From p-Boxes to p-Ellipsoids: Towards an Optimal Representation of Imprecise Probabilities

Konstantin K. Semenov
Saint-Petersburg State Polytechnical University
29, Polytechnicheskaya str.
Saint-Petersburg, 195251, Russia
Email: semenov.k.k@gmail.com

Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
Email: vladik@utep.edu

***Abstract*—One of the most widely used ways to represent a probability distribution is by describing its cumulative distribution function (cdf) $F(x)$. In practice, we rarely know the exact values of $F(x)$: for each x, we only know $F(x)$ with uncertainty. In such situations, it is reasonable to describe, for each x, the interval $[\underline{F}(x), \overline{F}(x)]$ of possible values of x. This representation of imprecise probabilities is known as a *p-box*; it is effectively used in many applications.**

Similar interval bounds are possible for probability density function, for moments, etc. The problem is that when we transform from one of such representations to another one, we lose information. It is therefore desirable to come up with a *universal* representation of imprecise probabilities in which we do not lose information when we move from one representation to another. We show that under reasonable objective functions, the optimal representation is an ellipsoid. In particular, ellipsoids lead to faster computations, to narrower bounds, etc.

I. Formulation of the Problem

Probabilistic information is important. In describing and processing uncertainty, it is very important to take into account information about the probabilities of different possible values [23]. This is especially true in many engineering applications, when we have a long history of similar situations, and we can use this history to estimate the probabilities of different scenarios. For example, for measurement uncertainty, it is important to use the available information about the probabilities of different possible values of the measurement error; see, e.g., [19].

Comment. There are many ways to represent probabilistic information. From the computational viewpoint, it is therefore important to select the most appropriate representation. In this section, we describe this problem in detail. Readers familiar with different uncertain representations of probabilistic information (such as p-boxes) and with the corresponding problem can skip this section and go directly to the next one.

How probability distributions are usually represented. There are many different ways of representing information about the probability distribution of a random variable X; see, e.g., [23]:

- we can use the *cumulative distribution function*

$$F(x) \stackrel{\text{def}}{=} \text{Prob}(x \le X);$$

- we can use the *probability density function* $\rho(x)$ which is defined as

$$\rho(x) \stackrel{\text{def}}{=} \lim_{\Delta x \to 0} \frac{\text{Prob}(x \le X \le x + \Delta x)}{\Delta x};$$

- we can use *moments*

$$M_k \stackrel{\text{def}}{=} E[X^k] = \int x^k \cdot \rho(x)\, dx;$$

  moments are often described in a slightly different but equivalent form: e.g., instead of describing the first two moments M_1 and M_2, we describe the mean M_1 and the variance $V = M_2 - M_1^2$;
- in some cases, it is convenient to use a *characteristic function* of the distribution, i.e., the expected values

$$E[\exp(\mathrm{i} \cdot \omega \cdot X)] = \int \exp(\mathrm{i} \cdot \omega \cdot x) \cdot \rho(x)\, dx$$

  corresponding to different values ω;
- in some practical applications, it is useful to consider the expected values

$$E[u(X)] = \int u(x) \cdot \rho(x)\, dx$$

  of the utility functions $u(x)$ that describe user preferences; see, e.g., [10], [16], [20];
- there are many other ways of representing the probability distribution.

All these representations are mathematically equivalent. All above ways of representing the probabilistic information are mathematically equivalent – in the sense that if know one of these representations, then we can uniquely determine all the other presentations as well. For example,

- once we know the cdf $F(x)$, we can reconstruct the pdf $\rho(x)$ as the derivative

$$\rho(x) = \frac{dF(x)}{dx};$$

- once we know the pdf $\rho(x)$, we can reconstruct the cdf $F(x)$ as an integral

$$F(x) = \int_{-\infty}^{x} \rho(t)\, dt,$$

etc.

In particular, if we start with $\rho(x)$, integrate to get $F(x)$, and then differentiate the resulting cdf $F(x)$, we get the same pdf $\rho(x)$ with which we started.

Need to take uncertainty into account. In practice, we rarely have full knowledge of the probability distribution: usually, several similar probability distributions are consistent with our knowledge. It is therefore desirable to take into account this uncertainty when we process and represent the corresponding probabilistic information; see, e.g., [26].

How probabilistic uncertainty is currently represented. In terms of cdf, uncertainty means that we do not know the exact values of $F(x)$. For some x, there are several possible values $F(x)$ which are consistent with our knowledge. Since we do not know the *exact value* of $F(x)$, it is reasonable to describe, for each x, the *interval* $[\underline{F}(x), \overline{F}(x)]$ of possible values of x. This representation of imprecise probabilities is known as a *p-box*; it is effectively used in many applications; see, e.g., [6], [7], [17].

Uncertainty can be similarly taken into account for other representations of probabilities; see, e.g., [17] and references therein:

- instead of the exact value $\rho(x)$ of the pdf, for each x, we can represent an interval $[\underline{\rho}(x), \overline{\rho}(x)]$ of possible values;
- instead of the exact values of the moments M_k, we can represent intervals $[\underline{M}_k, \overline{M}_k]$ of possible values;
- instead of the exact values of the characteristic function, we can represent intervals of possible values of the characteristic function, etc.

Problem: these representations are no longer mathematically equivalent. As we have mentioned, when we have the exact knowledge of the probabilities, all representations are mathematically equivalent. However, in the presence of uncertainty, these representations are no longer equivalent. For example, if we have a random variable which is located on a known interval $[x^-, x^+]$ with probability 1, and we know the bounds $\underline{\rho}(x)$ and $\overline{\rho}(x)$ on the corresponding pdf $\rho(x)$, then we can deduce the corresponding bounds on $F(x)$:

$$\underline{F}(x) = \int_{x^-}^{x} \underline{\rho}(t)\,dt \text{ and } \overline{F}(x) = \int_{x^-}^{x} \overline{\rho}(t)\,dt.$$

For example, when both bounds on $\rho(x)$ are constants, i.e., when $\underline{\rho}(x) = \underline{\rho}$ and $\overline{\rho}(x) = \overline{\rho}$, we get

$$\underline{F}(x) = (x - x^-) \cdot \underline{\rho} \text{ and } \overline{F}(x) = (x - x^-) \cdot \overline{\rho}.$$

Based on these bounds, however, we can no longer reconstruct the original bounds on $\rho(x)$: for example, these bounds contain a distribution for which first the cdf $F(x)$ is equal to $\underline{F}(x)$ and then at some point $x_0 \in [x^-, x^+]$, it jumps to $\overline{F}(x)$. For this distribution, the probability density is infinite at $x = x_0$.

Questions: how important is this problem and, if it is important, how to solve it. We have formulated the problem in mathematical terms. Natural questions are:

- how important is this problem for practical applications? do we really need different representations – or should we only use one of them?
- if this problem is practically important, then how can we solve it?

These are the problems that we will handle in this paper.

II. This Problem Is Practically Important

Let us first explain that different representations of probabilistic information are necessary in practical applications and therefore, that the above problem – that different representations are not equivalent in the presence of uncertainty – is practically important.

Which is the best way to describe the probabilistic information. One of the main objectives of data processing is to make decisions. A standard way of making a decision is to select the action a for which the expected utility (gain) is the largest possible. This is where probabilities are used: in computing, for every possible action a, the corresponding expected utility. To be more precise, we usually know, for each action a and for each actual value of the (unknown) quantity x, the corresponding value of the utility $u_a(x)$. We must use the probability distribution for x to compute the expected value $E[u_a(x)]$ of this utility.

In view of this application, the most useful characteristics of a probability distribution would be the ones which would enable us to compute the expected value $E[u_a(x)]$ of different functions $u_a(x)$.

Which representations are the most useful for this intended usage? General idea. Which characteristics of a probability distribution are the most useful for computing mathematical expectations of different functions $u_a(x)$? The answer to this question depends on the type of the function, i.e., on how the utility value u depends on the value x of the analyzed parameter.

Smooth utility functions naturally lead to moments. One natural case is when the utility function $u_a(x)$ is smooth. We have already mentioned, in the previous text, that we usually know a (reasonably narrow) interval of possible values of x. So, to compute the expected value of $u_a(x)$, all we need to know is how the function $u_a(x)$ behaves on this narrow interval. Because the function is smooth, we can expand it into Taylor series. Because the interval is narrow, we can consider only linear and quadratic terms in this expansion and safely ignore higher-order terms:

$$u_a(x) \approx c_0 + c_1 \cdot (x - x_0) + c_2 \cdot (x - x_0)^2,$$

where x_0 is a point inside the interval. Thus, we can approximate the expected value of this function by the expected value of the corresponding quadratic expression:

$$E[u_a(x)] \approx E[c_0 + c_1 \cdot (x - x_0) + c_2 \cdot (x - x_0)^2],$$

i.e., by the following expression:

$$E[u_a(x)] \approx c_0 + c_1 \cdot E[x - x_0] + c_2 \cdot E[(x - x_0)^2].$$

So, to compute the expectations of such utility functions, it is sufficient to know the first and second moments of the probability distribution.

Case of several variables. In the above text, we assumed that the situation is fully described by the value of a single random variable x. In practice, usually, we need several variables to describe the situation. For the case when we have several random variables $x_1, \ldots, x_n$, we can similarly expand the dependence of the smooth utility function $u_a(x_1, \ldots, x_n)$ in Taylor series and keep linear and quadratic terms in this expansion:

$$u_a(x_1, \ldots, x_n) \approx c_0 + \sum_{i=1}^{n} c_{1i} \cdot (x_i - x_{i0}) + \sum_{i=1}^{n} c_{2i} \cdot (x_i - x_{i0})^2 + \sum_{i=1}^{n} \sum_{j \neq i} c_{2ij} \cdot (x_i - x_{i0}) \cdot (x_j - x_{j0}).$$

Thus, we can approximate the expectation of this function by the expectation of the corresponding quadratic expression:

$$E[u_a(x)] \approx E\left[c_0 + \sum_{i=1}^{n} c_{1i} \cdot (x_i - x_{i0}) + \sum_{i=1}^{n} c_{2i} \cdot (x_i - x_{i0})^2 + \sum_{i=1}^{n} \sum_{j \neq i} c_{2ij} \cdot (x_i - x_{i0}) \cdot (x_j - x_{j0})\right],$$

i.e., by the following expression:

$$E[u_a(x)] \approx c_0 + \sum_{i=1}^{n} c_{1i} \cdot E[x_i - x_{i0}] + \sum_{i=1}^{n} c_{2i} \cdot E\left[(x_i - x_{i0})^2\right] + \sum_{i=1}^{n} \sum_{j \neq i} c_{2ij} \cdot E[(x_i - x_{i0}) \cdot (x_j - x_{j0})].$$

So, to compute the expectations of such utility functions, it is sufficient, in addition to the first and second moments of all the variables x_i, to also know the "mixed" moments

$$E[(x_i - x_{i0}) \cdot (x_j - x_{j0})],$$

which correspond, e.g., to covariance.

In decision making, non-smooth utility functions are common. In decision making, not all dependencies are smooth. There is often a threshold x_0 after which, say, a concentration of a certain chemical becomes dangerous.

This threshold sometimes comes from the detailed chemical and/or physical analysis. In this case, when we increase the value of this parameter, we see the drastic increase in effect and hence, the drastic change in utility value. Sometimes, this threshold simply comes from regulations. In this case, when we increase the value of this parameter past the threshold, there is no drastic increase in effects, but there is a drastic decrease of utility due to the necessity to pay fines, change technology, etc. In both cases, we have a utility function which experiences an abrupt decrease at a certain threshold value x_0.

Non-smooth utility functions naturally lead to cumulative distribution functions (cdfs). We want to be able to compute the expected value $E[u_a(x)]$ of a function $u_a(x)$ which

- changes smoothly until a certain value x_0,
- then drops its value and continues smoothly for $x > x_0$.

We usually know the (reasonably narrow) interval which contains all possible values of x. Because the interval is narrow and the dependence before and after the threshold is smooth, the resulting change in $u_a(x)$ before x_0 and after x_0 is much smaller than the change at x_0. Thus, with a reasonable accuracy, we can ignore the small changes before and after x_0, and assume that the function $u_a(x)$ is equal to a constant u^+ for $x < x_0$, and to some other constant $u^- < u^+$ for $x > x_0$.

The simplest case is when $u^+ = 1$ and $u^- = 0$. In this case, the desired expected value $E[u_a^{(0)}(x)]$ coincides with the probability that $x < x_0$, i.e., with the corresponding value $F(x_0)$ of the cumulative distribution function (cdf). A generic function $u_a(x)$ of this type, with arbitrary values u^- and u^+, can be easily reduced to this simplest case, because, as one can easily check, $u_a(x) = u^- + (u^+ - u^-) \cdot u^{(0)}(x)$ and hence,

$$E[u_a(x)] = u^- + (u^+ - u^-) \cdot F(x_0).$$

Thus, to be able to easily compute the expected values of all possible non-smooth utility functions, it is sufficient to know the values of the cdf $F(x_0)$ for all possible x_0.

Summarizing: which description should we select. Our analysis shows, depending on the application, different representation are optimal: it can be moments and covariances, it can be values of the cdf.

Since these two representations are not equivalent in the case of the usual (interval) approach to uncertainty, it is therefore desirable to come up with a new representation of uncertainty that would bring equivalence back.

III. Analysis of the Problem

Let us formulate the problem in precise mathematical terms. In all representations of a probability distribution, we represent this distribution by storing the values of different numerical characteristics of this distribution:

- for cdf, we store the values of the cdf $F(x)$ at different points x;
- for pdf, we store the values of the pdf $\rho(x)$ at different points x;
- for moments, we store the values of M_k for different $k = 1, 2, \ldots$;

- for characteristic functions, we store the values corresponding to different ω, etc.

To describe the distribution exactly, we need to know the values of infinitely many characteristics:

- we need to know the values $F(x)$ (correspondingly, $\rho(x)$) corresponding to infinitely many points x;
- we need to know the moments M_k corresponding to all infinitely many integers k, etc.

In practice, at any given moment of time, we can only store values of finitely many characteristics. Let us denote the total number of these values by n, and the values themselves by $v_1, \ldots, v_n$.

In the case of exact knowledge, we know the exact values of all n characteristics v_i, i.e., we know the exact point $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$. Under uncertainty, we do not know the exact point v; several such points are consistent with our knowledge, i.e., we have a *set* $V \subseteq \mathbb{R}^n$ of such points. What we need is a way to represent such sets.

Traditional way of representing uncertainty. The above way of representing uncertainty means that for each characteristic v_i, we find an interval $[\underline{v}_i, \overline{v}_i]$ of its possible values. In this case, the set of all possible combinations $v = (v_1, \ldots, v_n)$ forms a *box*

$$[\underline{v}_1, \overline{v}_1] \times \ldots \times [\underline{v}_n, \overline{v}_n];$$

(this is where the name of the p-box comes from). The problem is that to go from $F(x)$ to $\rho(x)$, we need linear transformations, and in general, a linear transformation transforms a box into a parallelepiped – and not into a box. In other words, what was a box in one representation becomes a different objects in another one, and so, different representations are not equivalent.

What we need. What we need is to come up with a different representation, i.e., with a different family of sets – not boxes, a family that would transform into a similar representation if we apply all appropriate transformations. In other words, we need a family F of sets so that if we start with a set V from this family, and apply an appropriate transformation T to each element $v \in V$ of this set, then the resulting new set $T(V) \stackrel{\text{def}}{=} \{T(v) : v \in V\}$ should also be an element of the family F.

What are appropriate transformations. The formulas for all the above characteristics are linear in $\rho(x)$, so all appropriate transformations are *linear.*

Need to speed up computations. The need to process uncertainty is ubiquitous. Often, these problems appear in situations like automatic control, when we need to make decision very fast. In such situations, the faster the data processing, the better. In general, the more parameters we need to process, the longer our computations; it is therefore desirable to select a family of sets which would need the smallest possible number of parameters to describe.

Comment. Of course, the computational complexity depends not only on the number of parameters, it also depends on the complexity of the corresponding computations. We will see that if we minimize the number of parameters, computations will become more efficient as well.

IV. Ellipsoids Are Optimal: First Result

Now, we are ready to formulate our main result.

Definition 1. *By a* closed domain*, we mean a closed set that is equal to the closure of the set of its interior points.*

Definition 2. *Let M and N be smooth manifolds.*

- *By a* multi-valued function $F : M \to N$ *we mean a function that maps each $m \in M$ into a discrete set $F(m) \subseteq N$.*
- *We say that a multi-valued function is* smooth *if for every point $m_0 \in M$ and for every value $f_0 \in F(m)$, there exists an open neighborhood U of m_0 and a smooth function $f : U \to N$ for which $f(m_0) = f_0$ and for every $m \in U$, $f(m) \subseteq F(m)$.*

Definition 3. *Let G be a Lie transformation group on a smooth manifold M.*

- *We say that a class A of closed subsets of M is G-*invariant *if for every set $X \in A$, and for every transformation $g \in G$, the set $g(X)$ also belongs to the class.*
- *If A is a $G-$invariant class, then we say that A is a* finitely parametric family of sets *if there exist:*
  - *a (finite-dimensional) smooth manifold V;*
  - *a mapping s that maps each element $v \in V$ into a set $s(v) \subseteq M$; and*
  - *a smooth multi-valued function $\Pi : G \times V \to V$*

  such that:
  - *the class of all sets $s(v)$ that corresponds to different $v \in V$ coincides with A, and*
  - *for every $v \in V$, for every transformation $g \in G$, and for every $\pi \in \Pi(g, v)$, the set $s(\pi)$ (that corresponds to π) is equal to the result $g(s(v))$ of applying the transformation g to the set $s(v)$ (that corresponds to v).*
- *Let $r > 0$ be an integer. We say that a class of sets B is a r-*parametric class *of sets if there exists a finite-dimensional family of sets A defined by a triple (V, s, Π) for which B consists of all the sets $s(v)$ with v from some $r-$dimensional sub-manifold $W \subseteq V$.*

Theorem 1. *Let $n > 0$ be an integer, $M = \mathbb{R}^n$, G_e be the group of all linear (affine) transformations*

$$v_i \to a_i + \sum_{j=1}^{n} a_{ij} \cdot v_j,$$

and B be a G_e-invariant r-parametric family of connected bounded closed domains from $\mathbb{R}^n$. Then:

- *$r \geq \dfrac{n(n+3)}{2}$; and*
- *if $r = \dfrac{n(n+3)}{2}$, then B coincides either with the family of all ellipsoids, or, for some $\lambda \in (0, 1)$, with the family*

of all regions obtained from ellipsoids by subtracting λ times smaller homothetic ellipsoids.

Comment. If we restrict ourselves to *convex* sets (or only to simply connected sets), we get ellipsoids only. So, the family of ellipsoids is indeed optimal – in the sense that it needs the smallest possible number of parameters to describe.

In view of this result, when we describe probabilistic uncertainty, instead of p-boxes, we should consider *p-ellipsoids*, i.e., ellipsoid-shaped regions in the linear space of all possible cdf functions $F(x)$.

Historical comment. Our proof is similar to the proofs of similar results presented in [9] and [15].

Proof.

1°. Let us first show that $r \geq \dfrac{n(n+3)}{2}$. Indeed, it is known (see, e.g., [3]) that for every open bounded set X, among all ellipsoids that contain X, there exists a unique ellipsoid E of the smallest volume. We will say that this ellipsoid E *corresponds* to the set X. Let us consider the set of ellipsoids $\mathcal{E}_c$ that correspond (in this sense) to all possible sets $X \in A$.

Let us fix a set $X_0 \in B$, and let E_0 denote an ellipsoid that corresponds to X_0.

An arbitrary ellipsoid E can be obtained from any other ellipsoid (in particular, from E_0) by an appropriate affine transformation g: $E = g(E_0)$. The ratio of volumes is preserved under arbitrary linear transformations g; hence, since the ellipsoid E_0 is the smallest volume ellipsoid that contains X_0, the ellipsoid $E = g(E_0)$ is the smallest volume ellipsoid that contains $g(X_0) = X$.

Hence, an arbitrary ellipsoid $E = g(E_0)$ corresponds to some set $g(X_0) \in B$. Thus, the family $\mathcal{E}_c$ of all ellipsoids that correspond to sets from A is simply equal to the set $\mathcal{E}$ of all ellipsoids. Thus, we have a (locally smooth) mapping from an r-dimensional set A onto the $\dfrac{n(n+3)}{2}$-dimensional set of all ellipsoids. Hence, $r \geq \dfrac{n(n+3)}{2}$.

2°. Let us now show that for $r = \dfrac{n(n+3)}{2}$, the only G_e-invariant families A are ellipsoids and "ellipsoid layers" (described in the formulation of the Theorem).

Indeed, let X_0 be an arbitrary set from the invariant family B, and let E_0 be the corresponding ellipsoid. Let $g_0 \in G_e$ be an affine transformation that transform E_0 into a ball $E_1 = g(E_0)$. This ball then contains the set $X_1 = g_0(E_0) \in B$.

Let us show, by reduction to a contradiction, that the set X_1 is invariant w.r.t. arbitrary rotations around the center of the ball E_1. Indeed, if it is not invariant, then the set R of all rotations that leave X_1 invariant is different from the set of all rotations $SO(n)$. Hence, R is a proper closed subgroup of $SO(n)$. From the structure of $SO(n)$, it follows that there exists a $1-$parametric subgroup R_1 of $SO(n)$ that intersects with R only in the identity transformation 1. This means that if $g \in R_1$ and $g \neq 1$, we have $g \notin R$, i.e., $g(X_1) \neq X_1$.

If $g(X_1) = g'(X_1)$ for some $g, g' \in R_1$, then we have $g^{-1}g'(X_1) = X_1$, where $g^{-1}g' \in R_1$. But such an equality is only possible for $g^{-1}g' = 1$, i.e., for $g = g'$. Thus, if $g, g' \in R_1$ and $g \neq g'$, then the sets $g(X_1)$ and $g'(X_1)$ are different. In other words, all the sets $g(X_1)$, $g \in R_1$, are different.

Since the family B is G_e-invariant, all the sets $g(X_1)$ for all $g \in R_1 \subseteq G_e$ also belong to A. For all these sets, the corresponding ellipsoid is $g(E_1)$, the result of rotating the ball E_1, i.e., the same ball $g(E_1) = E_1$. Hence, we have a $1-$parametric family of sets contained in the ball E_1.

By applying appropriate affine transformations, we will get 1-parametric families of sets from B in an arbitrary ellipsoid. So, we have an $\dfrac{n(n+3)}{2}$-dimensional family of ellipsoids, and inside each ellipsoid, we have a 1-dimensional family of sets from B. Thus, B would contain a $\left(\dfrac{n(n+3)}{2}+1\right)$-parametric family of sets, which contradicts to our assumption that the dimension r of the family B is exactly $\dfrac{n(n+3)}{2}$.

This contradiction shows that our initial assumption was false, and for $r = \dfrac{n(n+3)}{2}$, the set X_1 is invariant w.r.t. rotations. Hence, with an arbitrary point x, the set X_1 contains all the points that can be obtained from x by arbitrary rotations, i.e., the entire sphere that contains x. Since X_1 is connected, X_1 is either a ball, or a ball from which a smaller ball was deleted.

The original set $X_0 = g_0^{-1}(X_1)$ is an affine image of this set X_1, and therefore, X_0 is either an ellipsoid, or an ellipsoid with an ellipsoidal hole inside. The theorem is proven.

V. Ellipsoids Are Optimal: Second Result

In the previous section, we showed that ellipsoids are optimal in the sense that the family of ellipsoids requires the smallest possible number of parameters to describe. Let us show that ellipsoids are optimal with respect to many other optimality criteria as well. Moreover, we will prove that under reasonable conditions on an optimality criterion, a family of sets which is optimal with respect to this criterion consists of ellipsoids.

What is an optimality criterion? When we say "the best", we mean that on the set of all such families, there must be a relation $\succeq$ describing which family is better or equal in quality. This relation must be transitive (if B is better than B', and B' is better than B'', then B is better than B''). This relation is not necessarily asymmetric, because we can have two approximating families of the same quality. However, we would like to require that this relation be *final* in the sense that it should define a unique *best* family $B_{\rm opt}$ (i.e., the unique family for which $\forall B\,(B_{\rm opt} \succeq B)$). Indeed:

- If none of the families is the best, then this criterion is of no use, so there should be *at least one* optimal family.
- If *several* different families are equally best, then we can use this ambiguity to optimize something else: e.g., if we have two families with the same approximating quality,

then we choose the one which is easier to compute. As a result, the original criterion was not final: we get a new criterion ($B \succeq_{\rm new} B'$ if either B gives a better approximation, or if $B \sim_{\rm old} B'$ and B is easier to compute), for which the class of optimal families is narrower. We can repeat this procedure until we get a final criterion for which there is only one optimal family.

An optimality criterion should be invariant. It is reasonable to require that what is better in one representation should be better in another representation as well. In other words, it is reasonable to require the relation $B \succeq B'$ should be invariance relative to the affine transformations.

Definition 4. *Let $\mathcal{A}$ be a class of families of sets, and let G be a group of transformations defined on $\mathcal{A}$.*

- *By an* optimality criterion*, we mean a* pre-ordering *(i.e., a transitive reflexive relation) $\preceq$ on the class $\mathcal{A}$.*
- *We say that an optimality criterion is* G-invariant *if for all $g \in G$, and for all $B, B' \in \mathcal{A}$, $B \preceq B'$ implies*

$$g(B) \preceq g(B').$$

- *We say that an optimality criterion is* final *if there exists one and only one element $B_{\rm opt} \in \mathcal{A}$ that is preferable to all the others, i.e., for which $B \preceq B_{\rm opt}$ for all $B \neq B_{\rm opt}$.*

Theorem 2. *Let $n > 0$ be an integer, $M = \mathbb{R}^n$, G_e be the group of all affine transformations, and $\preceq$ be a G_e-invariant and final optimality criterion on the class $\mathcal{A}$ of all r-parametric families of connected bounded closed domains from $\mathbb{R}^n$. Then:*

- *$r \geq \dfrac{n(n+3)}{2}$; and*
- *if $r = \dfrac{n(n+3)}{2}$, then the optimal family coincides either with the family of all ellipsoids, or, for some $\lambda \in (0,1)$, with the family of all regions obtained from ellipsoids by subtracting λ times smaller homothetic ellipsoids.*

Comment. Similarly to Theorem 1, if we restrict ourselves to *convex* sets (or only to simply connected sets), we get ellipsoids only.

Proof of Theorem 2. Since the criterion $\preceq$ is final, there exists one and only one optimal family of sets. Let us denote this family by $B_{\rm opt}$.

Let us first show that this family $B_{\rm opt}$ is G_e-invariant, i.e., that $g(B_{\rm opt}) = B_{\rm opt}$ for every transformation $g \in G_e$.

Indeed, let $g \in G_e$. From the optimality of $B_{\rm opt}$, we conclude that for every $B \in \mathcal{A}$, $g^{-1}(B) \preceq B_{\rm opt}$. From the G_e-invariance of the optimality criterion, we can now conclude that $B \preceq g(B_{\rm opt})$. This is true for all $B \in \mathcal{A}$ and therefore, the family $g(B_{\rm opt})$ is optimal. But since the criterion is final, there is only one optimal family; hence, $g(B_{\rm opt}) = B_{\rm opt}$. So, $B_{\rm opt}$ is indeed invariant.

Now, the result follows from Theorem 1.

VI. Examples of How Ellipsoids Are Better Than Boxes

Historical comment. Before we start listing the way in which ellipsoids are better in *probability* representations, we should mention that *in general*, ellipsoids have been successfully used to represent uncertainty (and, more generally, to represent different sets); see, e.g., [2], [4], [5], [8], [11], [18], [21], [22], [24], [25]). Several other families of sets have been proposed to describe uncertainty, such as boxes, parallelepipeds, polytopes, etc. Experimental comparison of different families has lead to a conclusion that in many practical situations, *ellipsoids* indeed lead to the best results; see, e.g., [4], [5].

Ellipsoids are known to work better in *linear programming*, where we need to find minima or maxima of linear functions on a set defined by a system by linear inequalities (i.e., on a convex polytope). The traditionally used *simplex method* uses the original polytope; this method is, on average, very efficient, but in the worst case, it requires the unrealistic exponential number of computational steps ($\approx 2^n$, where n is the number of unknowns). For several decades, researchers have tried to find a polynomial time algorithm for linear programming. Success only came when they decided to approximate the original polytope with an ellipsoid; this lead to the well-known polynomial time algorithms of Khachiyan [13] and Karmarkar [12].

Ellipsoids also turned out to be better than polytopes or parallelepipeds (boxes) in many *pattern recognition* problems; see, e.g., [1].

Ellipsoids lead to faster computations. In many practical situations, based on a probability distribution $v = (v_1, \ldots, v_n)$, we need to estimate the value of a statistical characteristic $S(v_1, \ldots, v_n)$. In the case of uncertainty, we only know the range V of possible values of v. Different distributions $v \in V$ lead, in general, to different values of $S(v)$. It is therefore desirable to compute the *range* $S(V) \stackrel{\rm def}{=} \{S(v) : v \in V\}$ of possible values of the given characteristic.

The lower endpoint of this range is the minimum of the function $S(v)$ over the set V, while the upper endpoint of this range is the minimum of the function $S(v)$ over the set V. So, computing the range means computing the minimum and the maximum of a given function $S(v)$ over the set V.

In many cases, we have a reasonably good knowledge about the probability distribution, so we can expand the dependence $S(v)$ around an approximate estimate $\widetilde{v}$ and keep only terms which are linear and quadratic in $\Delta v \stackrel{\rm def}{=} v - \widetilde{v}$. In other words, instead of general functions $S(v)$, it is sufficient to consider functions which are quadratic in v. Thus, we face a problem of finding the minimum and the maximum of a quadratic function

$$S(v) = s_0 + \sum_{i=1}^{n} s_i \cdot v_i + \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} \cdot v_i \cdot v_j$$

on a given set V.

Under the traditional interval-type representation, the set V is the set of all $v = (v_1, \ldots, v_n)$ for which $\underline{v}_i \leq v_i \leq \overline{v}_i$. At

first glance, optimizing a quadratic function over such a box is an easy computational problem. Indeed, according to calculus, a function $S(\ldots,v_i,\ldots)$ attains its maximum or minimum on an interval $[\underline{v}_i,\overline{v}_i]$ either at one of the endpoints $\underline{v}_i$, $\overline{v}_i$ of this interval, or at a point where the derivative is equal to 0:

$$\frac{\partial S}{\partial v_i}=0.$$

The derivative of a quadratic function is linear, so for each i from 1 to n, we have one of the three linear equations:

- $v_i=\underline{v}_i$;
- $v_i=\overline{v}_i$; or
- $\dfrac{\partial S}{\partial v_i}=0$.

For each combination of these equations, we have an easy-to-solve system of n linear equations for finding n unknowns $v_1,\ldots,v_n$. However, while each system is easy to solve, there are 3^n possible combinations of these equations, which makes this approach computationally non-feasible.

Moreover, it has been shown that in general, the problem of computing the minimum or maximum of a quadratic function over a box is NP-hard; see, e.g., [14]. This means, crudely speaking, that under the hypothesis P≠NP (which most computer scientists believe to be true), it is not possible to have an algorithm for solving all particular cases of this problem in feasible time.

In contrast, if we optimize a quadratic function over an ellipsoid, i.e., under the condition

$$E(v)=e_0+\sum_{i=1}^{n}e_i\cdot v_i+\sum_{i=1}^{n}\sum_{j=1}^{n}e_{ij}\cdot v_i\cdot v_j\le 1,$$

then this minimum (maximum) is either attained inside the ellipsoid – in which case we need to solve a system of n linear equations $\dfrac{\partial S}{\partial v_i}=0$ – or this extremum is attained at the border of the ellipsoid, i.e., at a point where

$$E(v)=e_0+\sum_{i=1}^{n}e_i\cdot v_i+\sum_{i=1}^{n}\sum_{j=1}^{n}e_{ij}\cdot v_i\cdot v_j=1.$$

To maximize the objective function $S(v)$ under the constraint $E(v)=1$, we can use the Lagrange multiplier method and maximize the quadratic expression

$$S_\lambda(v)\stackrel{\rm def}{=}S(v)+\lambda\cdot(E(v)-1).$$

For each value λ, the conditions

$$\frac{\partial S_\lambda}{\partial v_i}=0$$

lead to an easy-to-solve system of linear equations. Out of all solutions v_λ, we select the one for which $E(v_\lambda)=1$. This equation with one unknown λ is computationally easy to solve, so the whole optimization is feasible. In other words, replacing p-boxes with p-ellipsoids indeed leads to faster computations.

Ellipsoids are in good agreement with additional probabilistic information. Often, in additional to *set* of possible values $v=(v_1,\ldots,v_n)$, we also have an information about which values v are more probably and which values v are less probable. In other words, we have a probability distribution on the set V of possible probability distributions.

There are usually many different reasons for the uncertainty with which we know v, i.e., for the difference between the actual (unknown) values of the parameters v_i and the estimates $\widetilde{v}_i$. Each of these reasons contributes to the difference $v_i-\widetilde{v}_i$, so this difference can be viewed as a sum of a large number of independent small contributions.

It is known that the distribution of such a sum is close to Gaussian; this fact follows from the Central Limit Theorem; see, e.g., [23]. Thus, it is reasonable to conclude that the probability distribution on the set $V\subseteq\mathbb{R}^n$ is Gaussian, with a Gaussian probability density $\rho_V(v)$.

Strictly speaking, a Gaussian distribution has positive density for all possible vector $v\in\mathbb{R}^n$. In practice, we dismiss values for which the probability is too small as not realistically possible. In mathematical terms, this means that we describe the set V of (practically) possible values v as the set of all the vectors v for which $\rho_V(v)\ge\rho_0$ for some threshold $\rho_0>0$. It is known that for a Gaussian distribution, this inequality describes an ellipsoid. Thus, ellipsoids are indeed in perfect agreement with the additional probabilistic information.

If we reconstruct a p-ellipsoid from data instead of a p-box, we get better estimates. Whether we use p-boxes or p-ellipsoids, we need to extract them from the observed data. When we use p-boxes, then, according to [7], the corresponding p-box can be extracted by using Kolmogorov-Smirnov criterion of a match between the empirical data $x_1,\ldots,x_n$ and the hypothetic probability distribution $F(x)$. This criterion is based on the bounds on the maximum $\max|F(x)-F_n(x)|$, where $F_n(x)$ is the cdf of the empirical distribution

$$F_n(x)\stackrel{\rm def}{=}\frac{\#\{i:x_i\le x\}}{n}.$$

Specifically, the Kolmogorov-Smirnov criterion produces upper bounds Δ for which $\max|F(x)-F_n(x)|\le\Delta$ with a given confidence level. Once the confidence level is selected and we have the bound Δ, we can then conclude that the actual (unknown) cdf $F(x)$ belongs to the interval $[F_n(x)-\Delta,F_n(x)+\Delta]$.

For p-ellipsoids, we can similarly use Cramer-von Mises ω^2 criterion for goodness of fit. This criterion is based on the bounds for the integral $\int(F(x)-F_n(x))^2\,dF_n(x)$. Once the bound Δ on this integral is calculated for a given confidence level, we can then conclude that the actual (unknown) cdf satisfies the inequality $\int(F(x)-F_n(x))^2\,dF_n(x)\le\Delta$. In terms of the ordered sample

$$x_{(1)}\le x_{(2)}\le\ldots\le x_{(n)},$$

this inequality has the form $\omega_n^2\le\Delta$, where

$$\omega_n^2=\frac{1}{n}\cdot\sum_{i=1}^{n}\left(F(x_{(i)})-\frac{2\cdot i-1}{2\cdot n}\right)^2+\frac{1}{12\cdot n^2}.$$

In geometric terms, this quadratic inequality describes an ellipsoid, so we get the desired p-ellipsoid.

It turns out that p-ellipsoids lead to more accurate estimates than p-boxes. As an example, we can consider computing the range for the mean $\int x \cdot \rho(x)\,dx = \int x\,dF(x)$.

The Cramer-von Mises criterion does not change if we apply an arbitrary non-linear transformation $x \to f(x)$. Since every distribution can be contained from a uniform one by using such a transformation, it is sufficient to use samples $x_1, \ldots, x_n$ which are distributed according to a uniform distribution on an interval $[0, 1]$. We use each such sample to extract a p-box and a p-ellipsoid (we use 95% confidence level in both cases). Then, we use the resulting p-box and the resulting p-ellipsoid to come up with two estimates for the range of the mean. For each n, we repeat the experiment 10^5 times.

For $n = 10$, 25, 50, 100, and 200, we compared the widths w_{KS} and w_{CvM} of the resulting interval estimates for the mean m with the width w_t estimated based on the t-test. In all the cases, the width w_{CvM} based on p-ellipsoids is smaller than the width w_{KS} based on p-boxes:

n	10	25	50	100	200
$\frac{w_{KS}}{w_t}$	1.46	2.05	2.17	2.25	1.88
$\frac{w_{CvM}}{w_t}$	1.21	1.60	1.65	1.67	1.49

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, by Grant 1 T36 GM078000-01 from the National Institutes of Health, and by a grant on F-transforms from the Office of Naval Research.

The authors would like to thank all the participants of the 15th GAMM – IMACS International Symposium on Scientific Computing, Computer Arithmetic, and Verified Numerical Computation SCAN'2012 (Novosibirsk, Russia, September 23–29, 2012), especially Sergey Shary, for inspiring discussions, and to the anonymous referees for valuable suggestions.

REFERENCES

[1] S. Abe and R. Thawonmas, "Fast training of a fuzzy classifier with ellipsoidal regions", *Proc. 1996 IEEE International Conference on Fuzzy Systems*, New Orleans, September 8–11, 1996, pp. 1875–1880.

[2] G. Belforte and B. Bona, "An improved parameter identification algorithm for signal with unknown-but-bounded errors", *Proceeding of the 7th IFAC Symposium on Identification and Parameter Estimation*, York, U.K., 1985.

[3] H. Busemann, *The Geometry of Geodesics*, Academic Press, N.Y., 1955.

[4] F. L. Chernousko, *Estimation of the Phase Space of Dynamic Systems*, Nauka publ., Moscow, 1988 (in Russian).

[5] F. L. Chernousko, *State Estimation for Dynamic Systems*, CRC Press, Boca Raton, FL, 1994.

[6] S. Ferson, *RAMAS Risk Calc 4.0*, CRC Press, Boca Raton, Florida, 2002.

[7] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf, and L. Ginzburg, *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*, Sandia National Laboratories, 2007, Publ. 2007-0939.

[8] A. F. Filippov, "Ellipsoidal estimates for a solution of a system of differential equations", *Interval Computations*, 1992, No. 2(4), pp. 6–17.

[9] A. Finkelstein, O. Kosheleva, and V. Kreinovich, "Astrogeometry, error estimation, and other applications of set-valued analysis", *ACM SIGNUM Newsletter*, 1996, Vol. 31, No. 4, pp. 3–25.

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

[11] E. Fogel and Y. F. Huang, "On the value of information in system identification. Bounded noise case", *Automatica*, 1982, Vol. 18, pp. 229–238.

[12] N. Karmarkar, "A new polynomial-time algorithm for linear programming", *Combinatorica*, 1984, Vol. 4, pp. 373–396.

[13] L. G. Khachiyan, "A polynomial-time algorithm for linear programming", *Soviet Math. Dokl.*, 1979, Vol. 20, pp. 191–194.

[14] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1997.

[15] S. Li, Y. Ogura, and V. Kreinovich, *Limit Theorems and Applications of Set Valued and Fuzzy Valued Random Variables*, Kluwer Academic Publishers, Dordrecht, 2002.

[16] R. D. Luce and R. Raiffa, *Games and Decisions: Introduction and Critical Survey*, Dover, New York, 1989.

[17] H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, 2012.

[18] J. P. Norton, "Identification and application of bounded parameter models", *Proceeding of the 7th IFAC Symposium on Identification and Parameter Estimation*, York, U.K., 1985.

[19] S. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer Verlag, New York, 2005.

[20] H. Raiffa, *Decision Analysis*, McGraw-Hill, Columbus, Ohio, 1997.

[21] F. C. Schweppe, "Recursive state estimation: unknown but bounded errors and system inputs", *IEEE Transactions on Automatic Control*, 1968, Vol. 13, p. 22.

[22] F. C. Schweppe, *Uncertain Dynamic Systems*, Prentice Hall, Englewood Cliffs, NJ, 1973.

[23] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2004.

[24] S. T. Soltanov, "Asymptotic of the function of the outer estimation ellipsoid for a linear singularly perturbed controlled system", In: S. P. Shary and Yu. I. Shokin (eds.), *Interval Analysis*, Krasnoyarsk, Academy of Sciences Computing Center, Technical Report No. 17, 1990, pp. 35–40 (in Russian).

[25] G. S. Utyubaev, "On the ellipsoid method for a system of linear differential equations", In: S. P. Shary (ed.), *Interval Analysis*, Krasnoyarsk, Academy of Sciences Computing Center, Technical Report No. 16, 1990, pp. 29–32 (in Russian).

[26] P. Walley, *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, New York, 1991.