resulting n values.

After that, we need to compare n values $S^{(k)}$ to select the maximum. This is also done in $O(\log n)$ steps on n processors ([7], Section 1.3). Totally, we thus need $O(\log n)$ steps on $O(n^2)$ processors. □

Proof of Theorem 3. For each probability distribution $p = \{p(W)\}$ that is consistent with the interval-valued knowledge base $\{(E_i, \mathbf{p}_i)\}$, there exist probabilities $p(E_i) \in \mathbf{p}_i$. Therefore, according to Definition 3, this probability distribution is consistent with the knowledge base $\left\{\left(E_i, p(E_i)\right)\right\}$ for some $p(E_i) \in \mathbf{p}_i$. On the other hand, if a probability distribution $\{p(W)\}$ is consistent with a knowledge base $\{(E_i, p_i)\}$ for some $p_i \in \mathbf{p}_i$, then, due to Definitions 3 and 4, it is consistent with the given interval-valued knowledge base.

So, a probability distribution is consistent with the interval-valued knowledge base $\{(E_i, \mathbf{p}_i)\}$ iff it is consistent with a knowledge base $\{(E_i, p_i)\}$ for some $p_i \in \mathbf{p}_i$. Therefore, to find the desired probability distribution p_{Maxent} with the largest entropy among all distributions consistent with $\{(E_i, \mathbf{p}_i)\}$, it is sufficient to compare the Maxent distributions $p_{\text{maxent}}(\vec{p})$, $\vec{p} = (p_1, \ldots, p_n)$ that correspond to $\{(E_i, p_i)\}$ for different $p_i \in \mathbf{p}_i$, and find the one with the largest entropy $S\left(p_{\text{maxent}}(\vec{p})\right)$.

According to Proposition 3, $S\left(p_{\text{maxent}}(\vec{p})\right) = \sum\left[-\left(p_i \log p_i + (1 - p_i)\log(1 - p_i)\right)\right]$. Therefore, we must maximize this sum under the conditions that $p_i \in [p_i^-, p_i^+]$. This sum is the largest iff each of the terms in the sum is the largest, i.e., if we choose p_i from the condition that $-\left(p_i \log p_i + (1 - p_i)\log(1 - p_i)\right) \to \max$. The function $-\left(p \log p + (1 - p)\log(1 - p)\right)$ is monotonically increasing for $p < 0.5$, and decreasing afterwards. Therefore:

- if $[p_i^-, p_i^+] \subseteq [0, 0.5]$ (i.e., if $p_i^+ \leq 0.5$), then we take $p_i = p_i^+$;
- if $[p_i^-, p_i^+] \subseteq [0.5, 1]$ (i.e., if $0.5 \leq p_i^-$), then we take $p_i = p_i^-$;
- finally, if $0.5 \in [p_i^-, p_i^+]$ (i.e., if $p_i^- \leq 0.5 \leq p_i^+$), then we take $p_i = 0.5$. □

Acknowledgments

This report was partially supported by NSF Grants No. CDA−9015006 and EEC−9322370, and by NASA Grant No. NAG 9−757. I am greatly thankful to Peter Cheeseman, André de Corvin, Imre Csiszár, Hung T. Nguyen, and Karen Villaverde for valuable discussions.

References

[1] Buchholtz, F.−I., Kessel, W., and Melchert, F. *Noise power measurements and measurement uncertainties.* IEEE Trans. on Instrumentation and Measurement **41** (4) (1992), pp. 476−481.

[2] Cheeseman, P. *In defense of probability.* In: "Proceedings of the 8-th International Joint Conference on AI", Los Angeles, 1985, pp. 1002−1009.

[3] Chokr, B. and Kreinovich, V. *How far are we from the complete knowledge: complexity of knowledge acquisition in Demspter-Shafer approach.* In: Yager, R. R., Kacprzyk, J., and Pedrizzi, M. (eds) "Advances in the Dempster-Shafer Theory of Evidence", Wiley, N.Y., 1994, pp. 555−576.

[4] Cormen, Th. H., Leiserson, C. E., and Rivest, R. L. *Introduction to algorithms.* MIT Press, Cambridge, MA, and Mc-Graw Hill Co., N.Y., 1990.

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

[6] Grandy, Jr., W. T. and Schick, L. H. *Maximum entropy and Bayesian methods.* Laramie, Wyoming, 1990, Kluwer Academic Publishers, Dordrecht, Boston, 1991.

[7] Jájá, J. *An introduction to parallel algorithms.* Addison-Wesley, Reading, MA, 1992.

[8] Jaynes, E. T. *Information theory and statistical mechanics.* Phys. Rev. **108** (1957), pp. 171−190.

[9] Jaynes, E. T. *Where do we stand on maximum entropy?* In: Levine, R. D. and Tribus, M. (eds) "The Maximum Entropy Formalism", MIT Press, Cambridge, MA, 1979.

[10] Kosheleva, O. M. and Kreinovich, V. *A letter on maximum entropy method.* Nature **281** (5733) (1979), pp. 708−709.

[11] Kreinovich, V. *Entropy approach for the description of uncertainty in knowledge bases.* Tech. Rep., Center for the New Informational Technology "Informatika", Leningrad, 1989 (in Russian).

[12] *The fifteenth international workshop on maximum entropy and Bayesian methods.* 30 July−4 August 1995, St.John's College Santa Fe, New Mexico, USA, Poster Abstracts.

[13] *The fifteenth international workshop on maximum entropy and Bayesian methods.* 30 July−4 August 1995, St.John's College Santa Fe, New Mexico, USA, Speaker Abstracts.

[14] *Maximum entropy and Bayesian methods. Santa Fe, New Mexico, 1995,* Kluwer Academic Publishers, Dordrecht, Boston, 1995 (to appear).

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

[16] Pearl, J. *Probabilistic reasoning in intelligent systems.* Morgan Kaufmann, San Mateo, CA, 1988.

[17] Ramer, A. and Kreinovich, V. *Maximum entropy approach to fuzzy control.* In: "Proceedings of the Second International Workshop on Industrial Applications of Fuzzy Control and Intelligent Systems", College Station, December 2–4, 1992, pp. 113–117.

[18] Ramer, A. and Kreinovich, V. *Maximum entropy approach to fuzzy control.* Information Sciences **81** (1994), pp. 3–4, 235–260.

[19] Ramer, A. and Kreinovich, V. *Information complexity and fuzzy control.* In: Kandel, A. and Langholtz, G. (eds) "Fuzzy Control Systems", CRC Press, Boca Raton, FL, 1994, pp. 75–97.

[20] *Recommendation INC–1 (1980).* In: Giacomo, P. "News from the BIPM", Metrologia **17** (1981), pp. 67–74.

[21] *Recommendation 1 (CI–1981).* In: Giacomo, P. "News from the BIPM", Metrologia **18** (1982), pp. 41–44.

[22] Tversky, A. and Kahneman, D. *Rational choice and the framing of decisions.* Journal of Business **59** (1986), pp. S251–S278; reprinted in Shafer, G. and Pearl, J. (eds) "Readings in Uncertain Reasoning", Morgan Kaufmann Pub., San Mateo, CA, 1990, pp. 91–104.

[23] Wadsworth, Jr., H. M. (ed.) *Handbook of statistical methods for engineers and scientists.* McGraw-Hill Publishing Co., N.Y., 1990.

[24] *Guidelines for the expression of the uncertainty of measurement in calibration.* WECC, Document No. 19–1990, 1990.

Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968
USA
E-mail: `vladik@cs.utep.edu`