

11-1-2001

# Representation, Elicitation, and Aggregation of Uncertainty in Risk Analysis - from Traditional Probabilistic Techniques to More General, More Realistic Approaches: A Survey

Scott Ferson

Vladik Kreinovich

*University of Texas at El Paso*, [vladik@utep.edu](mailto:vladik@utep.edu)

Follow this and additional works at: [http://digitalcommons.utep.edu/cs\\_techrep](http://digitalcommons.utep.edu/cs_techrep)



Part of the [Computer Engineering Commons](#)

Comments:

UTEP-CS-01-33.

---

## Recommended Citation

Ferson, Scott and Kreinovich, Vladik, "Representation, Elicitation, and Aggregation of Uncertainty in Risk Analysis - from Traditional Probabilistic Techniques to More General, More Realistic Approaches: A Survey" (2001). *Departmental Technical Reports (CS)*. Paper 420.

[http://digitalcommons.utep.edu/cs\\_techrep/420](http://digitalcommons.utep.edu/cs_techrep/420)

This Article is brought to you for free and open access by the Department of Computer Science at DigitalCommons@UTEP. It has been accepted for inclusion in Departmental Technical Reports (CS) by an authorized administrator of DigitalCommons@UTEP. For more information, please contact [lweber@utep.edu](mailto:lweber@utep.edu).

**Representation, Elicitation, and  
Aggregation of Uncertainty  
in Risk Analysis –  
From Traditional  
Probabilistic Techniques  
to More General,  
More Realistic Approaches:  
A Survey**

Scott Ferson and Vladik Kreinovich

**Outline:**

1. Introduction: Uncertainty in Risk Analysis	3
2. What Is a Natural Way of Representing Partial Information About Probabilities?	5
3. Elicitation of Uncertainty in Risk Analysis: Where Do p-Bounds Come From?	13
4. Aggregation of Uncertainty in Risk Analysis: Subjective vs. Mathematical Methods	36
5. Aggregation of Uncertainty in Risk Analysis: Outline	41
6. Aggregation of Uncertainty in Risk Analysis: Desirable Properties of Aggregation Operations	43
7. Aggregation of Uncertainty in Risk Analysis: Case of Real Numbers	50
8. Aggregation of Uncertainty in Risk Analysis: Case of Intervals	69
9. Aggregation of Uncertainty in Risk Analysis: Case of Probability Distributions	80
10. Aggregation of Uncertainty in Risk Analysis: General Case	113
References	127

# 1 Introduction: Uncertainty in Risk Analysis

## 1.1 Uncertainty in risk analysis: why

By definition, risk analysis deals with situations with uncertainty, i.e., with situations in which we do not have a complete and accurate knowledge about the state of the system. It is therefore very important that we be able to represent uncertainty in risk analysis as adequately as possible.

## 1.2 First component of uncertainty description: interval (set) uncertainty

In order to fully describe a system, we must know the exact values of all the physical quantities characterizing this system. For example, in environmental problems related to chemical pollution, a polluted system (e.g., a lake) can be fully described if we know the exact concentration of different pollutants in different parts of the lake.

Thus, to describe the uncertainty of our knowledge about a system, we must describe the uncertainty with which we know the values of each of the quantities (parameters) describing the system. Uncertainty means that we do not know the exact value of the quantity, several different values may be possible. For example, we may not know the exact value of the concentration but we may know that this concentration is between, say,  $10^{-5}$  and  $10^{-3}$ . In this case, any value from the interval  $[10^{-5}, 10^{-3}]$  is possible (Moore, 1979), (Hansen, 1992), (Hammer et al, 1993), (Kearfott, 1996), (Kearfott & Kreinovich, 1996), (Berleant & Kuipers, 1997), (Interval, 2001).

An important risk-related situation that leads to intervals is when a measurement does not detect any presence of a certain substance because its concentration  $x$  is below the detection limit  $D$ . In this case, the only information we have about  $x$  is that  $x$  belongs to the interval  $[0, D]$ .

In general, we usually know an interval  $\mathbf{x}$  of possible values of the unknown quantity  $x$  – or, sometimes, a more general set  $X$  of possible values of  $x$  (different from an interval, e.g., the union of two intervals).

## 1.3 Second components of uncertainty description: probabilistic uncertainty

The set  $X$  of possible values describes which values of the analyzed quantity are possible and which values are not. In addition to this information, we often know which values are more probable and which are less probable. In other words, we often have some information about the probability of different values  $x$  from the interval (set)  $\mathbf{x}$  of possible values.

## 1.4 Probabilistic uncertainty: traditional techniques

In some cases, we know the exact expression for this distribution. In these cases, we can use standard statistical techniques to represent, elicit, and aggregate uncertainty. A survey of the corresponding techniques as applied to risk analysis is given, e.g., in (Clemen & Winkler, 1999).

## 1.5 The need for techniques corresponding to partial information about probabilities

In many other real-life situations, however, we have only *partial* information about the probabilities. To handle such situations, it is necessary to expand known statistical techniques of representing, eliciting, and aggregating uncertainty to problems in which we only have partial information about the probabilities.

## 1.6 Section's conclusions

Uncertainty is very important in risk analysis. A natural way to describe this uncertainty is to describe a set of possible values of each unknown quantity (this set is usually an interval), plus any additional information that we may have about the probability of different values within this set. Traditional statistical techniques deal with the situations in which we have a complete information about the probabilities; in real life, however, we often have only partial information about them. We therefore need to describe methods of handling such partial information in risk analysis; see, e.g., (Goodman et al., 1997), (Gebhard & Kruse, 1998).

## 1.7 What we are planning to do

The main objective of this report is to present an overview of risk-oriented techniques for dealing with partial information about probabilities. Several such techniques have been presented, often on a heuristic basis, without a proper justification typical for traditional statistical techniques. Our goal is:

- to overview these techniques,
- to describe natural properties of these techniques,
- to provide the justification of each of these techniques, and,
- based on these justifications, to explain when each of these techniques should be used.

## 2 What Is a Natural Way of Representing Partial Information About Probabilities?

### 2.1 Introduction to the problem

#### 2.1.1 Which representation of probability distribution should we choose?

In probability theory, there are many different ways of representing a probability distribution. For example, one can use a probability density function (pdf), or a cumulative distribution function (CDF), or a probability measure, i.e., a function which maps different sets into a probability that the corresponding random variable belongs to this set. The reason why there are many different representations is that in different problems, different representations turned out to be the most useful.

We would like to select a representation which is the most useful for problems related to risk analysis. To make this selection, we must recall where the information about probabilities provided by risk analysis is normally used.

#### 2.1.2 How is the partial information about probabilities used in risk analysis?

The main objective of risk analysis 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  $u_a(x)$ .

#### 2.1.3 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.

## 2.2 Enter moments

### 2.2.1 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 Section 1, 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 safely consider only linear and quadratic terms in this expansion and 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 expectation of this function by the expectation 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.

In particular, if we use, as the point  $x_0$ , the average  $E[x]$ , the second moment turns into the variance of the original probability distribution. So, instead of the first and the second moments, we can use the mean  $E$  and the variance  $V$ .

### 2.2.2 From numerical moments to interval-valued moments

When we know the exact probability distribution, we must use the *exact* values of the first and the second moment (or mean and variance).

If we only have a *partial* information about the probability distribution, then we cannot compute the exact value of these moments; instead, we have *intervals* of possible values of these moments. So, from this viewpoint, a natural representation of the partial information about the probability distribution is given by intervals  $\mathbf{E}$  and  $\mathbf{V}$  of possible values of mean  $E$  and variance  $V$ .

## 2.3 Enter CDFs and p-bounds

### 2.3.1 In risk analysis, non-smooth utility functions are common

In engineering applications, most functions are smooth, so usually the Taylor expansion works pretty well. In risk analysis, however, 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  $x_0$ .

### 2.3.2 Non-smooth utility functions naturally lead to 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  $x_0$ .

### 2.3.3 From CDF to interval-valued CDF

When we know the exact probability distribution, we must use the exact values  $F(x)$  of the CDF. If we only have a *partial* information about the probability distribution, then we cannot compute the exact values  $F(x)$  of the CDF. Instead, for every  $x$ , we have an *interval*  $[F^-(x), F^+(x)]$  of possible values of the probability  $F(x)$ .

*Comment.* That in practice, we often do not have the exact values of the probabilities, only intervals of possible values of these probabilities, is a well-known fact in probability theory and practice, including risk-related practical applications; see, e.g., (Kuznetsov, 1991), (Walley, 1991), (Whalen, 1994), (Nguyen et al., 1997), (Levi, 2000), (de Cooman et al., 2001).

### 2.3.4 The notion of a p-bound

What are the properties of these bounds? From the definition of a CDF, it easily follows that  $x \leq y$  implies  $F(x) \leq F(y)$ . Is a similar property true for the bounds  $F^-(x)$  and  $F^+(x)$ ?

In principle, if we use different methods to compute these lower and upper bounds for different  $x$  and  $y$ , we may get values  $x$  and  $y$  for which  $x \leq y$  and  $F^+(x) > F^+(y)$ . However, this non-monotonicity can be easily corrected. Indeed, due to the fact that  $x \leq y$ , we have  $F(x) \leq F(y)$ . Thus, due to the fact that  $F^+(y)$  is an upper bound for  $F(y)$ , it is also an upper bound for  $F^+(x)$ , and a better upper bound than  $F^+(x)$ . By combining all the bounds corresponding to all the values  $y \geq x$ , we get a new (better) upper bound:

$$F_{\text{new}}^+(x) \stackrel{\text{def}}{=} \max_{y:x \leq y} F^+(y).$$

It is easy to check that the new upper bound is already monotonic, e.g., if  $x \leq y$ , then  $F_{\text{new}}^+(x) \leq F_{\text{new}}^+(y)$ .

Similarly, for every  $y \leq x$ , the lower bound  $F^-(y)$  for  $F(y)$  is also a lower bound for  $F(x)$ . By combining all the bounds corresponding to all the values  $y \leq x$ , we get a new (better) lower bound:

$$F_{\text{new}}^-(x) \stackrel{\text{def}}{=} \max_{y:y \leq x} F^-(y).$$

It is easy to check that the new upper bound is already monotonic, e.g., if  $x \leq y$ , then  $F_{\text{new}}^-(x) \leq F_{\text{new}}^-(y)$ .

In view of this “correcting” procedure, we can, without losing generality, assume that both bounds  $F^-(x)$  and  $F^+(y)$  are monotonic, i.e., that both bounds are CDFs. Such a pair of two CDFs which bounds the (unknown) actual CDF is called a *probability bound*, or a *p-bound*, for short.

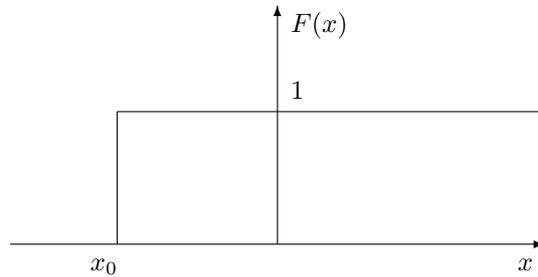
So, from this viewpoint, a natural representation of the partial information about the probability distribution is given by a p-bound.

### 2.3.5 Comment: real numbers, intervals, and probability distributions are particular cases of p-bound

It is worth mentioning that several other types of uncertainty can be viewed as particular cases of p-bound.

For example, the case of complete certainty, when we know the exact value  $x_0$  of the desired quantity, can be represented as a p-bound in which

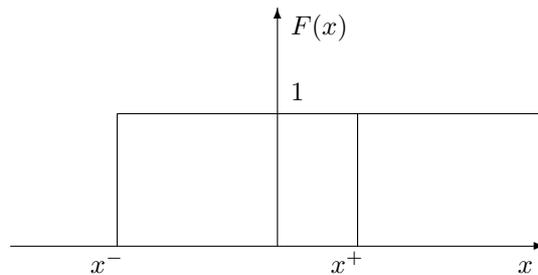
$$F^-(x) = F^+(x) = \begin{cases} 0 & \text{if } x \leq x_0, \\ 1 & \text{otherwise} \end{cases}$$



The case when our only information about  $x$  is that  $x$  belongs to the interval  $[x^-, x^+]$  can be represented by the following p-bound:

$$F^-(x) = \begin{cases} 0 & \text{if } x \leq x^+, \\ 1 & \text{otherwise} \end{cases}$$

$$F^+(x) = \begin{cases} 0 & \text{if } x \leq x^-, \\ 1 & \text{otherwise} \end{cases}$$



Finally, a probability distribution with a CDF  $F(x)$  can be represented as a p-bound with  $F^-(x) = F^+(x) = F(x)$ .

## 2.4 Beyond moments and p-bounds

### 2.4.1 What representation is the most appropriate for both types of decision making problems?

In the previous text, we have argued that:

- for decision problems with smooth utility functions, the best representation is by interval mean and interval variance, and
- for decision problems with discontinuous utility functions, the best representation of partial information is a p-bound.

Therefore, ideally, if we want to be able to use our partial information about the probabilities in all possible decision problems, we should represent this partial information by keeping *both* p-bounds and interval moments.

One of the ultimate objectives of our research is to be able to handle such combined data. We are working on it, but at present, we have not yet fully developed the corresponding techniques. So, instead of keeping both representations (p-bounds and interval moments), we must select one of them.

Of the two corresponding representations of a probability distribution, CDF is much more informative: if we know CDF, we can compute the moments, but if we only know the moments, we can have many different CDFs. Thus, because we want to make our representation as informative as possible, it makes sense to use CDFs and their interval analogues – p-bounds.

p-bounds are actively used in risk analysis; see, e.g., (Ferson et al., 2001), (Ramas, 2001). p-bounds is what we will be using in this survey.

#### 2.4.2 A brief comment: how to transform moments into p-bounds

Because we want to represent every piece of information about a probability distribution as a p-bound, we must, in particular, represent the information about the moments as a p-bound. The general way of representing such an information is given in (Ferson et al., 2001), (Ramas, 2001). For example, if we know the interval  $[x^-, x^+]$  on which the distribution is located, and if we know its mean  $E$ , then we can conclude that  $F(x) \in [F^-(x), F^+(x)]$ , where, e.g.,

$$F^+(x) = \min\left(1, \frac{x^+ - E}{x^+ - x}\right).$$

#### 2.4.3 What if we are not sure about the intervals either: from intervals to fuzzy numbers

Before we start the actual survey, let us make one important comment. In the above arguments, we assumed that, when we are given an interval  $\mathbf{x}$  of possible values of  $x$ , we are thus 100% guaranteed that the actual value  $x$  belongs to this given interval. In reality, often, there is a possibility that  $x$  is outside this interval. For example, the interval  $\mathbf{x}$  may come from statistical analysis, when it arises as confidence interval corresponding to a certain confidence level.

In this case, we know, e.g., that  $x$  belongs to  $\mathbf{x}$  with confidence 99%. In statistics, the probability of an error is usually used as a numerical characteristic of confidence; so, e.g., the case when  $x \in \mathbf{x}$  in 99% of the cases is described by confidence level  $\alpha = 0.01$ .

In principle, most statistical methods enable us to make conclusions of different levels of confidence. Therefore, in addition to the original confidence interval that corresponds to a confidence level  $\alpha$ , we can get another confidence level corresponding to a smaller confidence level  $\alpha' < \alpha$ . However, to decrease the confidence level, i.e., to increase our belief that the actual (unknown) value  $x$  belongs to the interval  $\mathbf{x}$ , we must widen the interval. So, the interval  $\mathbf{x}_{\alpha'}$  correspond to the new confidence level  $\alpha' < \alpha$  must contain the interval  $\mathbf{x}_{\alpha}$  corresponding to the original confidence level  $\alpha$ .

To describe uncertainty, it therefore makes sense to keep not just a *single* interval corresponding to a single confidence level, but to keep *several* intervals corresponding to different confidence levels. In more precise terms, we have several confidence values  $0 < \alpha_1 < \dots < \alpha_n < 1$ , and for each of these values, we have intervals  $\mathbf{x}_{\alpha_i}$  which are *nested* (like Russian matryoshka dolls), i.e., when  $\alpha_i < \alpha_j$ , then  $\mathbf{x}_{\alpha_i} \supseteq \mathbf{x}_{\alpha_j}$ .

In principle, we can have an interval  $\mathbf{x}_\alpha$  corresponding to every possible value  $\alpha \in (0, 1)$ . From the mathematical viewpoint, a sequence of nested intervals is exactly what is called a *fuzzy set* (Zadeh, 1965), (Zadeh, 1973), (Klir & Yuan, 1995), (Nguyen & Kreinovich, 1996), (Nguyen & Walker, 1999), (Dubois & Prade, 2000), or, to be more precise, a very specific type of a fuzzy set called a *fuzzy number*. Specifically, a fuzzy set (or, to be more precise, a fuzzy subset of the real line) is defined as a function  $\mu(x)$  from real numbers into the interval  $[0, 1]$ . A fuzzy set is called a *fuzzy number* if there exists a threshold value  $x_0$  such that the function  $\mu(x)$  is (non-strictly) increasing for  $x \leq x_0$  and (non-strictly) decreasing for  $x \geq x_0$ . Each fuzzy number can be characterized by its nested  $\alpha$ -cuts, i.e., intervals  $\mathbf{x}_\alpha = \{x \mid \mu(x) \geq \alpha\}$ .

Vice versa, if we know the nested sequence of intervals  $\mathbf{x}_\alpha$ , then we can easily find a fuzzy number  $\mu(x)$  for which these particular intervals are  $\alpha$ -cuts: indeed, for every  $x$ ,  $\mu(x)$  can be found as the largest  $\alpha$  for which  $x \in \mathbf{x}_\alpha$ .

Many interval operations have been extended to fuzzy numbers, and the resulting extended operations have been successfully applied in many practical problems, including problems related to decision making (Kacprzyk & Fedrizzi, 1990), (Slowinski, 1998) and risk analysis (Chang, 1974), including risk analysis related to nuclear energy (Uhrig & Tsoukalas, 1999); see also (Dubois et al., 1997), (Tsoukalas & Uhrig, 1997), (Nguyen & Sugeno, 1998), (Bezdek et al., 1999), (Bezdek et al., 1999a), (Zimmerman, 1999). However, in spite of the examples of successful applications of some fuzzy number techniques in risk analysis, the general techniques for such usage have not been fully developed. Development of such techniques is one of the important future directions of our research.

#### 2.4.4 What if we are not sure about the intervals: from p-values to hybrid numbers

In the previous subsection, we remarked that because we are never 100% confident in an interval, it makes sense, instead of using a single interval, to use several intervals corresponding to different confidence levels. In particular, this conclusion can be applied to intervals  $[F^-(x), F^+(x)]$  of possible values of CDF  $F(x)$ . So, instead of a single p-bound, we get several nested p-bounds corresponding to different confidence levels.

The resulting construction combines the ideas from two most frequently used techniques for describing uncertainty:

- probability – because we are combining bounds on CDFs, and

- fuzzy – because for every  $x$ , we have a nested sequence of intervals, i.e., a fuzzy number.

In view of this combination, the nested collection of p-bounds is called a *hybrid number* (Cooper et al., 1996).

From the viewpoint of risk-related applications, the situation with p-bounds is similar to the situation with fuzzy numbers: there are several successful applications, but overall, the development of the corresponding techniques is far from completion. Developing such techniques is one of the important future research directions.

## 2.5 Section's conclusions

A natural way to represent partial information about probabilities is by using a *p-bound*, i.e., a pair of CDFs  $F^-(x)$  and  $F^+(x)$  for which  $F^-(x) \leq F^+(x)$ . In particular, a real number, an interval, and a probability distribution are all particular cases of p-bounds. In this survey, we will mainly overview operations with p-bounds.

It is desirable, in the future, to extend these operations in two directions:

- to the cases when, in addition to a p-bound, we also know intervals for mean and variance; and
- to the case, when, instead of single p-bound, we have a *hybrid number*, i.e., a nested collection of p-bounds corresponding to different confidence levels.

Because fuzzy numbers are a particular case of hybrid numbers, operations with hybrid numbers will generalize not only our operations with p-bounds, but also known operations with fuzzy numbers.

### 3 Elicitation of Uncertainty in Risk Analysis: Where Do p-Bounds Come From?

#### 3.1 Where do p-bounds come from: a brief classification of sources

Where does uncertainty description come from? In the ideal case, when there is no uncertainty, we should be able to measure the exact values of all the quantities and thus, get a complete description of the system under study.

Measurements are never absolutely accurate, so, even when we are able to measure the values of the quantities of interest, the results of these measurements are not exact. Thus, *measurement results* – coming from (inevitably) uncertain measurements are the first source of our information about uncertainty.

Measurement results are the closest we can get to the ideal complete descriptions; so, it is desirable to have as many measurements as possible. However, in some real-life situations, it is very difficult (or even impossible) to directly measure the quantities of interest. For example, in geophysical applications, it is difficult to impossible to measure the characteristics of the deep geological layers. To compensate for the rarity of the corresponding measurement results, we can use the experience of human experts, e.g., geologists. Thus, *expert knowledge* is the second major source of our information about uncertainty; see, e.g., (Ayyub, 2001).

In some real-life situations, we do not have measurement results, and we do not have any specialists who possess a definite knowledge about this particular system. In some of such cases, we do not have a large amount of information about this specific system, but we know that this system belongs to a certain class, and we have a general information about systems of this class. In such situations, we can extract some information about uncertainty from this general information, i.e., from the *first principles*.

Finally, in some practical situations, we do not have any direct information about the quantity of interest  $y$ . This is a very typical situation in science and engineering: for example, we cannot directly measure a distance to a star, or the amount of oil in a given area. In this case, what we can do is find some other quantities  $x_1, \dots, x_n$  which can be directly measured (or at least estimated), and which are, in a known (or postulated) way, related to the desired quantity  $x$ . For example, to measure the distance to a nearby star, we can use the fact that the direction to the star differs while the Earth goes around the Sun. Based on the difference (“parallax”) between the corresponding angles  $x_1$  and the known radius of the Earth’s orbit  $x_3$ , we can reconstruct the desired distance by using known trigonometric formulas  $y = f(x_1, x_2, x_3)$ .

In general, if we know the relation  $y = f(x_1, \dots, x_n)$  between the quantity of interest  $y$  and the quantities  $x_1, \dots, x_n$  about which we can collect some

information, we gather the information  $X_1, \dots, X_n$  corresponding to  $x_i$ , and then use this information to get an (*indirect*) estimate of  $y$ .

Summarizing: there are four different sources of information about uncertainty: measurements, expert estimates, first principles, and indirect estimates. In this section, we will give a brief overview of these four sources of p-bounds.

## 3.2 p-bounds from measurements

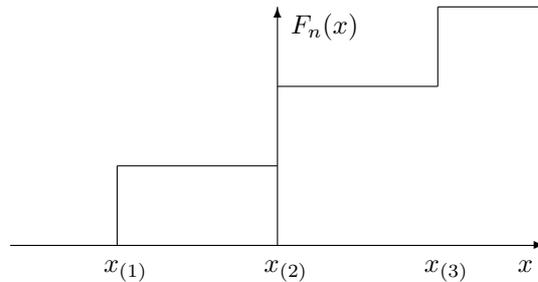
### 3.2.1 Case of exact measurements

We will start with the simplest case in which measurements are so accurate that they can be considered absolutely precise. We have  $n$  measurement results  $x_1, \dots, x_n$ , and we want to reconstruct the p-bound from these results.

After these measurements, we get  $n$  values, each of which occurred 1 out of  $n$  times. It is therefore natural to represent these measurement results by the following “empirical” distribution: we have  $x_1$  with probability  $1/n$ ,  $x_2$  with probability  $1/n$ , etc. Because we represent every probability distribution by its CDF, we should represent this distribution as a CDF  $F_n(x)$ . For that, we must first sort the values  $x_1, \dots, x_n$  in increasing order into a sequence  $x_{(1)} \leq \dots \leq x_{(n)}$ , and then, design the following CDF:

$$F_n(x) = \begin{cases} 0 & \text{if } x \leq x_{(1)}, \\ 1/n & \text{if } x_{(1)} < x \leq x_{(2)}, \\ 2/n & \text{if } x_{(2)} < x \leq x_{(3)}, \\ \dots & \\ k/n & \text{if } x_{(k)} < x \leq x_{(k+1)}, \\ 1 & \text{if } x_{(n)} < x \end{cases}$$

In particular, for  $n = 3$ , the distribution  $F_n(x)$  looks as follows:



The actual (unknown) CDF  $F(x)$  can differ from the above histogram-type distribution  $F_n(x)$ . The difference between  $F(x)$  and  $F_n(x)$  can be bounded if we use Kolmogorov-Smirnov criterion, according to which, for any given level of certainty, for each interval  $[x^-, x^+]$  (on which the variable is located), and for any number of measurements  $n$ , we can find the value  $D$  (which decreases to 0 as  $n$  increases) such that  $|F(x) - F_n(x)| \leq D$  for all  $x$ . Thus, for every

$x$ ,  $0 \leq F(x) \leq 1$  and  $F_n(x) - D \leq F(x) \leq F^+(x) + D$ . Combining these two inequalities, we conclude that

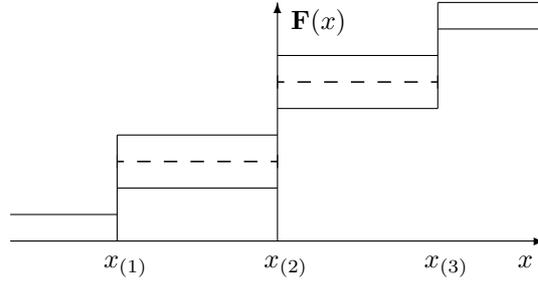
$$F(x) \in \mathbf{F}(x) = [F^-(x), F^+(x)],$$

where:

$$F^-(x) = \max(F_n(x) - D, 0); \quad F^+(x) = \min(F_n(x) + D, 1).$$

This is the desired p-bound.

An illustrative example is given by the following figure:



### 3.2.2 Case of measurements with known upper bounds on measurement errors

As we have mentioned, measurements are never absolutely accurate, there is always a measurement error, i.e., a non-zero difference between the actual and measured values. We have also mentioned that the simplest possible information about this error is the upper bound. Let us therefore consider a situation in which we have  $n$  measurement results  $x_1, \dots, x_n$ , and  $n$  upper bounds  $\Delta_i$  on the measurement errors. How to construct a p-bound in this case?

For every  $i$ , because the  $i$ -th measured value is  $x_i$ , and the upper bound of the measurement error of  $i$ -th measurement is  $\Delta_i$ , the only information that we have about the actual (unknown) value  $\tilde{x}_i$  of the  $i$ -th measured quantity is that this quantity belongs to the interval  $[x_i^-, x_i^+]$ , where  $x_i^- = x_i - \Delta_i$  and  $x_i^+ = x_i + \Delta_i$ .

If we knew the actual values  $\tilde{x}_i$ , then we could compute the histogram CDF  $F_n(x)$ , and then expand it by an appropriate value  $D$  to get the desired p-bound. In reality, we only know the intervals of possible values of each  $\tilde{x}_i$ . So, instead of finding the exact histogram, we should find, for each  $x$ , the interval of possible values of  $F_n(x)$  for different  $\tilde{x}_i \in [x_i^-, x_i^+]$ , and then increase the resulting interval by  $D$  in both up and down directions.

The value  $F_n(x)$  is the number of values  $\leq x$  divided by  $n$ . The larger the values from the intervals  $[x_i^-, x_i^+]$  we select, the smaller this probability. Thus, the value  $F_n(x)$  is the smallest if we take the values  $x_i^+$ . Similarly, the value  $F_n(x)$  is the largest if we build the histogram distribution by taking the smallest

values  $x_i^-$  from all the intervals  $[x_i^-, x_i^+]$ . As a result, we get the following algorithm for computing the corresponding p-bound:

- First, we sort the values  $x_1^+, \dots, x_n^+$  in increasing order into a sequence  $x_{(1)}^+ \leq \dots \leq x_{(n)}^+$ , and then, design the following CDF:

$$F_n^-(x) = \begin{cases} 0 & \text{if } x \leq x_{(1)}^+, \\ 1/n & \text{if } x_{(1)}^+ < x \leq x_{(2)}^+, \\ 2/n & \text{if } x_{(2)}^+ < x \leq x_{(3)}^+, \\ \dots & \\ k/n & \text{if } x_{(k)}^+ < x \leq x_{(k+1)}^+, \\ 1 & \text{if } x_{(n)}^+ < x \end{cases}$$

- Second, we sort the values  $x_1^-, \dots, x_n^-$  in increasing order into a sequence  $x_{(1)}^- \leq \dots \leq x_{(n)}^-$ , and then, design the following CDF:

$$F_n^+(x) = \begin{cases} 0 & \text{if } x \leq x_{(1)}^-, \\ 1/n & \text{if } x_{(1)}^- < x \leq x_{(2)}^-, \\ 2/n & \text{if } x_{(2)}^- < x \leq x_{(3)}^-, \\ \dots & \\ k/n & \text{if } x_{(k)}^- < x \leq x_{(k+1)}^-, \\ 1 & \text{if } x_{(n)}^- < x \end{cases}$$

- Finally, we conclude that

$$F(x) \in \mathbf{F}(x) = [F^-(x), F^+(x)],$$

where:

$$F^-(x) = \max(F_n^-(x) - D, 0); \quad F^+(x) = \min(F_n^+(x) + D, 1).$$

This is the desired p-bound.

### 3.2.3 Important comment: relationship between interval measurements and Dempster-Shafer techniques

For precise measurements, we had a direct interpretation of the histogram distribution  $F_n(x)$ : it describes the distribution in which we have  $x_1$  with probability  $1/n$ ,  $x_2$  with probability  $1/n$ , etc. The resulting p-bound is, therefore, a combination known as a *mixture*:

$$F_n(x) = \frac{1}{n} \cdot F_{x_1}(x) + \dots + \frac{1}{n} \cdot F_{x_n}(x),$$

where  $F_y(x)$  is a p-bound (actually, a CDF) that describes the exactly known real number (we have described this CDF in the previous section).

A similar interpretation makes sense for the mixture p-bound  $[F_n^-(x), F_n^+(x)]$ : with probability  $1/n$ , we have a real number from the interval  $\mathbf{x}_1 = [x_1^-, x_1^+]$ ; with probability  $1/n$ , we have a real number from the interval  $\mathbf{x}_2 = [x_2^-, x_2^+]$ ; etc. The resulting p-bound is, therefore, a combination

$$\mathbf{F}_n(x) = \frac{1}{n} \cdot \mathbf{F}_{\mathbf{x}_1}(x) + \dots + \frac{1}{n} \cdot \mathbf{F}_{\mathbf{x}_n}(x),$$

where  $\mathbf{F}_y(x) = [F_y^-, F_y^+]$  is a p-bound that describes the interval  $\mathbf{y}$  (we have described this p-bound in the previous section), and addition of two intervals is understood as a component-wise operation:

$$[F^-(x), F^+(x)] + [G^-(x), G^+(x)] \stackrel{\text{def}}{=} [F^-(x) + G^-(x), F^+(x) + G^+(x)].$$

In other words,

$$\begin{aligned} F_n^-(x) &= \frac{1}{n} \cdot F_{\mathbf{x}_1}^-(x) + \dots + \frac{1}{n} \cdot F_{\mathbf{x}_n}^-(x); \\ F_n^+(x) &= \frac{1}{n} \cdot F_{\mathbf{x}_1}^+(x) + \dots + \frac{1}{n} \cdot F_{\mathbf{x}_n}^+(x). \end{aligned}$$

The above representation, in which:

- with probability  $1/n$ , we have a real number from the interval  $\mathbf{x}_1$ ;
- with probability  $1/n$ , we have a real number from the interval  $\mathbf{x}_2$ ,
- etc.

is a particular case of so-called *Dempster-Shafer* representation, in which we have:

- $n$  sets (e.g., intervals)  $\mathbf{x}_1, \dots, \mathbf{x}_n$  (called *focal elements*),
- $n$  probabilities  $p_1, \dots, p_n$  that add up to 1 (these probabilities are called *masses*),

and we know that we can divide the general population into  $n$  groups, so that elements of  $i$ -th group all belong to the interval  $\mathbf{x}_i$  and the portion of elements which belongs to  $i$ -th group is equal to  $p_i$ .

The above interpretation corresponds to  $\mathbf{x}_i = [x_i^-, x_i^+]$  and  $p_1 = \dots = p_n = 1/n$ .

Likewise, every Dempster-Shafer “knowledge base” can be represented as a p-bound

$$\mathbf{F}_n(x) = p_1 \cdot F_{\mathbf{x}_1}(x) + \dots + p_n \cdot F_{\mathbf{x}_n}(x),$$

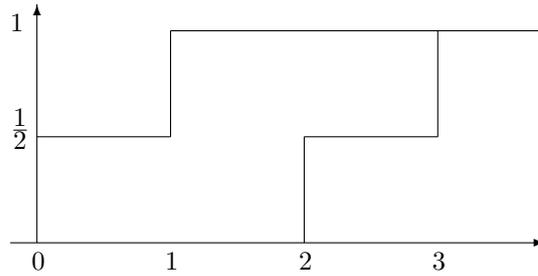
i.e., with

$$F_n^-(x) = \sum \{p_i \mid x_i^+ \leq x\}$$

and

$$F_n^+(x) = \sum \{p_i \mid x_i^- \leq x\}.$$

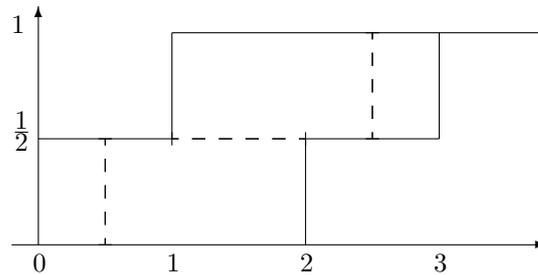
For example, for  $\mathbf{x}_1 = [0, 3]$ ,  $\mathbf{x}_2 = [1, 2]$ , and  $p_1 = p_2 = 1/2$ , we get the following p-bound:



### 3.2.4 Important comment: p-bound is not a complete description of uncertainty information, so “no aggregation” is sometimes a good policy

The above example is a good example to explain that although p-bound is a useful (and often sufficient) partial information about the distribution, it is not the complete information. Indeed, let us consider the probability that the value  $x$  is inside the interval  $[2, 3]$ .

- According to the original Dempster-Shafer representation, in which this interval is one of the focal elements with the probability  $1/2$ , this probability is at least as large as  $1/2$ .
- However, if we only use the p-bound, it is quite possible that the actual distribution includes  $0.5$  with probability  $1/2$  and  $2.5$  with probability  $0.5$ . Indeed, the corresponding CDF is within the above bounds:



In this case, the probability of  $x$  to be within the interval  $[1, 2]$  is actually equal to 0.

This example shows that when transforming the information into a p-bound and then aggregating these results into a single p-bound, we often lose some information. As a result, in situations where we do not have a lot of information, it may be beneficial, instead of (or at least in addition to) converting all the information into p-bounds and aggregating these p-bounds, to keep the original data. In short, “no aggregation” may sometimes be a good policy.

### 3.2.5 Relationship between interval measurements and Dempster-Shafer techniques (continued)

The above example also shows that although we can transform a Dempster-Shafer knowledge base into a p-bound, we cannot uniquely reconstruct the original knowledge base from the resulting p-bound.

Indeed, the exact same p-bound would have appeared if we started with a different Dempster-Shafer knowledge base, in which the probabilities (masses) are the same ( $p_1 = p_2 = 1/2$ ), but the focal elements are different:  $\mathbf{x}_1 = [0, 2]$  and  $\mathbf{x}_2 = [1, 3]$ .

### 3.2.6 Case of measurements with known probability distributions for measurement error

In addition to the upper bounds on the measurement errors, we sometimes know (or at least postulate) the probability distributions of these errors. In this case, for each measurement, we know the CDF  $F_i(x)$  corresponding to possible actual values  $\tilde{x}_i$  of the quantity measured in this particular measurement. Usually, this distribution is Gaussian, or sometimes uniform.

Because each of the measurement results occurred 1 time out of  $n$ , the CDF for describing all  $n$  measurement results can be obtained as a combination of  $n$  CDFs  $F_i(x)$  with the coefficients  $1/n$  (just like we did for real numbers and for intervals):

$$F_h(x) = \frac{1}{n} \cdot F_1(x) + \dots + \frac{1}{n} \cdot F_n(x).$$

This mixture distribution  $F_h(x)$  describes the results of the measurement. The actual CDF  $F(x)$  may be different from  $F_h(x)$ . To get the bounds for the actual distribution, we add and subtract the parameter  $D$  coming from the Kolmogorov-Smirnov criterion. As a result, we conclude that

$$F(x) \in \mathbf{F}(x) = [F^-(x), F^+(x)],$$

where:

$$F^-(x) = \max(F_h(x) - D, 0); \quad F^+(x) = \min(F_h(x) + D, 1).$$

This is the desired p-bound.

In some cases, we get only a partial information about the probability distribution of measurement error. In our approach, this partial information corresponding to  $i$ -th measurement is described by a p-bound  $\mathbf{F}_i(x) = [F_i^-(x), F_i^+(x)]$ . As a result of combining these p-bounds, we first get a mixture p-bound:

$$\mathbf{F}_h(x) = [F_h^-(x), F_h^+(x)] = \frac{1}{n} \cdot \mathbf{F}_1(x) + \dots + \frac{1}{n} \cdot \mathbf{F}_n(x),$$

i.e.,

$$\begin{aligned} F_h^-(x) &= \frac{1}{n} \cdot F_1^-(x) + \dots + \frac{1}{n} \cdot F_n^-(x); \\ F_h^+(x) &= \frac{1}{n} \cdot F_1^+(x) + \dots + \frac{1}{n} \cdot F_n^+(x); \end{aligned}$$

and then use this mixture p-bound to get a guaranteed enclosure for the actual (unknown) CDF  $F(x)$ :

$$F(x) \in \mathbf{F}(x) = [F^-(x), F^+(x)],$$

where:

$$F^-(x) = \max(F_h^-(x) - D, 0); \quad F^+(x) = \min(F_h^+(x) + D, 1).$$

### 3.3 p-bounds from expert estimates

#### 3.3.1 General overview

Information coming from measurements can be supplemented by expert knowledge. We have argued, in Section 2, that the most appropriate formalism for describing the expert information is p-bounds, i.e., estimates for the values of the CDF  $F(x)$ . However, the fact that p-bounds are a good representation for our problems does not mean that the expert knowledge is actually presented in this form. Our goal is to convert whatever the expert says into a p-bound. Depending on how difficult this translation is, we will consider three types of expert information:

First, it is possible that an expert's information is expressed directly in our desired form, as estimates for the values  $F(x)$ . This information is already in the form a p-bound, so no further transformation is needed.

Second, it is also possible that the expert provides us with numerical estimates for some other characteristics of the probability distribution, e.g., shape of the distribution, its moments (first and second), percentiles, mode (for a unimodal distribution), bounds on density, information about symmetry, etc. There exist algorithms (Ferson et al., 2001), (Ramas, 2001) that translate this information into p-bounds. (We have already mentioned some such algorithms

when we talked, in Section 2, about the necessity to describe moments in terms of p-bounds.)

From the viewpoint of translation into p-bounds, the most difficult third type of expert knowledge is when an expert does not provide us with any numerical information, but instead, describes his opinion by using words from natural language. For example, an expert can say that the value  $x$  is “small”, or “around 0.5”. This clearly is an additional information, but how can we describe it in terms of p-bounds?

### 3.3.2 Case of natural-language estimates

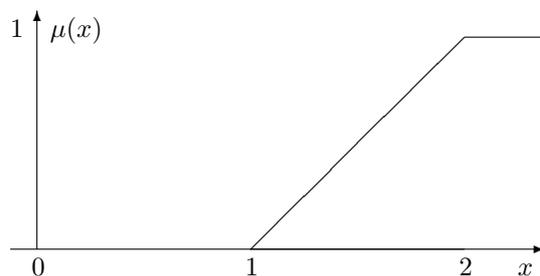
The problem with words from natural language is that they are usually vague. Let us take the word “small” as an example. When the value of, say, concentration, is really small, everyone would 100% agree that this value is small indeed. When the value is really large, everyone would agree that this value is not small. For intermediate values, however, we typically have some disagreement.

The need to translate expert knowledge from natural language to a computer-understandable language of real numbers was recognized as early as the 1960s, when the designs for the first expert systems were begun. A special formalism called *fuzzy logic* was created to help us capture the meaning of words for manipulation by software. In this formalism, to represent a meaning of a word like “small”, we assign, to every possible value  $x$ , a degree  $\mu_{\text{small}}(x)$  to which  $x$  is small. This degree is also called a *membership value* or a *subjective probability*. The dependence of this degree on  $x$  is called a *membership function*, or a *fuzzy set*.

Where do the values  $\mu(x)$  come from? There are several dozen different techniques for eliciting these values; see, e.g., (Klir & Yuan, 1995), (Nguyen & Walker, 1999). Sometimes, the experts can present these real numbers directly. If they cannot, then for every  $x$ , we can poll several ( $N$ ) experts on whether they believe that this particular values  $x$  is, say, small, and if  $M$  out of  $N$  experts answer “yes”, we take  $\mu(x) = M/N$ . This amounts to a social specification of a membership function. What is a natural way to translate these membership values into p-bounds?

We will answer this question on the example of membership functions of three most frequent types; for a general background on probabilistic interpretations of fuzzy, see also (Kreinovich, 1997), (Walley & de Cooman, 2001). The first type is a function which describes words like “large”, for which  $\mu(x)$  is increasing from 0 at  $x = 0$  to 1 for  $x \rightarrow \infty$ . Let us give a simple example of such function:

$$\mu_{\text{large}}(x) = \begin{cases} 0 & \text{if } x \leq 1, \\ x - 1 & \text{if } 1 \leq x \leq 2, \\ 1 & \text{if } x > 2 \end{cases}$$



Suppose that the expert tells us that the actual value of some quantity  $X$  is large. What does it say about the possible values of the probability  $F(x)$  (that  $X \leq x$ ) for different  $x$ ?

Let us start with a value  $x \leq 1$ . For this value,  $\mu_{\text{large}}(x) = 0$ . This means that the values below  $x$  cannot be large, so it is reasonable to take  $F(x) = 0$ .

Let us now take a value  $x \geq 2$ . For this value,  $\mu_{\text{large}}(x) = 1$ , which means that the value  $x$  is definitely large. Based on the expert opinion, we only know that the actual value  $X$  is large. It may be below  $x$  with probability 1; in this case  $F(x) = 1$ . It may be above  $X$  with probability 1; in this case,  $F(x) = 0$ . So, here, the corresponding value of the p-bound (i.e., the interval of possible values of  $F(x)$ ) is  $\mathbf{F}(x) = [0, 1]$ .

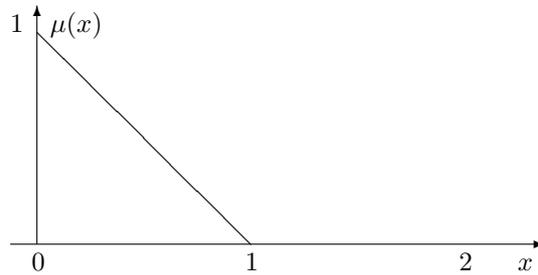
What if  $x$  is in between 1 and 2, e.g.,  $x = 1.6$ ? In this case, the probability  $\mu(x)$  that  $x$  is large is equal to 0.6. Because the function  $\mu(x)$  is increasing, the probability  $\mu(X)$  that  $X$  is large even smaller for  $X < x$ . Thus, out of all large values, values  $\leq 0.6$  should have a frequency  $\leq 0.6$ . So, because we know that actual value  $X$  is large, we conclude that the probability  $F(x)$  cannot exceed 0.6.

In general, the value  $F(x)$  cannot exceed the probability  $\mu(x)$ , i.e.,  $\mu(x)$  serves as the upper part  $F^+(x)$  of the p-bound. The lower part  $F^-(x)$  should be 0, because we may have  $X$  so large than it is much larger than 2.

Combining these three cases, we conclude that for increasing membership functions  $\mu(x)$  like “large”, a natural translation of the membership function is a p-bound  $[0, \mu(x)]$ .

The second type of membership functions that we will consider is a function which describes words like “small”, for which  $\mu(x)$  is decreasing from 1 at  $x = 0$  to 0 for  $x \rightarrow \infty$ . Let us give a simple example of such function:

$$\mu_{\text{small}}(x) = \begin{cases} 1 - x & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise} \end{cases}$$

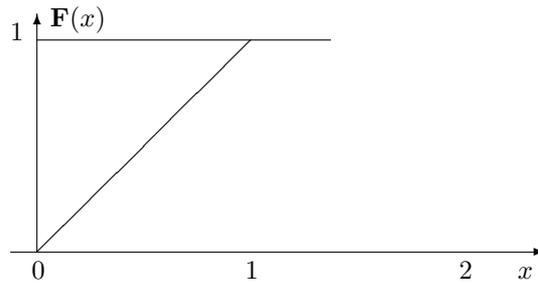


Suppose that the expert tells us that the actual value of some quantity  $X$  is small. What does it say about the possible values of the probability  $F(x)$  (that  $X \leq x$ ) for different  $x$ ?

Let us start with a value  $x \geq 1$ . For this value,  $\mu_{\text{small}}(x) = 0$ , which means that the value  $x$  is definitely not small. Based on the expert opinion, we only know that the actual value  $X$  is small. All values  $X$  which can be small (i.e., for which  $\mu(X) > 0$ ) are below 1, so they are all below  $x$ . Thus, all values of  $X$  are below  $x$  with probability 1, and  $F(x) = 1$ .

What if  $x$  is in between 0 and 1, e.g.,  $x = 0.2$ ? In this case, the probability  $\mu(x)$  that  $x$  is small is equal to 0.8. Hence, the probability that any larger value  $X > x$  is “small” also does not exceed 0.8. This means that if  $F(x)$  is smaller than  $1 - 0.8 = 0.2$  (e.g., equal to 0.1) then there will be more than  $> 0.8$  of values which are  $\geq x$ , and thus, some values  $X > x$  cannot be reasonably called small, in contradiction to the expert’s opinion. So, if the actual value  $X$  is small, the probability  $F(x)$  cannot exceed 0.2. In general, the value  $F(x)$  cannot be smaller than  $1 - \mu(x)$ , i.e.,  $1 - \mu(x)$  serves as the upper part  $F^-(x)$  of the p-bound. The upper part  $F^+(x)$  should be 1, because we may have  $X = 0$  with probability 1.

Combining these two cases, we conclude that for increasing membership functions  $\mu(x)$  like “small”, a natural translation of the membership function is a p-bound  $[1 - \mu(x), 1]$ :

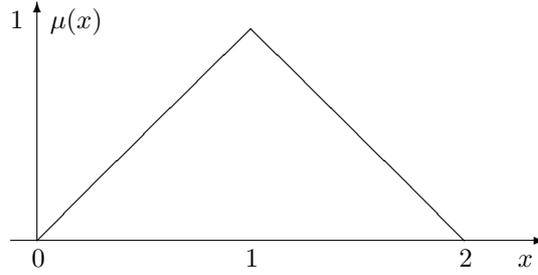


Finally, we can consider membership functions describing terms like “around  $x_0$ ”, which increase from 0 to 1 until they reach a certain value  $x_0$ , and then decrease from 1 to 0. For such membership functions, possible values (i.e.,

values for which the degree  $\mu(x)$  is large enough) are concentrated around the real number  $x_0$ , that is why such membership functions are called *fuzzy numbers*.

As an example, we will consider the following function corresponding to “around 1”:

$$\mu_{\approx 1}(x) = \begin{cases} x & \text{if } 0 \leq x \leq 1, \\ 2 - x & \text{if } 1 \leq x \leq 2, \\ 0 & \text{otherwise} \end{cases}$$



For a membership function of this type, with a maximum at some value  $x_0$ , similar arguments lead to the following p-bound  $[F^-(x), F^+(x)]$ :

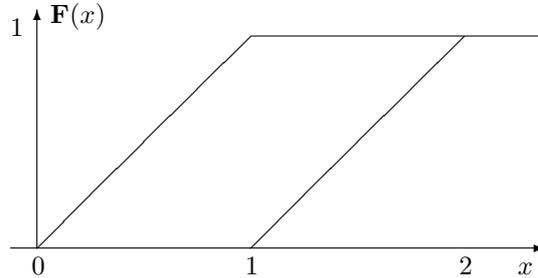
$$F^-(x) = \begin{cases} 0 & \text{if } x \leq x_0, \\ 1 - \mu(x) & \text{if } x > x_0 \end{cases}$$

$$F^+(x) = \begin{cases} \mu(x) & \text{if } x \leq x_0, \\ 1 & \text{if } x > x_0 \end{cases}$$

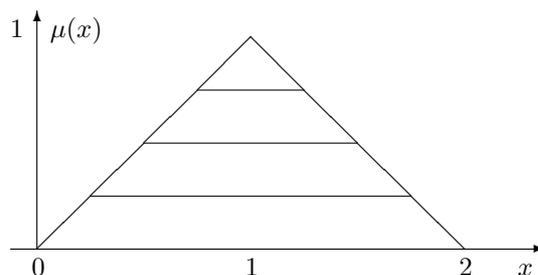
In particular, for the above membership function “around 1”, the corresponding p-bound has the following form:

$$F^-(x) = \begin{cases} 0 & \text{if } x \leq 1, \\ x - 1 & \text{if } 1 \leq x \leq 2, \\ 1 & \text{if } x > 2 \end{cases}$$

$$F^+(x) = \begin{cases} x & \text{if } x \leq 1, \\ 1 & \text{if } x > 1 \end{cases}$$



These three cases can be described in a way which is similar to our transformation of measurements into p-bounds. Indeed, how can we describe a fuzzy set that corresponds to a certain property like “around 1”? A natural way to characterize a fuzzy set is to describe, for every level  $\alpha$ , the set  $X_\alpha = \{x \mid \mu(x) \geq \alpha\}$  of all the values which have this property with degree at least  $\alpha$ . Such sets are called  $\alpha$ -cuts, because on the graph, they really correspond to horizontal cuts. For example, for the above membership function “around 1”, the  $\alpha$ -cuts are  $X_\alpha = [\alpha, 2 - \alpha]$ :



If we, e.g., have  $\alpha$ -cuts  $X_{0.1}$ ,  $X_{0.2}$ , etc., corresponding to  $\alpha = 0.1$ ,  $\alpha = 0.2$ , etc., this means, crudely speaking, that all experts agree that  $x \in X_0$ , that 90% of them agree that  $x \in X_{0.1}$ , that 80% of experts agree that  $x \in X_{0.2}$ , etc., until we reach we level  $X_{0.9}$  in which only 10% of the experts agree; see, e.g., (Nguyen & Kreinovich, 1996). So, we have a natural subdivision of experts into 10 groups: 10% believe that  $x$  is somewhere on the interval  $X_{0.1}$  and no narrower bounds are possible; 10% believe that  $x$  is somewhere on the interval  $X_{0.2}$  and no narrower bounds are possible, etc. We thus have a typical Dempster-Shafer knowledge base. One can easily see that if we use the above algorithm to transform this knowledge base into a p-bound, we get exactly the p-bound that we came up with.

Our description of transforming expert estimates into p-bounds may look like a success story, but the reader should be aware of the limitations and pitfalls associated with elicitation: the results may differ drastically if we select different set of experts, and even when two groups of experts agree, they still may be wrong, and the resulting p-bound may not contain the actual probability distribution.

### 3.4 p-bounds from first principles

#### 3.4.1 Normal distribution

The most well-known case of p-bounds coming from the first principles is the case of *normal (Gaussian) distribution*. Indeed, according to the central limit theorem, under some reasonable conditions, the sum of many small independent random variables tends to a normal distribution.

Thus, when the error in the desired quantity  $x$  can be represented as the sum  $x = x_1 + \dots + x_n$  of a large number of small independent error components  $x_i$ , the probability distribution for  $x$  is close to Gaussian. This is the reason why the Gaussian distribution is so frequently observed (and used) in practice; see, e.g., (Wadsworth, 1990). The normal distribution, like all other probability distributions, is a degenerate (“precise”) p-bound. Alternatively, we may only have interval of possible values for the parameters of a normal distribution. In this case, the corresponding CDFs form a non-degenerate p-bound.

### 3.4.2 Cauchy distribution

There are many similar situations which lead to different classes of distributions. For example, in a setting similar to the central limit theorem, if we keep the variables  $x_i$  small in some reasonable sense but allow them to have large (even infinite) standard deviations by allowing thick “tails”, we get distributions from the class of *infinitely divisible* distributions, the class that includes not only normal distribution, but also *Cauchy distribution*, with probability density function

$$\rho(x) = \frac{\Delta}{\pi \cdot (\Delta^2 + (x - a)^2)}.$$

### 3.4.3 Lognormal distribution

Other examples of distributions that can be derived from first principles come from the situations in which the error is caused by many small components  $x_i$ , but these components do not add up, they rather are combined in a more complex manner.

Adding up corresponds to the case of *additive* noise  $n$ , when each error is simply added to the actual value  $s$ , turning it into the sum  $s+n$ . In many real-life situations, we have *multiplicative* noise, in which  $s$  is multiplied by some value so that  $s$  becomes  $s \rightarrow s \cdot k$ . The difference  $n = k - 1$  between the value  $k$  and 1 constitutes the noise. For example, when a communication signal passes through the atmosphere, its amplitude changes depending on the specific properties of the medium. Suppose that we have several layers with independent noise values  $n_i$  and, correspondingly, independent multiplicative coefficients  $k_i = 1 + n_i$ . When a signal passes through each layer, it is multiplied by  $1 + n_i$ . By the time the signal passes through all the layers, it is multiplied by the product of many independent coefficients  $k_i = 1 + n_i$ .

This example shows that it is useful to analyze the probability distribution of the product of independent random variables whose values are close to 1. When we apply logarithms, the product turns into the sum, so we get a normal distribution. Thus, in general, the distribution for the product is *lognormal*.

### 3.4.4 Weibull distributions

This example can be extended to a more general case, when the desired quantity  $x$  describes the effect of a certain quantity  $y$ . For example, in risk analysis, the desired quantity  $x$  may represent the effect of a certain chemical. This effect is related to the concentration  $y$  of this chemical by some non-linear dependence  $x = f(y)$ . Addition does not make much sense for effects (this is why the distribution of  $x$  is not Gaussian), but it often makes perfect sense for concentrations, so it is reasonable to assume that the distribution for  $y$  is Gaussian. Then, the distribution of the desired quantity  $x$  can be described as the result of applying a non-linear transformation  $f(y)$  to a normally distributed variable  $y$ . In physiology, two types of non-linear transformations are most frequently used to describe the effect of a certain physical quantity; see, e.g., (Milner, 1970):

- Fechner scale, in which the perception corresponds to  $x = \ln(y)$ ; and
- Stevens' scale, in which the perception is best described by the value  $x = y^\alpha$  for an appropriate parameter  $\alpha$ .

There is also a general symmetry-based explanation of these scales, see, e.g., (Nguyen & Kreinovich, 1997). Fechner's law is a limit case of Stevens' law when  $\alpha \rightarrow 0$ . Thus, it is sufficient to only consider Stevens' law.

For Stevens' scale, the desired variable  $x$  is a power of a normally distributed variable. The corresponding distribution is called *Weibull* distribution, and it is indeed often a good fit for risk-related data. This justifies the use of Weibull distributions as degenerate p-bounds. Alternatively, we may only have interval of possible values for the parameters of a Weibull distribution. In this case, the corresponding CDFs form a non-degenerate p-bound.

### 3.4.5 Extreme statistics

Another situation which is very useful in risk analysis applications is the situation of the "weakest link", when a certain event happens if at least one of the numerous quantities  $x_1, \dots, x_n$  exceeds a certain threshold  $x_0$ . Thus, the event occurs if the largest  $x = \max(x_1, \dots, x_n)$  of these quantities exceeds  $x_0$ . To analyze such events, we therefore need to analyze the distribution of such maxima. It is known that under reasonable conditions, when  $n \rightarrow \infty$ , the distribution of the maximum tends to one of the standard distributions called *extreme statistics*, or *Gumbel-type distributions*. Thus, for large  $n$ , we can safely assume that the distribution of  $x = \max(x_1, \dots, x_n)$  can be described by one of these extreme statistics.

There are three types of extreme statistics (Galambos, 1978), (Wadsworth, 1990):

$$F_I(x) = 1 - \exp(-\exp(a \cdot x + b)),$$

$$F_{\text{I}}(x) = \begin{cases} 1 - \exp(1 - (-x)^{-\alpha}) & \text{if } x < 0, \\ 1 & \text{otherwise} \end{cases}$$

$$F_{\text{II}}(x) = \begin{cases} 1 - \exp(1 - x^\alpha) & \text{if } x \geq 0, \\ 0 & \text{otherwise} \end{cases}$$

### 3.4.6 Use of symmetry

Another use of “first principles” to derive p-bounds includes the use of symmetry and invariance. Let us give two examples:

- Often, we have no reason to assume that some values are more probable than others. In such situations, it is natural to select a distribution for which the value of the probability density  $\rho(x)$  is the same for all the points  $x$ ; thus, we get a *uniform distribution* on an interval  $[a, b]$ , for which  $\rho(x) = 1/(b - a)$  for all  $x \in [a, b]$ .
- Similarly, if there is no reason to believe that the probability of a correctly functioning system to become faulty between times  $t$  and  $t + \Delta$  depend on  $t$ , it is natural to assume that this probability is indeed constant. This assumption leads to the *exponential distribution*, in which  $F(t) = 1 - \exp(-\lambda \cdot t)$  for some constant  $\lambda > 0$ .

### 3.4.7 Relation with MaxEnt

It is important to make a comment here. Some of these “first principles” distributions are often justified by using the *Maximum Entropy principle* (MaxEnt). The entropy

$$S = - \int \rho(x) \cdot \ln(\rho(x)) \, dx$$

of a distribution with density  $\rho(x)$  describes its uncertainty, measured, e.g., by the average number of binary (“yes”-“no”) questions that one needs to ask to estimate the value of a randomly distributed variable with a given accuracy  $\varepsilon > 0$ .

When we only have a partial information about the distribution, then there are several distributions which are consistent with this information. Some of these distributions have a larger entropy, some have a smaller one.

As an example, let us consider a situation in which all we know about a probability distribution is that it is concentrated on the interval  $[0, 1]$ . In this case, we may have several possible distributions. We can have a uniform distribution, for which the entropy is high. We can also have a degenerate distribution in which the value of the quantity is equal to, say, 0.5 with probability 1. For the second distribution, we already know the value of  $x$  exactly, so no questions need to be asked, and  $S = 0$ .

When we want to select a single distribution representing this class of possible distributions with respect to uncertainty, then selecting a distribution in

which  $x = 0.5$  with probability 1 will lead to a wrong impression that there is no uncertainty at all. Every time we select, as a representative distribution, a distribution for which the entropy is smaller than for some other distributions for this class, we similarly create a misrepresentation of the level of uncertainty. From this viewpoint, it is better to select a representative distribution for which the corresponding uncertainty is as large as possible, i.e., for which  $S(\rho) \rightarrow \max$ . MaxEnt indeed leads to many of these distributions:

- If we consider all the distributions located on a given interval, then MaxEnt leads to a uniform distribution.
- If we consider all the distributions located on values  $x \geq 0$  with a given average  $E[x]$ , then MaxEnt selects the exponential distribution.
- If we consider all the distributions with a given average  $E[x]$  and standard deviation  $\sigma[x]$ , then MaxEnt selects the Gaussian (normal) distribution.

### 3.4.8 p-bounds are needed in some real-life risk applications: an example showing that MaxEnt-based techniques are not always adequate

We have seen that the Maximum Entropy principle is a reasonable way of selecting a representative probability distribution. However, we should caution readers that some analysts do not simply use this principle to select a *representative* of the class of distributions, they use it to *replace* the class of distributions with a single distribution. In some practical problems, this replacement is a reasonable idea, and MaxEnt has many successful practical applications. However, we would like to show that for risk problems, this idea can lead to mistaken conclusions.

Let us consider a simple example in which we are analyzing the value of the quantity  $x$ , e.g., a noise in some important transmission line, and we want to make sure that this value does not exceed the transmitted signal  $s$ , because if it does, the line becomes useless.

Let us assume that this noise  $x$  comes from a large number of possible sources, i.e.,  $x = x_1 + \dots + x_n$ , where  $x_i$  is the amount of noise which comes from the  $i$ th source. For each individual source of noise, we have succeeded in decreasing it to such an extent that it no longer exceeds the detection level  $d$ . In other words, the only information we have about each value  $x_i$  is that  $x_i$  belongs to the interval  $[-d, d]$ .

Because each of  $n$  components which form the noise does not exceed  $d$ , the total noise cannot exceed  $n \cdot d$ . In principle, it is quite possible that the noise generated by each source is actually deterministic and equal to exactly  $d$ . In this case, the total noise is actually equal to  $n \cdot d$ . So, if we want to guarantee that the noise stays below the signal level, we should require that  $n \cdot d < s$ .

What if we use a MaxEnt approach for this problem? For the variables  $x_1, \dots, x_n$ , the only information that we have is that the corresponding vector  $\vec{x} = (x_1, \dots, x_n)$  is located within a cube  $[-d, d]^n$ . Thus, according to MaxEnt, among all the probability distributions located on this cube, we should select the one for which the entropy is the largest. Similar to the 1-D case, this MaxEnt distribution is a uniform distribution on this cube. Therefore, if we replace the original class of possible distributions by a single MaxEnt distribution, we thus make an assumption that each of the variables  $x_i$  is uniformly distributed on the interval  $[-d, d]$ , and that these variables are independent.

For each uniform distribution of  $x_i$ , the expected value is  $E_i = 0$ , and the variance  $V_i = \sigma^2$  is equal to  $d^2/6$ . Because all the variables  $x_i$  are independent, the expectation  $E$  of the sum  $x = x_1 + \dots + x_n$  is equal to the sum of the expectations, i.e., to 0, and the variance  $V[x]$  is equal to the sum of  $n$  variances, i.e., to  $n \cdot d^2/6$ .

Due to the central limit theorem, for large  $n$ , the distribution of the sum  $x_1 + \dots + x_n$  is close to the normal distribution with the mean 0 and variance  $V$ . Therefore, with probability  $\geq 1 - 10^{-8}$ , the actual value of the quantity  $x$  lies within  $6\sigma$  from the mean. In other words, with a probability  $1 - 10^{-8}$ , the value of  $x$  cannot exceed  $6 \cdot \sqrt{n \cdot d^2/6} = \sqrt{6n} \cdot d$ . So, if we replace the original class with its MaxEnt representation, we conclude that  $|x| \leq \sqrt{6n} \cdot d$ . Hence, we would conclude that as long as  $\sqrt{6n} \cdot d < s$ , transmission is safe (with probability of its being unsafe  $\leq 10^{-8}$ ).

For large  $n$ , however,  $\sqrt{6n} \ll n$ . Thus, when we select a signal  $s$  which is slightly larger than  $\sqrt{6n} \cdot d$ , this signal will be much smaller than the signal level  $n \cdot d$  which is really needed for the safe transmission. So, the use of MaxEnt distribution instead of the whole distribution class will lead us to an erroneous unsafe choice.

There are many examples like this. What these example show is that in risk analysis, we cannot replace the class of distribution (as represented by a p-bound) by a single distribution, even by the most adequate one. In short, what this example shows is that in many risk analysis problems, traditional probabilistic techniques are not sufficient, we need more general, more realistic approaches.

## 3.5 p-bounds from indirect estimates (measurements)

### 3.5.1 General overview

Finally, let us consider indirect estimation (measurement), in which we know the relation  $y = f(x_1, \dots, x_n)$  between the quantity of interest  $y$  and the quantities  $x_1, \dots, x_n$  about which we can collect empirical information, we gather the information  $X_1, \dots, X_n$  corresponding to  $x_i$ , and then use this information to get an (*indirect*) estimate of  $y$ .

A natural way to represent uncertainty about  $X_i$  is by using p-bounds, so

we have  $n$  p-bounds  $X_1, \dots, X_n$ , and we must describe the p-bound  $Y$  corresponding to  $y = f(x_1, \dots, x_n)$ .

There are several methods for solving this problem. For example, when the p-bounds  $X_1, \dots, X_n$  are actually probability distributions, then we can use a Monte Carlo method to find  $Y$ , i.e., we can:

- simulate the random variables  $x_1, \dots, x_n$  distributed according to the corresponding probability laws,
- substitute the results  $x_1^{(k)}, \dots, x_n^{(k)}$  of the simulation into the function  $f$ , and then
- extract the actual distribution for  $y$  from the results  $y^{(k)} = f(x_1^{(k)}, \dots, x_n^{(k)})$  ( $1 \leq k \leq N$ ) of this substitution.

However, when we have only partial information about the probability distribution, the problem becomes much more complex.

In general, the problem of computing the best bounds for  $Y$  is provably very complex. Even in the case when all the inputs  $X_i$  are intervals, this problem is known to be NP-hard; see, e.g., (Kreinovich et al., 1997). This means, crudely speaking, that the effort needed to solve the problem increases as a function of the number of problem elements by a function that is bounded by no polynomial function (it grows exponentially). Consequently, although toy problems with very few elements may be solvable in practice, but large or even moderate-sized problems are effectively impossible to solve.

For those who are not familiar with this term, NP-hard is a computer science precise term formalizing the intuitive notion of “very difficult (practically impossible) to solve the corresponding general problem”.

Because we cannot have an algorithm that would generate the *exact* p-bound for  $Y$  in *all* possible situations, we have to do two things:

- First, because there is no “universal” algorithm that would find an exact p-bound for an *arbitrary* function  $f$ , we can try to find a *class* of functions for which the corresponding computation is possible.
- Second, because there is no algorithm that would always find an *exact* upper bound, we should develop algorithms which find an approximate (*enclosing*) p-bound. This enclosing p-bound may not be always exact, but we want to make it as close to the actual p-bound as possible.

Let us briefly overview the results of both efforts.

### 3.5.2 Case of simple functions $f(x_1, \dots, x_n)$

In risk analysis, often, the formulas are pretty simple. For example, the environmental effect  $y$  of a certain chemical can be simply proportional to the

product of the concentration  $x_1$  and the exposure time  $x_2$ . If we have p-bounds for the concentration  $x_1$ , for the exposure time  $x_2$ , and for the coefficient of proportionality  $x_3$ , then the problem becomes simply: given p-bounds  $X_1$ ,  $X_2$ , and  $X_3$  for  $x_1$ ,  $x_2$ , and  $x_3$ , find the p-bound  $Y$  for the product  $y = x_1 \cdot x_2 \cdot x_3$ .

For such simple functions, there are exact formulas describing the corresponding p-bounds; see, e.g., (Ferson et al., 2001), (Ramas, 2001). The resulting combination depends on whether the variables  $x_i$  are independent or correlated. How can we describe the corresponding degree of dependence?

We are describing uncertainty in terms of bounds on CDFs. So, if we have two variables  $x_1$  and  $x_2$ , then their uncertainty is described by bounds on the values  $F_1(x_1)$  and  $F_2(x_2)$ . Dependence or correlation relate to describing a joint distribution for  $x_1$  and  $x_2$  based on the marginal distributions  $F_1(x_1)$  and  $F_2(x_2)$ . For the 2-D joint distributions, the arguments similar to the ones that justified the selection of CDFs leads us to considering 2-D CDFs  $F(x_1, x_2) = P(X_1 \leq x_1 \& X_2 \leq x_2)$ .

In these terms, independence means that  $F(x_1, x_2) = F_1(x_1) \cdot F_2(x_2)$ . In other words, if we know the two probabilities  $p_1 = F_1(x_1)$  and  $p_2 = F_2(x_2)$ , then the joint probability  $F(x_1, x_2)$  is equal to  $p_1 \cdot p_2$ . It is natural to try to describe possible dependencies in terms of similar probability-combination functions. Specifically, for each values  $p_1 \in [0, 1]$  and  $p_2 \in [0, 1]$ , we describe  $p_1 * p_2$  as the the joint probability  $F(x_1, x_2)$  corresponding to the values  $x_1$  and  $x_2$  for which  $F_1(x_1) = p_1$  and  $F_2(x_2) = p_2$ . The resulting combination operation  $* : [0, 1] \times [0, 1] \rightarrow [0, 1]$  is called a *copula*; see, e.g., (Nelsen, 1999).

So, it is natural to use copulas to describe dependence. There are numerous reasonable reasonable copula operations. The most frequently used copulas are the product  $p_1 \cdot p_2$ , which corresponds to independence, and  $\min(p_1, p_2)$  and  $\max(p_1 + p_2 - 1, 0)$ , which describe the possible bounds for  $F(x_1, x_2)$  when we have no information about the possible dependence between  $x_1$  and  $x_2$ .

The existing algorithms (Ferson et al., 2001), (Ramas, 2001) describe the results of applying standard arithmetic operations ( $+$ ,  $-$ ,  $\cdot$ ,  $/$ ) and elementary functions under these and several other reasonable copulas.

### 3.5.3 Computing enclosures

Even for simple arithmetic operations, the exact formulas for p-bounds were only obtained in the 1980s. Not surprisingly, for many even slightly more complex functions  $f(x_1, \dots, x_n)$  – not to say anything about really complex ones – we do not know how to compute the exact p-bounds for  $Y$ .

The fact that we know the p-bounds for the results of arithmetic operations enables us to compute the *enclosure* for the resulting p-bound as follows:

- first, we *parse* the expression  $f(x_1, \dots, x_n)$ , i.e., represent computing  $f$  as a sequence of basic arithmetic operations;

- then, we replace each operation with the corresponding operation with p-bounds, and perform these operations in the original order.

For example, if  $f(x) = x \cdot (1 - x)$ , we represent  $f$  as a sequence of two elementary operations:

- $r := 1 - x$  ( $r$  denotes the 1st intermediate result);
- $y := x \cdot r$ .

If we know the interval  $\mathbf{x} = [0, 1]$  of possible values of  $x$ , then we perform the following computations:

- $\mathbf{r} := 1 - \mathbf{x}$ ;
- $\mathbf{y} := \mathbf{x} \cdot \mathbf{r}$ .

For intervals,

$$\mathbf{x}_1 - \mathbf{x}_2 = [x_1^- - x_2^+, x_1^+ - x_2^-],$$

$$\mathbf{x}_1 \cdot \mathbf{x}_2 =$$

$$[\min(x_1^- \cdot x_2^-, x_1^- \cdot x_2^+, x_1^+ \cdot x_2^-, x_1^+ \cdot x_2^+), \max(x_1^- \cdot x_2^-, x_1^- \cdot x_2^+, x_1^+ \cdot x_2^-, x_1^+ \cdot x_2^+)].$$

In particular, when  $\mathbf{x} = [0, 1]$ , we compute the intervals  $\mathbf{r} := [1, 1] - [0, 1] = [0, 1]$ , and

$$\mathbf{y} := [0, 1] \cdot [0, 1] = [\min(0 \cdot 0, 0 \cdot 1, 1 \cdot 0, 1 \cdot 1), \max(0 \cdot 0, 0 \cdot 1, 1 \cdot 0, 1 \cdot 1)] = [0, 1].$$

The interval  $[0, 1]$  is indeed an enclosure for the actual p-bound  $[0, 0.25]$ .

Sometimes it is possible by rearranging expressions to obtain narrower p-bounds. For example, in the above example, if the square of the quadratic equation is completed to yield:

$$f(x) = \frac{1}{4} - \left(x - \frac{1}{2}\right)^2,$$

then the exact range is returned if this expression is computed using interval arithmetic. There exist many methods for performing such computations, and several packages are available, the most actively used is GlobSol; these methods and packages (including GlobSol) are presented at the interval computations website (Interval, 2001).

These methods and packages are currently oriented towards interval computations, mainly because p-bounds were mostly used for simple operations. Because a p-bound is, in effect, a collection of intervals, we expect these methods to be naturally generalizable to arbitrary p-bounds.

### 3.5.4 An important case of p-bounds from indirect estimates (measurements): algorithms for “black-box” programs

In many real-life applications, an algorithm for computing  $f(x_1, \dots, x_n)$  may be written in a language for which a parser is not available, or a program is only available as an executable file, with no source code at hand. In such situations, when we have no easy way to analyze the code, the only thing we can do is to take this program as a *black box*: i.e., to apply it to different inputs and use the results of this application to compute the desired p-bound.

Such black-box algorithms are summarized in the survey (Trejo & Kreinovich, 2001). For example, a reasonable Monte Carlo type algorithm exists for computing the interval bound for the case when the inputs are intervals. Specifically, we know that  $x_i \in [\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]$ , and we want to compute the upper bound  $\Delta$  on the error  $\tilde{y} - y$ , where  $\tilde{y} = f(\tilde{x}_1, \dots, \tilde{x}_n)$  and  $y = f(x_1, \dots, x_n)$ . If we get this upper bound, we will then compute the interval  $Y$  for  $y$  as  $[\tilde{y} - \Delta, \tilde{y} + \Delta]$ .

This algorithm uses the Cauchy distribution. The algorithm: for  $k = 1, 2, \dots, N$ , repeat the following:

- use a standard random number generator to compute  $n$  real numbers  $r_i^{(k)}$ ,  $i = 1, 2, \dots, n$ , that are uniformly distributed on the interval  $[0, 1]$ ;
- compute  $\delta_i^{(k)} = \Delta_i \cdot \tan(\pi \cdot (r_i^{(k)} - 0.5))$ ;
- compute the (Euclidean) length  $\delta^{(k)} = \left\| \vec{\delta}^{(k)} \right\|$  of the vector  $\vec{\delta}^{(k)} = (\delta_1^{(k)}, \dots, \delta_n^{(k)})$ ;
- compute the normalized coefficient  $K_{\text{norm}}^{(k)} = \delta^{(k)} / \delta_0$  (for an appropriate small constant  $\delta_0$ );
- compute the auxiliary vector  $\vec{\beta}^{(k)} = \vec{\delta}^{(k)} / K_{\text{norm}}^{(k)}$  with components  $\beta_i^{(k)} = \delta_i^{(k)} / K_{\text{norm}}^{(k)}$ ;
- substitute  $\tilde{x}_i + \beta_i^{(k)}$  into the program  $f$  and compute

$$c^{(k)} = \frac{\delta^{(k)}}{\delta_0} \cdot \left( f \left( \tilde{x}_1 + \beta_1^{(k)}, \dots, \tilde{x}_n + \beta_n^{(k)} \right) - \tilde{y} \right);$$

- compute  $\Delta$  by applying the bisection method to solve the equation

$$\frac{1}{1 + \left( \frac{c^{(1)}}{\Delta} \right)^2} + \dots + \frac{1}{1 + \left( \frac{c^{(N)}}{\Delta} \right)^2} = \frac{N}{2}$$

on the interval  $[0, \max |c^{(k)}|]$ .

This method works well when the intervals are narrow enough relative to the curvature of the function evaluated.

It is worth mentioning that the use of Cauchy distribution in the above algorithm may seem somewhat counterintuitive. Indeed, in the interval setting, we do not know the exact probability distribution of each error  $\Delta x_i = \tilde{x}_i - x_i$ , but we do know that each error  $\Delta x_i$  belongs to the corresponding interval  $[-\Delta_i, \Delta_i]$ , so the actual (unknown) probability distribution for  $\Delta x_i$  must be located on this interval with probability 1. So, at first glance, if we want to design a simulation-type technique for computing  $\Delta$ , we should use one of such possible distributions in our simulations. Instead, we use a Cauchy distribution for which the probability to be outside the interval  $[-\Delta_i, \Delta_i]$  is non-zero. In other words, in order to make the simulations work, we use the distributions which are inconsistent with our knowledge. The reason why such impossible distributions are useful here is that it can be shown that if we select, for simulations, a distribution within the corresponding p-bound, we end up with a wrong estimate.

Because a p-bound is, in effect, a collection of intervals  $\mathbf{F}(x)$ , it is reasonable to expect that black-box algorithms like the one given above can be generalized to arbitrary p-bounds.

### 3.6 Section's conclusions

There are four different sources of information about a quantity: measurements, expert estimates, first principles, and indirect estimates. In this section, we gave a brief overview of these four sources as they can be used to generate p-bounds.

## 4 Aggregation of Uncertainty in Risk Analysis: Subjective vs. Mathematical Methods

### 4.1 Aggregation of uncertainty: an important problem

In risk analysis, we sometimes have information about a quantity coming from different sources. We must combine this information. The information may be redundant or mutually confirming, or the information may be partially or even totally contradictory. The operation by which one combines estimates obtained by different sources is called *aggregation*. What methods can be used for aggregation? What principles should govern which method of aggregation should be used in any particular situation?

### 4.2 Subjective (“behavioral”) methods are useful

A substantial part of information comes from experts. There are two basic ways of combining the expert information:

- A first natural idea is to take whatever information we got from the experts, real numbers, intervals, CDFs, p-bounds, etc., and apply some algorithm to combine these bounds. The corresponding aggregation methods are called *mathematical* because all we are doing is applying some mathematical formulas and expressions to the pieces of information that represent expert’s knowledge.
- A second natural idea is to ask the experts themselves to help us with aggregating their knowledge. In this case, the result of aggregation is determined by the experts’ behavior, by their subjective treatment of different uncertainties. So, such methods are called *behavioral* or *subjective*.

At first glance, one might get an impression that only mathematical methods are useful and subjective aggregation methods are not needed at all. Indeed, the experts’ time is extremely valuable. We have already used their time to get their estimates of the uncertainty, something that an automatic computer program cannot do. Why should we spend even more of their valuable time asking them to reconcile their different estimates if there exist numerous automatic procedures for such aggregation?

In order to answer this question, let us first ask ourselves: why do different experts come up with different estimates? Experts often differ because they are trying to express to different aspects of the knowledge about a quantity. As a result:

- When we apply mathematical aggregation techniques, we use only the values (intervals, CDFs, etc.) which resulted from analyzing different aspects of the situation, and we ignore the experts’ arguments leading to their estimates.

- When we bring experts together, the experts share not only the numerical results of their uncertainty analysis, they also share their arguments, their perspectives, etc. As a result, each expert learns more about the analyzed situation and is, therefore, able to use this new knowledge to update his or her original estimate of uncertainty.

The idea to make experts themselves reconcile their differences is the main idea behind such successful practices as doctors' conferences. Such subjective (behavioral) methods of combining uncertainty are widely believed to be very useful.

### 4.3 It is not easy to produce a good behavioral method

At first glance, a subjective method is easy: it would seem that all we would need to do is place several experts in one room and wait until they come out of this room with a combined consensus decision. Alas, the reality is much more difficult than that. Many known psychological phenomena of group interaction get into picture and need to be dealt with:

- There is often a certain social hierarchy among experts, as a result of which, some individuals will dominate the discussions and push their original viewpoint.
- When we only bring together experts of approximately the same social status, another phenomenon occurs: In order to reach a consensus, the group may simply follow the majority, in which case it may ignore an important idea provided by one or few experts and/or ignore an important piece of information whose importance is advocated by the minority of experts. As a result, in practice, the consensus will often be much more conservative and narrow than the information warrants.
- Alternatively, a group may arrive at a more extreme position than the information warrants. Indeed, usually, the expert's ideas are tempered by the implicit understanding of the overall consensus. However, when the experts get together, they can influence this consensus and this positive feedback can polarize the group's resulting opinion.

To avoid such problems, several techniques have been proposed; see, e.g., (Clemen & Winkler, 1999). The following subsections describe the most frequently used techniques.

### 4.4 Avoiding dominance: Delphi method

The first problem with subjective aggregation that we mentioned is that some experts are of higher social standing in the expert community, and their opinions may dominate other experts, to the extent that good ideas coming from less

prominent experts may be ignored. A reasonable way to avoid this domination is to provide every expert with all the estimates of uncertainty and arguments in favor of these estimates, but anonymously, without disclosing who exactly provided which estimate or which argument.

This anonymous presentation constitutes the main idea behind the so-called *Delphi method*. In this method, the opinions of different experts are collected by a facilitator and presented to all the experts anonymously. After getting this information, experts update their estimates and send the updated estimates to the facilitator. This procedure is repeated again and again until the experts stop changing their estimates.

#### 4.5 Potential drawbacks of Delphi method

The Delphi method is the safest against domination, but it drastically limits the interaction between the experts. There are many nuances in expert opinions and arguments that are lost in this indirect communication. As a result, in this method, there is a considerably less potential for a fruitful cross-pollination of ideas as compared to methods in which direct expert interaction is allowed.

#### 4.6 Aggregation techniques which involve direct expert interaction

We have mentioned that it is desirable to allow direct interaction between the experts. On the other hand, if we allow non-facilitated, unstructured discussions, we end up with the problems mentioned above. To avoid these problems, we must use a facilitator and structure the discussions. Depending on the level of structuring, there are three main techniques for such structuring:

- In *decision conferencing*, experts are given the freedom to follow any path to conciliation. The facilitator does not try to structure the discussion, and interferes only if he or she observes that the discussion is going into the direction which may cause one of the above three problems (dominance, ignoring minority opinions, or shifting towards an extreme stand).
- In *Nominal Group Techniques*, the discussion is more structured:
  - first, experts present their estimates;
  - then, they discuss these estimates and arguments in favor of these estimates; these discussions are done under the supervision of a skilled facilitator;
  - based on this discussion, each expert revises his/her estimate;
  - then, if necessary, a new round of discussions follows, etc.

- In *Kaplan's method*, experts first try to get a consensus on the evidence; then, after such a consensus is more or less reached, try to transform this body of evidence into a joint estimate. On both stages, Nominal Group Techniques can be used.

#### 4.7 When expert interaction is well organized, the resulting consensus is reasonable

In the previous subsections, we briefly described four subjective/behavioral methods: Delphi method, Decision Conferencing, Nominal Group Techniques, and Kaplan's method. There exist other subjective methods, but there four are the most prominent and the most useful. These subjective aggregation techniques are frequently used in risk analysis. The experience of using these techniques is in full accordance with the above arguments:

- On the one hand, in many cases, the phenomena of domineering, ignoring, and "extremizing" did occur, as a result of which, the resulting aggregated estimates were less adequate than a simple arithmetic average of experts' estimates.
- On the other hand, in the cases in which the expert interaction was well organized, the aggregation results turned out to be reasonable, often much better than a simple average or any other simple non-interactive aggregation.

#### 4.8 Mathematical aggregation methods are needed too

Because subjective (behavioral) aggregation methods are very useful, a natural question is: do we need any mathematical aggregation techniques at all? The answer is "yes", and the arguments in favor of these techniques are as follows:

- First of all, not all estimates come from experts. Many of them come from measurements. These estimates are "pure" real numbers, there are no arguments behind them, so there seems to be no advantage in letting experts combine these real numbers.
- Even if the estimates to be aggregated come from experts, it is not always possible to have the experts reconcile their differences. We have already mentioned that the experts' time is extremely valuable, so we may simply be unable to bring them together for a reconciliation session.
- Finally, even when we bring them together and they partially reconcile their differences – i.e., bring their estimates closer to each other – it is quite possible that there will still be some difference of opinion between different experts. As a result, the revised estimates – although closer to each other than the original ones – are still different and therefore, still

require aggregation of some kind. Because the experts failed to get a final aggregation of their own estimates, we have no other choice but to use some algorithms (mathematical aggregation techniques) to combine the resulting estimates.

In all three cases, we need mathematical techniques for aggregating estimates that involve uncertainty.

#### **4.9 Section's conclusions**

When estimates come from experts, it is beneficial to let the experts themselves discuss their differences and come out with the aggregated estimates. One needs to be very careful in arranging this discussion because, unless it follows a well-structured pattern with a skilled facilitator, the results may be worse than simple averaging of experts' estimates.

Moreover, even in the best situations, there may be some unresolved differences between the expert opinions. In other cases, it is not possible to get the experts together. It is also important to combine the experts' opinion with measurement results. In all these situations, we need mathematical techniques for aggregating uncertainty that can be conducted automatically. In the following sections, we will describe and analyze different mathematical techniques for aggregating uncertainty.

## 5 Aggregation of Uncertainty in Risk Analysis: Outline

### 5.1 Aggregation techniques will be presented in the order from the simplest to the most complex

One of the main goals of this survey is to provide explanations (justifications) for different heuristic aggregation techniques. The formulas corresponding to these techniques can be rather complicated, and thus, difficult to explain. To understand the different aggregation techniques, we start by explaining them on the simplest possible case, then we extend our understanding to more general situations, and from there, we move to explaining the formulas for the most general case.

Which cases are the simplest, which are next simplest? We have mentioned that real numbers, intervals, and probability distributions are particular cases of p-bounds. Which of these cases are the simplest? Second simplest? From the structural viewpoint, all representations of uncertainty consist of one or several real numbers. Specifically:

- a real number  $x$  is just a single real number;
- to represent an interval  $[x^-, x^+]$ , we need two real numbers: its lower endpoint  $x^-$  and its upper endpoint  $x^+$ ;
- to represent a CDF  $F(x)$ , we need to describe, for every value  $x$ , the corresponding probability  $F(x)$ ;
- to represent a p-bound  $[F^-(x), F^+(x)]$ , we need to describe, for every value  $x$ , two real numbers – the corresponding probability bounds  $F^-(x)$  and  $F^+(x)$ .

The more real numbers we use to represent uncertainty, the more processing we need to aggregate the corresponding representations of uncertainty. In view of this, we will

- start with describing different aggregation techniques for combining real numbers,
- then show how these techniques can be extended to intervals, and
- finally, explain how these techniques can be extended to CDFs and to the (general case of) p-bounds.

### 5.2 Desired properties of aggregation techniques

From the purely mathematical viewpoint, we could consider arbitrary operations for combining estimates involving uncertainty. However, from the common sense

viewpoint, we want to combine different estimates in a sensible way. There are some requirements that the aggregation operation should satisfy. These requirements are not absolutely necessary because, in addition to commonsense arguments in favor of these requirements, there are usually some counterarguments. However, it is desirable, when we analyze different aggregation operations, to consider to what extent these aggregation operations satisfy these commonsense requirements.

So, before we review different aggregation techniques, we will describe these natural requirements in detail. The outline of the survey is

- First, we describe different commonsense requirements that an aggregation operation should satisfy.
- Second, we describe the possible aggregation operations with real numbers, and show which commonsense requirements they satisfy.
- Third, we describe the possible aggregation operations with intervals, and show which commonsense requirements they satisfy.
- Fourth, we describe the possible aggregation operations with probability distributions, and show which commonsense requirements they satisfy.
- Fifth, we describe the possible aggregation operations with p-bounds, and show which commonsense requirements they satisfy.
- Finally, we select from among the methods considered those that, by their properties, can be recommended for general use for aggregating estimates involving uncertainty.

## 6 Aggregation of Uncertainty in Risk Analysis: Desirable Properties of Aggregation Operations

### 6.1 General outline

In this section, we will describe the properties of aggregation operations, starting from the simple properties which use only one estimate  $X$ , to more complex properties which deal with two, three, etc., different estimates, and finally, requirements which deal with the aggregation operation as a whole.

We have already mentioned, in Section 5, that the commonsense requirements are not absolutely necessary, because there are some counterarguments. So, after describing each commonsense requirement and listing arguments in favor of this requirement, we will also present some counterarguments which explain why it may make sense for an aggregation operation to violate the requirement.

### 6.2 Commonsense requirement relating a single estimate

#### 6.2.1 Idempotence

What if we have only one estimate  $X$ ? In other words, what if two experts come up with exactly the same  $X$ ; how can we combine their knowledge?

A natural idea is that if two experts came up with the same uncertainty, this means that this is the right representation of this uncertainty, so both experts are right. In other words, if we combine uncertainty  $X$  with itself, we should end up with exactly the same uncertainty  $X$ , i.e., we should have  $X * X = X$ . In mathematics, this “agreement preserving” property of an aggregation operation  $*$  is called *idempotence*. So, in mathematical terms, the first commonsense requirement is that the aggregation operation be idempotent.

#### 6.2.2 Idempotence: possible counterarguments

Suppose that two experts come up with exactly the same description of their uncertainty: that the (unknown) value  $x$  of the desired physical quantity belongs to the interval  $[0, 1]$  with probability  $\geq 90\%$ . What would the result of aggregating these uncertainties be?

If the two experts were using exactly the same sources of information and used the same arguments to process these sources, then the fact that these two experts came up with exactly the same conclusion simply confirms that they both did the correct computations. So, when we aggregate these two uncertainties, we should get the exact same uncertainty. In this case, idempotence is justified.

But what if the two experts used independent sources of information and end up with exactly the same conclusion – that  $x \in [0, 1]$  with probability 90%? In this case the fact that two experts, based on independent sources of information, came up with the same conclusion, increases the reliability of this conclusion. In this case, the result  $X * X$  of combining the two identical uncertainties  $X$  is that  $x$  belongs to the interval  $[0, 1]$  with some probability  $p > 90\%$ . In other words, in this case,  $X * X$  is different from  $X$  – so there is no idempotence.

Another example is stories told to a police officer or evidence about a historical event. If several independent witnesses tell exactly the same story, its reliability increases.

## 6.3 Commonsense requirement relating two estimates

### 6.3.1 Commutativity

The seemingly natural requirement is that if we have two sources of information about an uncertain quantity – e.g., two experts – then the result of aggregating their information  $X$  and  $Y$  should not depend on the order in which these two different pieces of information are presented. In other words, we should have  $X * Y = Y * X$ . In mathematical terms, this requirement is called *commutativity*.

### 6.3.2 Commutativity: possible counterarguments

Commutativity makes sense if there is no reason to prefer one of the two sources of information. In real life, often, one source of information is more reliable than the other. For example, when we combine information coming from two experts, it is normal to give more weight to the opinion of a more respected expert – who has a history of better estimates and better predictions. To handle such situations, we may set up an aggregation operation  $X * Y$  in such a way that  $X$  is the information coming from a more experienced expert, and  $Y$  is the information coming from a less experienced expert.

In this case, the result of combining uncertainties  $X$  and  $Y$  should depend on whether  $X$  comes from the more experienced expert and  $Y$  comes from the less experience one – in which case the result is  $X * Y$  – or whether  $Y$  comes from the more experienced expert and  $X$  comes from the less experience one – in which case the result is  $Y * X$ . In other words, in this case,  $X * Y \neq Y * X$ , and the aggregation operation is not commutative.

## 6.4 Commonsense requirement relating “slightly more than two” estimates

### 6.4.1 Continuity

What if we have “slightly more than two” different estimates, i.e., we have two estimates  $X$  and  $Y$ , and a third estimate  $X'$  which is very close to  $X$ , which

we symbolize as  $X' \approx X$ . Because  $X' \approx X$ , it is reasonable to require that  $X * Y$  is very close to  $X' * Y$ . Symbolically,  $X * Y \approx X' * Y$ . In other words, it is reasonable to require that a small change in one of the uncertainties  $X$  to be aggregated lead to only a small change in the result of the aggregation. In mathematics, such property is called *continuity*.

#### 6.4.2 Continuity: possible counterarguments

At first glance, continuity seems natural, but there are examples when it is counterintuitive. One such example is the case when each estimate is an interval of possible values of the desired quantity. In this case, if one piece of knowledge is that the quantity should be in the interval  $\mathbf{x} = [x^-, x^+]$ , and the other piece of knowledge is that this same quantity should be in the interval  $\mathbf{y} = [y^-, y^+]$ , this means that the actual value  $x$  should belong to both intervals. The set of all the values which belongs to both intervals  $\mathbf{x}$  and  $\mathbf{y}$  is the intersection  $\mathbf{x} \cap \mathbf{y} = [\max(x^-, y^-), \min(x^+, y^+)]$  of these intervals. So, in this case, the aggregation operation is simply an intersection.

From the purely mathematical viewpoint, the intersection seems to have the continuity property: indeed, both the lower endpoint  $\max(x^-, y^-)$  and the upper endpoint  $\min(x^+, y^+)$  of the intersection interval are continuous functions of the parameters  $x^-, x^+, y^-,$  and  $y^+$  that characterize the intervals to be aggregated. So, a small change in one of these four parameters leads to small changes in the endpoints of  $\mathbf{x} \cap \mathbf{y}$ .

However, from the commonsense viewpoint, the situation is not so simple. What happens if we slowly move the interval  $\mathbf{y}$  so that its intersection with  $\mathbf{x}$  becomes smaller and smaller and finally, empty? Before it becomes empty, because we assumed both uncertainty intervals to be 100% reliable, we simply conclude that the result of the aggregation is the intersection. However, when the intersection becomes empty, it clearly means that the two intervals cannot both reliably contain the (unknown) value of the estimated quantity, one of these two intervals is erroneous. If we do not know which of the two intervals is erroneous, then the only thing we can conclude about the actual value of  $x$  is that:

- either  $x$  belongs to the first interval  $\mathbf{x}$  (if the second interval is erroneous),
- or  $x$  belongs to the second interval  $\mathbf{y}$  (if the first interval is erroneous).

In this case, the set of all possible values of  $x$  is the *union*  $\mathbf{x} \cup \mathbf{y}$  of the two intervals to be aggregated. This union is not an interval, so, if we want an *interval* which is guaranteed to contain  $x$ , then we have to take the smallest interval that contains this union. Whether we take the union itself or the smallest interval containing this union, the aggregation result is not at all close to the intersection and so, the continuity property does not hold.

Let us give a simple example. Let  $\varepsilon > 0$  be a small real number, let  $Y = [1, 2]$ ,  $X = [0, 1 + \varepsilon]$ , and  $X' = [0, 1 - \varepsilon]$ . Here:

- Because  $X$  and  $Y$  have a non-empty intersection, the result  $X * Y$  of aggregating  $X$  and  $Y$  is their intersection:  $X * Y = [1, 1 + \varepsilon]$ .
- On the other hand, because  $X'$  and  $Y$  do not have any common points, then, depending on whether we only allow intervals or arbitrary sets, the result  $X' * Y$  of aggregating  $X'$  and  $Y$  is:
  - either the union  $[0, 1 - \varepsilon] \cup [1, 2]$ ,
  - or the smallest interval  $[0, 2]$  which contains this union.

In this case,  $X \approx X'$ , but  $X * Y \not\approx X' * Y$ , i.e., the continuity property is not satisfied.

## 6.5 Commonsense requirement relating three different estimates

### 6.5.1 Associativity

It is natural to require that we have three different sources of information  $X$ ,  $Y$ , and  $Z$ , then the result of aggregating the corresponding pieces of information should not depend on the order in which we aggregate these three pieces:

- If we first present  $X$  and  $Y$ , then we:
  - first combine  $X$  and  $Y$  into a combined knowledge  $X * Y$ , and
  - then, when  $Z$  is presented, we combine the resulting combined knowledge with  $Z$ , resulting in  $(X * Y) * Z$ .
- Alternatively, if we first present  $Y$  and  $Z$ , then we:
  - first combine  $Y$  and  $Z$  into a combined knowledge  $Y * Z$ , and
  - then, when  $X$  is presented, we combine the resulting combined knowledge with  $X$ , resulting in  $X * (Y * Z)$ .

The above requirement means that  $(X * Y) * Z = X * (Y * Z)$ . In mathematical terms, this requirement is called *associativity*.

### 6.5.2 Associativity: possible counterarguments

To show that the above argument is not always intuitively reasonable, let us consider two simple examples: the one in which it is reasonable and the one in which it is not.

Associativity is reasonable if each estimate is an interval of possible values for a quantity. In this case, as we have mentioned when describing continuity, one aggregation operation is simply an intersection. Intersection is, of course, an associative operation: for every three intervals  $\mathbf{x}$ ,  $\mathbf{y}$ , and  $\mathbf{z}$ , the two combinations

$(\mathbf{x} \cap \mathbf{y}) \cap \mathbf{z}$  and  $\mathbf{x} \cap (\mathbf{y} \cap \mathbf{z})$  are exactly the same: they both coincide with the intersection  $\mathbf{x} \cap \mathbf{y} \cap \mathbf{z}$  of all three intervals.

On the other hand, there are simple examples when associativity is not reasonable. One such example is the arithmetic average. When we combine two numerical values  $x$  and  $y$  and we want to get a new numerical estimate, it is reasonable to use the arithmetic average  $x * y \stackrel{\text{def}}{=} \frac{x+y}{2}$ . However, the arithmetic average is not associative: e.g.,

- $0 * 1 = \frac{0+1}{2} = 0.5$ , hence  $(0 * 1) * 2 = 0.5 * 2 = 1.25$ , although
- $0 * (1 * 2) = 0 * 1.5 = 0.75 \neq 1.25$ .

## 6.6 Commonsense requirement relating four estimates

### 6.6.1 Averaging property

We have already mentioned, in the previous subsections, that it is natural to require that the result of aggregating uncertainty information should not depend on the order in which we present different information to be aggregated. This informal requirement seems to justify associativity  $(X * Y) * Z = X * (Y * Z)$ , but, as we have shown, associativity is not always true for reasonable aggregation operations: e.g., for the arithmetic average  $X * Y = \frac{x+y}{2}$ ,  $(X * Y) * Z$  may be different from  $X * (Y * Z)$ . The reason for this possible non-associativity is very simple:

- when we combine  $X * Y$  and  $Z$ , we (kind of) assign “equal weight” to both combined uncertainties  $X * Y$  and  $Z$ ;
- however, in reality,  $X * Y$  is clearly preferable to  $Z$ , because:
  - $X * Y$  combines the expertise of two experts, and
  - $Z$  contains an experience of only one expert.

If we make sure that we always combine uncertainties of the same “strength” (e.g., coming from the same number of experts), then there is a better that the corresponding property will hold. We can formulate this property if we have *four* different estimates  $X$ ,  $Y$ ,  $Z$ , and  $T$ . For these four estimates, we have at least two different options:

- In the first alternative, we do the following:
  - first, we combine  $X$  with  $Y$  into  $X * Y$ , and  $Z$  with  $T$  into  $Z * T$ ;
  - then, we combine the results  $X * Y$  and  $Z * T$  of these combinations into  $(X * Y) * (Z * T)$ .
- In the second alternative, we do the following:

- first, we combine  $X$  with  $Z$  into  $X * Z$ , and  $Y$  with  $T$  into  $Y * T$ ;
- then, we combine the results  $X * Z$  and  $Y * T$  of these combinations into  $(X * Z) * (Y * T)$ .

The above informal requirement then means that

$$(X * Y) * (Z * T) = (X * Z) * (Y * T).$$

This property is (unlike associativity) true for the arithmetic average, and therefore, in mathematics, it is called the *averaging property*; see, e.g., (Suppes et al., 1989) where this property is called *bisymmetry*.

### 6.6.2 Averaging property: possible counterarguments

Possible counterarguments against the averaging property are the same as against commutativity. Often, one source of information is more reliable than the other, so we want to give it preferential treatment. If an aggregation operation  $X * Y$  is set up in such a way that  $X$  comes from a more experienced expert, then combinations  $(X * Y) * (Z * T)$  and  $(X * Z) * (Y * T)$  may correspond to different orderings of the experts' experience and thus, may lead to different aggregation results.

## 6.7 Commonsense requirement about the aggregation operation as a whole

### 6.7.1 Computational simplicity

The main reason why we look for aggregation operations is because we want to solve practical problems. If we have a mathematically perfect aggregation operation that needs years of computations on the fastest computers, then this operation is useless in practical problems that require a solution this year. It is therefore reasonable to consider only operations whose computations do not require too much time, i.e., in computer science terms, computations whose computational complexity is not too high. This requirement for computational simplicity accords with the well-known Occam's razor principle according to which, it is reasonable to select, among all possible hypotheses, the one which is the simplest. Algorithmic Information Theory has provided a theoretical justification for this principle; see, e.g., (Li & Vitányi, 1997) and (Fox et al., 1998). There is also empirical evidence that in general, the simplest techniques for aggregating estimates are indeed the best (Clemen & Winkler, 1999); see, however, (Cox, 1999) for counterexamples.

### 6.7.2 Computational simplicity: possible counterarguments

There may be situations when the above argument does not apply and computationally intensive aggregation operations are preferable. Indeed, we are talking

about the situations of risk analysis, so it all depends on what kind of risk we are talking about:

- If the risk includes a possible minor increase of a pollutant's level in a lake, a problem that can be corrected, then it is acceptable to use an approximate easier-to-compute aggregation technique when making a decision.
- On the other hand, if the risk includes a catastrophic nuclear explosion, then it may be preferable, when estimating the probability of this risk, to use methods which are as accurate as possible, even if it means spending much more effort on computing.

## 6.8 Section's conclusions

When analyzing different aggregation operations, it is reasonable to check whether these operations satisfy the following properties which are desirable: idempotence, commutativity, continuity, associativity, averaging property, and computational simplicity. These properties may not be essential because, as we have shown, there are reasonable examples when requiring these properties would be counterintuitive.

## 7 Aggregation of Uncertainty in Risk Analysis: Case of Real Numbers

### 7.1 General classification

#### 7.1.1 Incertitude vs. variability

Let us start the description of possible aggregation techniques with the simplest possible case in which each piece of information to be aggregated is a real number. Suppose we have two real number  $x_1$  and  $x_2$ . In order to find out the best way of aggregating these two real numbers, let us first ask a question: why are these two real numbers different? There can be two reasons for this difference:

- One possibility is that in both cases, we are measuring (or estimating) the same (unknown) quantity  $x$ . Because measurements (and estimates) are never absolutely accurate, each result of measuring  $x$  is different from the actual value  $x$ . So,  $x_1$  is different from  $x$ ,  $x_2$  is different from  $x$ , and these two measurement results  $x_1$  and  $x_2$  are different from each other. In such situations, the difference between  $x_1$  and  $x_2$  is caused by *incertitude*.
- Another possibility is that the quantity in which we are interested may change in time and/or it may vary slightly from one point to another. In this situation, the difference between the measurement results  $x_1$  and  $x_2$  can be explained simply by the fact these measurements were made at slightly different moments of time and in slightly different places. In such situations, the difference between  $x_1$  and  $x_2$  is caused by *variability*.

Of course, it is also possible that we have *both* incertitude and variability: the actual value of the measured quantity does vary from place to place and, on top of that, the measured value differs from the actual one.

#### 7.1.2 Averaging vs. enveloping

In case of incertitude, both values  $x_1$  and  $x_2$  are estimates for the unknown (actual) value  $x$ . There are two possible approaches to combining these two values. At present, a typical approach is to combine these two numerical estimates  $x_1$  and  $x_2$  into a (hopefully) better numerical estimate  $\tilde{x}$ . The resulting estimate combines (“averages”) the numbers  $x_1$  and  $x_2$ ; therefore, we will call such aggregation operations *averaging operations*. From the practical viewpoint, averaging has advantages and disadvantages.

- The main advantage of averaging is that it *compacts* the information, i.e., no matter how many numerical estimates we combine, the resulting estimate is a single number.

- The main disadvantage of averaging is related to its main advantage: it compacts the information and, as a result, the aggregated value lacks information on the original incertitude.

In the pre-computer era, when we had to process all the data manually, we could only process a small amount of numbers. In this situation, compactification is necessary, and so averaging was, in effect, the only choice. That is why averaging methods have been traditionally prevailing in data processing. Yes, averaging has a disadvantage, but since it was the only choice, we had to live with it.

With modern computers that can process (and do process) millions of data points, compactification is no longer necessary. Since we do not have to use averaging, we no longer have to tolerate its disadvantage, i.e., the loss of incertitude. It is therefore reasonable to consider aggregated estimates which preserve this incertitude, i.e., which “envelope” both values  $x_1$  and  $x_2$  to be aggregated. Such *enveloping* methods propagate incertitude.

In case of variability, the selection of an aggregation operation is simpler. Indeed, in this case, both values  $x_1$  and  $x_2$  are possible values of the analyzed quantity. Thus, a natural goal is to describe the *set* of possible value of this quantity, i.e., to provide an *envelope* for the values  $x_1$  and  $x_2$ . In other words, for variability, only enveloping operations make sense. In the following subsections, we will first describe more traditional averaging aggregation methods, and then enveloping methods.

## 7.2 Arithmetic average: its origin and properties

### 7.2.1 Averaging operation: loss minimization naturally leads to the arithmetic average

Let us start with describing averaging aggregation operations for the case of incertitude. Each of the two measurement results  $x_1$  and  $x_2$  is presumably close to  $x$ . The purpose of an averaging aggregation is to replace both  $x_1$  and  $x_2$  by a single value  $\tilde{x}$ . We want this aggregated value  $\tilde{x}$  to be as close to  $x$  as possible. How can we do that?

Let us first consider the simplified situation when there is only one measurement result:  $x_1$ . In this case, the only information that we have about the actual (unknown) value  $x$  is that  $x$  is close to  $x_1$ . Thus, in this simplified situation, the only possible way to make sure that  $\tilde{x}$  is close to the unknown value  $x$  is to make sure that  $\tilde{x}$  is close to  $x_1$ , i.e., that the deviation  $\Delta x_1 = \tilde{x} - x_1$  is small.

Similarly, in the simplified situation when there is only one measurement result  $x_2$ , the only possible way to make sure that  $\tilde{x}$  is close to the unknown value  $x$  is to make sure that  $\tilde{x}$  is close to  $x_2$ , i.e., that the deviation  $\Delta x_2 = \tilde{x} - x_2$  is small.

In our case, we know two approximations to  $x$ :  $x_1$  and  $x_2$ . Thus, to make absolutely sure that the aggregated value  $\tilde{x}$  is close to  $x$ , we should make sure

that  $\tilde{x}$  is close *both* to  $x_1$  and  $x_2$ , i.e., that both differences  $\Delta x_1$  and  $\Delta x_2$  be small.

The smaller the differences, the better. The larger the differences, the further  $\tilde{x}$  away from the actual value  $x$  and hence, the larger the loss caused by using, in decision making, the aggregated estimate  $\tilde{x}$  instead of the actual value  $x$ . To find the optimal value of  $\tilde{x}$ , we want to describe the objective function  $J(\Delta x_1, \Delta x_2)$  that describes this loss.

In describing this function  $J$ , we will use the same idea that we used when we described a natural way of representing partial information about probabilities. Specifically, we assume that this function  $J$  is smooth (infinitely differentiable). In this case, because the measurements errors are usually reasonably small, we can expand the function  $J$  into Taylor series and retain only linear and quadratic terms in its expansion and ignore cubic and higher-order terms. As a result, we get the general expression

$$J(\Delta x_1, \Delta x_2) = a_0 + a_1 \cdot \Delta x_1 + a_2 \cdot \Delta x_2 + a_{11} \cdot \Delta x_1^2 + a_{12} \cdot \Delta x_1 \cdot \Delta x_2 + a_{22} \cdot \Delta x_2^2.$$

Because we do not have any reason to prefer negative or positive values of the difference  $\Delta x_1$ , it makes sense to assume that the value of the loss function will not change if we replace  $\Delta x_1$  by  $-\Delta x_1$ :  $J(-\Delta x_1, \Delta x_2) = J(\Delta x_1, \Delta x_2)$ . Substituting these two requirements into the general quadratic expression, we conclude that

$$J(-\Delta x_1, \Delta x_2) =$$

$$a_0 + a_1 \cdot (-\Delta x_1) + a_2 \cdot \Delta x_2 + a_{11} \cdot (-\Delta x_1)^2 + a_{12} \cdot (-\Delta x_1) \cdot \Delta x_2 + a_{22} \cdot \Delta x_2^2 =$$

$$a_0 - a_1 \cdot \Delta x_1 + a_2 \cdot \Delta x_2 + a_{11} \cdot \Delta x_1^2 - a_{12} \cdot \Delta x_1 \cdot \Delta x_2 + a_{22} \cdot \Delta x_2^2 =$$

$$J(\Delta x_1, \Delta x_2) = a_0 + a_1 \cdot \Delta x_1 + a_2 \cdot \Delta x_2 + a_{11} \cdot \Delta x_1^2 + a_{12} \cdot \Delta x_1 \cdot \Delta x_2 + a_{22} \cdot \Delta x_2^2$$

for all possible values  $\Delta x_1$  and  $\Delta x_2$ . The two polynomials are equal for all possible values of the variables if and only if all their coefficients are equal, hence  $-a_1 = a_1$  and  $-a_{12} = a_{12}$ , i.e.,  $a_1 = a_{12} = 0$ . Similarly, it makes sense to require that the value of the loss function will not change if we replace  $\Delta x_2$  by  $-\Delta x_2$ :  $J(\Delta x_1, -\Delta x_2) = J(\Delta x_1, \Delta x_2)$ . Substituting this requirement into the general quadratic expression, we conclude that  $a_2 = 0$  and therefore

$$J(\Delta x_1, \Delta x_2) = a_0 + a_{11} \cdot \Delta x_1^2 + a_{22} \cdot \Delta x_2^2.$$

Because there is no reason to prefer one measurement over another, it makes sense to require that the loss function will not change if we swap the differences  $\Delta x_1$  and  $\Delta x_2$ , i.e., to require that  $J(\Delta x_2, \Delta x_1) = J(\Delta x_1, \Delta x_2)$ . This condition leads to  $a_{11} = a_{22}$ , i.e., that

$$J(\Delta x_1, \Delta x_2) = a_0 + a_{11} \cdot (\Delta x_1^2 + \Delta x_2^2).$$

Because we want the loss to be the smallest when  $\Delta x_1 = \Delta x_2 = 0$ , we can therefore conclude that  $a_{11} > 0$ . The resulting loss function

$$J(\tilde{x} - x_1, \tilde{x} - x_2) = a_0 + a_{11} \cdot ((\tilde{x} - x_1)^2 + (\tilde{x} - x_2)^2)$$

can be easily minimized by explicitly differentiating the loss function relative to  $\tilde{x}$  and equating this derivative to 0. The resulting minimum is the arithmetic average:

$$\tilde{x} = \frac{x_1 + x_2}{2}.$$

Similarly, when we have more than two measurements  $x_1, \dots, x_n$ , then, to make sure that the aggregate value  $\tilde{x}$  is close to the actual (unknown) value  $x$ , we must look for an aggregated value  $\tilde{x}$  that is close to all of these measurement results, i.e., for which all  $n$  differences  $\Delta x_i \stackrel{\text{def}}{=} \tilde{x} - x_i$ ,  $1 \leq i \leq n$ , are small. Similarly to the case of two measurements, we can expand the corresponding loss function  $J(\Delta x_1, \dots, \Delta x_n)$  into Taylor series and keep only linear and quadratic terms in this expansion. It also makes sense to require that the value of the loss function should not change if we simply change the sign of one of the differences  $\Delta x_i$ , or permute the values  $\Delta x_i$ . As a result, we conclude that:

$$J(\Delta x_1, \Delta x_2) = a_0 + a_{11} \cdot (\Delta x_1^2 + \dots + \Delta x_n^2).$$

In this case, the minimum is also attained for the arithmetic average:

$$\tilde{x} = \frac{x_1 + \dots + x_n}{n}.$$

Thus, loss minimization naturally leads to the use of arithmetic average as an aggregation operation.

From the commonsense viewpoint, arithmetic average is a very reasonable aggregation operation. The arithmetic average corresponds to a natural idea of “splitting the difference”. Indeed, when the first expert measures the value of the quantity as  $x_1$ , and the second expert measures the same quantity as  $x_2$ , with the difference  $x_2 - x_1$ , the natural way of reconciling this difference is to split it in half, i.e., to replace both estimates by the common estimate

$$x_1 + \frac{1}{2} \cdot (x_2 - x_1) = x_2 + \frac{1}{2} \cdot (x_1 - x_2) = \frac{x_1 + x_2}{2}.$$

### 7.2.2 An alternative derivation of arithmetic average of two measurement results

For two measurement results, there is simpler alternative way of showing that arithmetic average is indeed a natural aggregation operation. This explanation is based on the fact that we are dealing with the case of uncertainty, where the actual measured value  $x$  is the same, and the difference between the measurement results is caused by the fact that measurements are inaccurate.

A natural characteristic of the measurement inaccuracy is the upper bound  $\Delta$  on the measurement error  $x_i - x$ . If we know  $\Delta$ , then, from the fact the measurement result is  $x_1$ , we can conclude that the actual (unknown) value of the measured quantity  $x$  belongs to the interval  $[x_1 - \Delta, x_1 + \Delta]$ . Similarly, from the second measurement result  $x_2$ , we conclude that the actual value belongs to the interval  $[x_2 - \Delta, x_2 + \Delta]$ . Because the value  $x$  must belong to both intervals, these two intervals  $[x_1 - \Delta, x_1 + \Delta]$  and  $[x_2 - \Delta, x_2 + \Delta]$  must have a non-empty intersection. Because we know that the actual value  $x$  belongs to this intersection, we would like to select the aggregated value  $\tilde{x}$  in such a way that it should also belong to this intersection.

Because we do not know the actual value  $\Delta$ , we should select  $\tilde{x}$  that belongs to this intersection for all possible values  $\Delta$ . Which values  $\Delta$  are possible? From the fact that  $|x_1 - x| \leq \Delta$  and  $|x_2 - x| \leq \Delta$ , we can conclude, using triangle inequality, that  $|x_2 - x_1| \leq |x_1 - x| + |x_2 - x| \leq 2\Delta$ , therefore,  $\Delta \geq |x_1 - x_2|/2$ . When  $\Delta = |x_1 - x_2|/2$ , i.e., when  $\Delta$  is equal to exactly the half of the distance between  $x_1$  and  $x_2$ , the intersection of the corresponding intervals  $[x_1 - \Delta, x_1 + \Delta]$  and  $[x_2 - \Delta, x_2 + \Delta]$  consists of a single point: the midpoint  $(x_1 + x_2)/2$  between  $x_1$  and  $x_2$ . Thus, the only way to make sure that the aggregated value belongs to the intersection for all possible  $\Delta$  is to select this midpoint (arithmetic average) as the desired averaging operation. (One can show that the arithmetic average indeed belongs to the intersection for all possible  $\Delta$ .)

### 7.2.3 Warning: this alternative derivation does not lead to arithmetic average for more than two measurement results

The above derivation is simpler than analyzing a loss function, but it has a problem: it does not scale to aggregation of three or more measurement results  $x_1, \dots, x_n$ . In this case, the only value that belongs to the intersection of all the intervals  $[x_i - \Delta, x_i + \Delta]$  for all possible values  $\Delta$  – i.e., for all the values  $\Delta$  for which these  $n$  intervals have a non-empty intersection – is the midpoint  $0.5 \cdot x_{(1)} + 0.5 \cdot x_{(n)}$  between the smallest  $x_{(1)}$  and the largest  $x_{(n)}$  of the  $n$  measurement results.

This formula is different from the arithmetic average. It is rather a *weighted* average (see below).

### 7.2.4 Properties of arithmetic average

In the previous section, we enumerated reasonable properties of aggregation operations. Let us check which of these properties hold for arithmetic average:

- Arithmetic average is clearly idempotent:  $x * x = \frac{x + x}{2} = x$ .
- Arithmetic average is also clearly commutative:  $x * y = y * x$ .
- Arithmetic average is continuous: the function  $\frac{x + y}{2}$  is continuous in both  $x$  and  $y$  and thus, small changes in  $x$  and  $y$  lead to a small change in  $x * y$ .

- We have already mentioned that the arithmetic average is not associative.
- Arithmetic average satisfies the averaging property.
- Finally, the arithmetic average is very easy to compute, so it clearly satisfies the condition of computational simplicity.

## 7.3 Weighted average

### 7.3.1 From arithmetic average to weighted average

When we derived the formula for the arithmetic average, we assumed that all measurements (estimates)  $x_i$  are of equal quality. In reality, some measurements may be more accurate than others, some estimates may be done by experts who are more reliable than others, etc. How can we take such differences into consideration?

In some sense, non-associativity is a blessing in disguise, because it enables us to describe combinations of measurement results of different quality. Indeed, why is  $(x * y) * z$  different from  $x * (y * z)$ ? As we have discussed in our analysis of the averaging property,

- the arithmetic average combination rule  $*$  corresponds to the case when both values to be aggregated are of equal accuracy, but
- $x * y$  is more accurate than  $z$ , because:
  - $x * y$  combines the results of two measurements (or, in case of expert estimates, the expertise of two experts), and
  - $z$  contains only a single measurement (or the experience of only one expert).

We can use this idea to describe a general combination of measurements or expert estimates of different accuracy. Let us start with a simple physical example. Let us assume that we have two rocks of different weight. If the weight of the second rock is exactly double the weight of the first one, we can say that the second rock is equivalent to two rocks of the same weight as the first one. If the weight of the second rock is  $5/4$  of the weight of the first rock, we can say that the first rock is equivalent to 4 pieces of the same weight, and the second rock is equivalent to 5 pieces of the same weight. The cases when the weight ratio is irrational can be approximated by rational ratios and thus, also described in this form.

Similarly, if the sources to be aggregated are of different accuracy, we can assume that the first value is equivalent to a combination of  $k$  measurement results of equal accuracy, and the second value is equivalent to a combination of  $l \neq k$  measurement results of the same accuracy.

In other words, the corresponding aggregation operation can be described as an operation which transforms the result  $\bar{x}_1$  of aggregating  $k$  measurements  $x_1, \dots, x_k$  of equal accuracy and the result  $\bar{x}_2$  of aggregating  $l$  measurements  $x_{k+1}, \dots, x_{k+l}$  of the same accuracy into the result  $\bar{x}$  of aggregating all  $k+l$  values  $x_1, \dots, x_k, x_{k+1}, \dots, x_{k+l}$ . Due to the above arithmetic average formula,

$$\bar{x}_1 = \frac{x_1 + \dots + x_k}{k}$$

and

$$\bar{x}_2 = \frac{x_{k+1} + \dots + x_{k+l}}{l}.$$

Thus,  $x_1 + \dots + x_k = k \cdot \bar{x}_1$ ,  $x_{k+1} + \dots + x_{k+l} = l \cdot \bar{x}_2$ , hence,

$$x_1 + \dots + x_{k+l} = (x_1 + \dots + x_k) + (x_{k+1} + \dots + x_{k+l}) = k \cdot \bar{x}_1 + l \cdot \bar{x}_2,$$

and

$$\bar{x} = \frac{x_1 + \dots + x_{k+l}}{k+l} = w_1 \cdot \bar{x}_1 + w_2 \cdot \bar{x}_2,$$

where we denoted  $w_1 = k/(k+l)$  and  $w_2 = l/(k+l)$ .

In other words, the resulting aggregation operation is the *weighted average*  $x_1 * x_2 = w_1 \cdot x_1 + w_2 \cdot x_2$  for some weights  $w_1$  and  $w_2$  for which  $w_1 \geq 0$ ,  $w_2 \geq 0$ , and  $w_1 + w_2 = 1$ .

In the general case, when we combine  $n$  measurement results, we get a general formula for the weighted average:

$$\tilde{x} = w_1 \cdot x_1 + \dots + w_n \cdot x_n,$$

where  $w_i \geq 0$  and  $w_1 + \dots + w_n = 1$ .

### 7.3.2 Where do we get the weights for a weighted average?

- If we have some information about the relative accuracy of different estimates, we can use this information to determine the weights:
  - If each measurement result  $x_i$  comes from averaging a sample of size  $N_i$ , then, as we have mentioned, we should take

$$w_i = \frac{N_i}{N_1 + \dots + N_n}.$$

- If the values  $x_i$  represent expert estimates, then the weights  $w_i$  are proportional to trustworthiness of different experts.
- The weights could also depend on the order of the measurement results. In this case, we have an expression of the type

$$\tilde{x} = w_1 \cdot x_{(1)} + \dots + w_n \cdot x_{(n)},$$

where

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

is the result of sorting the original measurement results in increasing order. Yager calls these *Ordered Weighted Averages* (OWA) (Yager, 1988), (Yager & Kelman, 1996), (Yager & Kacprzyk, 1997), (Nguyen & Kreinovich, 1997), (Grabisch et al., 1998), (Kelman & Yager, 1998), (Fodor & Yager, 2000). When  $w_i = 1/n$ , these degenerate to the simple arithmetic average. There are several non-degenerate examples of such combinations:

- When we explained the arithmetic average for  $n = 2$ , we derived a formula  $0.5 \cdot x_{(1)} + 0.5 \cdot x_{(n)}$  which corresponds to choosing  $w_1 = w_n = 0.5$  and  $w_2 = w_3 = \dots = w_{n-1} = 0$ .
- Formulas of this type are used in *robust statistics* (Wadsworth, 1990), i.e., in statistical processing in which we do not know the exact shape of the probability distribution, only a general class to which this shape belongs.
- Similar formulas are used not only in more mathematical applications, but in common sense as well. For example, when a panel of judges from different countries judge, e.g., figure skating, then, due to highly politicized nature of sports, a judge from the same country as the athlete is usually biased in favor of this athlete, and a judge from this country's main rival is usually biased against the athlete. These biased estimates have little to do with the actual performance of the athlete, so it is desirable to eliminate them. To eliminate these biased estimates, we delete the highest grade (which most probably comes from the positively biased judge) and the lowest grade (which most probably comes from the negatively biased judge). The remaining grades are averaged. This procedure corresponds to selecting  $w_1 = w_n = 0$  and  $w_2 = w_3 = \dots = w_{n-1} = 1/(n-2)$ .
- The weight of the evidence may also depend on how recent this evidence is. The exact dependence of the weight on the date depends on the problem:
  - If we are estimating the current value of the quantity, then, due to possible change in time, we should put more weight to more recent measurements and less weight to more distant ones.
  - On the other hand, if we are estimating the value of the quantity at a certain time in the past (e.g., if we are trying to reconstruct a past event), then, vice versa, we should put more weight on past measurements and less weight to more recent ones.

### 7.3.3 Properties of weighted average

Weighted average is idempotent but not commutative, and continuous but not associative. It is computationally simple but does not satisfy the averaging property.

To be more precise, weighted average is almost associative; we just simply have to change the weights. For example, if we start with an arithmetic average operation  $x * y = \frac{x+y}{2}$ , then the arithmetic average

$$x * y * z = \frac{x + y + z}{3}$$

of three values  $x$ ,  $y$ , and  $z$  is different from  $(x * y) * z$ , but we can get  $x * y * z$  as a weighted average of  $x * y = \frac{x+y}{2}$  and  $z$ :

$$x * y * z = \frac{2}{3} \cdot (x * y) + \frac{1}{3} \cdot z.$$

## 7.4 Weighted average with interval weights

### 7.4.1 Motivations

In some cases, we do not know the exact values of the weights  $w_i$ , only *intervals*  $\mathbf{w}_i = [w_i^-, w_i^+]$  of possible values of these weights. These interval weights must be consistent in the sense that there should exist values  $w_i \in \mathbf{w}_i$  for which  $w_1 + \dots + w_n = 1$ . One can check that this consistency is equivalent to the following inequality:

$$w_1^- + \dots + w_n^- \leq 1 \leq w_1^+ + \dots + w_n^+.$$

There are several different values of the weights, so instead of a single weighted average, we have an *interval* of possible values of the weighted average:

$$\mathbf{x} = \left\{ \sum w_i \cdot x_i \mid w_i \in \mathbf{w}_i, \sum w_i = 1 \right\}.$$

### 7.4.2 Algorithm

The endpoints  $x^-$  and  $x^+$  of this interval  $\mathbf{x}$  can be computed as follows:

- First, we sort the values  $x_i$  in increasing order. Without losing generality, we can assume that the values  $x_i$  are already sorted, i.e., that

$$x_1 \leq x_2 \leq \dots \leq x_n.$$

- Second, we find the value  $k$  from 1 to  $n$  for which

$$w_1^- + \dots + w_{k-1}^- + w_k^- + w_{k+1}^+ + \dots + w_n^+ \leq 1 \leq$$

$$w_1^- + \dots + w_{k-1}^- + w_k^+ + w_{k+1}^+ + \dots + w_n^+.$$

When  $n$  is small, this value can be found by simply trying all  $k$  from 1 to  $n$ ; when  $n$  is large, we can use bisection by first trying  $k = 1$  and  $k = n$ , then trying a midpoint  $(1 + n)/2$ , etc.

- Then, we compute

$$w_k = 1 - (w_1^- + \dots + w_{k-1}^- + w_{k+1}^+ + \dots + w_n^+)$$

and take

$$x^+ = w_1^- \cdot x_1 + \dots + w_{k-1}^- \cdot x_{k-1} + w_k \cdot x_k + w_{k+1}^+ \cdot x_{k+1} + \dots + w_n^+ \cdot x_n.$$

- Next, we find the value  $l$  from 1 to  $n$  for which

$$w_1^+ + \dots + w_{l-1}^+ + w_l^- + w_{l+1}^- + \dots + w_n^- \leq 1 \leq$$

$$w_1^+ + \dots + w_{l-1}^+ + w_l^+ + w_{l+1}^- + \dots + w_n^-.$$

When  $n$  is small, this value can be found by simply trying all  $l$  from 1 to  $n$ ; when  $n$  is large, we can use bisection by first trying  $l = 1$  and  $l = n$ , then trying a midpoint  $(1 + n)/2$ , etc.

- Then, we compute

$$w_l = 1 - (w_1^+ + \dots + w_{l-1}^+ + w_{l+1}^- + \dots + w_n^-)$$

and take

$$x^- = w_1^+ \cdot x_1 + \dots + w_{l-1}^+ \cdot x_{l-1} + w_l \cdot x_l + w_{l+1}^- \cdot x_{l+1} + \dots + w_n^- \cdot x_n.$$

### 7.4.3 Examples

Let us give two examples.

The first example is the simpler: we have two values  $x_1$  and  $x_2$  to be aggregated, and we have no information about the corresponding weights  $w_1$  and  $w_2$ . This means that each of the weights can take any value from the interval  $[0, 1]$ , i.e., that  $\mathbf{w}_1 = \mathbf{w}_2 = [0, 1]$ . In terms of bounds,  $w_1^- = w_2^- = 0$  and  $w_1^+ = w_2^+ = 1$ . To find the result of the corresponding aggregation, we first sort the values; let us assume that they are already sorted, i.e., that  $x_1 \leq x_2$ . In this case,  $w_1^- + w_2^- \leq 1 = w_1^- + w_2^+$ , so  $k = 2$ ,  $w_2 = 1 - w_1^- = 1 - 0 = 1$ , and  $x^+ = w_1^- \cdot x_1 + w_2 \cdot x_2 = x_2$ . Similarly,  $x^- = x_1$ , so the resulting interval is  $\mathbf{x} = [x_1, x_2]$ . In other words, the aggregated interval is the simply the smallest interval containing the values  $x_1$  and  $x_2$  – i.e., the *envelope* of the values.

This conclusion makes perfect sense because for different values of  $w_1 \geq 0$  and  $w_2 = 1 - w_1$ , the combination  $w_1 \cdot x_1 + w_2 \cdot x_2$  ranges from  $x_1$  (when  $w_1 = 1$ )

to  $x_2$  (when  $w_1 = 0$  and  $w_2 = 1$ ), covering all the intermediate values for  $w_1$  between 0 and 1.

Similarly, if we have  $n$  measurement results  $x_1, \dots, x_n$ , and we do not have any information about the weights  $w_i$  (i.e.,  $\mathbf{w}_i = [0, 1]$ ), then the aggregated interval is  $\mathbf{x} = [x_{(1)}, x_{(n)}]$ .

The second example is about the case when we have some information about the weights. Let  $x_1 = 1$ ,  $x_2 = 2$ ,  $x_3 = 3$ , and let  $\mathbf{w}_i = [0.25, 0.35]$  for all  $i$ . In his case,

$$0.95 = w_1^- + w_2^+ + w_3^+ \leq 1 \leq w_1^+ + w_2^+ + w_3^+ = 1.05,$$

so  $k = 1$ ,  $w_1 = 1 - (w_2^+ + w_3^+) = 1 - 0.7 = 0.3$ , and

$$x^+ = w_1 \cdot x_1 + w_2^+ \cdot x_2 + w_3^+ \cdot x_3 = 0.3 \cdot 1 + 0.35 \cdot 2 + 0.35 \cdot 3 = 2.05.$$

Similarly,

$$0.95 = w_1^+ + w_2^+ + w_3^- \leq 1 \leq w_1^+ + w_2^+ + w_3^+ = 1.05,$$

so  $l = 3$ ,  $w_3 = 1 - (w_1^+ + w_2^+) = 1 - 0.7 = 0.3$ , and

$$x^- = w_1^+ \cdot x_1 + w_2^+ \cdot x_2 + w_3 \cdot x_3 = 0.35 \cdot 1 + 0.35 \cdot 2 + 0.3 \cdot 3 = 1.95.$$

Thus, the aggregated interval is  $\mathbf{x} = [1.95, 2.05]$ .

#### 7.4.4 Properties

Weighted average with interval weights is idempotent but not commutative, continuous but not associative. It does not satisfy the averaging property but it is computationally simple.

### 7.5 More general averaging operations

#### 7.5.1 Problems with arithmetic average and weighted average

At first glance, arithmetic average and its minor modification – weighted average – sound reasonable, and for many practical problems, these methods are very useful. However, in some practical situations, we get somewhat counterintuitive results when aggregating expert estimates in risk analysis. Sometimes, we need to estimate the value of a small quantity, e.g., concentration  $x$  of a certain pollutant in a lake. Experts may give answers which differ by orders of magnitude: e.g., three experts can give estimates as  $10^{-5}$ ,  $10^{-6}$ , and  $10^{-7}$ . When the experts give these estimates, of course, they do not mean that, in their opinion, the concentration is, say, *exactly*  $10^{-7}$ . Rather, an expert means that the actual value is *around*  $10^{-7}$ , i.e., of order  $10^{-7}$ . What happens if we use arithmetic average to combine these estimates? As a result, we get

$$\tilde{x} = \frac{10^{-5} + 10^{-6} + 10^{-7}}{3} \approx \frac{1}{3} \cdot 10^{-5}.$$

In other words, as a result of the aggregation, instead of some meaningful “average” of the three expert estimates, we get, as an aggregated value, the answer which is, in effect, the opinion of the first expert. In general, when we use the arithmetic average to aggregate values of different orders of magnitude, we get, in effect, the *largest* of the estimates. The situation is not helped much if we allow weights, unless, of course, we assign a tiny weight (about 0.1 or 0.01) to the expert whose estimates is the largest. This may not sound so bad if we aggregate the opinions of only three experts, but what if we had dozens of them and one expert has an orders of magnitude higher value, arithmetic average will be dominated by the opinion of this particular expert.

This problem is not limited to risk analysis; it is well known in statistics, where, in the presence of outliers, the arithmetic average stops making sense. Within robust statistics (Wadsworth, 1990), several modifications of arithmetic average have been developed to solve this problem. Let us show that similar modifications naturally appear in our risk-analysis problems.

### 7.5.2 From arithmetic average to geometric average and generalized average

In the above text, we derived the arithmetic average formula from several assumptions which, at that time, seemed to be pretty natural (and which are, in many applications, quite natural). Because in some practical situations, the resulting formulas turn out to be counterintuitive, this means that in these situations, some of our original assumptions are not as intuitively reasonable as we originally thought.

Basically, we had four assumptions:

- that the objective function is smooth – this still seems like a reasonable assumption;
- that there is no reason to prefer positive or negative errors – this also seems reasonable;
- that there is no prior reason to prefer one expert to another – this is also still reasonable;
- and, finally, the assumption that the quality of an approximation depends on the difference between the original and the approximated values.

This last assumption is what is suspect in our current example. Let us use this assumption to compare the following two situations:

- in the first situation, the actual value is  $10^{-7}$  and an expert estimates it as  $10^{-6}$ ;
- in the first situation, the actual value is  $10^{-6}$  and an expert estimates it as  $10^{-5}$ .

In both cases, an expert errs by an order of magnitude, so intuitively, the errors in these two situations should be of similar graveness. However, based on the above assumption, we end up with a completely different conclusion. In fact,

- in the first situation, the difference between the actual and the estimated values is  $0.9 \cdot 10^{-6}$ , although
- in the second situation, the difference between the actual and the estimated values is  $0.9 \cdot 10^{-5}$ .

In the second situation, the difference is 10 times larger than in the first one, so, according to the above assumption, the expert's estimate in the first situation is much more accurate than in the second one.

This example shows that the above assumption is not intuitively reasonable. How can we modify this assumption so that it will become intuitively reasonable again? The main problem with the above assumption is that it compared the *values* of the estimated quantity. From the intuitive viewpoint, we are interested not in the values themselves, but rather in their *effect*, or their *perception*. Researchers in psychology and psychophysiology have discovered that if we want to capture difference in perception and/or difference in effect, we must use not the original physical scale, but its non-linear rescaling  $X = f(x)$  for some function  $f(x)$ . The meaning of this scale is that the difference between the values  $x$  and  $x'$  is perceived to be the same as the difference between the values  $y$  and  $y'$  if and only if  $X - X' = Y - Y'$ , where  $X = f(x)$ ,  $X' = f(x')$ ,  $Y = f(y)$ , and  $Y' = f(y')$  represent the values  $x$ ,  $x'$ ,  $y$ , and  $y'$  in the new scale.

Two types of rescaling are most frequently used; see, e.g., (Milner, 1970):

- Fechner scale, in which the perception corresponds to  $X = \ln(x)$ ; and
- Stevens' scale, in which the perception is best described by the value  $X = x^\alpha$  for an appropriate parameter  $\alpha$ .

There is also a general symmetry-based explanation of these scales, see, e.g., (Nguyen & Kreinovich, 1997).

To describe the approximation accuracy, we must use the difference not between the values expressed in the original scale, but between the values expressed in a new scale. In the new scale, all four above assumptions are valid and therefore, in the new scale, the appropriate aggregation operation is the arithmetic average.

How will this aggregation operation look in the original scale? If we start with two values  $x$  and  $y$  in the original scale, then to find the aggregation result  $x * y$  in the original scale, we must do the following:

- first, transform  $x$  and  $y$  into the new scale, thus computing  $X = f(x)$  and  $Y = f(y)$ ;
- compute the arithmetic average  $Z = \frac{X + Y}{2}$  of the values  $X$  and  $Y$ ;

- transform the value  $Z$  into the original scale by applying the inverse function  $f^{-1}$ :

$$x * y = f^{-1}(Z) = f^{-1}\left(\frac{X + Y}{2}\right) = f^{-1}\left(\frac{f(x) + f(y)}{2}\right).$$

Specifically:

- for Fechner law,  $f(x) = \ln(x)$ , and  $x * y = \sqrt{x \cdot y}$ , i.e., we get the *geometric average*;
- for Stevens' law,  $f(x) = x^\alpha$ , and

$$x * y = \left(\frac{x^\alpha + y^\alpha}{2}\right)^{1/\alpha}.$$

Similarly, for aggregating  $n$  values  $x_1, \dots, x_n$ , we get the formula

$$x_1 * \dots * x_n = f^{-1}\left(\frac{f(x_1) + \dots + f(x_n)}{n}\right).$$

Specifically:

- for Fechner law,  $f(x) = \ln(x)$ , and

$$x_1 * \dots * x_n = \sqrt[n]{x_1 \cdot \dots \cdot x_n};$$

- for Stevens' law,  $f(x) = x^\alpha$ , and

$$x_1 * \dots * x_n = \left(\frac{x_1^\alpha + \dots + x_n^\alpha}{n}\right)^{1/\alpha}.$$

In particular:

- when  $\alpha \rightarrow -\infty$ , the corresponding operation tends to the minimum  $\min(x_1, \dots, x_n)$ ;
- when  $\alpha \rightarrow 0$ , the corresponding operation tends to the geometric average  $\sqrt[n]{x_1 \cdot \dots \cdot x_n}$ ;
- when  $\alpha = 1$ , the corresponding operation coincides with the original arithmetic average;
- when  $\alpha \rightarrow \infty$ , the corresponding operation tends to the maximum  $\max(x_1, \dots, x_n)$ .

In particular, if we apply the geometric average to three estimates  $10^{-7}$ ,  $10^{-6}$ , and  $10^{-5}$ , we get a reasonable aggregation result  $10^{-6}$ .

### 7.5.3 Properties of geometric average and generalized average

Because geometric average and generalized average (with finite  $\alpha$ ) are, in fact, arithmetic average described in a different scale, these aggregation operations have exactly the same properties as the arithmetic average: they are idempotent, commutative, continuous, non-associative, satisfy the averaging property, and are computationally simple.

The operations min and max corresponding to the limit cases  $\alpha = \pm\infty$  have an additional property of associativity.

### 7.5.4 Weighted geometric and generalized average

If we apply weighted average  $Y = w_1 \cdot X_1 + \dots + w_n \cdot X_n$  ( $\sum w_i = 1$ ) to the values  $X_i$  expressed in the new scales, then, in the original scales, we get the following operations:

- for Fechner law,  $f(x) = \ln(x)$ , and

$$x_1 * \dots * x_n = x_1^{w_1} \cdot \dots \cdot x_n^{w_n};$$

- for Stevens' law,  $f(x) = x^\alpha$ , and

$$x_1 * \dots * x_n = (w_1 \cdot x_1^\alpha + \dots + w_n \cdot x_n^\alpha)^{1/\alpha}.$$

### 7.5.5 Properties of weighted geometric and generalized average

Similar to simple weighted average, weighted geometric and generalized average are idempotent, not commutative, continuous, not associative, not satisfying the averaging property, and computationally simple.

### 7.5.6 Weighted geometric and generalized average with interval weights

If we only know the intervals  $\mathbf{w}_i$  of possible values of the weights, then we can compute the weighted geometric and generalized average of the values  $x_1, \dots, x_n$  as follows:

- first, transform the values  $x_1, \dots, x_n$  into the new scale, thus computing  $X_1 = f(x_1), \dots, X_n = f(x_n)$ ;
- second, we use the above algorithm to compute the interval  $[X^-, X^+]$  corresponding to the weighted average of  $X_i$  with interval weights  $\mathbf{w}_i$ .
- finally, we transform the interval  $[X^-, X^+]$  into the original scale by applying the inverse function  $f^{-1}$ , i.e., by computing  $\mathbf{x} = \{f(X) \mid X \in [X^-, X^+]\}$ . Because each rescaling function  $f(x)$  is monotonic, to compute the range  $\mathbf{x} = f([X^-, X^+])$  of the function  $f(x)$  on the interval  $[X^-, X^+]$ ,

it is sufficient to compute the values  $f(X^-)$  and  $f(X^+)$  of the function  $f(x)$  on the endpoints of this interval. The resulting values  $f(X^-)$  and  $f(X^+)$  form the desired interval.

Of course, similarly to the case of weighted geometric and generalized average with numerical weights, we can describe this algorithm more directly:

- First, we sort the values  $x_i$  in increasing order. Without losing generality, we can assume that the values  $x_i$  are already sorted, i.e., that

$$x_1 \leq x_2 \leq \dots \leq x_n.$$

- Second, we find the value  $k$  from 1 to  $n$  for which

$$\begin{aligned} w_1^- + \dots + w_{k-1}^- + w_k^- + w_{k+1}^+ + \dots + w_n^+ &\leq 1 \leq \\ w_1^- + \dots + w_{k-1}^- + w_k^+ + w_{k+1}^+ + \dots + w_n^+ &. \end{aligned}$$

- Then, we compute

$$w_k = 1 - (w_1^- + \dots + w_{k-1}^- + w_{k+1}^+ + \dots + w_n^+)$$

and take either

$$x^+ = x_1^{w_1^-} \cdot \dots \cdot x_{k-1}^{w_{k-1}^-} \cdot x_k^{w_k} \cdot x_{k+1}^{w_{k+1}^+} \cdot \dots \cdot x_n^{w_n^+},$$

or

$$x^+ = (w_1^- \cdot x_1^\alpha + \dots + w_{k-1}^- \cdot x_{k-1}^\alpha + w_k \cdot x_k^\alpha + w_{k+1}^+ \cdot x_{k+1}^\alpha + \dots + w_n^+ \cdot x_n^\alpha)^{1/\alpha}.$$

- Next, we find the value  $l$  from 1 to  $n$  for which

$$\begin{aligned} w_1^+ + \dots + w_{l-1}^+ + w_l^- + w_{l+1}^- + \dots + w_n^- &\leq 1 \leq \\ w_1^+ + \dots + w_{l-1}^+ + w_l^+ + w_{l+1}^- + \dots + w_n^- &. \end{aligned}$$

- Then, we compute

$$w_l = 1 - (w_1^+ + \dots + w_{l-1}^+ + w_{l+1}^- + \dots + w_n^-)$$

and take

$$x^- = x_1^{w_1^+} \cdot \dots \cdot x_{l-1}^{w_{l-1}^+} \cdot x_l^{w_l} \cdot x_{l+1}^{w_{l+1}^-} \cdot \dots \cdot x_n^{w_n^-},$$

or

$$x^- = (w_1^+ \cdot x_1^\alpha + \dots + w_{l-1}^+ \cdot x_{l-1}^\alpha + w_l \cdot x_l^\alpha + w_{l+1}^- \cdot x_{l+1}^\alpha + \dots + w_n^- \cdot x_n^\alpha)^{1/\alpha}.$$

The desired interval is formed by the endpoints  $x^-$  and  $x^+$ . To be more precise:

- for geometric average and for  $\alpha > 0$ , the desired interval is  $[x^-, x^+]$ ;
- for  $\alpha < 0$ , the desired interval is  $[x^+, x^-]$ .

### 7.5.7 Weighted geometric and generalized average with interval weights: Properties

Weighted geometric and generalized average operations with interval weights are idempotent but not commutative, continuous but not associative. They don't satisfy the averaging property, and computationally simple.

## 7.6 Envelope

### 7.6.1 Case of variability: envelope

In the above text, we dealt with the averaging operations. These operations correspond to the case when the values  $x_1, \dots, x_n$  to be aggregated come from measuring and/or estimating the same (unknown) quantity  $x$  (i.e. to the case of incertitude), and our goal is to combine these estimates into a single (better) estimate for the same quantity  $x$ .

In some practical situations, however, the difference between the measurement results  $x_1, \dots, x_n$  is due not to incertitude, but to *variability*. All  $n$  values are correct, but they represent the values of the measured property in different zones or at different moments of time.

In this case, there is no single “actual” value of the measured quantity. The adequate description of the quantity is not by a single value, but by a *set* of possible values. Because the quantity usually changes continuously from place to place or from time to time, we can use the “mean value” property of continuous functions – that a continuous function attains all intermediate values – and conclude that the set of possible values of this quantity is an *interval*.

In this case, the goal of the aggregation is to provide an estimate for the “variability” interval of possible values of  $x$ . We know  $n$  measured values  $x_1, \dots, x_n$ . These values have been measured and thus, they are clearly possible. Therefore, the desired interval of possible values must contain all these values. The smallest interval that contains all these values is the interval

$$[x_{(1)}, x_{(n)}] = [\min(x_1, \dots, x_n), \max(x_1, \dots, x_n)]$$

between the smallest and the largest of the  $n$  values.

Values from outside this interval may also be possible. However, because we do not have any reason to add any specific value from outside the interval, the best we can do is to present the interval  $[x_{(1)}, x_{(n)}]$  as the aggregation result. The resulting aggregation operation

$$x_1 * \dots * x_n = [x_{(1)}, x_{(n)}] = [\min(x_1, \dots, x_n), \max(x_1, \dots, x_n)]$$

is called an *envelope* operation.

Envelope operation can be also used in the case of incertitude, when we want to produce not a single numerical estimate, but an aggregated estimate that combines both values  $x_1$  and  $x_2$ .

### 7.6.2 Relation with weighted average

We have already mentioned, when analyzing weighted average, that exactly the same formula appears if we consider weighted average with unknown weights, i.e., with weight intervals all equal to  $\mathbf{w}_i = [0, 1]$ .

### 7.6.3 Properties of the envelope operation

The envelope operation is clearly idempotent, commutative, and continuous.

From the commonsense viewpoint, it is also associative (i.e.,  $(x_1 * x_2) * x_3 = x_1 * (x_2 * x_3)$ ) and satisfies the averaging property, because the result of applying this operation does not depend on the order in which the values  $x_1, \dots, x_n$  are presented. However, because the result of aggregating two real numbers  $x_1 * x_2$  is an interval, we cannot even formalize these two properties without first extending this operation to the case when one of the estimates to be aggregated is an interval in order to formulate this property. We will do that in the section devoted to operations with intervals, and the resulting operation indeed turns out to be associative.

Finally, computing the envelope is clearly a computationally simple operation.

## 7.7 Mixing operation

### 7.7.1 Description

In accordance with our original analysis of different ways of describing uncertainty, in addition to the interval of possible values, we may want to describe the frequency of different values within this interval.

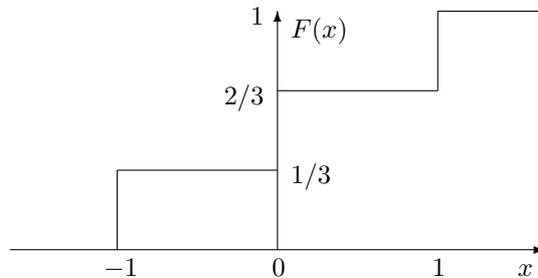
In particular, when we have  $n$  measurement results  $x_1, \dots, x_n$ , we can say that each of them occurs with a frequency  $1/n$ . In this result, the result of aggregating  $n$  values is a probability distribution in which each of  $n$  values occurs with exactly the same probability  $1/n$ . This aggregation operation is called *mixing*.

When describing different ways to represent uncertainty, we have concluded that for generic risk analysis problems, CDF is the best representation for a probability distribution. To represent the result of mixing as a CDF, we first sort the values  $x_1, \dots, x_n$ , and then the resulting ordered sequence  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  to produce the resulting CDF.

Mixing operation can be also used in the case of incertitude, when we want to produce not a single numerical estimate, but an aggregated estimate that combines both values  $x_1$  and  $x_2$ .

### 7.7.2 Mixing operation: example

Let  $x_1 = 1$ ,  $x_2 = -1$ , and  $x_3 = 0$ . In this case,  $x_{(1)} = -1$ ,  $x_{(2)} = 0$ ,  $x_{(3)} = 1$ , and the resulting CDF looks as follows:



### 7.7.3 Properties of the mixing operation

The mixing operation is clearly idempotent, commutative, and continuous.

The result of applying this operation does not depend on the order in which the values  $x_1, \dots, x_n$  are presented. However, we cannot yet formalize and check associativity and the averaging property, because the result  $x_1 * x_2$  of mixing two real numbers  $x_1$  and  $x_2$  is a probability distribution, and the operation has been so far defined only for real numbers.

Finally, mixing is clearly a computationally simple operation.

## 7.8 Section's conclusions

When the values  $x_1, \dots, x_n$  to be aggregated represent incertitude, we can either use more traditional averaging techniques which “erase” the incertitude, or we can use enveloping techniques which represent and propagate the incertitude. When the values  $x_1, \dots, x_n$  to be aggregated represent variability, enveloping techniques are the only choice.

Reasonable averaging techniques include the arithmetic average, the generalized averages, and the weighted versions of these operations. Reasonable enveloping techniques lead either to an envelope  $[\min(x_i), \max(x_i)]$ , or to a probability distribution in which each value  $x_i$  occurs with the same probability  $1/n$ .

## 8 Aggregation of Uncertainty in Risk Analysis: Case of Intervals

### 8.1 General classification

#### 8.1.1 Incertitude vs. variability

How can we aggregate  $n$  intervals  $\mathbf{x}_1, \dots, \mathbf{x}_n$ ? As in aggregation of real numbers, the choice of the aggregation operation depends on whether we have incertitude (when there is only one actual value  $x$ ) or we have variability (when we have several possible values).

For real numbers, this was the main issue. As soon as we agreed on whether real numbers represent incertitude or variability, we get a well-formulated problem for which we can use several reasonable methods. For intervals, the situation is somewhat more complicated because even if we select incertitude or variability, we still have several choices depending on what exactly the intervals represent.

In the following text, we first describe all the possibilities, and then enumerate aggregation techniques corresponding to these possibilities.

#### 8.1.2 Incertitude: two options

Let us first consider the case of incertitude, when there is a single (unknown) value of the desired quantity  $x$ , and each interval  $\mathbf{x}_i$  is a result of measuring or estimating this unknown quantity. What operation we should choose to aggregate these intervals depends on the relation between the intervals  $\mathbf{x}_i$  and the unknown value  $x$ .

In order to describe different relationships, let us recall that a measuring instrument does not usually return an interval, it returns a *real number*  $\tilde{x}$ . An interval is usually obtained if we take into consideration that the measurement result is never absolutely accurate, as a result of which the actual measured value  $x$  usually differs from the measured result  $\tilde{x}$ . If we know the upper bound  $\Delta$  on the measurement result, then we can conclude that the actual value  $x$  belongs to the interval  $[\tilde{x} - \Delta, \tilde{x} + \Delta]$ . This interval is then returned as the result of the measurement.

Within this general scheme, there are two possible choices of  $\Delta$ . In view of the above description, a natural idea is to choose the value  $\Delta$  that bounds the *total* error. In this case, each of  $n$  intervals  $\mathbf{x}_i$  is guaranteed to contain the actual (unknown) value  $x$ .

However, the selection of total error bound for  $\Delta$  is not the only possibility. There is another very natural possibility for selecting  $\Delta$ . Indeed, engineers and scientists often do not use intervals to describe their uncertainty. They get the numerical measurement results, and then they apply statistical techniques (e.g., computing the arithmetic average) to process the results  $\tilde{x}_1, \dots, \tilde{x}_n$

of these measurements. In these statistical methods, it is usually assumed that the measurement error  $\tilde{x}_i - x$  is purely random. In reality, in addition to random measurement error, there are additional (presumably small) error components such as discretization error, un-resolved systematic error, etc. These components are not distributed according to standard distributions, they are usually not independent, and so treating these components by normal statistical techniques can lead to inaccuracy of data processing results.

For example, if all measurement errors are independent, then, when we repeat the measurement sufficiently many ( $N$ ) times, the standard deviation of the average drops as  $1/\sqrt{N}$ , so we would think that with  $N$  increasing, this error will eventually tend to 0. However, if all the measurement results have the same remaining systematic error component, then no amount of averaging will decrease this error.

So, to avoid this problem and to make the the results of statistical methods more reliable, it makes sense to replace the original values  $\tilde{x}_i$  by intervals  $[\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]$ , where  $\Delta_i$  is the upper bound on the systematic error component. In this representation, the actual value  $x$  of the measured quantity does not, in general, belong to this interval  $\mathbf{x}_i$ , but the difference between  $x$  and one of the values from  $\mathbf{x}_i$  is already a purely random error and therefore, standard statistical methods can be applied.

Summarizing: in case of *incertitude*, we have two options:

- The first option is that the intervals  $\mathbf{x}_i$  contain the bounds for the total error. In this case, the actual value  $x$  of the measured quantity belongs to each of these intervals.
- The second option is that the intervals  $\mathbf{x}_i$  contain the bounds only for some error components. In this case, the actual value  $x$  does not, in general, belong to any of these intervals  $\mathbf{x}_i$ .

### 8.1.3 Aggregation for the second option: averaging vs. enveloping

In the second option, when each of the intervals  $\mathbf{x}_1, \dots, \mathbf{x}_n$  to be aggregated is an estimate for the actual (unknown) value  $x$  and does not necessarily contain this value  $x$ , how can we aggregate these intervals? We already know, from Section 7, how to aggregate numerical estimates. Each interval  $\mathbf{x}_i$  represents the set of possible numerical estimates  $x_i$  for  $x$ . Therefore, to describe different possible operations  $\mathbf{x} * \mathbf{y} = [x^-, x^+] * [y^-, y^+]$  of aggregating these intervals, it is reasonable, for each operation  $*$  that aggregates real numbers, to take, as  $\mathbf{x} * \mathbf{y}$ , the set of possible values of  $x * y$  when  $x \in \mathbf{x}$  and  $y \in \mathbf{y}$ .

Depending on whether we start with an averaging or with an enveloping operation for aggregating real numbers, we end up with, correspondingly, an averaging or an enveloping operation for aggregating intervals.

#### 8.1.4 Variability: two options

Similarly, in the variability case, when, say, an expert presents an interval, it may mean two different things:

- It may mean that we have observed sufficiently many values within the presented interval, so we claim that *all* values within this interval are possible; some values from outside this interval may also be possible, but we are not sure about them.
- It may also mean that we present an interval that *contains* all possible values of  $x$  (e.g., its range), but we are not sure that all values from this interval are actually possible. For example, we may say that the variable should be between 0 and 1, but in reality, it may only take values from 0 to 0.3.

These two cases can be easily described in terms of the relation between the presented interval  $\mathbf{x}$  and the actual (unknown) interval of possible values  $\mathbf{X}$ . In the first case,  $\mathbf{x} \subseteq \mathbf{X}$ , and in the second case,  $\mathbf{X} \subseteq \mathbf{x}$ .

## 8.2 Intersection

### 8.2.1 Aggregation operation corresponding to the case when intervals contain total error: intersection

Now that we have enumerated several possible meanings of interval data, let us start describing the corresponding aggregation techniques. We will start with the first option of both uncertainty and variability, when each of the intervals  $\mathbf{x}_1, \dots, \mathbf{x}_n$  to be aggregated contain all possible values of the desired quantity. So, we have  $n$  pieces of information, according to which  $x$  belongs to the interval  $\mathbf{x}_1$ , to the interval  $\mathbf{x}_2$ , etc. This information can be easily aggregated;  $x$  belongs to several intervals if and only if it belongs to their intersection. Thus, the corresponding aggregation operation is *intersection*  $\mathbf{x} * \mathbf{y} = \mathbf{x} \cap \mathbf{y}$ .

This operation is easy to compute: if  $\mathbf{x} = [x^-, x^+]$  and  $\mathbf{y} = [y^-, y^+]$ , then the intersection is equal to  $\mathbf{x} \cap \mathbf{y} = [\max(x^-, y^-), \min(x^+, y^+)]$ .

For example, if  $\mathbf{x} = [1, 3]$  and  $\mathbf{y} = [2, 4]$ , we get  $\mathbf{x} \cap \mathbf{y} = [2, 3]$ .

### 8.2.2 Properties of intersection

Intersection is idempotent, commutative, associative, satisfies the averaging property, and is easy to compute.

From our list of properties, the only property that may not be fully satisfied for intersection is continuity. We have discussed this very example when we introduced and analyzed the continuity property:

- If we are 100% sure that all the intervals are correct, then, because  $x$  belongs to all of them, their intersection is non-empty. In this case, the intersection operation is continuous.
- On the other hand, realistically, mistakes happen. What if one of the intervals is erroneous: e.g., the researcher who did this measurement did not take into account one of the error components (a very realistic situation, by the way). In this case, the intersection may turn out to be empty, and in this case, the aggregation operation stops being continuous.

This analysis of properties of intersection leads to a natural practical question: what should we do if the intersection is empty? This issue commonly arises (Kosko, 1986); for example, it arises in the analysis of expert estimates related to computer simulations of atomic bomb detonation (Ross et al., 2000). In general, this situation should precipitate a more careful review of the data. Something is clearly wrong. Perhaps some of the estimates are answering different questions or are the result of some fundamental misunderstanding. This inconsistency provides an analyst with an opportunity to reconsider the problem. Given that this sort of inconsistency can arise in situations where no obvious errors have been made, we would like to proceed even though the intersection is empty.

### 8.2.3 What if intersection is empty?

If the intersection of all  $n$  intervals is empty, this means that at least one of the intervals is erroneous and should be dismissed. Which interval(s) should we dismiss?

In principle, if, say, three intervals almost coincide, and the fourth does not have any point in common with any of the first three, there are two options:

- we can dismiss the fourth interval, or
- we can dismiss the first three intervals.

If the intervals are the only information that we have, then it might be reasonable to prefer dismissing one interval.

In general, we would like to dismiss the smallest possible number of intervals. If we can achieve consistency by dismissing only one interval, then, because we do not know which of the intervals is erroneous, we could do the following:

- try all intervals whose dismissal leads to consistency,
- compute the intersection for each resulting set, and
- then (because we are not sure which is correct) take the union of the resulting intersections.

If we cannot achieve consistency by dismissing only one interval, but we can achieve consistency by dismissing two intervals, then we must do the following:

- try all pairs of intervals whose dismissal leads to consistency,
- compute the intersection for each set resulting from the dismissal of each such pair, and
- then take the union of the resulting intersections.

The resulting algorithms are pretty straightforward, but when the number of intervals grows, these algorithms require too much time. Indeed, if we have, say, 1,000 intervals, and 5% of them are erroneous (a reasonable rate in real-life measurements), this means that we have to dismiss 50 intervals out of 1,000. In this case, to apply the above straightforward algorithm, we need to try all possible subsets of 50 intervals from 1,000. The corresponding number of combinations is larger than the number of particles in the universe, and cannot be tested. We therefore need a better indirect algorithm. Such an algorithm has been used in (Ross et al., 2000). It consists of the following steps.

1. First, we sort all  $2n$  endpoints

$$x_1^-, \dots, x_n^-, x_1^+, \dots, x_n^+$$

into an ordered sequence of real numbers

$$x_{(1)} \leq x_{(k)} \leq \dots \leq x_{(2n)}.$$

2. Then, for each  $k = 1, \dots, 2n - 1$ , we consider the interval  $[x_{(k)}, x_{(k+1)}]$ . For this interval, we check, for every  $i = 1, \dots, n$ , whether this interval belongs to  $\mathbf{x}_i$  or not. Thus, we form the set  $S_{(k)}$  of all indices  $i$  for which  $[x_{(k)}, x_{(k+1)}] \subseteq \mathbf{x}_i$ , and compute the total number  $|S_{(k)}|$  of elements in this set.
3. As a result of the second stage, we get  $2n - 1$  different values of  $|S_{(k)}|$ . To find the largest value (or values), we compare them with each other: we start with the value corresponding to the leftmost interval  $[x_{(1)}, x_{(2)}]$ , mark it as the largest-so-far, and then go over other values, replacing the largest-so-far with the next one if the next value of  $|S_{(k)}|$  is indeed larger.
4. Then, we take all the intervals for which  $|S_{(k)}|$  attains the largest possible value, and take the interval envelope of the corresponding intervals  $[x_{(k)}, x_{(k+1)}]$ . To be more precise, it is sufficient to take the first such interval  $[x_{(k)}, x_{(k+1)}]$  and the last such interval  $[x_{(l)}, x_{(l+1)}]$ , then the resulting aggregation is  $[x_{(k)}, x_{(l+1)}]$ .

If the intersection of  $n$  input intervals is not empty, then this algorithm returns the intersection. Let us give a simple example how this algorithm works when the intersection is empty. Let

$$\mathbf{x}_1 = [0.0, 1.0], \quad \mathbf{x}_2 = [0.5, 1.5], \quad \mathbf{x}_3 = [0.7, 1.2], \quad \mathbf{x}_4 = [1.3, 1.6].$$

These four intervals have an empty intersection. According to our algorithm, first we sort all 8 values  $x_i^\pm$  into a sequence

$$x_{(1)} = 0.0 \leq x_{(2)} = 0.5 \leq x_{(3)} = 0.7 \leq x_{(4)} = 1.0 \leq \\ x_{(5)} = 1.2 \leq x_{(6)} = 1.3 \leq x_{(7)} = 1.5 \leq x_{(8)} = 1.6.$$

Then, for each of seven intervals  $[x_{(k)}, x_{(k+1)}]$ , we compute the corresponding value  $|S_{(k)}|$ :

$$|S_{(1)}| = 1, \quad |S_{(2)}| = 2, \quad |S_{(3)}| = 3, \quad |S_{(4)}| = 2, \quad |S_{(5)}| = 1, \quad |S_{(6)}| = 2, \quad |S_{(7)}| = 1,$$

and select the set  $S_{(3)}$  for which the corresponding real number is the largest. The result of aggregation is therefore  $[0.7, 1.0]$ .

In general, for two intervals  $\mathbf{x}_1 = [x_1^-, x_1^+]$  and  $\mathbf{x}_2 = [x_2^-, x_2^+]$  with an empty intersection, this *dismissive intersection* leads to an interval envelope of their union, i.e., to the interval  $[\min(x_1^-, x_2^-), \max(x_1^+, x_2^+)]$ . Let us analyze the properties of the dismissive intersection operation. This operation is idempotent and commutative. It is not continuous because, e.g., when a real number  $\alpha$  goes from positive values to 0, the dismissive intersection between the intervals  $[-1, 0]$  and  $[\alpha, 1 + \alpha]$  changes from an interval  $[-1, 1 + \alpha]$  whose width is close to 2 to a degenerate interval  $[0, 0]$  of width 0. The dismissive intersection is not associative; for example,

$$[0, 1] * ([1, 2] * [2, 3]) = [0, 1] * [2, 2] = [0, 2],$$

although

$$([0, 1] * [1, 2]) * [2, 3] = [1, 1] * [2, 3] = [1, 3] \neq [0, 2].$$

This operation does not satisfy the averaging property; for example,

$$([0, 1] * [0, 4]) * ([1, 2] * [2, 3]) = [0, 1] * [2, 2] = [0, 2],$$

although

$$([0, 1] * [1, 2]) * ([0, 4] * [2, 3]) = [1, 1] * [2, 3] = [1, 3] \neq [0, 2].$$

Dismissive intersection is easy to compute.

## 8.3 Averaging

### 8.3.1 General description

In the previous subsection, we described aggregation operations corresponding to the case when each of the intervals to be aggregated is known to contain all possible values of  $x$ . Let us now consider the second option, when intervals to be aggregated do not necessarily contain all possible values of the quantity  $x$ .

In Section 8.1, we have shown that the corresponding aggregation operations can be obtained by using the corresponding operations for aggregating real numbers. Namely, for each operation  $*$  that aggregates real numbers, we take, as  $\mathbf{x} * \mathbf{y}$ , the set of possible values of  $x * y$  when  $x \in \mathbf{x}$  and  $y \in \mathbf{y}$ .

Let us start with averaging aggregation operations. Because most averaging operations are monotonic, from  $x^- \leq x \leq x^+$  and  $y^- \leq y \leq y^+$ , we conclude that  $x^- * y^- \leq x * y \leq x^+ * y^+$ . Therefore, all possible values  $x * y$  belong to the interval  $[x^- * y^-, x^+ * y^+]$  which is, therefore, the desired aggregation:

$$[x^-, x^+] * [y^-, y^+] = [x^- * y^-, x^+ * y^+].$$

Let us consider all real-number averaging aggregation operations one by one and see what interval operations they lead to.

### 8.3.2 Arithmetic average

For arithmetic average, the above formula leads to

$$[x_1^-, x_1^+] * [x_2^-, x_2^+] = \left[ \frac{x_1^- + x_2^-}{2}, \frac{x_1^+ + x_2^+}{2} \right];$$

$$[x_1^-, x_1^+] * \dots * [x_n^-, x_n^+] = \left[ \frac{x_1^- + \dots + x_n^-}{n}, \frac{x_1^+ + \dots + x_n^+}{n} \right].$$

Similar to the case of real numbers, this operation is idempotent, commutative, continuous, not associative, satisfies the averaging property, and is easy to compute.

### 8.3.3 Weighted average

For weighted average, the above formula leads to

$$[x_1^-, x_1^+] * \dots * [x_n^-, x_n^+] = [w_1 \cdot x_1^- + \dots + w_n \cdot x_n^-, w_1 \cdot x_1^+ + \dots + w_n \cdot x_n^+].$$

Similar to the case of real numbers, this operation is idempotent, non-commutative, continuous, not associative, satisfies the averaging property, and is easy to compute.

An interesting new aspect of weighted average with respect to interval data is that with intervals instead of real numbers, we now have new criteria for selecting weights depending on the widths of the corresponding intervals. Specifically, there are two possible criteria:

- If we are sure that each interval correctly describes the corresponding systematic error, then wider intervals correspond to worse measurements, with larger systematic error. In this case, it makes sense to assign smaller weights to these bad measurements, thus decreasing their impact on the aggregation result.

- What shall we do in the opposite situation, when we are not sure that each intervals correctly describes the corresponding systematic error? We have already mentioned, when discussing the possibility of an empty intersection, that a typical mistake of a measurer is to underestimate the measurement error. So, if all the intervals come from the sources of approximately the same measurement capability, but one of the corresponding intervals is wider than the others, this probably means that this wider interval describes the corresponding systematic error much more accurately than the narrower ones. Because we trust wider intervals more than narrower ones, we should assign larger weights to wider intervals and smaller weights to narrower intervals.

### 8.3.4 Weighted average with interval weights

In accordance with the above expression, the endpoints  $x^-$  and  $x^+$  of the aggregated interval  $\mathbf{x}$  can be computed as follows:

- First, we sort the values  $x_i^+$  in increasing order:

$$x_{(1)}^+ \leq x_{(2)}^+ \leq \dots \leq x_{(n)}^+.$$

Let  $w_{+(i)}^-$  and  $w_{+(i)}^+$  be the lower and upper endpoints of the weight intervals that corresponds to  $x_{(i)}^+$ .

- Second, we find the value  $k$  from 1 to  $n$  for which

$$\begin{aligned} w_{+(1)}^- + \dots + w_{+(k-1)}^- + w_{+(k)}^- + w_{+(k+1)}^+ + \dots + w_{+(n)}^+ &\leq 1 \leq \\ w_{+(1)}^- + \dots + w_{+(k-1)}^- + w_{+(k)}^+ + w_{+(k+1)}^+ + \dots + w_{+(n)}^+ &\cdot \end{aligned}$$

- Then, we compute

$$w_{+k} = 1 - (w_{+(1)}^- + \dots + w_{+(k-1)}^- + w_{+(k+1)}^+ + \dots + w_{+(n)}^+)$$

and take

$$\begin{aligned} x^+ = w_{+(1)}^- \cdot x_{(1)}^+ + \dots + w_{+(k-1)}^- \cdot x_{(k-1)}^+ + w_{+k} \cdot x_{(k)}^+ + \\ w_{+(k+1)}^+ \cdot x_{(k+1)}^+ + \dots + w_{+(n)}^+ \cdot x_{(n)}^+. \end{aligned}$$

- Next, we sort the values  $x_i^-$  in increasing order:

$$x_{(1)}^- \leq x_{(2)}^- \leq \dots \leq x_{(n)}^-.$$

Let  $w_{-(i)}^-$  and  $w_{-(i)}^+$  be the lower and upper endpoints of the weight intervals that corresponds to  $x_{(i)}^-$ .

- Then, we find the value  $l$  from 1 to  $n$  for which

$$w_{-(1)}^+ + \dots + w_{-(l-1)}^+ + w_{-(l)}^- + w_{-(l+1)}^- + \dots + w_{-(n)}^- \leq 1 \leq w_{-(1)}^+ + \dots + w_{-(l-1)}^+ + w_{-(l)}^+ + w_{-(l+1)}^- + \dots + w_{-(n)}^-.$$

- Finally, we compute

$$w_{-l} = 1 - (w_{-(1)}^+ + \dots + w_{-(l-1)}^+ + w_{-(l+1)}^- + \dots + w_{-(n)}^-)$$

and take

$$x^- = w_{-(1)}^+ \cdot x_{(1)}^- + \dots + w_{-(l-1)}^+ \cdot x_{(l-1)}^- + w_{-l} \cdot x_{(l)}^- + w_{-(l+1)}^- \cdot x_{(l+1)}^- + \dots + w_{-(n)}^- \cdot x_{(n)}^-.$$

In particular, when  $\mathbf{w}_i = [0, 1]$  for all  $i$ , the aggregation result is the interval

$$[\min(x_1^-, \dots, x_n^-), \max(x_1^+, \dots, x_n^+)]$$

– the smallest interval which contains all the intervals  $\mathbf{x}_i$ . This interval is the *envelope* of the intervals  $\mathbf{x}_1, \dots, \mathbf{x}_n$ .

### 8.3.5 Geometric average and generalized average

For geometric average:

$$x^- = \sqrt[n]{x_1^- \cdot \dots \cdot x_n^-};$$

$$x^+ = \sqrt[n]{x_1^+ \cdot \dots \cdot x_n^+}.$$

For generalized average:

$$x^- = \left( \frac{(x_1^-)^\alpha + \dots + (x_n^-)^\alpha}{n} \right)^{1/\alpha};$$

$$x^+ = \left( \frac{(x_1^+)^\alpha + \dots + (x_n^+)^\alpha}{n} \right)^{1/\alpha}.$$

These operations are idempotent, commutative, continuous, non-associative, satisfy averaging properties, and are computationally simple.

### 8.3.6 Weighted geometric average and generalized average

For geometric average:

$$x^- = (x_1^-)^{w_1} \cdot \dots \cdot (x_n^-)^{w_n};$$

$$x^+ = (x_1^+)^{w_1} \cdot \dots \cdot (x_n^+)^{w_n}.$$

For generalized average:

$$x^- = (w_1 \cdot (x_1^-)^\alpha + \dots + w_n \cdot (x_n^-)^\alpha)^{1/\alpha};$$

$$x^+ = (w_1 \cdot (x_1^+)^\alpha + \dots + w_n \cdot (x_n^+)^\alpha)^{1/\alpha}.$$

Similar formulas can be written for geometric and generalized averages with interval weights.

## 8.4 Enveloping

In the previous subsection, we considered interval aggregation operations that result from averaging of real numbers. Let us now describe what interval aggregation operations emerge if we start with enveloping operations.

### 8.4.1 Envelope

In the envelope aggregation of  $n$  real numbers  $x_1, \dots, x_n$ , the result is the interval  $[\min(x_1, \dots, x_n), \max(x_1, \dots, x_n)]$ . So, if we want to use this operation to combine  $n$  intervals  $\mathbf{x}_1 = [x_1^-, x_1^+], \dots, \mathbf{x}_n = [x_n^-, x_n^+]$ , we must take all possible values  $x_1 \in \mathbf{x}_1, \dots, x_n \in \mathbf{x}_n$ , aggregate these  $n$  values, and then combine the resulting aggregations. One can easily show that the union of all the corresponding aggregation intervals is the smallest interval that contains the union of the  $n$  intervals to be aggregated, i.e., the interval

$$\mathbf{x}_1 * \dots * \mathbf{x}_n \stackrel{\text{def}}{=} [\min(x_1^-, \dots, x_n^-), \max(x_1^+, \dots, x_n^+)].$$

This formula makes perfect sense; indeed, e.g., in the case of variability, all the values from each interval  $\mathbf{x}_i$  are known to be possible. Thus, a natural aggregation operation  $\mathbf{x}_1 * \dots * \mathbf{x}_n$  is the envelope, the smallest interval that contains the union of these  $n$  intervals.

Similarly to the case of aggregating point estimates, the envelope is equal to the weighted average with interval weights  $\mathbf{w}_i = [0, 1]$ .

The envelope operation is idempotent, commutative, continuous, associative, satisfies the averaging property, and is computationally simple.

### 8.4.2 Mixing

Another possibility is to consider the *mixing* operation. As a result, we combine each interval with probability  $1/n$ . Mixing operation is exactly what we used when we transformed interval measurement results into a p-bound.

## 8.5 Section's conclusions

When each of the intervals  $\mathbf{x}_1, \dots, \mathbf{x}_n$  to be aggregated contain all possible values of a quantity, then a natural aggregation operation is the intersection. In other cases, we can use either averaging operations which “erase” the uncertainty or enveloping operations which preserve and propagate the uncertainty. For enveloping, aggregation leads to either the envelope  $[\min(x_i^-), \max(x_i^+)]$ , or to a p-bound in which each interval  $\mathbf{x}_i$  occurs with the same probability  $1/n$ . When the intervals  $\mathbf{x}_1, \dots, \mathbf{x}_n$  to be aggregated represent variability, enveloping techniques are the only choice.

## 9 Aggregation of Uncertainty in Risk Analysis: Case of Probability Distributions

### 9.1 General classification

#### 9.1.1 Incertitude vs. variability, averaging vs. enveloping

Let us now consider probability distributions which are, of course, a special case of p-bounds. If we have  $n$  CDFs  $F_1(x), \dots, F_n(x)$ , how can we aggregate them? Similarly to aggregation of real numbers and intervals, the choice of the aggregation operation depends on:

- whether we have incertitude – i.e., whether there is only one actual CDF – or
- whether we have variability – when several different probability distributions are possible in different situations, and thus, we need a p-bound to describe the resulting uncertainty.

As in the case of intervals, even after we choose incertitude or variability, we still have several choices depending on where exactly the CDFs to be aggregated come from, and whether we want to use averaging or enveloping. In the following text, we first describe all the possibilities, and then enumerate aggregation techniques corresponding to these possibilities.

#### 9.1.2 Incertitude: two subcases

Let us first consider the case of incertitude, when there is a single (unknown) CDF  $F(x)$ , and each CDF  $F_i(x)$  to be aggregated is a result of measuring or estimating this unknown CDF. What operation we should choose to aggregate these CDFs depends on what causes the difference among the different CDFs, or, equivalently, what causes the known CDFs  $F_i(x)$  to be different from the actual probability distribution  $F(x)$ . We have already described, in Section 3, where different p-bounds (and CDFs are examples of p-bounds) come from. Based on this analysis, we can see that depending on where the CDFs come from, there are two main sources of the differences between  $F_i(x)$  and  $F(x)$ .

- One possibility is that the CDFs come from *expert estimation*, when an expert estimates the values of  $F(x)$  for each  $x$ . In this case, the main source of the difference between  $F_i(x)$  and  $F(x)$  is the estimation error, due to which, for every  $x$ , the expert's estimate  $F_i(x)$  of the CDF is different from the estimated (unknown) value  $F(x)$  of the CDF. In this case, each CDF  $F_i(x)$  is based on the data points in which the value  $x$  is known exactly, and the probability  $F(x)$  is estimated approximately. For short, we can say that in this case, we have an uncertainty in probability, i.e., *p-uncertainty*.

- Another possibility is that CDFs come from *measurement*. In this case, a CDF is simply a mixture distribution representing measurement results. For such measured CDFs, the main source of the difference between  $F_i(x)$  and  $F(x)$  is the measurement error, due to which the measurement results are different from the actual values of the measured quantity. In this case, each CDF  $F_i(x)$  is based on the data points in which the probabilities  $1/n$  of each measured value are known exactly, but the measured value  $x$  is estimated approximately. For short, we can say that in this case, we have an uncertainty in  $x$ , i.e.,  $x$ -uncertainty.

Please note that we listed these possibilities in the order which is opposite to the order in which we presented them in describing where CDFs come from. This change of order is not accidental. In Section 3, when we described possible sources of p-bounds, we started with the situation of complete information, then described different possible situations in the order of increasing incompleteness.

Expert estimates are usually less accurate than measurements, hence we start with expert estimates and consider CDFs coming from measurements later. In the following text, when analyzing and describing different aggregation operations, we will follow this same order.

### 9.1.3 Variability: two subcases

Similarly to two subcases within the incertitude case, there are also two subcases within the variability case: the case when the CDFs come from expert estimates, and the case when CDFs come from measurements.

## 9.2 Averaging operations for the case when CDFs come from expert estimates: general overview

As in the cases of real numbers and intervals, for incertitude, we can use either averaging or enveloping techniques. Let us first consider averaging techniques for the case when CDFs  $F_i(x)$  come from expert estimates. In this case, for every  $x$ , we have  $n$  estimates  $F_i(x)$  for the desired probability  $F(x)$ . For every  $x$ , we need to aggregate  $n$  real numbers, so we can use all the operations that we described for aggregating (averaging) real numbers. To distinguish between averaging the measured values  $x_i$  and averaging probabilities  $p_i$ , we will call these operations, correspondingly, *x-averaging* and *p-averaging*. In these terms, aggregation operations for the case when:

- CDFs represent incertitude and
- CDFs come from expert estimates,

correspond to *p-averaging*. Let us describe how different averaging operations (described above for the general problem of aggregating real numbers) work for *p-averaging*.

### 9.3 Arithmetic $p$ -average

#### 9.3.1 General description and properties

The simplest averaging operation is the arithmetic average, when

$$F_1(x) * F_2(x) = \frac{F_1(x) + F_2(x)}{2},$$

and

$$F_1(x) * \dots * F_n(x) = \frac{F_1(x) + \dots + F_n(x)}{n}.$$

This formula is also called *linear opinion pool* (Clemen & Winkler, 1999).

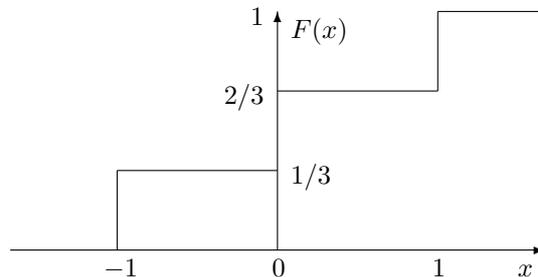
As in the cases of real numbers and intervals, this operation is idempotent, commutative, continuous, not associative, satisfies the averaging property, and is easy to compute.

#### 9.3.2 Case of real numbers

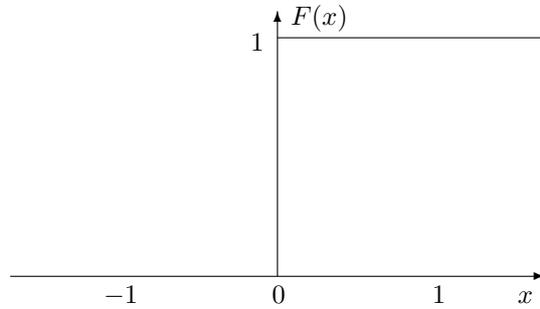
Because real numbers are a particular case of CDFs, we will try, for this and for other aggregation operations, to analyze what these operations lead to when applied to real numbers. Specifically, when each CDF represents an exact real number  $x_i$ , i.e., if

$$F_i(x) = \begin{cases} 0 & \text{if } x \leq x_i, \\ 1 & \text{if } x > x_i, \end{cases}$$

then the  $p$ -average corresponds to what we called, in Section 7, the “mixing” of these real numbers, when we have each real number  $x_i$  with the same probability  $1/n$ . (This is also the same operation used to generate a histogram  $F_n(x)$  from measurement results.) Thus, every example of mixing or histogram generation can also serve as an example of arithmetic  $p$ -averaging. It is important to emphasize that the arithmetic  $p$ -average of  $n$  real numbers is different from the arithmetic ( $x$ -)average of these numbers. The arithmetic  $x$ -average is a single number (degenerate distribution), while the arithmetic  $p$ -average is a non-degenerate probability distribution. For example, for  $x_1 = 1$ ,  $x_2 = -1$ , and  $x_3 = 0$ , the arithmetic  $p$ -average looks as follows



and the corresponding arithmetic  $x$ -average is the number  $x = 0$ , which corresponds to the following CDF

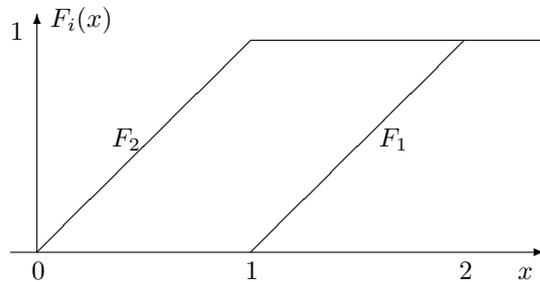


### 9.3.3 Example

In addition to these degenerate examples, let us present a simple numerical example of arithmetic  $p$ -averaging in which the distributions to be aggregated are not real numbers. Let

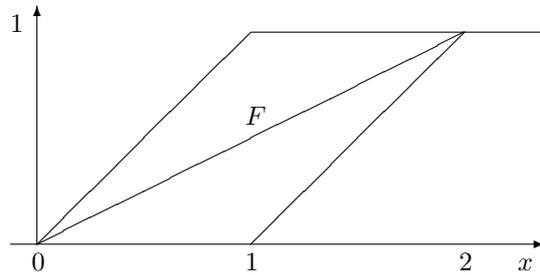
$$F_1(x) = \begin{cases} 0 & \text{if } x \leq 1, \\ x - 1 & \text{if } 1 \leq x \leq 2, \\ 1 & \text{if } x > 2 \end{cases}$$

$$F_2(x) = \begin{cases} x & \text{if } x \leq 1, \\ 1 & \text{if } x > 1 \end{cases}$$

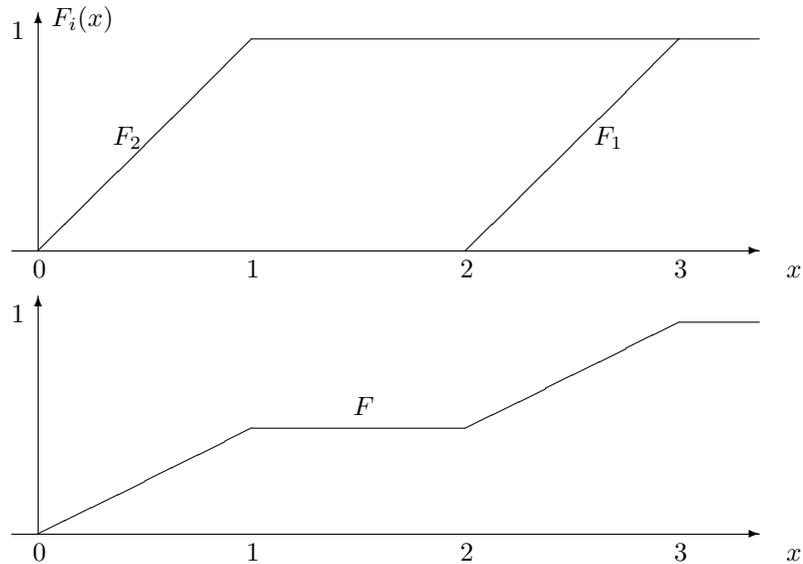


Then, the result  $F = F_1 * F_2$  of  $p$ -averaging is the following CDF:

$$F(x) = \begin{cases} x/2 & \text{if } x \leq 2, \\ 1 & \text{if } x > 2 \end{cases}$$



Another example is this



#### 9.3.4 New motivations

It is worth mentioning that for  $p$ -averaging, the use of arithmetic average is very natural. Let us explain what we mean. In the general situation of aggregation operations with real numbers (general in the sense that these real numbers are not necessarily probabilities, may be measured values), we presented, in Section 7, several motivations for using arithmetic average.

For the specific case of  $p$ -averaging, when we are averaging probabilities, arithmetic average has a very natural probabilistic meaning. To provide this meaning, let us approach the aggregation problem from a slightly different angle. Two experts provided us with two different estimates  $F_1(x)$  and  $F_2(x)$  for the same probability  $F(x)$ .

- If we trust the first expert 100%, then we simply take his/her estimate  $F_1(x)$  as the estimate for the same probability, and completely dismiss the estimate  $F_2(x)$  provided by the second expert.

- If we trust the second expert 100%, then we simply take his/her estimate  $F_2(x)$  as the estimate for the same probability, and completely dismiss the estimate  $F_1(x)$  provided by the first expert.

If we do not any reason to prefer the first or the second expert, then it makes sense to trust either the result of the first expert or the result of the second expert with the same probability 0.5. If we do that, what will be the resulting probability  $F(x)$  that  $X \leq x$ ? This probability can be computed by using the formula of full probability:

$$P(X \leq x) = P(1) \cdot P(X \leq x | 1) + P(2) \cdot P(X \leq x | 2),$$

where:

- $P(1) = 0.5$  is the probability that the first expert is correct;
- $P(2) = 0.5$  is the probability that the second expert is correct;
- $P(X \leq x | 1) = F_1(x)$  is the conditional probability that  $X \leq x$  under the condition that the first expert is correct;
- $P(X \leq x | 2) = F_2(x)$  is the conditional probability that  $X \leq x$  under the condition that the second expert is correct.

As a result, we get the formula  $F(x) = 0.5 \cdot F_1(x) + 0.5 \cdot F_2(x)$ , which is exactly the arithmetic p-average.

Similarly, if we have  $n$  expert estimates, then the corresponding full probability formula

$$P(X \leq x) = P(1) \cdot P(X \leq x | 1) + \dots + P(n) \cdot P(X \leq x | n),$$

with  $P(i) = 1/n$ , also leads to the arithmetic  $p$ -average of  $n$  expert estimates  $F_1(x), \dots, F_n(x)$ .

### 9.3.5 Relation to averaging densities

In Section 2, we argued that from the viewpoint of risk applications, CDFs are the best characteristics to use. In view of this conclusion, we used CDFs (and bounds for CDFs) in our survey.

However, in statistical applications, often, probability densities  $\rho(x)$  are often used instead of CDFs  $F(x)$ . In this case, it is natural to aggregate densities  $\rho_1(x), \dots, \rho_n(x)$  corresponding to the CDFs  $F_i(x)$  – and not CDFs themselves. Several researchers have proposed to use the density-aggregation operations in risk analysis as well; see, e.g., (Clemen & Winkler, 1999). A natural question is: how are these operations related to CDF-aggregation?

In particular, for the arithmetic average, the corresponding density-aggregation operations are

$$\rho(x) = \frac{\rho_1(x) + \rho_2(x)}{2}$$

and

$$\rho(x) = \frac{\rho_1(x) + \dots + \rho_n(x)}{n}.$$

One can show that because the aggregation operation is linear, the arithmetic average of densities leads to exactly the same resulting CDF as arithmetic  $p$ -average of CDFs.

## 9.4 Weighted arithmetic $p$ -average

### 9.4.1 General description and properties

If we have reasons to believe that some experts are more trustworthy than others, then (as we have mentioned when discussing aggregation of real numbers) we should use *weighted* arithmetic  $p$ -average instead of the arithmetic  $p$ -average:

$$F_1(x) * \dots * F_n(x) = w_1 \cdot F_1(x) + \dots + w_n \cdot F_n(x).$$

This formula is also sometimes called a *linear opinion pool* (Clemen & Winkler, 1999). The same methods as we outlined when discussing aggregation of real numbers can be used to determine these weights.

Similar to the cases of real numbers and intervals, this operation is idempotent, non-commutative, continuous, not associative, does not satisfy the averaging property, but is easy to compute.

### 9.4.2 Case of real numbers

When CDFs are actually real numbers  $x_i$ , the aggregation result coincides with the result of translating the Dempster-Shafer knowledge base, in which we have  $x_i$  with probability  $w_i$ , into a  $p$ -bound (which, for this particular knowledge base, turns out to be a CDF). Here also, the arithmetic  $p$ -average of  $n$  real numbers is different from their arithmetic  $x$ -average.

### 9.4.3 New motivation

The new motivation for the arithmetic  $p$ -average can be naturally extended to the weighted average. Indeed, when we combine the opinions of  $n$  experts, then the formula of full probability takes the following form:

$$P(X \leq x) = P(1) \cdot P(X \leq x | 1) + \dots + P(n) \cdot P(X \leq x | n),$$

where:

- $P(i)$  is the probability that  $i$ th expert is correct; and
- $P(X \leq x | i) = F_i(x)$  is the conditional probability that  $X \leq x$  under the condition that the  $i$ th expert is correct.

This is exactly the weighted average formula for  $w_i = P(i)$ .

This new derivation of weighted average means that we can interpret the weights  $w_i$  as the probabilities that  $i$ th expert is correct. This is in complete accordance with the general interpretation of the weight  $w_i$  as describing trustworthiness of  $i$ th expert.

#### 9.4.4 Relation to the average of densities

Similar to the case of arithmetic  $p$ -average, for the weighted  $p$ -average, we get exactly the same result whether we combine the CDFs  $F_i(x)$ , or whether we combine the corresponding densities  $\rho_i(x)$ . In precise terms, the weighted average density

$$\rho(x) = w_1 \cdot \rho_1(x) + \dots + w_n \cdot \rho_n(x)$$

leads to the weighted arithmetic  $p$ -average for the corresponding CDFs:

$$F(x) = w_1 \cdot F_1(x) + \dots + w_n \cdot F_n(x).$$

#### 9.4.5 Weighted arithmetic $p$ -average with interval weights

As in the cases of real numbers and intervals, we can consider weighted arithmetic  $p$ -average with interval weights. The algorithm for computing the weighted average was given when we described aggregation of real numbers in Section 7.

This weighted average operation can be interpreted via a formula for full probability, in which, instead of knowing the exact probability  $P(i) = w_i$  that  $i$ th expert is correct, we know only the intervals  $\mathbf{w}_i$  of possible values of these probabilities.

In particular, when we have no prior information about the weights and therefore, take  $\mathbf{w}_i = [0, 1]$ , the aggregation results in the following interval:

$$[\min(F_1(x), \dots, F_n(x)), \max(F_1(x), \dots, F_n(x))].$$

It is worth mentioning that this is the  $p$ -bound envelope of the probability distributions  $F_1(x), \dots, F_n(x)$ .

### 9.5 Geometric $p$ -average and generalized $p$ -average

#### 9.5.1 General description and properties

When we discussed operations with real numbers in Section 7, we noticed that in many practical cases, the arithmetic average leads to counterintuitive conclusions. As an example, we used the average of three small concentration estimates

$10^{-5}$ ,  $10^{-6}$ , and  $10^{-7}$ , or, more generally, aggregating several values of different orders of magnitude. For such values, the arithmetic average leads, in effect, to selecting the largest of the values to be aggregated. This is not exactly what we intuitively understand by averaging. We showed that a natural remedy for this situation is the use of geometric average and generalized average.

For probability, this same example is also valid. It is quite possible that three different experts estimate the same small probability as, correspondingly,  $10^{-5}$ ,  $10^{-6}$ , and  $10^{-7}$ , and we do not want the result of aggregation to always coincide with the largest estimate. So, in  $p$ -averaging, it also make sense to consider the geometric average and the generalized average. The corresponding formulas are:

$$F(x) = \sqrt[n]{F_1(x) \cdot \dots \cdot F_n(x)};$$

$$F(x) = \left( \frac{F_1(x)^\alpha + \dots + F_n(x)^\alpha}{n} \right)^{1/\alpha}.$$

Because, as we have mentioned, the geometric average comes from the fact that effects are often logarithmically dependent on the actual values, the geometric  $p$ -average is often called a *logarithmic opinion pool* (Clemen & Winkler, 1999). One can show that when all functions to be aggregated are CDFs, the result of this aggregation is also a CDF. Similar to the cases of real numbers and intervals, this operation is idempotent, commutative, continuous, not associative, satisfies the averaging property, and is easy to compute.

When  $\alpha \rightarrow -\infty$ , the generalized  $p$ -average tends to  $\min(F_1(x), \dots, F_n(x))$ ; when  $\alpha \rightarrow 0$ , it tends to geometric  $p$ -average; when  $\alpha \rightarrow \infty$ , it tends to  $\max(F_1(x), \dots, F_n(x))$ . These operations min and max are implemented in RiskCalc.

It is worth mentioning that min and max are among the main operations used to combine uncertainty in fuzzy logic; thus, we get a new justification for these operations. These operations also have a natural meaning in risk analysis. For example, if we are looking for the concentration  $x$  of a certain pollutant, then  $F(x)$  describes the probability that this concentration does not exceed  $x$ . Several estimates give different estimates  $F_1(x), \dots, F_n(x)$  for this probability. Some of these estimates are larger, some of them are smaller.

- If we want to be extra cautious, then we take the largest of these probabilities as the resulting estimate. In other words, the aggregation operation max can be viewed as a conservative choice in risk analysis.
- Alternatively, we can believe the most optimistic estimate and take  $\min(F_1(x), \dots, F_n(x))$  as the aggregation result. Thus, min corresponds to an “optimistic” choice in risk analysis.

### 9.5.2 Case of real numbers

For the case when each distribution  $F_i(x)$  actually represents a real number  $x_i$ , the geometric  $p$ -average and the generalized  $p$ -average with  $\alpha < 0$  lead to the function which is equal to 0 when  $x \leq \max(x_1, \dots, x_n)$  and to 1 after that. In other words, the geometric  $p$ -average and the generalized  $p$ -average with  $\alpha < 0$  lead to a real number which is equal to the maximum of real numbers to be aggregated. So, for real numbers, geometric  $p$ -average and generalized  $p$ -average for  $\alpha < 0$  are equivalent to taking a maximum.

The generalized  $p$ -average for  $\alpha > 0$  is also reasonably easy to describe when the CDFs  $F_i(x)$  to be aggregated really are real numbers  $x_i$  (expressed in the CDF form). To get this description, we must first sort these real numbers into an increasing sequence  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . Then, the aggregation result has the following form:

$$F(x) = \begin{cases} 0 & \text{if } x \leq x_{(1)}, \\ \left(\frac{1}{n}\right)^{1/\alpha} & \text{if } x_{(1)} < x \leq x_{(2)}, \\ \dots & \\ \left(\frac{k}{n}\right)^{1/\alpha} & \text{if } x_{(k)} < x \leq x_{(k+1)}, \\ \dots & \\ 1 & \text{if } x > x_{(n)} \end{cases}$$

One can show that this distribution is equivalent to a weighted  $p$ -average of the CDFs which correspond to the real numbers  $x_{(1)}, \dots, x_{(n)}$ , with the weights:

$$w_i = \left(\frac{i}{n}\right)^{1/\alpha} - \left(\frac{i-1}{n}\right)^{1/\alpha}.$$

Let us give two simple numerical examples corresponding to  $n = 2$ :

- when  $\alpha = 1/2$ , we have  $w_1 = 1/4$  and  $w_2 = 3/4$ ;
- when  $\alpha/2$ , we have  $w_1 = 1/\sqrt{2} \approx 0.71$ , and  $w_2 = 1 - w_1 \approx 0.29$ .

In general, for  $\alpha = 1$  (arithmetic average), the weights are equal. For  $\alpha > 1$ , larger weights are assigned to smaller values. For  $\alpha < 1$ , larger weights are assigned to larger values  $x_i$ . So,

- if we want the aggregated values to be more conservative, i.e., if we give more weight to larger values of  $x$ , then we should use the generalized  $p$ -average with  $\alpha < 1$ ;
- if we want the aggregated values to be more optimistic, i.e., if we give more weight to smaller values of  $x$ , then we should use the generalized  $p$ -average with  $\alpha > 1$ .

### 9.5.3 Relation to the average of densities

A similar aggregation operation can be proposed for probability densities; see, e.g., (Clemen & Winkler, 1999):

$$\begin{aligned}\tilde{\rho}(x) &= \sqrt[n]{\rho_1(x) \cdot \dots \cdot \rho_n(x)}; \\ \tilde{\rho}(x) &= \left( \frac{\rho_1(x)^\alpha + \dots + \rho_n(x)^\alpha}{n} \right)^{1/\alpha}.\end{aligned}$$

The resulting functions  $\tilde{\rho}(x)$  are not necessarily pdfs, because they do not necessarily integrate to 1. To get a pdf, we must normalize these functions by dividing them by an appropriate constant:

$$\rho(x) = \frac{\tilde{\rho}(x)}{\int \tilde{\rho}(y) dy}.$$

Let us show that the geometric average of densities and generalized average with  $\alpha < 0$  lead to counterintuitive conclusions. This will be an additional argument in favor of selecting CDFs (and not pdfs) as the appropriate representation for risk analysis.

We combine two pieces of knowledge, each collected from an expert. Let us assume that according to the first expert, a variable  $X$  is uniformly distributed on the interval  $[0, 1]$ . The corresponding pdf has the form

$$\rho_1(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

The second expert has different information about  $X$ . He believes that  $X$  is close to 0.5, or, more precisely, that  $X$  is uniformly distributed on the narrow interval  $[0.5 - \varepsilon, 0.5 + \varepsilon]$  for some small  $\varepsilon > 0$

$$\rho_2(x) = \begin{cases} \frac{1}{2\varepsilon} & \text{if } 0.5 - \varepsilon \leq x \leq 0.5 + \varepsilon, \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the function  $\tilde{\rho}(x)$  is constant on almost the entire interval  $[0, 1]$  except for the zone  $[0.5 - \varepsilon, 0.5 + \varepsilon]$  where the value is larger. When we normalize and compute the probability of being within this zone, this probability tends to 0 as  $\varepsilon \rightarrow 0$ .

So, in the limit  $\varepsilon \rightarrow 0$ , if one expert is absolutely sure that  $X = 0.5$ , and another thinks that it is uniformly distributed on the entire interval  $[0, 1]$ , the corresponding density-aggregation operation completely ignores the expert who predicts  $X = 0.5$  and reproduces the opinion of the more cautious expert. This is not what we intuitively expect from aggregation.

#### 9.5.4 Weighted geometric $p$ -average and generalized $p$ -average: general description and properties

As in the cases of real numbers and intervals, we can have weighted geometric and generalized  $p$ -averages:

$$F(x) = F_1(x)^{w_1} \cdot \dots \cdot F_n(x)^{w_n};$$

$$F(x) = (w_1 \cdot F_1(x)^\alpha + \dots + w_n \cdot F_n(x)^\alpha)^{1/\alpha}.$$

The weighted geometric  $p$ -average is also called a *logarithmic opinion pool* (Clemen & Winkler, 1999). One can show that when all functions to be aggregated are CDFs, the result of this aggregation is also a CDF. Similar to the cases of real numbers and intervals, this operation is idempotent, non-commutative, continuous, non-associative, does not satisfy the averaging property, but is easy to compute.

#### 9.5.5 Weighted geometric $p$ -average and generalized $p$ -average: case of real numbers

For the case when each distribution  $F_i(x)$  actually represents a real number  $x_i$ , the weighted geometric  $p$ -average and generalized  $p$ -average with  $\alpha < 0$  is also equivalent to taking a maximum. This is just like the corresponding un-weighted operations.

The weighted generalized  $p$ -average for  $\alpha > 0$  can be explicitly described when the CDFs  $F_i(x)$  really are real numbers  $x_i$  (expressed in the CDF form). To get this description, we must first sort these real numbers into an increasing sequence  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . If we denote the corresponding weights by  $w_{(i)}$ , we get the following formula:

$$F(x) = \begin{cases} 0 & \text{if } x \leq x_{(1)}, \\ w_{(1)}^{1/\alpha} & \text{if } x_{(1)} < x \leq x_{(2)}, \\ \dots & \\ (w_{(1)} + \dots + w_{(k)})^{1/\alpha} & \text{if } x_{(k)} < x \leq x_{(k+1)}, \\ \dots & \\ 1 & \text{if } x > x_{(n)} \end{cases}$$

This distribution is equivalent to a weighted  $p$ -average of the CDFs which correspond to the real numbers  $x_{(1)}, \dots, x_{(n)}$ , with the new weights

$$w_{(i)}^{\text{new}} = (w_{(1)} + \dots + w_{(i)})^{1/\alpha} - (w_{(1)} + \dots + w_{(i-1)})^{1/\alpha}.$$

#### 9.5.6 Weighted geometric $p$ -average and generalized $p$ -average: relation to the average of densities

When formulated in terms of probability density functions instead of CDFs, the weighted geometric  $p$ -average and weighted generalized  $p$ -average have exactly the same problem as the un-weighted ones.

## 9.6 The use of copulas

In the above aggregation operations, we only used the marginal CDFs  $F_i(x)$  corresponding to each expert's estimates. If we have an additional information about the dependence between expert estimates, we can improve the estimates. For example, if we know that two experts are highly correlated, and that the third one is completely independent from the first two, then, instead of combining the estimates of all three estimates with equal weights, we should probably just combine the estimates of the first and the third experts.

As we have mentioned in Section 3, when we describe the probability distributions in terms of CDFs, the possible independence or dependence between the expert opinions can be described in terms of copulas. If we know the copulas describing the relation between the experts, we can therefore get better estimates for  $F(x)$ ; see, e.g., (Clemen & Winkler 1999) and (Jouini & Clemen, 1996). It is desirable to combine copula techniques with other aggregation methods. This is one of the directions of our current research.

## 9.7 Averaging operations for the case when CDFs come from measurements

### 9.7.1 General discussion

Let us now consider the case when the CDFs  $F_i(x)$  estimating the actual (unknown) CDF  $F(x)$  come from measurements. We already know, from Section 7, how to aggregate real numbers, i.e., individual measurement results. Therefore, a natural way to aggregate the corresponding CDFs is as follows:

- first, we extract the corresponding measurement results from the CDFs;
- then, we average the extracted measurement results into aggregated estimates;
- finally, we convert the resulting estimates into a new CDF.

Because in this case, we are averaging the values  $x$  of the measured quantity, the corresponding methods are called methods of *x-averaging*.

Let us try to flesh out the above idea. We know how to aggregate individual measurement results from the survey of the methods in Section 7 devoted to aggregating numerical estimates. We know how to convert the resulting estimates into a single CDF. We described it in Section 3, when we surveyed different methods of getting p-bounds (and a CDF is, of course, a particular case of a p-bound). So, the only step in this description for which we have not yet explained the algorithm is the first step: extracting measurement results from the CDF.

The problem of extracting measurement results from a CDF is an inverse problem to the problem of combining several measurement results into a single

CDF. So, in order to find out how we can extract measurement results from a CDF, let us recall how measurement results lead to a CDF. When we have  $N$  measurement results  $x_1, \dots, x_N$ , then, to combine them into a CDF, we first *sort* them into a sequence  $x_{(1)} \leq \dots \leq x_{(N)}$ . Then, we design a histogram CDF  $F_h(x)$ , for which the value is 0 before  $x_{(1)}$ ,  $1/N$  between  $x_{(1)}$  and  $x_{(2)}$ ,  $2/N$  between  $x_{(2)}$  and  $x_{(3)}$ , etc. Thus, if we know the CDF  $F_h(x)$ , we can reconstruct each value  $x_{(k)}$  as the first value  $x$  for which  $F_h(x) = k/N$ . This notion can be describe in analytical terms if we use the notion of an inverse function  $F_h^{-1}(p)$ . In these terms, out of  $n$  measurement results, the value  $x_{(k)}$  can be reconstructed as the *percentile*  $F_h^{-1}(k/N)$ .

We consider the situation when the difference between the estimates CDFs  $F_i(x)$  and the actual (unknown) CDF  $F(x)$  is caused by measurement error. In other words, we consider the situation in which the measurement results  $x_1^{(i)}, \dots, x_n^{(i)}$  used to construct each distribution  $F_i(x)$  are somewhat different from the actual values  $x_1, \dots, x_n$  of the corresponding quantities. We must therefore aggregate values coming from different measurements to improve our estimates of  $x_j$ .

In the ideal case, if we have a probability distribution  $F(x)$ , then, from this probability distribution, we could extract the actual values  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  of the measured quantities. From each measured CDF  $F_i(x)$ , we can actually extract the corresponding sorted values  $x_{(1)}^{(i)} \leq x_{(2)}^{(i)} \leq \dots \leq x_{(n)}^{(i)}$ . We know that each of these measurement results is an approximation to one of the actual values  $x_k$ . Depending on how small the measurement error is, we have two possible situations here.

If the measurement error is *small*, then from  $x_{(k)} < x_{(l)}$ , it follows that the result of measuring  $x_{(k)}$  is still smaller than the result of measuring  $x_{(l)}$ . Thus, in this case, the order between the measurement results mirrors the original order between the values. In particular, the first of the sorted measurement results  $x_{(1)}^{(i)}$  is an approximation to  $x_{(1)}$ , the second of the sorted measurement results  $x_{(2)}^{(i)}$  is an approximation to  $x_{(2)}$ , etc. As a result, for every  $k$ , the values  $x_{(k)}^{(1)}, \dots, x_{(k)}^{(n)}$  approximate the exact same value  $x_{(k)}$ . So, we can use the known techniques for aggregating numerical values to combine these  $n$  values  $x_{(k)}^{(1)} = F_1^{-1}(k/N), \dots, x_{(k)}^{(n)} = F_n^{-1}(k/N)$  into a single estimate for  $x_{(k)}$  – which will then serve as the value  $F^{-1}(k/N)$  (percentile) for the new (aggregated) probability distribution.

In other words, if the measurement errors are small, then for every number  $p = k/N$ , the percentile  $F^{-1}(p)$  can be obtained by aggregating the percentiles  $F_1^{-1}(p), \dots, F_n^{-1}(p)$ . Every rational number from the interval  $[0, 1]$ , by definition, can be represented as  $k/N$  for some  $k$  and  $N$ . Every real number from the interval  $[0, 1]$  can be approximated, with arbitrary accuracy, by rational numbers. Hence, the above conclusion holds for every real number  $p$ .

In short, if measurement errors are small, then for every  $p \in [0, 1]$ , to get the corresponding percentile  $F^{-1}(p)$ , we aggregate  $n$  percentiles  $F_1^{-1}(p), \dots, F_n^{-1}(p)$ . In geometric terms, if we draw CDFs in the usual way, with  $x$  horizontal and  $p$  vertical, this means that we do the averaging in each horizontal section.

This conclusion holds if measurement errors are small. The operation probably makes intuitive sense also when the measurement errors are not small, but our formal justification only works when the measurement errors are small. Since we are describing justified methods in our survey, we will therefore restrict  $x$ -averaging techniques to the case when the errors are small. How can we check whether these errors are small or not? In the situation under consideration, when the CDFs  $F_i(x)$  come from measurements, the difference between different CDFs is caused by measurement errors:

- when the measurement errors are small, then the CDFs  $F_i(x)$  are close to each other;
- when the measurement errors are large, the difference between different CDFs also become large.

So, the measurement errors are small if and only if the CDFs to be aggregated are close to each other. Thus, we can apply the above idea of horizontal  $x$ -averaging if the CDFs to be aggregated are close enough to each other.

If the measurement errors are not small, i.e., if the CDFs to be aggregated are not close to each other, then the approximation may change the order between the values  $x_{(k)}$  and  $x_{(l)}$ . In this case, the value which is first in the ordering corresponding to  $F_1(x)$  may actually correspond to measuring the quantity which is last in the ordering corresponding to  $F_2(x)$ . In this case, it does not make much sense to match the measurement results placed on exactly the same place. So, instead of matching each result with a precisely determined one, we can match it with randomly selected values corresponding to  $F_2(x)$ . In other words, in this case, a natural thing to do it to take a random variable  $x_1$  distributed according to the distribution  $F_1(x)$ , another random variable  $x_2$  distributed according to the distribution  $F_2(x)$ , aggregate these two random variables, and then take, as the aggregated CDF, the CDF describing the distribution of the result of this aggregation.

Let us discuss the corresponding aggregation methods in detail. We will start with the methods of horizontal aggregation, corresponding to close CDFs  $F_i(x)$ .

### 9.7.2 Aggregating close CDFs: arithmetic $x$ -average. General description and properties

In accordance with the above arguments, when CDFs are close, we should aggregate, for each  $p$ , the percentiles  $F_i^{-1}(p)$  corresponding to this particular  $p$ . In short, we must aggregate horizontally.

In particular, when we apply the simplest possible aggregation procedure – arithmetic average – we conclude that for every level  $p$ , we must take an arithmetic average of the corresponding percentiles of the CDFs  $F_i$  – i.e., points in which the graphs of the CDFs  $F_i(x)$  intersect with this  $p$ :

$$F^{-1}(p) = \frac{F_1^{-1}(p) + \dots + F_n^{-1}(p)}{n}.$$

As in the cases of real numbers, intervals, and arithmetic  $p$ -average, this operation is idempotent, commutative, continuous, not associative, satisfies the averaging property, and is easy to compute.

### 9.7.3 Case of real numbers

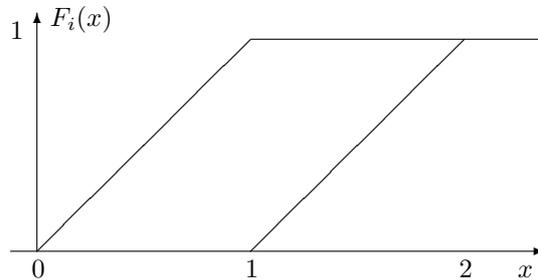
If each of the CDFs to be aggregated represents an exact real number  $x_i$ , then the arithmetic  $x$ -average of these CDFs is simply a CDF corresponding to the arithmetic average  $x = (x_1 + \dots + x_n)/n$  of these real numbers. We can immediately see that this is different from the arithmetic  $p$ -average where the result was a mixture. Similarly, if we apply an arbitrary close-CDF aggregation operation  $*$  to CDFs  $F_i(x)$  representing exact real numbers  $x_i$ , we end up with a CDF that also represents an exact real number  $x$ , specifically, the result  $x = x_1 * \dots * x_n$  of applying the same aggregation operation to these real numbers. For this reason, we see that  $x$ -averaging rather than  $p$ -averaging is the natural generalization of averaging for real numbers.

### 9.7.4 Example

To illustrate this formula, we can use the same example as we used to illustrate arithmetic  $p$ -average. In this example,

$$F_1(x) = \begin{cases} 0 & \text{if } x \leq 1, \\ x - 1 & \text{if } 1 \leq x \leq 2, \\ 1 & \text{if } x > 2 \end{cases}$$

$$F_2(x) = \begin{cases} x & \text{if } x \leq 1, \\ 1 & \text{if } x > 1 \end{cases}$$

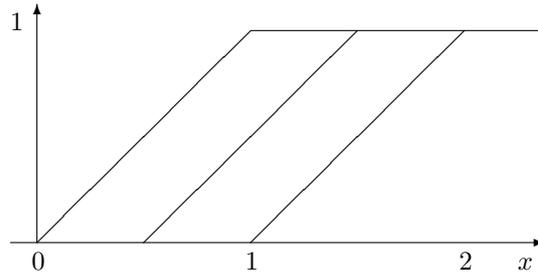


Therefore,  $F_1^{-1}(p) = 1 + p$ ,  $F_2^{-1}(p) = p$ , and:

$$F^{-1}(p) = \frac{(1+p) + p}{2} = \frac{1}{2} + p.$$

Hence,

$$F(x) = \begin{cases} 0 & \text{if } x \leq 1/2, \\ x - 1/2 & \text{if } 1/2 \leq x \leq 3/2, \\ 1 & \text{if } x > 3/2 \end{cases}$$



The resulting CDF is different from the result of the arithmetic  $p$ -average shown in the previous section.

### 9.7.5 NUREG controversy

In the 1980s, the arithmetic  $x$ -average was used in risk analysis as part of the Nuclear Regulations Committee document NUREG-1150 (Hora & Iman, 1989). This use was deprecated because

- first, it was claimed that this rule is not well justified;
- second, when we have outliers or other drastically different CDFs, it does not work.

As we can see from the above text, both criticisms were true, but this does not preclude us from recommending the use of this rule. Indeed, this rule was not sufficiently justified at the time when it was adapted, but now it is: above, we outlined the justification for using this rule. The only remaining problem is that our justification only recommends the usage of this rule for combining close CDFs.

### 9.7.6 Aggregating close CDFs: other $x$ -averaging operations

In addition to arithmetic average, we can use all other averaging operations to combine percentiles: weighted arithmetic average, geometric and generalized average, and weighted geometric and generalized average.

The resulting  $x$ -averaging operations have exactly the same properties as the corresponding operations with real numbers. Let us briefly describe the corresponding formulas.

### 9.7.7 Weighted arithmetic $x$ -average

In this aggregation, the new percentile is:

$$F^{-1}(p) = w_1 \cdot F_1^{-1}(p) + \dots + w_n \cdot F_n^{-1}(p).$$

As in the cases of real numbers, intervals, and  $p$ -average, this operation is idempotent, non-commutative, continuous, not associative, does not satisfy the averaging property, and is easy to compute.

We can also consider interval weights, in which case we get interval percentiles, i.e.,  $p$ -bounds.

### 9.7.8 Geometric and generalized $x$ -average

In this aggregation, the new percentile is either

$$F^{-1}(p) = \sqrt[n]{F_1^{-1}(p) \cdot \dots \cdot F_n^{-1}(p)}$$

or

$$F^{-1}(p) = \left( \frac{F_1^{-1}(p)^\alpha + \dots + F_n^{-1}(p)^\alpha}{n} \right)^{1/\alpha}.$$

As in the cases of real numbers, intervals, and to  $p$ -average, this operation is idempotent, commutative, continuous, not associative, satisfies the averaging property, and is easy to compute.

### 9.7.9 Weighted geometric and generalized $x$ -average

In this aggregation, the new percentile is either

$$F^{-1}(p) = F_1^{-1}(p)^{w_1} \cdot \dots \cdot F_n^{-1}(p)^{w_n}$$

or

$$F^{-1}(p) = (w_1 \cdot F_1^{-1}(p)^\alpha + \dots + w_n \cdot F_n^{-1}(p)^\alpha)^{1/\alpha}.$$

As in the cases of real numbers, intervals, and  $p$ -averaging, this operation is idempotent, non-commutative, continuous, not associative, does not satisfy the averaging property, and is easy to compute.

We can also consider interval weights.

### 9.7.10 Aggregating drastically different CDFs: arithmetic $x$ -average

Let us now consider the case of drastically different CDFs  $F_1(x), \dots, F_n(x)$ . As we have mentioned in the general discussion subsection of this section, to get the aggregation of such CDFs, we do the following:

- First, we select a numerical aggregation operation  $*$  (e.g.,  $x*y = (x+y)/2$ ).

- Next, we solve the following problem:
  - we have a variable  $x_1$  which is distributed according to a CDF (p-bound)  $F_1(x)$ ;
  - we have a variable  $x_2$  which is distributed according to a CDF (p-bound)  $F_2(x)$ ;
  - ...
  - we have a variable  $x_n$  which is distributed according to a CDF (p-bound)  $F_n(x)$ ;
  - we want to find a CDF (or p-bound) which corresponds to the quantity  $x = x_1 * \dots * x_n$ .

This is exactly the problem that we mentioned when we described indirect measurement/estimation, and all methods for solving that problem (Ferson et al., 2001), (Ramas, 2001) are applicable here as well.

If we assume that the variables are independent (or if we assume that we know an exact copula describing their dependence), then, as a result of the aggregation, we get a CDF. If we do not make this assumption, then, as a result of the aggregation, we get a p-bound.

What are the properties of the resulting aggregation operation? Even for the simplest operation of the arithmetic average, this operation (for the case of independent variables) is not idempotent, commutative, continuous, not associative, satisfies the averaging property, and is computable.

An example that this operation is not idempotent is easy to generate. Let us take  $F(x)$  in which we have 0 and 1 with equal probability 1/2, and show that  $F * F \neq F$ . Here,  $x_1$  is equal to 0 and 1 with probability 1/2, and  $x_2$  is another variable with the same distribution which is independent from  $x_1$ . By combining 2 possible values (0 and 1) of the variable  $x_1$  with 2 possible values (0 and 1) of the variable  $x_2$ , we conclude that there are four possible values of the pair  $(x_1, x_2)$ : (0,0), (0,1), (1,0), and (1,1). Each of these 4 combinations has the same probability 1/4. As a result, the average  $(x_1 + x_2)/2$  takes the value 0 with probability 1/4, 1 with probability 1/4, and 0.5 with probability 1/2. The resulting probability distribution is clearly different from the original distribution  $F(x)$ .

The fact that this operation is not idempotent seems troubling, because idempotence is what we naturally require of all aggregation operations. However, we might not worry about non-idempotence of this particular operation. Although from the purely mathematical viewpoint, we do have  $F * F \neq F$ , from the practical viewpoint, this operation was only justified for the case when the CDFs are drastically different, so we should not apply it to the case when  $F_1 = F_2$ .

## 9.8 Dempster-Shafer aggregation methods

### 9.8.1 General description

In addition to the well justified methods described above, there are other methods for aggregating CDFs and p-bounds which are of a more heuristic nature. In particular, such a combination method exists for Dempster-Shafer formalism.

This combination method is a natural generalization of the intersection operation, according to which, if we know that  $x \in \mathbf{x}$  and we know that  $x \in \mathbf{y}$ , then we can conclude that  $x$  belongs to the intersection  $\mathbf{x} \cap \mathbf{y}$ .

In general, if we have two pieces of knowledge, one in which we have focal elements  $\mathbf{x}_i$  with masses  $p_i$ , and another in which we have focal elements  $\mathbf{y}_j$  with masses  $q_j$ , then Dempster's rule (Shafer, 1984), (Shafer, 1986) tells us to combine these two pieces of knowledge as follows. We consider all possible pairs of intervals  $\mathbf{x}_i$  and  $\mathbf{y}_j$  which have non-empty intersection, and assign to every such intersection  $\mathbf{x}_i \cap \mathbf{y}_j$  the probability  $p_i \cdot q_j$ . These intersections will be the focal elements of the new knowledge bases. The use of the products of the probabilities corresponds to an assumption of independence between the two estimates.

Because some pairs have an empty intersection, the values  $p_i \cdot q_j$  corresponding to these focal elements do not add up to 1, so we cannot just take these products as the masses, we must first normalize them by dividing by their sum. What we described is the original Dempster's combination rule. There are many modifications of this rule; see, e.g., (Yager, 1983), (Yager, 1985), (Zadeh, 1986), (Yager, 1987), (Halpern & Fagin, 1992), (Baldwin, 1994a), (Chateaufneuf, 1994), (Dubois & Prade, 1994), (Kreinovich et al., 1994), (Kruse & Klawonn, 1994), (Saffiotti, 1994), (Spies, 1994), (Yager et al., 1994), (Zhang, 1994), (Mahler, 1995), (Srivastava & Shenoy, 1995), (Yager, 2001). For example, instead of taking a product of probabilities (which corresponds to independence), we can use another t-norm. This modification will be useful for our purposes.

Alternatively, instead of normalizing the values of the product, we could keep the same as they are, and assign the rest to the entire real line. This would mean that in the case of inconsistency (non-empty intersection), we have no information at all. One can show that this particular modification does not work for our problems.

Dempster's rule and its modifications have been successfully used in many real-life applications (Yager et al., 1994), including such various areas as

- water resource distribution (Caselton & Luo, 1992);
- databases (Baldwin, 1994);
- decision making (Jaffray, 1994);
- target identification and localization (Schubert, 1994), (Appriou, 1998), (Jousselme et al., 2001);

- climate change analysis (Luo & Caselton, 1997);
- audit decisions (Srivastava, 1997);
- image processing (Block & Maitre, 1998).

### 9.8.2 How can Dempster's rule be applied to combining CDFs?

When we justified the use of p-bounds, we showed that an arbitrary Dempster-Shafer (DS) knowledge base can be actually represented as a p-bound. So, if we want to apply the Dempster's combination rule to CDFs, we must do the following:

- First, we transform CDF into a DS knowledge base.
- Second, we apply Dempster's combination rule to get a new DS knowledge base.
- Finally, we translate the resulting DS knowledge base into a p-bound.

We know, from Section 3, how to perform the second and the third steps. The first step is somewhat ambiguous because, as we have mentioned in Section 3, there are many ways to reconstruct a DS knowledge base from a p-bound. As in the above general discussions, a natural way to perform the first step depends on whether the CDFs to be aggregated come from expert estimation or from measurements.

Let us first consider the case when the CDFs come from expert estimation. In this case, for each given value, we estimate the probability  $F(x)$ . From the purely mathematical viewpoint, we thus have infinitely many values  $F(x)$  for all possible values of  $x$ . In practice, of course, we discretize  $x$ ; this idea goes back to (Start, 1984). So, we consider small intervals  $[0, \Delta x]$ ,  $[\Delta x, 2\Delta x]$ . These intervals will be our focal points. The probability (mass) assigned to each interval  $[x, x + \Delta x]$  is equal to the probability to be within this interval, i.e., to the difference  $F(x + \Delta x) - F(x)$ . Because the intervals are narrow, this probability is practically equal to  $\rho(x) \cdot \Delta x$ , where  $\rho(x)$  is the pdf corresponding to this probability distribution.

When the CDFs come from measurement, then CDF is actually the result of combining several ( $N$ ) measurement results, with equal probabilities  $1/N$ , into a single CDF. In this case, a natural way is to represent the CDF as a combination of different measurement results with probability  $1/N$ . In other words, we subdivide the interval  $[0, 1]$  into small subintervals  $[0, \Delta p]$ ,  $[\Delta p, 2\Delta p]$ , etc. To each of these narrow intervals, we assign the corresponding  $x$ -interval  $[F^{-1}(p), F^{-1}(p + \Delta p)]$ . These narrow intervals are our focal points, and each has the probability  $\Delta p$ .

We will see that for these two different DS knowledge bases, the combination rule leads to two different aggregation operations for CDF.

### 9.8.3 Dempster-Shafer methods for the case when CDFs come from expert estimates: main formula

In this case, the standard Dempster's rule results in a new probability distribution whose density  $\rho_{\text{new}}(x)$  is proportional to the product of the densities  $\rho_1(x)$  and  $\rho_2(x)$  of the two probability distributions to be aggregated:

$$\rho_{\text{new}}(x) = \frac{1}{N} \cdot \rho_1(x) \cdot \rho_2(x),$$

where  $N$  is the normalizing factor guaranteeing that  $\int \rho_{\text{new}}(x) dx = 1$ , i.e.,

$$N = \int_{-\infty}^{\infty} \rho_1(x) \cdot \rho_2(x) dx.$$

### 9.8.4 Example

For example, if both  $\rho_i(x)$  are identical Gaussian distributions with standard deviation  $\sigma$ , i.e., if:

$$\rho_1(x) = \rho_2(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{x^2}{2\sigma^2}\right),$$

then the result of the Dempster's rule aggregation is a new Gaussian distribution with a different standard deviation  $\sigma_{\text{new}} = \sigma/\sqrt{2}$ .

### 9.8.5 Properties

The above example shows that this aggregation operation is not idempotent – which is not a good thing. It does, however, possess other desired quantities: it is commutative, continuous, associative, satisfies the averaging property, and is rather easy to compute.

### 9.8.6 Relation with Bayesian aggregation

It is worth mentioning that the formula resulting from Dempster's rule can be naturally interpreted in Bayesian terms. Indeed, the general Bayes formula describes how to update probabilities of different hypotheses. If we start with prior probabilities  $P_0(H_i)$  of different hypotheses  $H_1, \dots, H_n$ , then, after the observation  $E$ , we replace the original prior probabilities with the new values:

$$P(H_i) = \frac{P(E | H_i) \cdot P_0(H_i)}{P(E | H_1) \cdot P_0(H_1) + \dots + P(E | H_n) \cdot P_0(H_n)},$$

where  $P(E | H_i)$  is the conditional probability of the observation  $E$  under the hypothesis  $H_i$ . A natural way to apply this formula to the aggregation of several probability distributions is as follows. As hypotheses  $H_i$ , we take different values

of the quantity  $x$ . To be more precise, we can divide the real axis into small intervals of length  $\Delta x$ , and consider belonging to these subintervals as different hypotheses. As prior probabilities, we take the probabilities of different values of  $x$  according to the first of the distributions to be aggregated.

The probability of having a value equal to  $x$  (or, to be more precise, to have a value in the interval  $[x, x + \Delta x]$ ) is proportional to  $\rho_1(x)$  (to be more precise, this probability is equal to  $\rho_1(x) \cdot \Delta x$ ). Correspondingly, the resulting probabilities  $P(H_i)$  are proportional to the corresponding values  $\rho(x)$  of the aggregated density function:  $P(H_i) \sim \rho(x)$ .

As the observation  $E$ , we take the opinion of the second expert, as described by the probability distribution with the probability density  $\rho_2(x)$ . To apply Bayes formula, we must estimate the conditional probability  $P(E | H_i)$  that the second expert is right, under the condition  $H_i$  that the actual value of the quantity is within the interval  $[x, x + \Delta x]$ . Intuitively, this value depends on the probability  $\rho_2(x) \cdot \Delta x$  that the second expert assigns to this interval. Indeed:

- If the second expert predicts a high probability  $\rho_2(x)$  for  $x$ , and we do observe this  $x$ , then this observation confirms that the second expert is right. In other words, in this case, the corresponding conditional probability  $P(E | H_i)$  is (relatively) large.
- On the other hand, if the second expert predicts a very low probability  $\rho_2(x)$  for this value  $x$  – e.g., equal to 0 – and we do observe this value, then this is a strong argument that the second expert is wrong. In other words, in this case, the corresponding conditional probability  $P(E | H_i)$  is (relatively) small.

Thus, the larger  $\rho_2(x)$ , the larger the conditional probability. In mathematical terms, it means that the value  $P(E | H_i)$  the conditional probability is a monotonic function of  $\rho_2(x)$ , i.e.,  $P(E | H_i) = f(\rho_2(x))$  for some increasing function  $f(x)$ .

When we substitute  $P_0(H_i) \sim \rho_1(x)$  and  $P(E | H_i) = f(\rho_2(x))$  into the Bayes formula, we conclude that the resulting aggregated density function is proportional to  $\rho(x) \sim \rho_1(x) \cdot f(\rho_2(x))$ . This is not yet the final formula, because we do not know yet what the function  $f(x)$  is.

To determine the function  $f(x)$ , we can use the fact that we are interested in the case when both distributions to be aggregated are equally trusted. In this case, the aggregation result should not depend on the order in which we present the distributions. In other words, the functions  $\rho_1(x) \cdot f(\rho_2(x))$  and  $f(\rho_1(x)) \cdot \rho_2(x)$  should differ only by a multiplicative constant. One can show that this condition is only satisfied for a linear function  $f(x) = k \cdot x$ , in which case, the Bayesian combination leads to  $\rho(x) \sim \rho_1(x) \cdot \rho_2(x)$ .

Thus, we get exactly the same aggregation formula as the Dempster-Shafer method.

### 9.8.7 Alternative formulas

If, instead of the product, we use a different t-norm (aggregation operation)  $F(x, y)$ , then we get an alternative formula:

$$\rho_{\text{new}}(x) = \frac{1}{N} \cdot F(\rho_1(x), \rho_2(x)),$$

where

$$N = \int_{-\infty}^{\infty} F(\rho_1(x), \rho_2(x)) \, dx.$$

Unless  $F(x, y)$  is an idempotent operation (e.g.,  $F(x, y) = \min(x, y)$ ), this aggregation operation is still not idempotent.

### 9.8.8 Dempster-Shafer methods for the case when CDFs come from measurements: formula

In this case, DS-aggregation results in a new probability distribution whose density  $\rho_{\text{new}}(x)$  is proportional to the *maximum* of the densities  $\rho_1(x)$  and  $\rho_2(x)$  of the two probability distributions to be aggregated:

$$\rho_{\text{new}}(x) = \frac{1}{N} \cdot \max(\rho_1(x), \rho_2(x)),$$

where

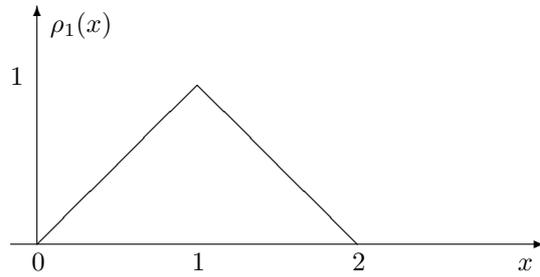
$$N = \int_{-\infty}^{\infty} \max(\rho_1(x), \rho_2(x)) \, dx.$$

### 9.8.9 Dempster-Shafer methods for the case when CDFs come from measurements: examples

In view of the fact that the previous Dempster-Shafer operation turned out to be not idempotent, let us first check this new one on the example when  $\rho_1(x) = \rho_2(x)$ . In this case, we get  $\max(\rho_1(x), \rho_2(x)) = \rho(x)$ , hence  $N = 1$  and  $\rho_{\text{new}}(x) = \rho_1(x)$ . This example shows that, unlike the expert-originated DS aggregation, this aggregation operation is idempotent.

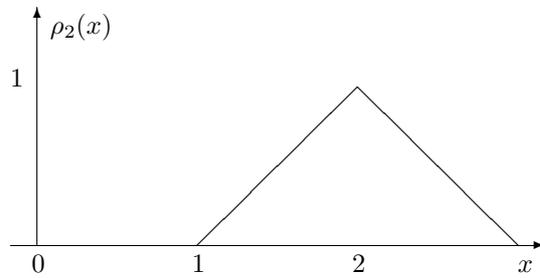
Let us now give one non-trivial example of this operation. Let  $\rho_1(x)$  be a unimodal distribution with a triangular density function:

$$\rho_1(x) = \begin{cases} x & \text{if } 0 \leq x \leq 1, \\ 2 - x & \text{if } 1 \leq x \leq 2, \\ 0 & \text{otherwise} \end{cases}$$



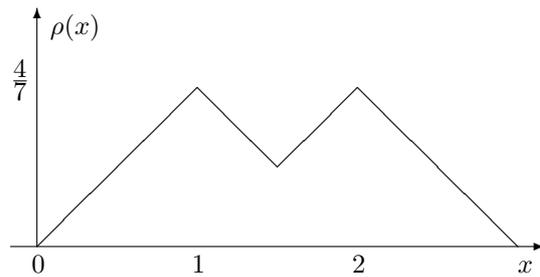
As  $\rho_2(x)$ , let us take a similar distribution, but shifted by 1:

$$\rho_2(x) = \begin{cases} x - 1 & \text{if } 1 \leq x \leq 2, \\ 3 - x & \text{if } 2 \leq x \leq 3, \\ 0 & \text{otherwise} \end{cases}$$



Here,  $N = 7/4$ , and the aggregation result  $\rho(x)$  is the following bi-modal distribution:

$$\rho(x) = \begin{cases} (4/7) \cdot x & \text{if } 0 \leq x \leq 1, \\ (4/7) \cdot (2 - x) & \text{if } 1 \leq x \leq 1.5, \\ (4/7) \cdot (x - 1) & \text{if } 1.5 \leq x \leq 2, \\ (4/7) \cdot (3 - x) & \text{if } 2 \leq x \leq 3, \\ 0 & \text{otherwise} \end{cases}$$

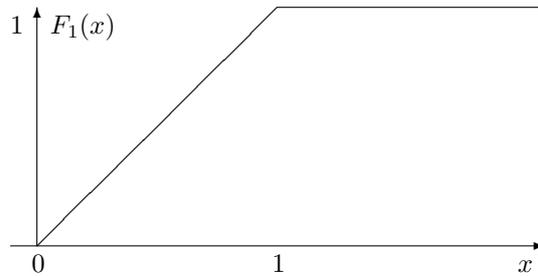


### 9.8.10 How can we describe this operation in terms of CDF?

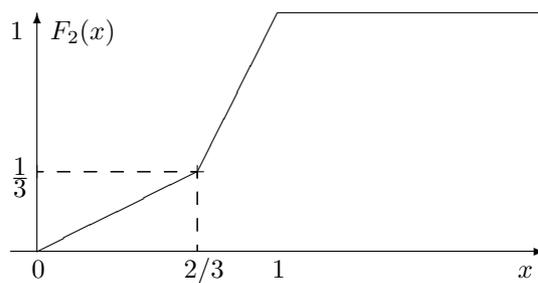
Because p-bounds are described in terms of the cumulative distribution function (CDF), not probability density function  $\rho(x)$ , it is desirable to describe this combination operation in terms of CDF. For that, we can use the fact that the probability density is a slope of the CDF. So, if we start with the two CDFs, then, in essence, on each subinterval of the real line, we pick the shape corresponding to the steepest of the two CDFs.

Let us give a simple example. Let us consider the following two CDFs:  $F_1(x)$  corresponds to a uniform distribution on the interval  $[0, 1]$ , and  $F_2(x)$  is a combination of two uniform distributions: on the interval  $[0, 2/3]$  (with probability  $1/3$ ) and on the interval  $[2/3, 1]$  (with probability  $2/3$ ). The corresponding CDFs are:

$$F_1(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ x & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x \geq 1 \end{cases}$$



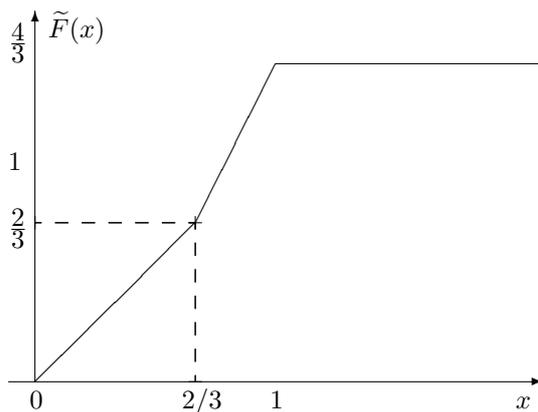
$$F_2(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ (1/2) \cdot x & \text{if } 0 \leq x \leq 2/3, \\ 2x - 1 & \text{if } 2/3 \leq x \leq 1, \\ 1 & \text{if } x \geq 1 \end{cases}$$



In this example, for  $x \leq 2/3$ ,  $F_1(x)$  is steeper, and for  $x \geq 2/3$ ,  $F_2(x)$  is steeper. Thus, for  $x \leq 2/3$ , we copy the CDF  $F_1(x)$ , and for  $x \geq 2/3$ , we copy the CDF

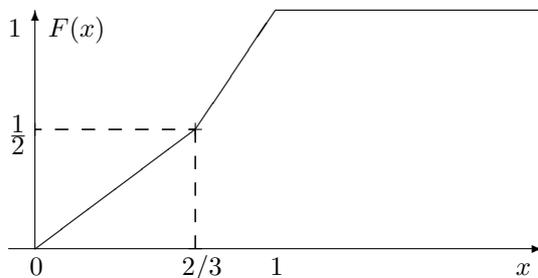
$F_2(x)$ . As a result, we get the following “CDF”  $\tilde{F}(x)$ :

$$\tilde{F}(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ x & \text{if } 0 \leq x \leq 2/3, \\ 2x - 2/3 & \text{if } 2/3 \leq x \leq 1, \\ 4/3 & \text{if } x \geq 1 \end{cases}$$



Finally, the resulting “CDF” needs to be normalized, so we get the following aggregated CDF  $F(x)$ :

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ (1/2) \cdot x & \text{if } 0 \leq x \leq 2/3, \\ 2x - 1 & \text{if } 2/3 \leq x \leq 1, \\ 1 & \text{if } x \geq 1 \end{cases}$$



### 9.8.11 Relation with previously described methods

This aggregation operation corresponds to the  $\alpha = \infty$  case of power “averaging” of probability densities.

This new operation is also related to the idea of “enveloping”, combining together ranges provided by different sources. For distributions with finite support, this statement has direct sense: indeed, the support of this new probability

distribution, i.e., the set of all the values  $x$  for which  $\rho_{\text{new}}(x) > 0$ , is equal to the *union* of the supports of two combined ones. Indeed,  $\rho_{\text{new}}(x) > 0$  if and only if  $\max(\rho(x), \rho'(x)) > 0$  if and only if  $(\rho(x) > 0 \text{ and } \rho'(x) > 0)$ .

For distributions like Gaussian for which the probability density function is always positive, the support is the entire real line. For such distributions, a relation to “enveloping” is slightly more indirect. Indeed, in practical applications of such distributions, it is usually assumed that values  $x$  corresponding to very low probability density are impossible. For example, in normal statistical applications of Gaussian distributions, a measurement result which differs from the mean by more than, say, six standard deviations, is considered impossible: a normal conclusion is that either this result is erroneous, or the distribution is not really Gaussian.

In other words, there is a certain threshold  $\rho_0$  such that only values  $x$  for which  $\rho(x) \geq \rho_0$  are considered practically possible. The exact value of  $\rho_0$  depends on the practical situation: for example, for normal distributions, people use  $2\sigma$ ,  $3\sigma$ , and  $6\sigma$  as a cut-off. If we select  $\rho_0$  as a cut-off level for the aggregated distribution with the probability density function  $\rho_{\text{new}}(x) = (1/N) \cdot \max(\rho(x), \rho'(x))$ , then the set

$$S_{\text{new}} = \{x \mid \rho_{\text{new}}(x) \geq \rho_0\}$$

of all the *possible* values, i.e., values for which  $\rho_{\text{new}}(x) \geq \rho_0$ , is equal to the union of the two cut-off sets corresponding to the original distributions, albeit with different cut-off values:  $S_{\text{new}} = S \cup S'$ , where

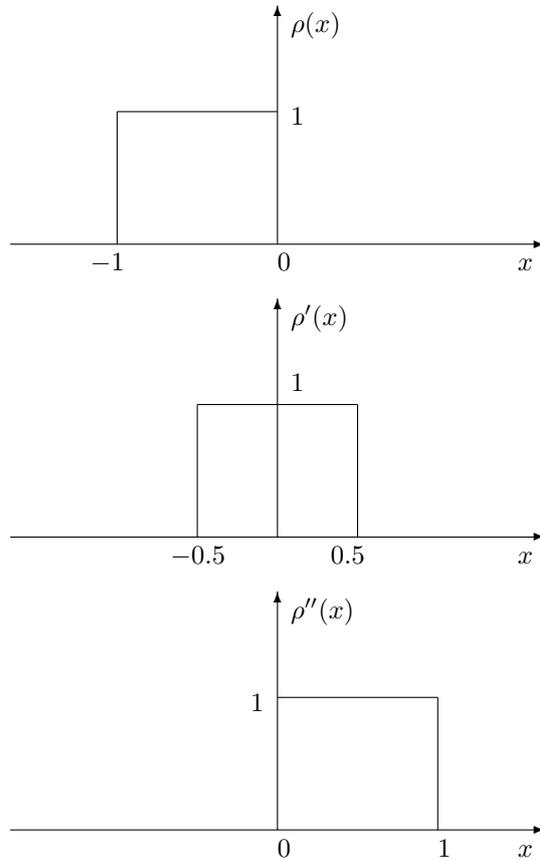
$$S = \{x \mid \rho(x) \geq \rho_1\},$$

$$S' = \{x \mid \rho'(x) \geq \rho_1\},$$

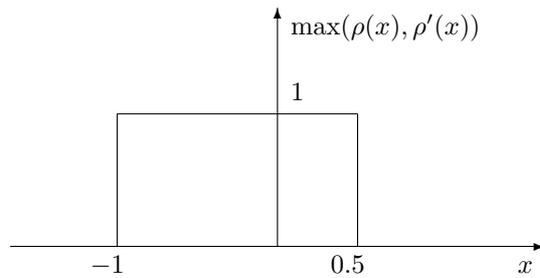
and  $\rho_1 = N \cdot \rho_0$ .

### 9.8.12 Dempster-Shafer methods for the case when CDFs come from measurements: properties

This operation is idempotent, commutative, continuous, and rather easy to compute. However, one property does not hold: this operation is not associative, i.e., it is not true that  $(\rho * \rho') * \rho'' = \rho * (\rho' * \rho'')$  for all possible distributions  $\rho$ ,  $\rho'$ , and  $\rho''$ . As an example of non-associativity, we take three uniform distributions:  $\rho(x)$  is a uniform distribution on the interval  $[-1, 0]$ ,  $\rho'(x)$  is a uniform distribution on the interval  $[-0.5, 0.5]$ , and  $\rho''(x)$  is a uniform distribution on the interval  $[0, 1]$ . Each of these three intervals has a unit length, so each probability density function has a value 1 within this interval:

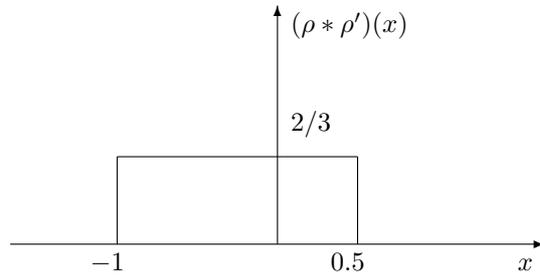


In this example,  $\max(\rho(x), \rho'(x))$  is equal to 1 on the interval  $[-1, 0.5]$ :



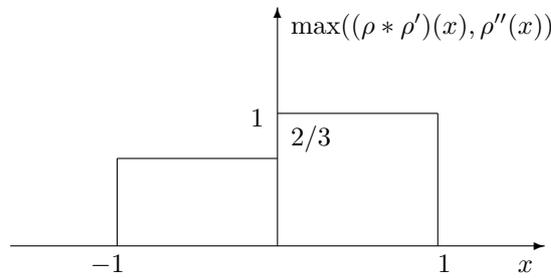
The integral  $N$  of the corresponding function  $\max(\rho(x), \rho'(x))$  is  $3/2$ , hence the normalized function  $(\rho * \rho')(x)$  has the following form:

$$(\rho * \rho')(x) = \begin{cases} 2/3 & \text{if } -1 \leq x \leq 0.5, \\ 0 & \text{otherwise} \end{cases}$$



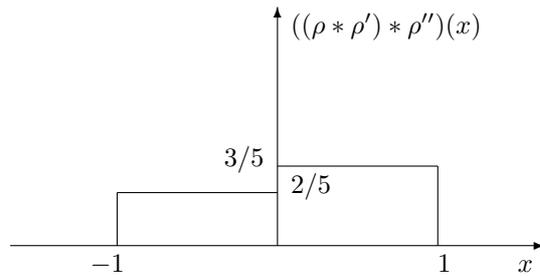
Now,

$$\max((\rho * \rho')(x), \rho''(x)) = \begin{cases} 2/3 & \text{if } -1 \leq x \leq 0, \\ 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise} \end{cases}$$



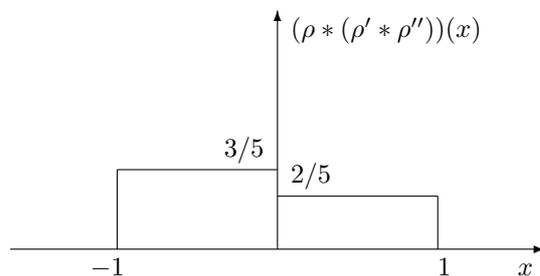
Here,  $N = 5/3$ , hence:

$$((\rho * \rho') * \rho'')(x) = \begin{cases} 2/5 & \text{if } -1 \leq x \leq 0, \\ 3/5 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise} \end{cases}$$



Similarly, we conclude that:

$$(\rho * (\rho' * \rho''))(x) = \begin{cases} 3/5 & \text{if } -1 \leq x \leq 0, \\ 2/5 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise} \end{cases}$$



Hence,  $(\rho * \rho') * \rho'' = \rho * (\rho' * \rho'')$ . So, the aggregation operation is indeed non-associative.

*Comment.* In terms of densities, the DS aggregation operation for probability distributions consists of two steps: taking the maximum of the values of the probability density functions and normalization. Because  $\max(a, b)$  is clearly an associative operation, non-associativity comes from normalization.

## 9.9 Enveloping aggregation operations for the case when CDFs come from expert estimates

In the previous subsections, we considered averaging aggregation operations. Let us now consider enveloping operations. These operations are necessary in case of variability and recommended in case of incertitude. We will first consider the case when all CDFs to be aggregated come from expert estimates.

For expert estimates, for every  $x$ , each expert provides an estimate  $F_i(x)$ . So, to aggregate the CDFs, we should combine, for each  $x$ , the corresponding  $n$  numbers  $F_1(x), \dots, F_n(x)$  into an estimate for  $F(x)$ . Since in this survey, we use p-bounds as a means of describing uncertainty, we want the resulting estimates to form a p-bound. For a p-bound, for each  $x$ , the corresponding estimate for  $F(x)$  is an interval. Thus, we want to combine  $n$  numerical values  $F_1(x), \dots, F_n(x)$  into an interval estimate for  $F(x)$ .

In Section 7, we described two enveloping methods of combining  $n$  numbers: the envelope method leads to an interval, and the mixture method leads to a probability distribution. Since we want to produce an interval, we have to use the envelope method, i.e., generate a p-bound

$$F_1(x) * \dots * F_n(x) = [\min(F_1(x), \dots, F_n(x)), \max(F_1(x), \dots, F_n(x))].$$

This formula makes perfect sense in case of variability. Indeed, variability means that all probability distributions  $F_i(x)$  are possible. This means that for every  $x$ , all  $n$  values  $F_1(x), \dots, F_n(x)$  are possible values of the probability  $F(x)$  (that the actual value  $X$  is  $\leq x$ ). As in the aggregation of numerical estimates (see Section 7), it is therefore reasonable to return, as a result of this aggregation, the smallest interval that combines these  $n$  values, i.e., the interval

$[\min(F_1(x), \dots, F_n(x)), \max(F_1(x), \dots, F_n(x))]$ . As a result, we get the above *envelope* operation which transforms CDFs into a p-bound.

As in the cases of aggregating real values and intervals, this envelope is equal to the weighted average with interval weights  $\mathbf{w}_i = [0, 1/n]$ . The envelope operation is idempotent, continuous, and computationally simple. The notions of commutativity, associativity, and averaging property are not formally applicable because we have defined this operation only for CDFs and the result of this operation is a p-bound which is not a CDF. However, when we actually extend this operation to p-bounds, we will see that the resulting operation is indeed commutative, associative, and satisfies the averaging property.

In the particular case when the CDFs are real numbers  $x_1, \dots, x_n$ , the resulting p-bound describes the envelope interval  $[\min(x_1, \dots, x_n), \max(x_1, \dots, x_n)]$ .

### 9.10 Enveloping aggregation operations for the case when CDFs come from measurements

As in the discussion of averaging CDFs coming from measurements, we can conclude that we should take, for every  $p$ , the corresponding  $n$  percentiles  $F_1^{-1}(p), \dots, F_n^{-1}(p)$ , and then apply an enveloping operation to these  $n$  values. As in the previous subsection, we can conclude that the envelope is the only reasonable operation for enveloping the corresponding numbers (percentiles). Therefore, for each  $p$ , we should take the envelope of these percentiles, i.e.,

$$[\min(F_1^{-1}(p), \dots, F_n^{-1}(p)), \max(F_1^{-1}(p), \dots, F_n^{-1}(p))].$$

Once can show that the resulting “ $x$ -envelope” p-bound is exactly the same as the above p-bound corresponding to the  $p$ -envelope.

### 9.11 Section’s conclusions

For aggregating several probability distributions  $F_1(x), \dots, F_n(x)$ , the selection of an appropriate aggregation operation depends on whether we are in the situation of incertitude, i.e., whether there is an actual (unknown) probability distribution  $F(x)$ , or whether we have variability, when several different probability distributions are possible in different situation, and thus, we need a p-bound to describe the resulting uncertainty.

When the distributions  $F_1(x), \dots, F_n(x)$  to be aggregated represent incertitude, traditionally, averaging aggregation techniques are used. The averaging aggregation depends on whether these CDFs come from expert estimates or from measurements. In this first case, we get aggregation corresponds to averaging probabilities ( $p$ -averaging), in the second case, it corresponds to averaging values of  $x$  ( $x$ -averaging).

Averaging operations “erase” the incertitude, so we recommend using enveloping operations that preserve and propagate the incertitude. Enveloping leads to the envelope p-bound  $[F^-(x), F^+(x)]$ , with  $F^-(x) =$

$\min(F_1(x), \dots, F_n(x))$  and  $F^+(x) = \max(F_1(x), \dots, F_n(x))$ . When the distributions  $F_1(x), \dots, F_n(x)$  to be aggregated represent variability, enveloping techniques are the only choice.

## 10 Aggregation of Uncertainty in Risk Analysis: General Case

### 10.1 Enclosures vs. estimates

How can we aggregate several p-bounds  $\mathbf{F}_1(x) = [F_1^-(x), F_1^+(x)], \dots, \mathbf{F}_n(x) = [F_n^-(x), F_n^+(x)]$  into a single p-bound? As in the aggregation of intervals and CDFs, we have several possible situations here, and the appropriate aggregation operation depends on the situation.

First, as in the case of intervals, we have two cases depending on:

- whether the corresponding p-bounds are *enclosures* for the actual CDFs or p-bounds, i.e., whether we are sure that the actual CDF or the actual CDFs (in case of variability) are within each of  $n$  given p-bounds,
- or whether the p-bounds are *estimates* which do not necessarily enclose the actual CDF (CDFs).

### 10.2 Enclosures lead to intersection

Let us first consider the case when each of  $n$  p-bounds  $\mathbf{F}_i$  is an enclosure, meaning that for every  $x$ , the actual (unknown) value of every possible CDF  $F(x)$  belongs to the interval  $[F_i^-(x), F_i^+(x)]$ . We want to generate an aggregated p-bound which would also serve as an enclosure for all possible CDFs  $F(x)$ .

The above p-bounds can be easily aggregated. Indeed, by definition of an intersection, knowing that  $F(x)$  belong to all  $n$  p-bounds is equivalent to knowing that  $F(x)$  belongs to the intersection of these  $n$  p-bounds. So, the natural aggregation operation  $*$  here is the *intersection*

$$\mathbf{F}_1 * \dots * \mathbf{F}_n = \mathbf{F},$$

where

$$\mathbf{F}(x) = [F^-(x), F^+(x)] = \mathbf{F}_1(x) \cap \dots \cap \mathbf{F}_n(x),$$

i.e.,

$$F^-(x) = \max(F_1^-(x), \dots, F_n^-(x));$$

$$F^+(x) = \min(F_1^+(x), \dots, F_n^+(x)).$$

Just like for intervals, this operation is idempotent, commutative, associative, satisfies the averaging property, and is easy to compute.

Just like for intervals, the only property this operation does not always satisfy is continuity. To be more precise, if all experts are correct, then the intersection is non-empty, and the corresponding aggregation operation is continuous. However, experts can sometimes be wrong, as a result, the intersection may turn out to be empty. What shall we do then?

If the intersection of all  $n$  p-bounds indeed turns out to be empty, this means that at least one of the p-bounds is wrong. Because these p-bounds come either from experts or from measurements, we have a certain trust in these p-bounds. We cannot avoid marking some of these p-bounds as erroneous, but it is natural to assume that as few of them as possible are erroneous. So, a natural thing to do is to start with an assumption that only one p-bound is wrong. So, we sequentially check all  $i$ 's from 1 to  $n$ . For each  $i$ , we check whether the intersection

$$\mathbf{I}_i(x) = \mathbf{F}_1(x) \cap \dots \cap \mathbf{F}_{i-1}(x) \cap \mathbf{F}_{i+1}(x) \cap \dots \cap \mathbf{F}_n(x)$$

of all p-bounds except for  $i$ th one is non-empty. If one of these intersections is actually non-empty, then, because we do not know which p-bound is actually erroneous, we should take, as the desired aggregated enclosure  $\mathbf{F}(x)$  for the unknown (actual) CDF  $F(x)$ , the *union* of all the non-empty intersections  $\mathbf{I}_i$ .

If all the intersections  $\mathbf{I}_i(x)$  are empty, this means that at least two p-bounds are erroneous. In this case, we need to check all the pairs  $i_1 < i_2$ . For each such pair, we check whether the intersection

$$\mathbf{I}_{i_1 i_2}(x) = \mathbf{F}_1(x) \cap \dots \cap \mathbf{F}_{i_1-1}(x) \cap \mathbf{F}_{i_1+1}(x) \cap \dots \cap \mathbf{F}_{i_2-1}(x) \cap \mathbf{F}_{i_2+1}(x) \cap \dots \cap \mathbf{F}_n(x)$$

of all p-bounds except for  $i_1$ th and  $i_2$ th ones is non-empty. If one of these intersections is actually non-empty, then, because we do not know which p-bounds are actually erroneous, we should take, as the desired aggregated enclosure  $\mathbf{F}(x)$  for the unknown (actual) CDF  $F(x)$ , the *union* of all the non-empty intersections  $\mathbf{I}_{i_1 i_2}$ .

If all the intersections  $\mathbf{I}_{i_1 i_2}(x)$  are empty, this means that at least three p-bounds are erroneous. In this case, we need to check all the triples  $i_1 < i_2 < i_3$ , etc.

### 10.3 Estimates: incertitude vs. variability, expert estimates vs. measurements

We have described how to aggregate enclosing p-bounds. When p-bounds are estimates, then the appropriate aggregation selection of an aggregation operation depends on whether we have incertitude or variability:

- We may have incertitude, when there is only one actual CDF, and a p-bound reflects incomplete knowledge about the CDF.
- Alternatively, we may have variability, when several different distributions are possible, and a p-bound combines the CDFs corresponding to different distributions.

In case of incertitude, we have two further subcases:

- One possibility is that the p-bounds come from *expert estimation*. In this case, each expert estimates the values of the CDF  $F(x)$  for each  $x$  and produces the interval  $[F_i^-(x), F_i^+(x)]$ . So, to aggregate the p-bounds, we should, for every  $x$ , combine the corresponding intervals.
- Another possibility is that p-bounds come from *measurements*. In this case, each p-bound  $F_i(x)$  is simply a “histogram” distribution representing interval measurement results. As in the case of CDFs, a natural way to extract these intervals from a p-bound  $[F_i^-(x), F_i^+(x)]$  is to consider, for every probability  $p \in [0, 1]$ , the corresponding percentile intervals  $[(F_i^+)^{-1}(p), (F_i^-)^{-1}(p)]$ . In this case, to aggregate the p-bounds, we should, for every  $p$ , aggregate these percentile intervals.

As in the cases of real numbers, intervals, and CDFs, in each of these cases, we can consider either more traditional averaging operations, or more recommended enveloping ones. Let us describe the resulting subcases one by one.

## 10.4 $p$ -averaging

### 10.4.1 Incertitude, averaging, p-bounds come from experts: $p$ -averaging

Let us first consider the case when the p-bounds come from experts, represent incertitude, and we apply averaging methods to aggregate them. In this case, for each  $x$ ,  $n$  experts provide  $n$  probability intervals  $[F_i^-(x), F_i^+(x)]$ , so we should use interval averaging techniques to aggregate these intervals. Because we are aggregating probabilities, we can call these techniques  $p$ -averaging. Let us briefly overview these methods and their properties.

### 10.4.2 Arithmetic $p$ -average: general description and properties

The simplest averaging operation is the arithmetic average, when

$$F^-(x) = \frac{F_1^-(x) + F_2^-(x)}{2}, \quad F^+(x) = \frac{F_1^+(x) + F_2^+(x)}{2},$$

of, for the general case of  $n \geq 2$ :

$$F^-(x) = \frac{F_1^-(x) + \dots + F_n^-(x)}{n}; \quad F^+(x) = \frac{F_1^+(x) + \dots + F_n^+(x)}{n}.$$

As in the case of CDFs, this operation can be naturally interpreted via the formula of full probability.

As in the cases of real numbers, intervals, and CDFs, this operation is idempotent, commutative, continuous, not associative, satisfies the averaging property, and is easy to compute.

### 10.4.3 Weighted arithmetic $p$ -average: general description and properties

If we have reasons to believe that some experts are more trustworthy than others, then (as we have mentioned when discussing aggregation of real numbers in Section 7) we should use *weighted* arithmetic  $p$ -average instead of the arithmetic  $p$ -average:

$$\begin{aligned} F^-(x) &= w_1 \cdot F_1^-(x) + \dots + w_n \cdot F_n^-(x); \\ F^+(x) &= w_1 \cdot F_1^+(x) + \dots + w_n \cdot F_n^+(x). \end{aligned}$$

The same methods as we outlined when discussing aggregation of real numbers can be used to determine these weights.

As in the case of CDFs, this operation can be naturally interpreted via the formula of full probability; here, the weight  $w_i$  becomes the probability that  $i$ th expert is right.

As in the cases of real numbers, intervals, and CDFs, this operation is idempotent, non-commutative, continuous, not associative, does not satisfy the averaging property, and is easy to compute.

### 10.4.4 Weighted arithmetic $p$ -average with interval weights

As in the cases of real numbers and intervals, we can consider weighted arithmetic  $p$ -average with interval weights. The algorithm for computing the weighted average was given when we described aggregation of real numbers.

This weighted average operation can be interpreted via a formula for full probability, in which, instead of knowing the exact probability  $P(i) = w_i$  that  $i$ th expert is correct, we know only the intervals  $\mathbf{w}_i$  of possible values of these probabilities.

In particular, when we have no prior information about the weights and therefore, take  $\mathbf{w}_i = [0, 1]$ , the aggregation results in the following  $p$ -bound

$$[\min(F_1^-(x), \dots, F_n^-(x)), \max(F_1^+(x), \dots, F_n^+(x))].$$

### 10.4.5 Geometric $p$ -average and generalized $p$ -average: general description and properties

When we discussed operations with CDFs in Section 9, we remarked that in many practical cases, the arithmetic average leads to counterintuitive conclusions: it is quite possible that three different experts estimate the same small probability as, correspondingly,  $\approx 10^{-5}$ ,  $\approx 10^{-6}$ , and  $\approx 10^{-7}$ , and we do not want the result of aggregation to always coincide with the largest estimate. So, in  $p$ -averaging, it also make sense to consider the geometric average and the generalized average. For the geometric average, the corresponding formulas are

$$F^-(x) = \sqrt[n]{F_1^-(x) \cdot \dots \cdot F_n^-(x)};$$

$$F^+(x) = \sqrt[n]{F_1^+(x) \cdot \dots \cdot F_n^+(x)}.$$

For the generalized average, we get

$$F^-(x) = \left( \frac{(F_1^-(x))^\alpha + \dots + (F_n^-(x))^\alpha}{n} \right)^{1/\alpha};$$

$$F^+(x) = \left( \frac{(F_1^+(x))^\alpha + \dots + (F_n^+(x))^\alpha}{n} \right)^{1/\alpha}.$$

As in the cases of real numbers, intervals, and CDFs, this operation is idempotent, commutative, continuous, not associative, satisfies the averaging property, and is easy to compute.

When  $\alpha \rightarrow -\infty$ , the generalized  $p$ -average tends to

$$[\min(F_1^-(x), \dots, F_n^-(x)), \min(F_1^+(x), \dots, F_n^+(x))];$$

when  $\alpha \rightarrow 0$ , it tends to geometric  $p$ -average; when  $\alpha \rightarrow \infty$ , it tends to

$$[\max(F_1^-(x), \dots, F_n^-(x)), \max(F_1^+(x), \dots, F_n^+(x))].$$

#### 10.4.6 Weighted geometric $p$ -average and generalized $p$ -average: general description and properties

As in the cases of real numbers, intervals, and CDFs, we can have weighted geometric  $p$ -average:

$$F^-(x) = (F_1^-(x))^{w_1} \cdot \dots \cdot (F_n^-(x))^{w_n};$$

$$F^+(x) = (F_1^+(x))^{w_1} \cdot \dots \cdot (F_n^+(x))^{w_n};$$

and weighted generalized  $p$ -average:

$$F^-(x) = (w_1 \cdot (F_1^-(x))^\alpha + \dots + w_n \cdot (F_n^-(x))^\alpha)^{1/\alpha};$$

$$F^+(x) = (w_1 \cdot (F_1^+(x))^\alpha + \dots + w_n \cdot (F_n^+(x))^\alpha)^{1/\alpha}.$$

As in the cases of real numbers, intervals, and CDFs, this operation is idempotent, non-commutative, continuous, non-associative, does not satisfy the averaging property, and is easy to compute.

#### 10.4.7 The use of copulas

As in the case of CDFs, if we know the copula  $C(x, y)$  that describes the dependence between different expert opinions, we can use these copulas to combine the probability intervals  $\mathbf{F}_1(x) = [F_1^-(x), F_1^+(x)]$  and  $\mathbf{F}_2 = [F_2^-(x), F_2^+(x)]$ . Because copulas are increasing functions, to find the interval of possible values

of  $C(F_1(x), F_2(x))$  when  $F_1(x) \in \mathbf{F}_1$  and  $F_2(x) \in \mathbf{F}_2(x)$ , it is sufficient to apply the copula operations to the smallest and largest values from these intervals. As a result, we get the following aggregation operation:

$$\mathbf{F}(x) = [C(F_1^-(x), F_2^-(x)), C(F_1^+(x), F_2^+(x))].$$

As we have mentioned in our description of the use of copulas in aggregating CDFs, it is desirable to combine copula techniques with other aggregation methods. This is one of the directions of our current research.

It is worth mentioning that when for p-bounds, when probabilities are known with an interval uncertainty, there are alternative definitions of independence and dependence. Different definitions may be more appropriate in different practical situations involving risk analysis; see, e.g., (Couso et al., 2000).

## 10.5 $x$ -averaging

### 10.5.1 Incertitude, averaging, p-bounds come from measurements: $x$ -averaging

Let us now consider the case when the p-bounds  $\mathbf{F}_i(x)$  come from measurements, represent incertitude, and we apply averaging methods to aggregate them. We already know, from Section 8, how to aggregate intervals, i.e., individual interval measurement results. Therefore, a natural way to aggregate the corresponding p-bounds is as follows:

- first, we extract the corresponding interval measurement results from the p-bounds;
- then, we combine (aggregate) the extracted interval measurement results into aggregated estimates;
- finally, we transform the resulting interval estimates into a new p-bound.

Because in this case, we are aggregating (averaging) intervals of values  $x$  of the measured quantity, the corresponding methods are called methods of *x-averaging*.

How can we extract interval measurement results from a p-bound? As in the case of CDFs, a natural way to do that is to take the percentile intervals  $[(F_i^+)^{-1}(p), (F_i^-)^{-1}(p)]$  corresponding to each p-bound  $[F_i^-(x), F_i^+(x)]$ .

As in the case of CDFs, if the measurement error is *small*, i.e., if all p-bounds are reasonably narrow and close to each other, it is reasonable to aggregate, for every  $p$ , the percentile interval  $[(F^+)^{-1}(p), (F^-)^{-1}(p)]$  corresponding to the aggregated p-bound can be obtained by aggregating the percentile intervals  $[(F_1^+)^{-1}(p), (F_1^-)^{-1}(p)], \dots, [(F_n^+)^{-1}(p), (F_n^-)^{-1}(p)]$ . In short, if measurement errors are small, then for every  $p \in [0, 1]$ , to get the corresponding percentile interval, we aggregate  $n$  percentiles intervals corresponding to p-bounds to be

aggregated. In geometric terms, if we draw CDFs in the usual way, with  $x$  horizontal and  $p$  vertical, this means that we do the averaging in each horizontal section.

If the measurement errors are not small, i.e., if the p-bounds to be aggregated are not close to each other and/or not small, then, as in the case of CDFs, a natural thing to do it to take a random variable  $x_1$  distributed according to some distribution from the p-bound  $\mathbf{F}_1(x)$ , another random variable  $x_2$  distributed according to some distribution from the p-bound  $\mathbf{F}_2(x)$ , aggregate these two random variables, and then take, as the aggregated p-bound, the p-bound describing all possible distributions resulting from this aggregation.

Let us discuss the corresponding aggregation methods in detail. We will start with the methods of horizontal aggregation corresponding to close p-bounds  $\mathbf{F}_i$ .

### 10.5.2 Arithmetic $x$ -average, general description and properties

In accordance with the above arguments, when p-bounds are close, we should aggregate, for each  $p$ , the percentile intervals corresponding to this particular  $p$ . In short, we must aggregate horizontally.

In particular, when we apply the simplest possible aggregation procedure – the arithmetic average – we conclude that for every level  $p$ , we must take an arithmetic average of the corresponding percentile intervals of the p-bound  $\mathbf{F}_i$ , i.e.,

$$(F^-)^{-1}(p) = \frac{(F_1^-)^{-1}(p) + \dots + (F_n^-)^{-1}(p)}{n};$$

$$(F^+)^{-1}(p) = \frac{(F_1^+)^{-1}(p) + \dots + (F_n^+)^{-1}(p)}{n};$$

As in the cases of real numbers, intervals, CDFs (and as for the arithmetic  $p$ -average), this operation is idempotent, commutative, continuous, not associative, satisfies the averaging property, and is easy to compute.

### 10.5.3 Weighted arithmetic $x$ -average

In this averaging, the new percentile interval is:

$$(F^-)^{-1}(p) = w_1 \cdot (F_1^-)^{-1}(p) + \dots + w_n \cdot (F_n^-)^{-1}(p);$$

$$(F^+)^{-1}(p) = w_1 \cdot (F_1^+)^{-1}(p) + \dots + w_n \cdot (F_n^+)^{-1}(p).$$

As in the cases of real numbers, intervals, CDFs (and as for  $p$ -averaging), this operation is idempotent, non-commutative, continuous, not associative, does not satisfy the averaging property, and is easy to compute.

We can also consider interval weights, by using the algorithm for aggregating with interval weights presented in the section for aggregating real numbers.

#### 10.5.4 Geometric and generalized $x$ -average

In this averaging, the new percentile interval is either

$$(F^-)^{-1}(p) = \sqrt[n]{(F_1^-)^{-1}(p) \cdot \dots \cdot (F_n^-)^{-1}(p)},$$

$$(F^+)^{-1}(p) = \sqrt[n]{(F_1^+)^{-1}(p) \cdot \dots \cdot (F_n^+)^{-1}(p)},$$

or

$$(F^-)^{-1}(p) = \left( \frac{((F_1^-)^{-1}(p))^\alpha + \dots + ((F_n^-)^{-1}(p))^\alpha}{n} \right)^{1/\alpha}.$$

As in the cases of real numbers, intervals, CDFs (and as in  $p$ -averaging), this operation is idempotent, commutative, continuous, not associative, satisfies the averaging property, and is easy to compute.

#### 10.5.5 Weighted geometric and generalized $x$ -average

In this aggregation, the new percentile interval is either

$$(F^-)^{-1}(p) = ((F_1^-)^{-1}(p))^{w_1} \cdot \dots \cdot ((F_n^-)^{-1}(p))^{w_n},$$

$$(F^+)^{-1}(p) = ((F_1^+)^{-1}(p))^{w_1} \cdot \dots \cdot ((F_n^+)^{-1}(p))^{w_n},$$

or

$$(F^-)^{-1}(p) = (w_1 \cdot ((F_1^-)^{-1}(p))^\alpha + \dots + w_n \cdot ((F_n^-)^{-1}(p))^\alpha)^{1/\alpha}.$$

As in the cases of real numbers, intervals, CDFs (and as in  $p$ -averaging), this operation is idempotent, non-commutative, continuous, not associative, does not satisfy the averaging property, and is easy to compute. We can also consider interval weights.

#### 10.5.6 Averaging drastically different and/or wide p-bounds: arithmetic $x$ -average

Let us now consider the case of drastically different and/or wide p-bounds  $\mathbf{F}_1(x), \dots, \mathbf{F}_n(x)$ . As we have mentioned, to get the aggregation of such p-bounds, we do the following:

- First, we select a numerical aggregation operation  $*$  (e.g.,  $x*y = (x+y)/2$ ).
- Next, we solve the following problem:
  - we have a variable  $x_1$  which is distributed according to some distribution from a p-bound  $\mathbf{F}_1(x)$ ;
  - we have a variable  $x_2$  which is distributed according to some distribution from a p-bound  $\mathbf{F}_2(x)$ ;

- ...
- we have a variable  $x_n$  which is distributed according to some distribution from a p-bound  $\mathbf{F}_n(x)$ ;
- we want to find a p-bound which corresponds to all resulting distributions of the quantity  $x = x_1 * \dots * x_n$ .

This is exactly the problem that we mentioned when we described indirect measurement/estimation, and all methods for solving that problem (Ferson et al., 2001), (Ramas, 2001) are applicable here as well.

Because CDFs are particular cases of p-bounds, the non-idempotence CDF example from Section 9 shows that even in the simple case of arithmetic average and independent  $x_i$ , the corresponding aggregation operation is not idempotent. In general, for this case (arithmetic average and independence), this aggregation operation is commutative, continuous, not associative, satisfies the averaging property, and is computable.

As in the case of CDFs, we should not worry about non-idempotence  $\mathbf{F}_1 * \mathbf{F}_1 \neq \mathbf{F}_1$ , because this operation is mainly justified when the p-bounds are drastically different, so we should not apply it to the case when  $\mathbf{F}_1 = \mathbf{F}_2$ .

## 10.6 Enveloping

What are reasonable enveloping operations for aggregating  $n$  p-bounds  $\mathbf{F}_1(x), \dots, \mathbf{F}_n(x)$ ? These p-bounds come either from expert estimating the probabilities, or from measurements. If p-bounds to be aggregated come from experts, then, for each  $x$ , we have  $n$  intervals  $\mathbf{F}_i(x)$ , and we want to aggregate these  $n$  intervals into a single estimate for  $F(x)$ . We want these estimates to form a p-bound, hence, for every  $x$ , we need an interval  $\mathbf{F}(x)$ . The only enveloping operation that combines  $n$  intervals into a single interval is the *envelope* operation, for which

$$\mathbf{F}(x) = [\min(F_1^-(x), \dots, F_n^-(x)), \max(F_1^+(x), \dots, F_n^+(x))].$$

Similarly, when p-bounds come from measurements, enveloping means applying the envelope operation to the corresponding percentile intervals  $\mathbf{F}^{-1}(p)$ . One can show that the resulting enveloping aggregation is the same as enveloping the intervals  $\mathbf{F}_i(x)$ .

This operation makes perfect sense for the case of variability. In this case, we are given  $n$  p-bounds  $\mathbf{F}_i$ , and we know that the actual CDF  $F(x)$  can be within each of these  $n$  p-bounds. Thus, CDFs within  $\mathbf{F}_1$  are possible, and CDFs within  $\mathbf{F}_2$  are possible, etc. Hence, we can have the CDF within any of these  $n$  p-bounds. We want to aggregate this information. We know that CDFs from  $\mathbf{F}_1$  are possible, we know that CDFs from  $\mathbf{F}_2$  are possible, etc. Thus, we can conclude that all CDFs from the union of these  $n$  p-bounds are possible. The union is not necessarily a p-bound itself: e.g., the union of the probability

intervals  $[0, 1/3]$  and  $[2/3, 1]$  is not an interval. Because we want to describe the aggregation result as a p-bound, it is natural to describe it as the smallest p-bound that contains this union, i.e., their *envelope*

$$\mathbf{F}(x) = [\min(F_1^-(x), \dots, F_n^-(x)), \max(F_1^+(x), \dots, F_n^+(x))].$$

As in the cases of aggregating real numbers, intervals, and CDFs, the envelope coincides with the weighted average with interval weights  $\mathbf{w}_i = [0, 1]$ . The envelope operation is idempotent, commutative, continuous, associative, satisfies averaging property, and is computationally simple.

## 10.7 Dempster-Shafer approach to aggregating p-bounds

### 10.7.1 Formulas

When we discussed aggregation of CDFs, we have mentioned that every CDF can be represented as a Dempster-Shafer knowledge base, with percentiles values  $F_i^{-1}(p)$  (or, to be more precise, with small intervals around percentile values), as focal elements.

For a p-bound, we do not have a single percentile value, we have an *interval*  $[(F_i^+)^{-1}(p), (F_i^-)^{-1}(p)]$  of possible percentile values. As in the case of CDFs, every p-bound can be represented as Dempster-Shafer knowledge base, with the corresponding percentile intervals as focal elements. If we apply Dempster-Shafer combination rule to combine these percentiles, we get the following formula for the result  $[F^-(x), F^+(x)]$  of aggregating two p-bounds  $[F_1^-(x), F_1^+(x)]$  and  $[F_2^-(x), F_2^+(x)]$ . First, we compute pre-normalized ‘‘CDF’’s as follows:

$$\tilde{F}^-(x) = \int \rho_1^-(x) \cdot (F_2^+(x) - F_2^-(x)) dx + \int \rho_2^-(x) \cdot (F_1^+(x) - F_1^-(x)) dx;$$

$$\tilde{F}^+(x) = \int \rho_1^+(x) \cdot (F_2^+(x) - F_2^-(x)) dx + \int \rho_2^+(x) \cdot (F_1^+(x) - F_1^-(x)) dx.$$

where  $\rho_i^-(x)$  and  $\rho_i^+(x)$  are probability density functions corresponding to CDFs  $F_i^-(x)$  and  $F_i^+(x)$ . We can rewrite these formulas exclusively in terms of CDFs if we use the notion of a Stieltjes integral  $\int f(x) dF(x)$  (which is equivalent to  $\int f(x) \cdot \rho(x) dx$ ):

$$\tilde{F}^-(x) = \int (F_2^+(x) - F_2^-(x)) dF_1^-(x) + \int (F_1^+(x) - F_1^-(x)) dF_2^-(x);$$

$$\tilde{F}^+(x) = \int (F_2^+(x) - F_2^-(x)) dF_1^+(x) + \int (F_1^+(x) - F_1^-(x)) dF_2^+(x).$$

After we compute each ‘‘CDF’’s  $\tilde{F}(x)$ , we normalize it by dividing by the normalizing constant, which happens to be the value  $\tilde{F}^+(\infty)$  of the pre-normalized

“CDF”  $\tilde{F}^+(x)$  when  $x \rightarrow \infty$ :

$$F^-(x) = \frac{\tilde{F}^-(x)}{\tilde{F}^+(\infty)}; \quad F^+(x) = \frac{\tilde{F}^+(x)}{\tilde{F}^+(\infty)}.$$

These formulas can be naturally generalized to the case when we aggregate an arbitrary number  $n$  of p-bounds. In this case, we get:

$$\begin{aligned} \tilde{F}^-(x) &= \sum_{i=1}^n \int \rho_i^-(x) \cdot \prod_{j \neq i} (F_j^+(x) - F_j^-(x)) dx; \\ \tilde{F}^+(x) &= \sum_{i=1}^n \int \rho_i^+(x) \cdot \prod_{j \neq i} (F_j^+(x) - F_j^-(x)) dx. \end{aligned}$$

Alternatively:

$$\begin{aligned} \tilde{F}^-(x) &= \sum_{i=1}^n \int \prod_{j \neq i} (F_j^+(x) - F_j^-(x)) dF_i^-(x); \\ \tilde{F}^+(x) &= \sum_{i=1}^n \int \prod_{j \neq i} (F_j^+(x) - F_j^-(x)) dF_i^+(x). \end{aligned}$$

This aggregation operation sounds somewhat complex, but it leads to a simple formula for the interval width  $w(x) = F^+(x) - F^-(x)$  of the resulting p-bound:

$$w(x) = k \cdot w_1(x) \cdot \dots \cdot w_n(x),$$

where  $k$  is a normalizing constant, and  $w_i(x) = F_i^+(x) - F_i^-(x)$  are the widths of the p-bounds to be aggregated.

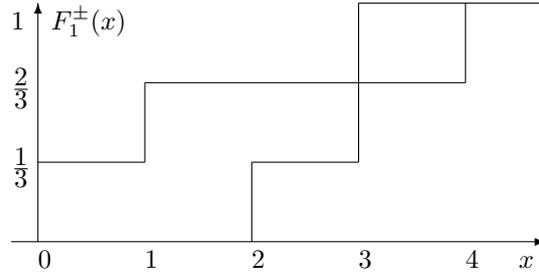
### 10.7.2 Properties and examples

Let us start with checking idempotence. When we combine a p-bound  $\mathbf{F}_1(x)$  with itself, we get a new p-bound with the width  $w(x) = k \cdot w_1(x)^2$ . The only way for the p-bound to stay the same is when  $w(x) = k \cdot w_1^2(x) = w_1(x)$ , i.e., when  $w_1(x) \equiv \text{const}$ . For CDFs, a similar operation was idempotent, because a CDF can be viewed as a CDF with a contact (0) width. However, as soon as the width stops being constant, we lose the idempotence property.

Let us give a simple example of why Dempster’s combination rule is not idempotent. Let us take, as the aggregated p-bound, a DS knowledge base with three focal elements  $\mathbf{x}_1 = [0, 2]$ ,  $\mathbf{x}_2 = [1, 3]$ , and  $\mathbf{x}_3 = [3, 4]$ , to each of which we assign the same mass  $p_1 = p_2 = p_3 = 1/3$ . This DS knowledge base corresponds to the following CDF:

$$F_1^-(x) = \begin{cases} 0 & \text{if } x < 2, \\ 1/3 & \text{if } 2 \leq x < 3, \\ 2/3 & \text{if } 3 \leq x < 4, \\ 1 & \text{if } x \geq 4 \end{cases}$$

$$F_1^+(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1/3 & \text{if } 0 \leq x < 1, \\ 2/3 & \text{if } 1 \leq x < 3, \\ 1 & \text{if } x \geq 3 \end{cases}$$

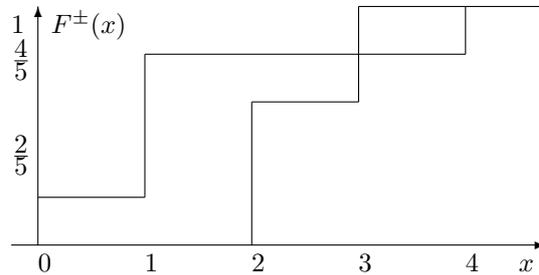


In accordance with the Dempster-Shafer combination rule, we take all intervals  $\mathbf{x}_i$  from the first knowledge case, all intervals  $\mathbf{y}_j$  from the second knowledge base (which in this case is the same, i.e.,  $\mathbf{y}_j = \mathbf{x}_j$ ), and assign the mass  $p_i \cdot q_j$  (in this case,  $1/9$ ) to all non-empty intersections. Because these masses do not add up to 1, we normalize them so that they will.

In this case, the following intersections are non-empty:  $\mathbf{x}_1 \cap \mathbf{y}_1 = [0, 2]$ ;  $\mathbf{x}_1 \cap \mathbf{y}_2 = [1, 2]$ ;  $\mathbf{x}_2 \cap \mathbf{y}_1 = [1, 2]$ ;  $\mathbf{x}_2 \cap \mathbf{y}_2 = [1, 3]$ ;  $\mathbf{x}_3 \cap \mathbf{y}_3 = [3, 4]$ . These five intervals get the same mass, so after normalization, they are each assigned the same mass  $1/5$ . The resulting p-bounds are as follows:

$$F^-(x) = \begin{cases} 0 & \text{if } x < 2, \\ 3/5 & \text{if } 2 \leq x < 3, \\ 4/5 & \text{if } 3 \leq x < 4, \\ 1 & \text{if } x \geq 4 \end{cases}$$

$$F^+(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1/5 & \text{if } 0 \leq x < 1, \\ 4/5 & \text{if } 1 \leq x < 3, \\ 1 & \text{if } x \geq 3 \end{cases}$$



On this example, we see that the width of the aggregated p-bound is indeed proportional to the square of the original one:

- In the original p-bound, the width was  $2/3$  on  $[1, 2]$  and twice smaller ( $1/3$ ) elsewhere on  $[0, 4]$ .
- In the aggregated p-bound, the weight is  $4/5$  on  $[1, 2]$ , and four times smaller elsewhere on  $[0, 4]$ .

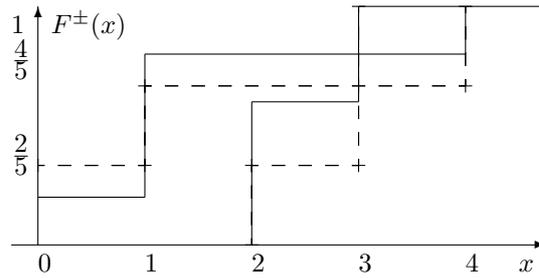
Comparing the original p-bound with the aggregated one, we can see that not only the aggregated p-bound is different: it is neither enclosed in the original one, not enclosing the original one. Indeed:

- for  $x \in (0, 1)$ , we have:

$$[F^-(x), F^+(x)] = \left[0, \frac{1}{5}\right] \subset [F_1^-(x), F_1^+(x)] = \left[0, \frac{1}{3}\right].$$

- On the other hand, for  $x \in (1, 2)$ , we have:

$$[F^-(x), F^+(x)] = \left[0, \frac{4}{5}\right] \supset [F_1^-(x), F_1^+(x)] = \left[0, \frac{2}{3}\right].$$



Summarizing: this operation is not idempotent. It is commutative, continuous, rather easy to compute, but (as in the case of CDFs) not associative.

## 10.8 Section's conclusions

When each of the p-bounds

$$\mathbf{F}_1(x) = [F_1^-(x), F_1^+(x)], \dots, \mathbf{F}_n(x) = [F_n^-(x), F_n^+(x)]$$

to be aggregated is an enclosure, i.e., if we are sure that the actual CDF (CDFs) are within each of these  $n$  p-bounds, then a natural aggregation operation is the intersection

$$\mathbf{F}(x) = [\max(F_1^-(x), \dots, F_n^-(x)), \min(F_1^+(x), \dots, F_n^+(x))].$$

When p-bounds are not enclosures but estimates which do not necessarily enclose the actual CDF (CDFs), then we can use either averaging or enveloping operations.

When the p-bounds  $F_1(x), \dots, F_n(x)$  to be aggregated represent incertitude, averaging depends on whether these CDFs come from expert estimates or from measurements. In this first case, we get aggregation corresponds to averaging probability intervals  $[F_i^-(x), F_i^+(x)]$  ( $p$ -averaging), in the second case, it corresponds to averaging the corresponding intervals of  $x$  ( $x$ -averaging).

Averaging operations “erase” the incertitude, so we recommend using enveloping operations that preserve and propagate the incertitude. Enveloping leads to the envelope p-bound

$$\mathbf{F}(x) = [\min(F_1^-(x), \dots, F_n^-(x)), \max(F_1^+(x), \dots, F_n^+(x))].$$

When the estimates  $\mathbf{F}_1(x), \dots, \mathbf{F}_n(x)$  to be aggregated represent variability, enveloping techniques are the only choice.

## References

- Appriou, A. (1998). Uncertain Data Aggregation in Classification and Tracking Processes. Aggregation and Fusion of Imperfect Information. B. Bouchon-Meunier, Physica-Verlag Heidelberg.
- Ayyub, B. M. (2001). Elicitation of Expert Opinions for Uncertainty and Risks, CRC Press.
- Baldwin, J. F. (1994). Mass assignments and fuzzy sets for fuzzy databases. Advances in the Dempster-Shafer Theory of Evidence, John Wiley & Sons, Inc.
- Baldwin, J. F. (1994a). A calculus for mass assignments in evidential reasoning. Advances in the Dempster-Shafer Theory of Evidence, John Wiley & Sons, Inc.
- Berleant, D. and B. J. Kuipers (1997). "Qualitative and quantitative simulation: bridging the gap." Artificial Intelligence 95(2): 215–255.
- Bezdek, J. C., D. Dubois, et al. (1999). Fuzzy Sets in Approximate Reasoning and Information Systems, Kluwer Academic Publishers.
- Bezdek, J. C., J. Keller, et al. (1999a). Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Toulouse, France, Kluwer Academic Publishers.
- Bloch, I. and H. Maitre (1998). Fusion of Image Information under Imprecision. Aggregation and Fusion of Imperfect Information. B. Bouchon-Meunier, Physica-Verlag Heidelberg.
- Caselton, W. F. and W. Luo (1992). "Decision Making With Imprecise Probabilities: Dempster-Shafer Theory and Application." Water Resources Research 28(12): 3071–3083.
- Chang, S. L. (1974). On Risk and Decision Making in a Fuzzy Environment. U.S.-Japan Seminar on Fuzzy Sets and Their Applications, University of California, Berkeley, 1974, Academic Press, Inc.
- Chateauneuf, A. (1994). Combination of compatible belief functions and relation of specificity. Advances in the Dempster-Shafer Theory of Evidence, John Wiley & Sons, Inc.
- Clemen, R. T. and R. L. Winkler (1999). "Combining probability distributions from experts in risk analysis". Risk Analysis 19(2): 187–203.

- Cooper, J. A., S. Ferson, and L. R. Ginzburg (1996). “Hybrid processing of stochastic and subjective uncertainty data”. *Risk Analysis* 16: 785–791.
- Couso, I., S. Moral, et al. (2000). “A survey of concepts of independence for imprecise probabilities.” *Risk Decision and Policy* 5: 165–181.
- Cox, L. A., Jr. (1999). “Internal Dose, Uncertainty Analysis, and Complexity of Risk Models.” *Environment International* 25(6/7): 841–852.
- de Cooman, G., T. L. Fine, et al. (2001). *ISIPTA '01. Second International Symposium on Imprecise Probabilities and Their Applications*, Robert Purcell Community Center of Cornell University, Ithaca, NY, USA, Shaker Publishing BV.
- Dubois, D. and H. Prade (1994). Focusing versus updating in belief function theory. *Advances in the Dempster-Shafer Theory of Evidence*, John Wiley & sons, Inc.
- Dubois, D. and H. Prade (2000). *Fundamentals of Fuzzy Sets*, Kluwer Academic Publishers.
- Dubois, D., H. Prade, et al., Eds. (1997). *Fuzzy Information Engineering: A Guided Tour of Applications*. New York, John Wiley & Sons, Inc.
- Ferson, S., L. Ginzburg, and R. Akçakaya (2001). “Whereof one cannot speak: when input distributions are unknown”. *Risk analysis* (to appear).
- Fodor, J. and R. R. Yager (2000). Fuzzy Set-Theoretic Operators and Quantifiers. *Fundamentals of Fuzzy Sets*. D. Dubois and H. Prade, Kluwer Academic Publishers.
- Fox, D., M. Schmidt, M. Koshelev, V. Kreinovich, L. Longpré, and J. Kuhn (1998). “We must choose the simplest physical theory: Levin-Li-Vitányi theorem and its potential physical applications”. In: G. J. Erickson, J. T. Rychert, and C. R. Smith (eds.), *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht, 238–251.
- Galambos, J. (1978). *The asymptotic theory of extreme order statistics*. Wiley, N.Y.
- Gebhardt, J. and R. Kruse (1998). Parallel Combination of Information Sources. *Handbook of Defeasible Reasoning and Uncertainty Management Systems*. D. M. Gabbay and P. Smets. Dordrecht, The Netherlands, Kluwe Academic Publishers. 3: 393–439.

- Goodman, I. R., R. P. S. Mahler, et al. (1997). *Mathematics of Data Fusion*. Dordrecht, The Netherlands, Kluwer Academic Publishers.
- Grabisch, M., S. A. Orlovski, et al. (1998). *Fuzzy Aggregation of Numerical Preferences*. *Fuzzy Sets in Decision Analysis, Operations Research and Statistics*. D. Dubois and H. Prade, Kluwer Academic Publishers.
- Halpern, J. Y. and R. Fagin (1992). “Two views of belief: belief as generalized probability and belief as evidence.” *Artificial Intelligence* 54: 275–317.
- Hammer, R., M. Hocks, U. Kulisch, D. Ratz (1993). *Numerical toolbox for verified computing. I. Basic numerical problems*. Springer Verlag, Heidelberg, N.Y.
- Hansen, E. R. (1992). *Global optimization using interval analysis*. Marcel Dekker, N.Y.
- Hora, S. C. and R. L. Iman (1989). “Expert Opinion in Risk Analysis: The NUREG-1150 Methodology.” *Nuclear Science and Engineering* 102: 323–331.
- Interval (2001). Interval computations website  
<http://www.cs.utep.edu/interval-comp>
- Jaffray, J.-Y. (1994). *Dynamic decision making with belief functions*. *Advances in the Dempster-Shafer Theory of Evidence*, John Wiley & Sons, Inc.
- Jouini, M. N. and R. T. Clemen (1996). “Copula models for aggregating expert opinions”. *Operations Research*, 44(3): 444–457.
- Jousselme, A.-L., D. Grenier, et al. (2001). “A new distance between two bodies of evidence.” *Information Fusion* 2: 91–101.
- Kacprzyk, J. and M. Fedrizzi (1990). *Multiperson Decision Making Models Using Fuzzy Sets and Possibility Theory*, Kluwer Academic Publishers.
- Kearfott, R. B. (1996). *Rigorous global search: continuous problems*. Kluwer, Dordrecht.
- Kearfott, R. B. and V. Kreinovich (1996). *Applications of Interval Computations*. Dordrecht, The Netherlands, Kluwer Academic Publishers.
- Kelman, A. and R. R. Yager (1998). *Using Priorities in Aggregation Connectives*. *Aggregation and Fusion of Imperfect Information*. B. Bouchon-Meunier, Physica-Verlag Heidelberg.

- Klir, G. and B. Yuan (1995). *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall, Upper Saddle River, NJ.
- Kosko, B. (1986). "Fuzzy Knowledge Combination." *International Journal of Intelligent Systems* 1: 293–320.
- Kreinovich, V. (1997). *Random Sets Unify, Explain, and Aid Known Uncertainty Methods in Expert Systems. Random Sets, Theory and Applications*. A. Friedman and R. Gulliver. New York, Springer. 97.
- Kreinovich, V., A. Bernat, et al. (1994). Monte-Carlo methods make Dempster-Shafer formalism feasible. *Advances in the Dempster-Shafer Theory of Evidence*, John Wiley & Sons, Inc.
- Kreinovich, V., A. Lakeyev, J. Rohn, and P. Kahl (1997). "Computational complexity and feasibility of data processing and interval computations". Kluwer, Dordrecht, 1997.
- Kruse, R. and F. Klawonn (1994). Mass distributions on L-fuzzy sets and families of frames of discernment. *Advances in the Dempster-Shafer Theory of Evidence*, John Wiley & Sons, Inc.
- Kuznetsov, V. P. (1991). *Interval statistical models*. Moscow, Radio i Svyaz Publ. (in Russian).
- Levi, I. (2000). "Imprecise and indeterminate probabilities." *Risk Decision and Policy* 5: 111–122.
- Li, M. and P. Vitányi (1997). *An Introduction to Kolmogorov Complexity and its Applications* Springer-Verlag, N.Y.
- Luo, W. B. and B. Caselton (1997). "Using Dempster-Shafer Theory to Represent Climate Change Uncertainties." *Journal of Environmental Management* 49: 73–93.
- Mahler, R. P. S. (1995). "Combining ambiguous evidence with respect to ambiguous a priori knowledge. Part II: Fuzzy Logic." *Fuzzy Sets and Systems* 75: 319–354.
- Milner, P. M. (1970). *Physiological psychology*. Holt, N.Y.
- Moore, R. (1979). *Methods and Applications of Interval Analysis*. SIAM, Philadelphia.
- Nelsen, R. B. (1999). *Introduction to copulas*. Springer Verlag, New York.

Nguyen, H. T. and V. Kreinovich (1996). “Nested Intervals and Sets: Concepts, Relations to Fuzzy Sets, and Applications”. In: R. B. Kearfott and V. Kreinovich (eds.), *Applications of Interval Computations*, Kluwer, Dordrecht, 245–290.

Nguyen, H. T. and V. Kreinovich (1997). *Applications of continuous mathematics to computer science*. Kluwer, Dordrecht.

Nguyen, H. T. and V. Kreinovich (1997a). “Kolmogorov’s Theorem and its impact on soft computing”, In: Yager, R. R. and J. Kacprzyk. *The Ordered Weighted Averaging Operators: Theory and Applications*, Kluwer, Boston, MA, 1–17.

Nguyen, H. T., V. Kreinovich, and Q. Zuo (1997). “Interval-valued degrees of belief: applications of interval computations to expert systems and intelligent control”. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems (IJUFKS)* 5(3): 317–358.

Nguyen, H. T. and M. Sugeno (1998). *Fuzzy Systems Modeling and Control*, Kluwer Academic Publishers.

Nguyen, H. T. and E. A. Walker (1999). *First Course in Fuzzy Logic*. CRC Press, Boca Raton, FL.

Ramas (2001). Ramas website <http://www.ramas.com>

Ross, T. J., B. Wu, and V. Kreinovich (2000). “Optimal Elimination of Inconsistency in Expert Knowledge: Formulation of the Problem, Fast Algorithms”. *Proc. of International Conference on Intelligent Technologies*, Bangkok, Thailand, December 13–15, 2000, 450–458.

Saffiotti, A. (1994). *Issues of knowledge representation in Dempster-Shafer’s theory*. *Advances in the Dempster-Shafer Theory of Evidence*, John Wiley & Sons, Inc.

Schubert, J. (1994). *Cluster-based specification techniques in Dempster-Shafer theory for an evidential intelligence analysis of multiple target tracks*. Ph.D. Dissertation, Royal Institute of Technology, Department of Numerical Analysis and Computer Science, Stockholm, Sweden.

Shafer, G. (1984). *A Mathematical Theory of Evidence*. *The AI Magazine*: 81–83.

Shafer, G. (1986). “The Combination of Evidence.” *International Journal of Intelligent Systems* 1: 155–179.

- Slowinski, R. (1998). *Fuzzy Sets in Decision Analysis, Operations Research and Statistics*, Kluwer Academic Publishers.
- Spies, M. (1994). Evidential reasoning with conditional events. *Advances in the Dempster-Shafer Theory of Evidence*, John Wiley & Sons, Inc.
- Srivastava, R., P. (1997). "Audit decisions using belief functions: a review." *Control and Cybernetics* 26(2): 135–160.
- Srivastava, R. P., P. P. Shenoy, et al. (1995). "Propagating Belief Functions in AND-Trees." *International Journal of Intelligent Systems* 10: 647–664.
- Strat, T. M. (1984). Continuous Belief Functions for Evidential Reasoning. *Proceedings of the National Conference on Artificial Intelligence*, University of Texas at Austin.
- Suppes, P., D. M. Krantz, R. D. Luce, and A. Tversky (1989). *Foundations of measurement*. Vol. I–III, Academic Press, San Diego, CA.
- Trejo, R. and V. Kreinovich (2001). "Error Estimations for Indirect Measurements: Randomized vs. Deterministic Algorithms For 'Black-Box' Programs". In: S. Rajasekaran, P. Pardalos, J. Reif, and J. Rolim (eds.), *Handbook on Randomized Computing*, Kluwer, 673–729.
- Tsoukalas, L. H. and R. E. Uhrig (1997). *Fuzzy and Neural Approaches in Engineering*, John Wiley & Sons, Inc.
- Uhrig, R. E. and L. H. Tsoukalas (1999). "Soft Computing Technologies in Nuclear Engineering Applications." *Progress in Nuclear Energy* 34(1): 13–75.
- Wadsworth, H. M., ed. (1990). *Handbook of statistical methods for engineers and scientists*, McGraw-Hill Publishing Co., N.Y.
- Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman and Hall, N.Y.
- Walley, P. and G. de Cooman (2001). "A behavioral model for linguistic uncertainty." *Information Sciences* 134: 1–37.
- Whalen, T. (1994). Interval probabilities induced by decision problems. *Advances in the Dempster-Shafer Theory of Evidence*, John Wiley & Sons, Inc.
- Yager, R. R. (1983). "Hedging in the Combination of Evidence." *Journal of Information & Optimization Sciences* 4(1): 73–81.

- Yager, R. R. (1985). "On the Relationship of Methods of Aggregating Evidence in Expert Systems." *Cybernetics and Systems: An International Journal* 16: 1–21.
- Yager, R. (1987). "On the Dempster-Shafer Framework and New Combination Rules." *Information Sciences* 41: 93–137.
- Yager, R. R. (1988). "On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decisionmaking." *IEEE Transactions on Systems, Man, and Cybernetics* 18(1): 183–190.
- Yager, R. R. (2001). "Dempster-Shafer Belief Structures with Interval Valued Focal Weights." *International Journal of Intelligent Systems* 16: 497–512.
- Yager, R. R., J. Kacprzyk, et al., Eds. (1994). *Advances in the Dempster-Shafer Theory of Evidence*, John Wiley & Sons, Inc.
- Yager, R. R. and J. Kacprzyk (1997). *The Ordered Weighted Averaging Operators: Theory and Applications*, Kluwer, Boston, MA, 1997.
- Yager, R. and A. Kelman (1996). "Fusion of Fuzzy Information." *International Journal of Approximate Reasoning* 15: 93–122.
- Zadeh, L. A. (1965). "Fuzzy Sets." *Information and Control* 8: 338–353.
- Zadeh, L. A. (1973). "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes." *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3(1): 28–44.
- Zadeh, L. A. (1986). A Simple View of the Dempster-Shafer Theory of Evidence and its Implication for the Rule of Combination. *The AI Magazine*: 85–90.
- Zhang, L. (1994). Representation, independence, and combination of evidence in the Dempster-Shafer theory. *Advances in the Dempster-Shafer Theory of Evidence*. R. R. Yager, J. Kacprzyk and M. Fedrizzi, John Wiley & Sons, Inc.
- Zimmerman, H.-J. (1999). *Practical Applications of Fuzzy Technologies*, Kluwer Academic Publishers.