PROCEEDINGS OF THE 3RD INTERNATIONAL WORKSHOP ON RELIABLE ENGINEERING COMPUTING

NSF WORKSHOP ON IMPRECISE PROBABILITY IN ENGINEERING ANALYSIS AND DESIGN

FEBRUARY 20-22, 2008 | SAVANNAH, GEORGIA USA

EDITORS
Rafi L. Muhanna
Robert L. Mullen

3RD INTERNATIONAL WORKSHOP ON RELIABLE ENGINEERING COMPUTING
NSF WORKSHOP ON IMPRECISE PROBABILITY IN ENGINEERING ANALYSIS & DESIGN

PROCEEDINGS: Workshop Organization

EDITORS
Rafi L. Muhanna, Georgia Institute of Technology
Robert L. Mullen, Case Western Reserve University

WORKSHOP SPONSORS
National Science Foundation
Sun Microsystems
The Society for Imprecise Probability: Theories and Applications
CASE (Computer Aided Structural Engineering) Center (GT STRUDL)
Georgia Institute of Technology

HONORARY WORKSHOP CHAIR
Ivo Babuška

WORKSHOP CHAIR
Rafi L. Muhanna, Georgia Institute of Technology

WORKSHOP CO-CHAIR
Robert L. Mullen, Case Western Reserve University

WORKSHOP SCIENTIFIC COMMITTEE
Michael Beer, National University of Singapore
Daniel Berleant, Iowa State University
George Corliss, Marquette University
William Edmonson, North Carolina State University
Scott Ferson, Applied Biomathematics
Raphael Haftka, University of Florida
Baker Kearfott, University of Louisiana at Lafayette
Vladik Kreinovich, University of Texas at El Paso
Mehdi Modares, Tufts University
Bernd Möller, Dresden University of Technology
Zissimos Mourelatos, Oakland University
Arnold Neumaier, University of Vienna, Austria
Efstratios Nikolaidis, University of Toledo
Andrezej Pownuk, University of Texas at El Paso
Sigfried Rump, Technical University of Hamburg
Pol Spanos, Rice University
Mark Stadtherr, University of Notre Dame
William Walster, Consultant

LOCAL WORKSHOP ORGANIZING COMMITTEE
Natalie Cosner
Kimberly Gaither
Jillison Parks
David Tucker

PROCEEDINGS: Table of Contents

PROCEEDINGS: Table of Contents

PROCEEDINGS: Preface

The Center for Reliable Engineering Computing (REC) at Georgia Tech Savannah has, as part of its mission, organized several international workshops that involve the investigation and advancement of different aspects of reliable engineering computing. This NSF workshop focuses on Imprecise Probability in Engineering Analysis and Design.

Probability-based methods for uncertainty treatment have been under development for about 50 years. While there has been much progress over that time, there remains a lack of widespread use of probabilistic methods by designers. The difficulty of acquiring the needed information, specifically the Probability Density Functions (PDF's) for risk based design, and the lack of viable engineering tools allowing for imprecise or incomplete information to be employed are among the main difficulties with the methodology of probability-based design approaches. Similar concerns arise when dealing with utility. In many practical cases, a complete ranking over all rewards is unrealistic. Imprecise utility aims to represent and reason with such incomplete preferences over rewards.

The objective of this workshop is to bring together researchers from various engineering fields as well as from mathematics and computer science to share, discuss and lay ground for the development of novel methods and tools for ensuring reliability of engineering models with incomplete information. In addition, the workshop is looking for integrating the individual advancement of the various disciplines into a general approach of imprecise probabilistic methodology for engineering analysis and design, allowing smooth transition between probabilistic and non-probabilistic approaches.

The topics of the workshop include:
1. Uncertainty modeling with incomplete information
2. Analysis and design of engineering systems with imprecise parameters
3. Design-based decision making under imprecise information

While some aspects of this workshop's focus are included in conferences on general numerical methods, computer science, and engineering, to our knowledge, none have united all these disciplines with a focus on engineering analysis and design calculations. This workshop is unique in combining computer science, mathematics, and engineering analysis to discuss the integration of the treatment of modeling errors and uncertainty into engineering computations.

The work presented represent a significant step towards achieving the goal of true reliability in engineering calculations.

The sponsors of this workshop are:
- National Science Foundation
- Sun Microsystems
- The Society for Imprecise Probability: Theories and Applications
- CASE (Computer Aided Structural Engineering) Center (GT STRUDL)
- Georgia Institute of Technology

The organizers appreciate the support of the sponsors: this workshop would not have occurred without their contributions and commitment.

Rafi L. Muhanna and Robert L. Mullen
Editors

Uncertainty modeling with clouds in autonomous robust design optimization

Martin Fuchs, Arnold Neumaier

University of Vienna
Faculty of Mathematics
Nordbergstr. 15
1090 Wien
Austria
email: martin.fuchs@univie.ac.at

Abstract. The task of autonomous and robust design cannot be regarded as a single task, but consists of two tasks that have to be accomplished concurrently. First, the design should be found autonomously; this indicates the existence of a method which is able to find the optimal design choice automatically. Second, the design should be robust; in other words: the design should be safeguarded against uncertain perturbations.

Traditional modeling of uncertainties faces several problems. The lack of knowledge about distributions of uncertain variables or about correlations between uncertain data, respectively, typically leads to underestimation of error probabilities. Moreover, in higher dimensions the numerical computation of the error probabilities is very expensive, if not impossible, even provided the knowledge of the multivariate probability distributions.

Based on the *clouds* formalism we have developed new methodologies to gather all available uncertainty information from expert engineers, process it to a reliable worst-case analysis and finally optimize the design seeking the optimal robust design.

The new methods are applied to problems for autonomous optimization in robust spacecraft system design at the European Space Agency (ESA).

Keywords: uncertainty modeling, robust design, clouds, autonomous design, design optimization

Acknowledgements

I would like to acknowledge the support of the people from the ESA Advanced Concepts Team who contributed to this paper in various ways, in particular I would like to thank Daniela Girimonte and Dario Izzo.

1. Introduction

In general terms, uncertainty handling for design optimization has the goal to safeguard reliably against uncertain perturbations while seeking an optimal design. The achieved design can thus be qualified as robust.

An engineer who designs a structure faces the task to develop a product which satisfies given requirements formulated as design constraints. Output of the engineer's work should be an optimal design with respect to a certain design objective. In many cases this is the cost or the mass of the designed product. An algorithmic method for design optimization functions as decision making support for engineers. In the last years, much research has been dedicated to the achievement of decisions support systems. Even the attempt of autonomous design has been made trying to capture the reasoning of the system experts. For more complex kinds of structures, e.g., a spacecraft component or a whole spacecraft, the design process involves several different engineering fields, so the design optimization becomes multidisciplinary, and an interaction between the comprised disciplines is necessary. The resulting overall optimization process is known as multidisciplinary design optimization (MDO). Design related uncertainties are handled to safeguard against failures of the design, i.e., a violation of the design requirement constraints, caused by uncertain errors.

In many cases, in particular for early design phases, it is common engineering practice to handle uncertainties by assigning intervals, or safety margins, to the uncertain variables, usually combined with an iterative process of refining the intervals while converging to a robust optimal design. The refinement of the intervals is done by experts who assess whether the worst-case scenario, that has been determined for the design at the current stage of the iteration process, is too pessimistic or too optimistic. How to assign the intervals and how to choose the endpoint of the assigned intervals to get the worst-case scenario is usually not computed but assessed by an expert. The goal of the whole iteration includes both optimization of the design and safeguarding against uncertainties. Apart from interval assignments there are further ways to handle uncertainties in design processes, e.g., methods from probability theory or fuzzy theory like fuzzy clustering, portfolio theory, or simulation techniques like Monte Carlo.

Real life applications of uncertainty methods disclose various problems. The dimension of many uncertain real life scenarios is very high which causes severe computational problems, famous as the curse of dimensionality, see, e.g., (Koch et al., 1999). Even given the knowledge of the multivariate probability distributions the numerical computation of the error probabilities becomes very expensive, if not impossible. Moreover, the available uncertainty information in early design phases is often very limited, mostly there are only interval bounds on the uncertain variables, sometimes probability distributions for single variables without correlation information. When the amount of uncertainty information available is small, traditional methods face additional problems. To make use of well-known current methods from probability or fuzzy theory more such information would be required. Simulation techniques also require a larger amount of information to be reliable, or unjustified assumptions on the uncertainties have to be made. The lack of information typically causes these methods to underestimate the effects of the uncertain tails of the probability distribution, cf. (Ferson, 1996). Similarly, a reduction of the problem to an interval analysis after assigning intervals to the uncertain variables as described before (e.g., 3 σ boxes) entails a loss of valuable uncertainty

information which would actually be available, maybe unformalized, but is not at all involved in the uncertainty model.

Many previous works are dedicated to MDO or robust design. In a classical approach to MDO, cf. (Alexandrov and Hussaini, 1997), (Roy, 1996), (Belton and Stewart, 2002), each specialist would prepare a subsystem design rather independently, using stand-alone tools. Design iterations among the different discipline experts would take place in meetings at certain time intervals. This well-established approach reduces the opportunity to find interdisciplinary solutions and to create system awareness in the specialists. A considerable step forward in MDO for early design phases has been achieved by concurrent engineering where a sequential iterative routine is replaced by a parallel and cooperative procedure. Facilities where these methodologies are implemented for the special case of spacecraft design are, among others, the ESA Concurrent Design Facility (Bandecchi et al., 1999), the NASA Goddard Integrated Mission Design Center (Karpati et al., 2003) and the Concept Design Center at The AeroSpace Corporation (Aguilar et al., 1998). An approach to MDO via game theory can be found, e.g., in (Lewis and Mistree, 1997). To improve the robustness in the process of design optimization there are various approaches dealing with uncertainty modeling. In (Pate-Cornell and Fischbeck, 1993) probability risk analysis is applied to the uncertainties in space shuttle design; an approach from fuzzy theory can be found, e.g., in (Ross, 1995); in (Thunnissen, 2005) a general qualitative and quantitative investigation of uncertainties in space design is given. The work by (Amata et al., 2004) presents studies harmonizing the interests from different disciplines in multidisciplinary design optimization. The attempt to incorporate both uncertainty and autonomy in the design process was made, e.g., in (McCormick and Olds, 2002), using Monte-Carlo simulation techniques, or in (Lavagna and Finzi, 2002), with a fuzzy logic approach.

The ESA Advanced Concepts Team in cooperation with the University of Vienna performed an Ariadna study on the application of the clouds theory in space design optimization, cf. (Neumaier et al., 2007). This study presented an initial step on how clouds could be applied to handle uncertainties in spacecraft design. A significant further step is given in (Fuchs et al., 2007).

Deepening the understanding of the latter studies, we here focus on the theory of clouds and emphasize the capability of an adaptive processing of unformalized uncertainty information with clouds. Clouds allow the representation of incomplete stochastic information in a clearly understandable and computationally attractive way, mediating between aspects of fuzzy set theory and probability distributions, cf. (Dubois and Prade, 2005). The use of clouds permits an adaptive worst-case analysis without losing track of important probabilistic information. At the same time, all computed probabilities, and hence the resulting designs, are reasonably safeguarded against perturbations due to unmodeled and possibly unavailable information. For given confidence levels, the clouds provide regions of relevant scenarios affecting the worst-case for a given design. We have the ambitious goal to achieve a quantification of reliability close to classical probability theory methods, but in higher dimensional spaces of uncertain scenarios so that we can deal with real-life design problems. To find a reliable robust and optimal design autonomously, we have additionally developed heuristic optimization methods.

Figure 1 illustrates the basic concept of our approach. The expert provides the underlying model, given as a black-box model, and all currently available uncertainty information on the model inputs. The information is processed to generate a cloud that provides a nested collection of regions of relevant scenarios parameterized by a confidence level α, and thus produces safety constraints for

the optimization. The optimization minimizes a certain objective function (e.g., cost, mass) subject to the safety constraints and to the functional constraints which are represented by the underlying model. The results of the optimization are returned to the expert, who is given an interactive possibility to provide additional uncertainty information afterwards and rerun the procedure.

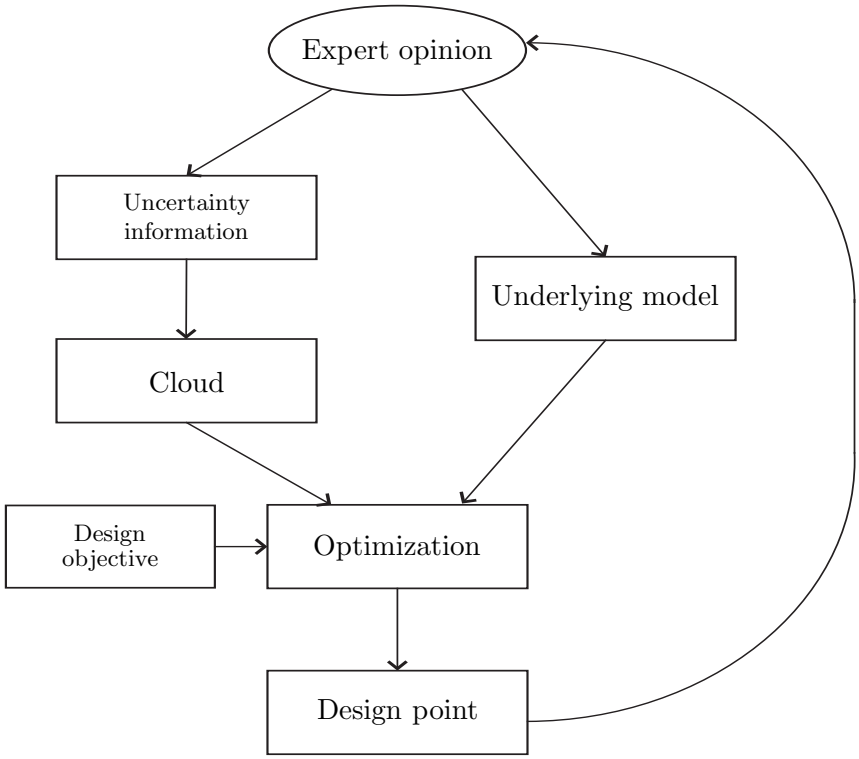


Figure 1. Basic concept.

Focussing on application examples from early phase spacecraft design, we will deal with a limited amount of uncertainty information, provided on the one hand as bounds or marginal probability distributions on the uncertain variables, without any formal correlation information. On the other hand, the engineers can adaptively improve the uncertainty model, even if their expert knowledge is only little formalized, by adding correlation constraints to exclude scenarios deemed irrelevant. The information can also be provided as real sample data, if available.

This paper is organized as follows. In Section 2 we present a more detailed study of uncertainty modeling with clouds. This is used to investigate robust design optimization, cf. Section 3. The techniques are applied to an example from spacecraft system design, described in Section 4. In Section 5 we discuss general and detailed aspects of our approach and conclude with a summary of our results.

2. Uncertainty modeling with clouds

The clouds formalism will serve as the central theoretical background for our uncertainty handling. Clouds will allow us an interpretation of uncertainties in terms of safety constraints. An important additional aspect of clouds is the ability to deal with high dimensional and non-formalized uncertainties.

This section starts with the definition of clouds in Section 2.1. The special case of potential clouds will be introduced as particularly interesting in Section 2.2, Section 2.3 will give an introduction about potential cloud generation.

2.1. Theoretical background

We start with the formal definition of clouds and introduce the notations. Let $\varepsilon \in \mathbb{M} \subseteq \mathbb{R}^n$ be an n-dimensional vector of uncertainties, we call ε an uncertain scenario. A *cloud* is a mapping $\chi(\varepsilon) = [\underline{\chi}(\varepsilon), \overline{\chi}(\varepsilon)]$, where $\chi(\varepsilon)$ is a nonempty, closed and bounded interval $\in [0,1]$ for all $\varepsilon \in \mathbb{M}$, and $]0,1[\subseteq \bigcup_{\varepsilon \in \mathbb{M}} \chi(\varepsilon) \subseteq [0,1]$. We call $\overline{\chi}(\varepsilon) - \underline{\chi}(\varepsilon)$ the width of the cloud χ. A cloud is called thin if it has width 0, and continuous if the lower level $\underline{\chi}$ and the upper level $\overline{\chi}$ are continuous functions of ε.

There exists a close relationship between thin continuous 1-dimensional clouds and cumulative distribution functions (CDFs) of real univariate random variables ε which is stated in Proposition 4.1 in (Neumaier, 2004): Let $F_\varepsilon(x) = \Pr(\varepsilon \leq x)$ be the CDF of ε, then $\chi(x) := F_\varepsilon(x)$ defines a thin cloud and $\Pr(\chi(\varepsilon) \leq y) = y, y \in \mathbb{M}$. The latter refers just to the fact that $F_\varepsilon(x)$ is uniformly distributed.

CDFs are well known from probability theory. Especially the 1-dimensional case is computationally unproblematic and intuitively understandable. However, we want to deal with significantly higher dimensions than 1. This leads to the idea to construct continuous clouds from user-defined potential functions $V : \mathbb{M} \to \mathbb{R}$.

2.2. Potential clouds

As we learned in the last section potential function based clouds, in short *potential clouds*, are a special class of continuous clouds supposed to help to cope with high dimensional uncertainties. The idea is to construct a cloud from an interval-valued function χ of a user-defined potential function V, i.e., $\chi \circ V : \mathbb{M} \to [a,b]$, where $[a,b]$ is an interval in $[0,1]$.

Define the mapping

$$\chi(x) := [\underline{\alpha}(V(x)), \overline{\alpha}(V(x))], \tag{1}$$

where $\underline{\alpha}(y) := \Pr(V(\varepsilon) < y)$, $\overline{\alpha}(y) := \Pr(V(\varepsilon) \leq y)$, $y \in \mathbb{R}$, and $\varepsilon \in \mathbb{M}$ a random variable. Then we get from Theorem 4.3 in (Neumaier, 2004) that we thus constructed a cloud χ that gives us an important interpretation in terms of confidence regions for ε.

Let $\alpha \in [0,1]$ be a given confidence level. The remarks to Theorem 4.3 in (Neumaier, 2004) tell us that if we choose $\underline{\alpha}(y)$ as a lower bound for $\Pr(V(\varepsilon) < y)$ and $\overline{\alpha}(y)$ as an upper bound for $\Pr(V(\varepsilon) \leq y)$, $\underline{\alpha}$, $\overline{\alpha}$ smooth and monotone, then χ as defined above is still a cloud. An appropriate

bounding $\underline{\alpha}$, $\overline{\alpha}$ can be found, e.g., by Kolmogoroff-Smirnov (KS) statistics (Kolmogoroff, 1941). Then we define

$$\underline{C}_\alpha := \{\varepsilon \mid V(\varepsilon) \le \underline{V}_\alpha\}, \tag{2}$$

if a solution $\underline{V}_\alpha$ of $\overline{\alpha}(\underline{V}_\alpha) = \alpha$ exists and $\underline{C}_\alpha := \emptyset$ otherwise; analogously

$$\overline{C}_\alpha := \{\varepsilon \mid V(\varepsilon) \le \overline{V}_\alpha\}, \tag{3}$$

if a solution $\overline{V}_\alpha$ of $\underline{\alpha}(\overline{V}_\alpha) = \alpha$ exists and $\overline{C}_\alpha := \mathbb{M}$ otherwise. These are nested families of confidence regions parameterized by α: The region $\underline{C}_\alpha$ contains at most a fraction of α of all scenarios in $\mathbb{M}$, since $\Pr(\varepsilon \in \underline{C}_\alpha) \le \Pr(\overline{\alpha}(V(\varepsilon)) \le \alpha) \le \Pr(F(V(\varepsilon)) \le \alpha) = \alpha$; analogously $\overline{C}_\alpha$ contains at least a fraction of α of all scenarios in $\mathbb{M}$.

2.3. Potential cloud generation

Let's summarize what is needed to generate a potential cloud: a potential function V has to be chosen, then appropriate bounds on the CDF F of $V(\mathbb{M})$ must be found. We will investigate how to find these bounds. But first we consider the question how to choose the potential function. There are endless possibilities (see, e.g., Figure 2) to make the choice.

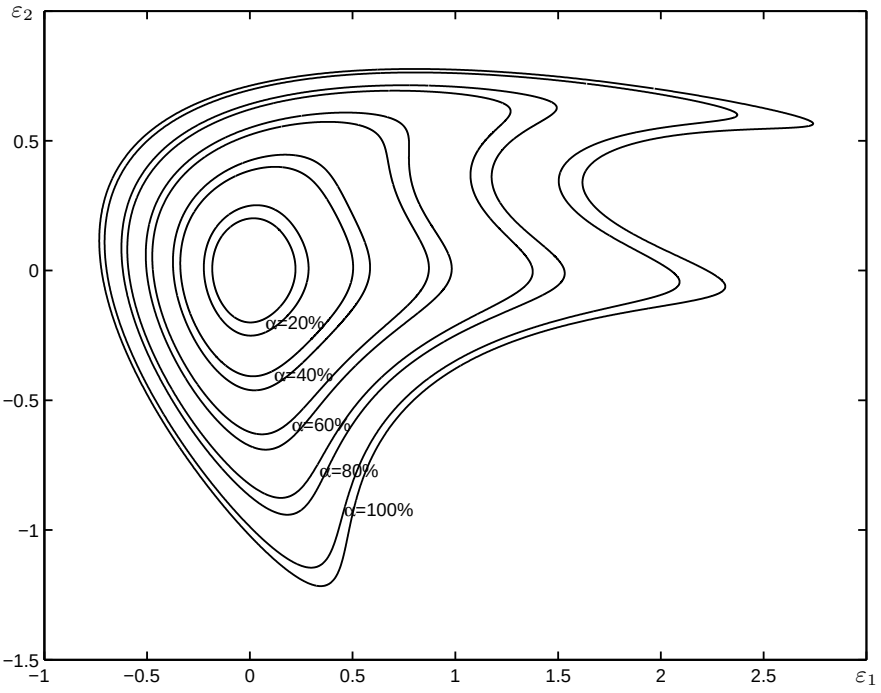


Figure 2. Nested confidence regions for the example of a 2-dimensional potential cloud, $\alpha = 0.2, 0.4, 0.6, 0.8, 1$.

Two special cases for choices of the potential function are

$$V(\varepsilon) := \max_k \frac{|\varepsilon^k - \mu^k|}{r^k}, \tag{4}$$

where $\varepsilon, \mu, r \in \mathbb{R}^n$, $\varepsilon^k, \mu^k, r^k$ are the k^{th} components of the vectors, defines a box-shaped potential.

$$V(\varepsilon) := \|A\varepsilon - b\|_2^2, \tag{5}$$

where $\varepsilon, b \in \mathbb{R}^n, A \in \mathbb{R}^{n\times n}$, defines an ellipsoid-shaped potential.

A good choice of the potential should allow for a simple computational realization of the confidence regions, e.g., by linear constraints represented by $A\varepsilon \leq b$. This leads us to the investigation of polyhedron-shaped potentials, a generalization of box-shaped potentials. A polyhedron potential can be defined as:

$$V(\varepsilon) := \max_k \frac{(A\varepsilon)^k}{b^k}, \tag{6}$$

where $(A\varepsilon)^k, b^k$ are the k^{th} components of the vectors $(A\varepsilon)$ and b, respectively.

But how to achieve a polyhedron that reflects the given uncertainty information in the best way? As mentioned we assume the uncertainty information to consist of given samples, boxes or marginal distributions, and unformalized correlation constraints. After generation of a sample S as described later we define a box b_0 containing all sample points, and we define our potential $V_0(\varepsilon)$ box-shaped taking the value 1 on the margin of b_0.

Based on expert knowledge, a user-defined variation of V_0 can be performed by cutting off sample points deemed irrelevant for the worst-case. The exclusion of sample points is given by linear constraints $A\varepsilon \leq b$. Thus an expert can specify the uncertainty information in the form of linear correlation bounds adaptively resulting in a polyhedron shaped potential (6), even if the expert knowledge is only little formalized.

The adaptive exclusion of irrelevant scenarios, cf. Figure 3, can be realized in a graphical user interface (GUI). This procedure imitates iterative improvement in common real life MDO.

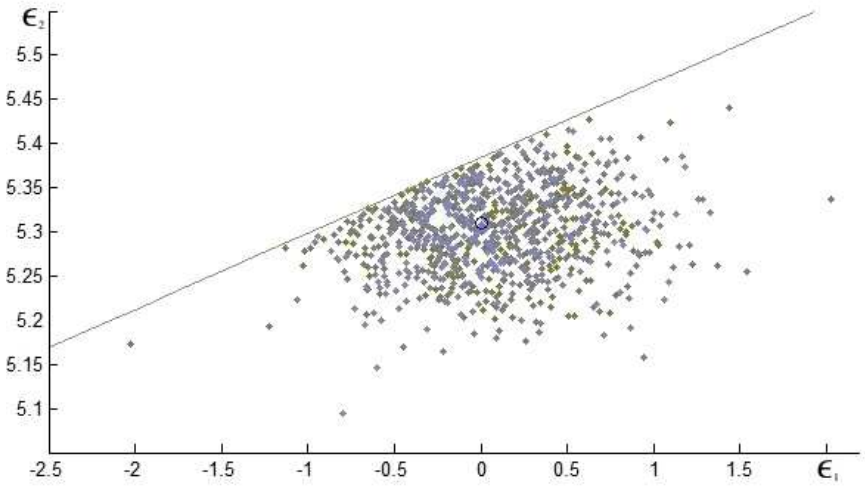


Figure 3. Exclusion of irrelevant scenarios by correlation bounds.

Now we turn to the investigation on how to find appropriate bounds on the CDF $F(V(\varepsilon))$. As we do not have the knowledge of F we have to approximate it before we can assign bounds on it.

To this end we will make use of KS statistics as suggested before. That means we approximate F by an empirical distribution $\widetilde{F}$. The generation of an empirical distribution requires the existence of a sample S representing our uncertainties.

It depends on the given uncertainty information whether a sample already exists. In case there is no sample provided or the given sample is very small, a sample has to be generated. For these cases we first use a Latin hypercube sampling, cf. (McKay et al., 1979), inspired method to generate the sample $S = \{x_1, \ldots, x_{N_S}\}$ of N_S sample points. The sample points are chosen from a grid fulfilling the well-known Latin hypercube condition. If only boxes are given, then the grid is equidistant, if marginal distributions are given the grid is transformed with respect to them to ensure that each grid interval has the same marginal probability. Thus the generated sample represents the marginal distributions. However after a modification of S, e.g., by cutting off sample points as described, an assignment of weights to the sample is necessary to preserve the marginal CDFs.

In order to do so the weights $\omega_1, \ldots, \omega_{N_S} \in [0,1]$ are required to satisfy the following conditions: Let π_j be a sorting permutation of $\{1, \ldots, N_S\}$, such that $x^j_{\pi_k(1)} \leq \ldots \leq x^j_{\pi_k(N_S)}$. Let I be the index set of those entries of the uncertainty vector ε where a marginal CDF $F_i, i \in I \subseteq \{1, \ldots, n\}$ is given. Then the weights should satisfy (7) $\forall i \in I, k = 1, \ldots, N_S$

$$\sum_{j=1}^{k} \omega_{\pi_i(j)} \ \in [F_i(x^i_{\pi_i(k)}) - d, F_i(x^i_{\pi_i(k)}) + d], \ \sum_{k=1}^{N_S} \omega_k = 1. \tag{7}$$

The function

$$\widetilde{F}_i(\xi) \ := \textstyle\sum_{\{j | x^i_j \leq \xi\}} \omega_j \tag{8}$$

is a weighted marginal empirical distribution. For trivial weights, $\omega_1 = \ldots = \omega_{N_S} = \frac{1}{N_S}$, $\widetilde{F}_i$ is a standard empirical distribution. The constraints (7) require the weights to represent the marginal CDFs with some reasonable margin d. In other words, the weighted marginal empirical distributions $\widetilde{F}_i, i \in I$ should not differ from the given marginal CDF F_i by more than d. In practice, one chooses $d = d_{\mathrm{KS}}$ with KS statistics:

$$d_{\mathrm{KS}} = \frac{\phi^{-1}(\alpha_{\mathrm{KS}})}{\sqrt{N_S} + 0.12 + \frac{0.11}{\sqrt{N_S}}}, \tag{9}$$

where ϕ is the Kolmogoroff function, α_{KS} the confidence in the KS theorem, cf. (Kolmogoroff, 1941), (Press et al., 1992).

Assume we have achieved weights satisfying (7), this yields a weighted empirical distribution

$$\widetilde{F}(\xi) \ := \textstyle\sum_{\{j | V(x_j) \leq \xi\}} \omega_j \tag{10}$$

approximating the CDF of $V(\varepsilon)$. If weights satisfying (7) can only be achieved with $d > d_{\mathrm{KS}}$, the relaxation d gives us an indicator for the quality of the approximation which will be useful to construct bounds on the CDF $F(V(\varepsilon))$.

After the approximation of $F(V(\varepsilon))$ with $\widetilde{F}$ we are just one step away from generating a potential cloud. Remember that we seek an appropriate bounding on $F(V(\varepsilon))$. We define $\overline{F} := \min(\widetilde{F} + D, 1)$

and $\underline{F} := \max(\widetilde{F} - D, 0)$, where D is computed with help of the KS approach (9), and fit these two step functions to smooth, monotone lower bounds $\underline{\alpha}(V(\varepsilon))$ and upper bounds $\overline{\alpha}(V(\varepsilon))$. If the the quality of our approximation with $\widetilde{F}$ or the sample size N_S is decreased, the width of the bounds is increased correspondingly.

Thus we have found an appropriate bounding of the CDF $F(V(\varepsilon))$ and according to the remarks to Theorem 4.3. in (Neumaier, 2004) mentioned we have generated a potential cloud that fulfills the conditions that define a cloud via the mapping $\chi : \varepsilon \to [\underline{\alpha}(V(\varepsilon)), \overline{\alpha}(V(\varepsilon))]$.

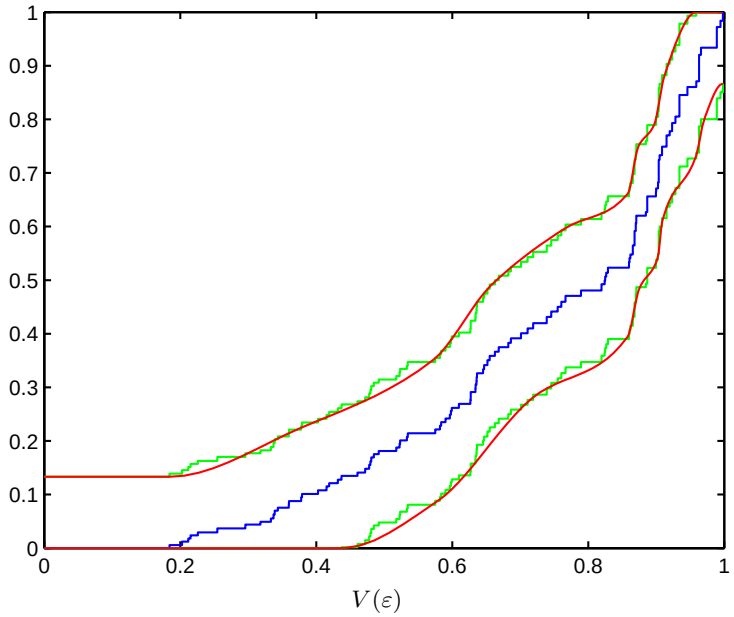


Figure 4. The smooth lower bounds $\underline{\alpha}(V(\varepsilon))$ and upper bounds $\overline{\alpha}(V(\varepsilon))$ for a potential cloud.

The cloud represents the given uncertainty information and now enables us to interpret the potential level maps $\{\varepsilon \mid V(\varepsilon)) \leq \underline{V}_\alpha\} = \underline{C}_\alpha$ as *confidence regions* for our uncertain vector ε. They are the worst-case relevant regions.

Hence the clouds give an intuition and guideline how to construct confidence regions for safety constraints. To this end we have combined several different theoretical means: potential functions, CDF approximations with empirical distributions, KS statistics to estimate bounds, sample generation methods, and weighting techniques.

3. Robust design optimization

A classic approach to design optimization, without taking uncertainties into account, leads to decision support for engineers, but to a design which completely lacks robustness. We want to safe-

guard the design against uncertain errors. That will involve the methods for uncertainty modeling we introduced in the last section.

First we give a formal statement of the optimization problem in Section 3.1. Afterwards we point out the difficulties related in Section 3.2 and finally present a solution approach in Section 3.3.

3.1. Problem formulation

Provided an underlying model of a given structure like a spacecraft component, with several inputs and outputs, we denote as x the vector containing all output variables, and as z the vector containing all input variables.

The inputs contained in z can be divided into global input variables u and design variables v. The design variables are determined by the so called design choice variables. A choice variable is a univariate variable controllable for the design. The choice variables can be continuous, e.g., the diameter of an antenna, or discrete, e.g., the choice of a thruster from a set of different thruster types. Let θ be the vector of design choice variables $\theta^1, \ldots, \theta^{n_o}$. Let I_d be the index set of choice variables which are discrete and I_c be the index set of choice variables which are continuous, $I_d \cup I_c = \{1, \ldots, n_o\}$, $I_d \cap I_c = \emptyset$. In the discrete case, $i \in I_d$, the choice variable θ^i determines the value of n_i design variables. For example, if θ^i was the choice of a thruster, each choice could be specified by the thrust and specific impulse of the thruster. Thrust and specific impulse would be design variables v_1^i and v_2^i, and $n_i = 2$ in this example. Let $1, \ldots, N_i$ be the possible choices for θ^i, $i \in I_d$, then the discrete choice variable θ^i corresponds to a finite set of N_i points $(v_1^i, \ldots, v_{n_i}^i) \in \mathbb{R}^{n_i}$. Usually this set is provided in a $N_i \times n_i$ table (see, e.g., Table II, $N_i = 30$, $n_i = 3$). In the continuous case, $i \in I_c$, the choice variable θ^i can be regarded as a design variable in a given interval $[\underline{\theta^i}, \overline{\theta^i}]$. A global input variable is an external input with a nominal value that cannot be controlled for the underlying model, this could be, e.g., a specific temperature. Let $Z(\theta)$ be a mapping assigning an input vector z to the design choice θ. We call Z a table mapping as the nontrivial parts of Z consist of tables.

Both design and global input variables contained in z can be uncertain, ε denotes the related vector of uncertainties. We assume that the optimization problem can be formulated as a mixed-integer, bi-level problem of the following form:

$$\begin{array}{lll} \min\limits_{\theta} \ \max\limits_{x,z,\varepsilon} & g(x) & \text{(objective functions)} \\ \text{s.t.} & z = Z(\theta) + \varepsilon & \text{(table constraints)} \\ & G(x,z) = 0 & \text{(functional constraints)} \\ & \theta \in T & \text{(selection constraints)} \\ & V(\varepsilon) \le \underline{V}_\alpha & \text{(cloud constraint)} \end{array} \tag{11}$$

where the design objective $g(x)$ is a function of the output variables of the underlying model. The table constraints assign to each choice θ a vector z of input variables whose value is the nominal entry from $Z(\theta)$ plus its error ε with uncertainty specified by the cloud. The functional constraints express the functional relationships defined in the underlying model. It is assumed that the number of equations and the number of output variables is the same (i.e., $\dim G = \dim x$), and that the equations are (at least locally) uniquely solvable for x. The selection constraints specify which

choices are allowed for each choice variable, i.e., $\theta^i \in \{1, \ldots, N_i\}$ if $i \in I_d$ and $\theta^i \in [\underline{\theta^i}, \overline{\theta^i}]$ if $i \in I_c$. The cloud constraint involves the potential function V as described in the Section 2 and models the worst-case relevant region $\{\varepsilon \mid V(\varepsilon) \leq \underline{V}_\alpha\} = \underline{C}_\alpha$.

3.2. Difficulties

The problem formulated in the last section features several difficulties of most complex nature. The variable types can be both continuous and integer, so the problem comes as a mixed integer nonlinear program (MINLP). MINLP is still a recent research direction which has not yet matured. Profound difficulties arise from the fact that the functional constraints, represented by G, can have strong nonlinearities and can contain branching decisions such as case differentiation (implemented as, e.g., if-structures in the code) which leads to discontinuities. Additionally we face a bi-level structure imposed by the uncertainties, which is already a nontrivial complication in the traditional situation where all variables are continuous. The current methods for handling such problems require at least that the objective and the functional constraints are continuously differentiable. Standard optimization tools cannot be used to tackle problem (11).

In view of these difficulties we are limited to the use of heuristic methods, i.e., we treat the functional constraints of the underlying model as a black-box function $x = G_{\mathrm{bb}}(z)$ and make use of specific strategies to sample from the set of allowed inputs $z = Z(\theta)$, $\theta \in T$.

3.3. Solution approach

We will first reformulate the problem incorporating the objective function and functional constraints for the underlying model in the black-box function $G_{\mathrm{bb}}(z)$.

$$\begin{aligned} \min_{\theta} \ \max_{z,\varepsilon} \quad & G_{\mathrm{bb}}(z) \\ \text{s.t.} \quad & z = Z(\theta) + \varepsilon \\ & \theta \in T \\ & V(\varepsilon) \leq \underline{V}_\alpha \end{aligned} \tag{12}$$

We start with a look at the inner level of the problem, i.e., for a fixed $\theta \in T$

$$\begin{aligned} \max_{z,\varepsilon} \quad & G_{\mathrm{bb}}(z) \\ \text{s.t.} \quad & z = Z(\theta) + \varepsilon \\ & V(\varepsilon) \leq \underline{V}_\alpha \end{aligned} \tag{13}$$

Because of the polyhedral structure of our clouds, the cloud constraint $V(\varepsilon) \leq \underline{V}_\alpha$ can be written as a collection of linear inequalities parameterized by the confidence level α. We approximate G_{bb} in a small box containing the region $\{\varepsilon \mid V(\varepsilon) \leq \underline{V}_\alpha\}$ linearly. Thus problem (13) becomes an LP solved by an LP solver, cf. (Grant and Boyd, 2007). The maximizer $\widehat{\varepsilon}$, $\widehat{z} = Z(\theta) + \widehat{\varepsilon}$ for the fixed design choice θ corresponds to the worst-case objective function value $\widehat{G}_{\mathrm{bb}}(\theta) := G_{\mathrm{bb}}(Z(\theta) + \widehat{\varepsilon})$. The function $\theta \to \widehat{G}_{\mathrm{bb}}(\theta)$ implicated by the solution of problem (13) is now used to get rid of the

bi-level structure in problem (12):

$$\begin{aligned} \min_{\theta} \quad & \widehat{G}_{\mathrm{bb}}(\theta) \\ \text{s.t.} \quad & \theta \in T \end{aligned} \tag{14}$$

The method we develop to solve this 1-level problem, and to seek the robust, optimal design, is based on separable underestimation. It exploits the characteristics of the problem, takes advantage of the discrete nature of many of the choice variables involved in real life design, supporting, at the same time, continuous choice variables. Remember θ is the vector of design choice variables $\theta^1, \ldots, \theta^{n_o}$. We look for a separable underestimator $q(\theta)$ for the objective function of the form:

$$q(\theta) := \sum_{i=1}^{n_o} q_i(\theta^i). \tag{15}$$

Let $\theta \in T$, $z = Z(\theta)$. Assume the black-box G_{bb} has been evaluated N_o times resulting in the function evaluations $G_{\mathrm{bb}_1}, \ldots, G_{\mathrm{bb}_{N_o}}$ for the design choices $\theta_1, \ldots, \theta_{N_0}$. Let $l \in \{1, \ldots, N_o\}$. For a discrete choice θ_l^i, $i \in I_d$, we define $q_i(\theta_l^i) := q_{i,\theta_l^i}$, $\theta_l^i \in \{1, \ldots, N_i\}$, simply as a constant. For a continuous choice θ_l^i, $i \in I_c$, we define $q_i(\theta_l^i) := q_{i1} \cdot \theta_l^i + q_{i2} \cdot {\theta_l^i}^2$ by a quadratic expression with the two constants q_{i1} and q_{i2}. If $I_d = \emptyset$ we add an integer choice θ^i with $N_i = 1$ artificially to represent the constant part which is missing in the definition of q_i, $i \in I_c$. The vectors q_i of constants have the length N_i for $i \in I_d$, and 2 for $i \in I_c$. They are treated as variables q_i in a linear optimization program (LP) satisfying the constraints

$$\sum_{i=1}^{n_o} q_i(\theta_l^i) \ \leq G_{\mathrm{bb}_l} \quad l = 1, \ldots, N_o \tag{16}$$

and ensuring that many constraints in (16) will be active. The underestimator $q(\theta)$ is separable and can be easily minimized.

Apart from the method of separable underestimation we also make use of further strategies to find a solution of the optimization problem (14). The first one fits a quadratic model for the G_{bb} which is minimized afterwards, cf. (Huyer and Neumaier, 2006). Integers are treated as continuous variables and rounded to a grid with step width 1. Another method is based on evolution strategy with covariance matrix adaptation, cf. (Hansen and Ostermeier, 2001). It is a stochastic method to sample the search space. Integers are also treated as continuous variables rounded to the next integer value.

Finally the minimizers that result from all methods used are starting points for a limited global search that consists of an integer line search for the discrete choice variables and multilevel coordinate search (Huyer and Neumaier, 1999) for the continuous choice variables. Thus we hope to find the global optimal solution, but as we are using heuristics there is no guarantee.

Remark. For the implementation of our methods we formulated them as MATLAB code. The following is a summary of all external routines we use in our methods: we make use of the Statistics Toolbox of MATLAB to evaluate probability distributions; we use CVX (Grant and Boyd, 2007) to solve linear programs; SNOBFIT (Huyer and Neumaier, 2006) and MCS (Huyer and Neumaier,

1999) as external optimization routines; NLEQ (Deuflhard, 2004), (Nowak and Weimann, 1990) to solve systems of nonlinear equations.

4. Application example

Here we apply our methods for robust and autonomous design to a case study of early phase spacecraft engineering, i.e., the Attitude Determination and Control Subystem (ADCS) for the NASA's Mars Exploration Rover (MER) mission cf. (MER, 2003), (Erickson, 2004) whose scientific goal is to investigate the history of water on Mars. The ADCS is composed by eight thrusters aligned in two clusters. Onboard the spacecraft there is no main propulsion subsystem. The mission sequence after orbit injection includes a number of spin maneuvers and slew maneuvers. Spin maneuvers are required for keeping the gyroscopic stability of the spacecraft, whereas slew maneuvers serve to control the direction of the spacecraft and to fight effects of solar torque. Fault protection is considered to correct possible errors made when performing nominal maneuvers.

Our goal is to select the type of thrusters (from a set of possible candidates as listed in Table II) considering both minimization of the total mass m_{tot}, and assessment of the worst possible performance of a thruster with respect to m_{tot}. That corresponds to finding the thruster with the minimal worst-case scenario. The total mass consists of the fuel needed for attitude control (computed as the sum of the fuel needed for each maneuver) plus the mass of the eight thrusters that need to be mounted on the spacecraft. According to the notations introduced, the choice variable θ, i.e., the type of thruster, can be selected as an integer between 1 and 30.

Uncertainty specifications, variable structure, the MER mission maneuver sequence, and system model equations to compute the total mass m_{tot} are taken from (Thunnissen, 2005). The uncertainty specification for the model variables are reported in Table III of Appendix C. The number of uncertain global input variables (dimension of u) in this application example is 33 plus 1 uncertain design variable. The variable structure is summarized in Appendix A. Moreover, a survey on the system model equations and the MER mission sequence can be found in the Appendices of (Fuchs et al., 2007).

4.1. Results

The cloud constraints for the optimization are generated for a confidence level of $\alpha = 95\%$ and a generated sample size $N_S = 1000$. The results for optimization are divided into four different configurations of uncertainty handling and specifications:

a. The uncertainties are as specified in Table III. Here we treat them in a classical engineering way, assigning 3 σ boxes to the uncertain variables which is supposed to correspond to a 99.7% confidence interval for a single variable. Then the optimal design choice is $\theta = 9$ with an objective function value of $m_{tot} = 3.24$ kg in the nominal case and $m_{tot} = 5.56$ kg in the worst case.

b. The uncertainties are again as in Table III. With our methods we find the optimal design choice $\theta = 9$ as in Configuration a. However, if we compare the worst-case analysis of b and a,

it is apparent that the results for the 3 σ boxes are far too optimistic to represent a reliable worst-case scenario, the value of m_{tot} is now 8.08 kg instead of 5.56 kg for the 3 σ boxes.

c. In this configuration we do not take any uncertainties into account, generally assuming the nominal case for all uncertain input variables. The optimal design choice then is $\theta = 3$ with a value of $m_{tot} = 2.68$ kg in the nominal, but $m_{tot} = 8.75$ kg in the worst case, which is significantly worse than in Configuration b.

d. The uncertainties are obtained by taking the values from Table III and doubling the standard deviation of the normally distributed variables. It is interesting to report that if we increase the uncertainty in the normally distributed uncertain variables simply in this way, the optimal design choice changes to $\theta = 17$ with a value of $m_{tot} = 3.38$ kg in the nominal and $m_{tot} = 9.49$ kg in the worst case.

The results are summarized in Table I, showing the optimal design choice for each configuration and the corresponding value of the objective function m_{tot} for the nominal case and for the worst case, respectively.

Table I. Nominal and worst-case values of m_{tot} for different design choices obtained by the four different configurations.

Configuration	**Design Choice** θ	**Nominal value** m_{tot}	**Worst-case** m_{tot}
a	9	3.24	5.56
b	9	3.24	8.08
c	3	2.68	8.75
d	17	3.38	9.49

The results show a number of important facts related to spacecraft design. The comparison between the configurations b and d suggests that in a preliminary stage of the spacecraft systems modeling the optimal design point θ is quite sensitive to the uncertainty description, a fact well-known to the system engineers who see their spacecraft design changing frequently during preliminary phases when new information becomes continuously available. Our method captures this important dynamics and processes it in rigorous mathematical terms.

The comparison between the configurations b and c suggests that the uncertainties need to be accounted for in order not to critically overestimate the spacecraft performances.

Finally, the comparison between the configurations b and a suggests that the simple 3 σ analysis of uncertainties, frequent in real engineering practice, produces a quite different estimation of the spacecraft performances with respect to a more rigorous accounting of the uncertainty information.

5. Discussion & Conclusions

The importance of robustness in design optimization has been the starting point and main motivation of our research work, and our results on a case study confirm that the optimal spacecraft design is strongly sensitive to uncertainties. At the present stage we can clearly state that neglecting uncertainties results in a design that completely lacks robustness and a simplified uncertainty model (like a 3 σ approach) may yield critical underestimations of worst-case scenarios.

When trying to collect the uncertainty information, it turned out to be very difficult to get useful information directly from expert engineers. To collect the information, an interactive dialogue between the experts and the computer can be realized by a GUI where the engineers can specify uncertainties, provide sample data, cut off worst-case irrelevant scenarios, and adjust the quality of the uncertainty model. We expect that this kind of interaction is an inevitable next step in design processes, especially spacecraft design. We continue the discussion with more detailed considerations on the study.

- In the theory of clouds, cf. Section 2 and (Neumaier, 2004), there is a distinction between the confidence regions of α-relevant scenarios $\underline{C}_\alpha$, α-reasonable scenarios $\overline{C}_\alpha$ and borderline cases (which is the set difference of the α-reasonable and the α-relevant regions). In robust design the possibly uncertain scenarios are required to satisfy safety constraints. With respect to our terminology the regions above have the following interpretation: if at least one of the α-relevant scenarios fails to satisfy the safety constraints, the design is unsafe; if all of the α-reasonable scenarios satisfy the safety constraints, the design is safe. Between these two cases there is the borderline region where no precise statement can be made without additional uncertainty information. The volume of the borderline region is increasing if the width of the cloud increases and vice versa. So widening the cloud enlarges the borderline region, corresponding to a lack of uncertainty information. This fact is reflected in our approach as both a smaller sample size and an increased dimension of the uncertainty result in a wider cloud.
- The width of the cloud is defined as the difference between the mappings $\underline{\alpha}$ and $\overline{\alpha}$ (cf. Section 2). We constructed the mappings to fulfill the conditions that define a cloud with an algorithm which is non-rigorous, but has a high, adjustable reliability. Thus the user of the algorithm is able to control the desired level of reliability.
- As mentioned before the reliability of our worst-case analysis with clouds is determined by user-defined parameters, i.e., the size of the generated sample S and confidence levels for sample generation, CDF bounding and approximation. Concerning the sample size: if we increase the size of S we artificially refine the uncertainty model and get more reliability of the worst-case analysis. A larger sample is computationally more expensive, in particular the weight computation, so the reliability is also a trade-off with performance.
- The choice of the potential function is arbitrary. Different shapes of the cloud (i.e., shapes of the potential) can make the worst-case analysis more pessimistic or optimistic. We point out that a poor choice of the potential makes the worst-case analysis more pessimistic, but will still result in a valid robust design. We allow a variation of the potential by switching from a box-shaped to a polyhedron-shaped potential to enable the experts to improve the uncertainty model iteratively.

- A good weight computation (cf. Section 2.3) is the key to a good uncertainty representation with clouds. In higher dimensions the weight computation is very expensive. To overcome this problem and to allow the adjustment of the computation time, the relaxation radius d must be increased carefully. In our algorithm we respect the relaxation property, widening the cloud by the amount of relaxation after evaluating the quality of the weights as described in Section 2.3.
- As mentioned before, we are limited to the use of heuristic methods since the design problem (11) is highly complex and not suitable for standard optimization methods. In our problem formulation we seek the design with the optimal worst-case scenario. It is possible to trade off between the worst-case scenario and the nominal case of a design, but this would lead to a multi-objective optimization problem formulation.
- The number 34 of uncertain variables in our case study is large enough to make our problem representative for uncertainty handling in real-life applications.
- Though global optimality for the solution in our application example is very likely, as the choice variable is 1-dimensional and discrete, in general the heuristical methods cannot guarantee global optimality of the problem solution.
- The approach with separable underestimation introduced in this chapter takes advantage of inherent characteristics of spacecraft design problems, i.e., the discrete nature of many of the variables involved, supporting, at the same time, continuous choice variables. Details on our heuristic methods for design optimization introduced in Section 3 will be published elsewhere.

5.1. Conclusions and Future Work

In this chapter we presented a new approach to autonomous robust design optimization. Starting from the background of the cloud theory we developed methodologies to process the uncertainty information from expert knowledge towards a reliable worst-case analysis and an optimal and robust design. Our approach is applicable to real-life problems such as, e.g., early phase spacecraft system design. In the example of the community of spacecraft engineers, at present, in most instances of the design process, reliability is only assessed qualitatively by the experts. We present a step forward towards quantitative statements about the design reliability.

The adaptive nature is one of the key features of our uncertainty model as it imitates real-life design strategies. The iteration steps significantly improve the uncertainty information and we are able to process the new information to an improved uncertainty model.

The presented approach is generally applicable to problems of robust design optimization, not only spacecraft design. In particular problems with discrete design choices can be tackled. The advantages of achieving the optimal design autonomously is undeniable. Though we already applied the new methods to different design problems, cf. (Neumaier et al., 2007), one future goal is to apply them to more problem classes in order to learn from new challenges.

With our approach we can process the available uncertainty information to perform a reliable worst-case analysis linked to an adjustable confidence level. An additional value of the uncertainty model is the fact that one can capture various forms of uncertainty information, even those less formalized. There is no loss of valuable information, and the methods are capable of handling the uncertainties reliably, even if the amount of information is very limited.

Summing up, the presented methods offer an exciting novel approach to face the highly complex problem of autonomous robust design optimization, an approach which is easily understandable, reliable and computationally realizable.

Appendix

A. Model Variable Structure

Remark. Do not confuse the notations in these appendices with our notation of the main sections.

The 47 variables involved in the model fall into the following four categories:

– **5 constant parameters**.
Input variables for the model with fixed values and no uncertainty.

Constant parameter	**Description**	**Value**
c_0	speed of light in a vacuum	299792458 m/s
d	average distance from the spacecraft to the sun in AU	1.26 AU
g_0	gravity constant	9.8 m/s^2
t	total mission time	216 days
θ_i	sunlight angle of incidence	0°

– **33 Uncertain input variables**.
The uncertainties are specified by probability distributions for each of these variables (cf. Appendix C).

Variable	**Description**
A_{max}	maximal cross-sectional area
J_{xx}, J_{zz}	moments of inertia
R	engine moment arm
δ_1, δ_2	engine misalignment angle
g_s	solar constant at 1 AU
κ	distance from the center of pressure to the center of mass
ω_{spin_i}	spin rates, $i = 0...3$, given in rpm
ψ_{slew_i}	slew angles, $i = 1...19$, given in °
q	spacecraft surface reflectivity
$uncfuel$	additive uncertain constant that represents inaccuracies in the equations used for the calculation of the fuel masses

– **3 Design variables**.
Thruster specifications relevant for the model. There is uncertainty information given on one of them (the thrust).

Variable	Description
F	thrust
I_{sp}	specific impulse
m_{thrust}	mass of a thruster

– **6 Result variables**.
Result variables containing the objective for the optimization m_{tot}.

Variable	Description
m_{fp}	fuel mass needed for fault protection maneuvers
m_{fuel}	total fuel mass needed for all maneuvers
m_{slew}	fuel mass needed for slew maneuvers
m_{slew_s}	fuel mass needed for slew maneuvers fighting solar torque
m_{spin}	fuel mass needed for spin maneuvers
m_{tot}	total mass of the subsystem

B. Thruster specification

Table II shows the thruster specifications and the linked choice variable θ. The table entries are sorted by the thrust F. The difference between the so called design and choice variables can be seen easily in this table: the table represents 30 discrete choices in $\mathbb{R}^3$. The 3 design variables are the 3 components of these points in $\mathbb{R}^3$. The choice variable θ is 1-dimensional and has an integer value between 1 and 30. The various sources for the data contained in Table II are (EADS, 2007), (Thunnissen, 2005), (Purdue School of Aeronautics and Astronautics, 1998), (Zonca, 2004), (Personal communication, 2007).

C. Uncertainty specification

All uncertainty specifications taken from (Thunnissen, 2005) are reported in Table III. The notation used for the probability distributions is:

Notation	Distribution
$U(a,b)$	uniform distribution in (a,b)
$N(\mu,\sigma)$	normal distribution with mean μ and variance σ^2
$\Gamma(\alpha,\beta)$	gamma distribution with mean $\alpha\beta$ and variance $\alpha\beta^2$
$L(\mu,\sigma)$	lognormal distribution, distribution parameters μ and σ (mean and standard deviation of the associated normal distribution)

The uncertainty information on the design variable F should be interpreted as follows: The actual thrust of a thruster is normally distributed, has the mean F_{table} (:= the nominal value for F specified in Table II) and standard deviation $\frac{7}{300}F_{table}$.

Table II. Thruster specifications and the linked choice variable θ.

θ	**Thruster**	F/N	I_{sp}/s	m_{thrust}/kg
1	Aerojet MR-111C	0.27	210	0.2
2	EADS CHT 0.5	0.5	227.3	0.195
3	MBB Erno CHT 0.5	0.75	227	0.19
4	TRW MRE 0.1	0.8	216	0.5
5	Kaiser-Marquardt KMHS Model 10	1	226	0.33
6	EADS CHT 1	1.1	223	0.29
7	MBB Erno CHT 2.0	2	227	0.2
8	EADS CHT 2	2	227	0.2
9	EADS S4	4	284.9	0.29
10	Kaiser-Marquardt KMHS Model 17	4.5	230	0.38
11	MBB Erno CHT 5.0	6	228	0.22
12	EADS CHT 5	6	228	0.22
13	Kaiser-Marquardt R-53	10	295	0.41
14	MBB Erno CHT 10.0	10	230	0.24
15	EADS CHT 10	10	230	0.24
16	EADS S10 - 01	10	286	0.35
17	EADS S10 - 02	10	291.5	0.31
18	Aerojet MR-106E	12	220.9	0.476
19	SnM 15N	15	234	0.335
20	TRW MRE 4	18	217	0.5
21	Kaiser-Marquardt R-6D	22	295	0.45
22	Kaiser-Marquardt KMHS Model 16	22	235	0.52
23	EADS S22 - 02	22	290	0.65
24	ARC MONARC-22	22	235	0.476
25	ARC Leros 20	22	293	0.567
26	ARC Leros 20H	22	300	0.4082
27	ARC Leros 20R	22	307	0.567
28	MBB Erno CHT 20.0	24	234	0.36
29	EADS CHT 20	24.6	230	0.395
30	Daimler-Benz CHT 400	400	228.6	0.325

References

J. A. Aguilar, A. B. Dawdy, and G. W. Law. The aerospace corporations concept design center. In *8th Annual International Symposium of the International Council on Systems Engineering*, 1998.

N. M. Alexandrov and M. Y. Hussaini. Multidisciplinary design optimization: State of the art. In *Proceedings of the ICASE/NASA Langley Workshop on Multidisciplinary Design Optimization*, 1997.

Table III. ADCS uncertainty specifications.

Variable	Probability Distribution	Variable	Probability Distribution
A_{max}	$N(5.31, 0.053)$	ψ_{slew5}	$N(2.76, 0.2)$
J_{xx}	$U(300, 450)$	ψ_{slew6}	$N(8.51, 0.4)$
J_{zz}	$U(450, 600)$	ψ_{slew7}	$N(9.88, 0.5)$
R	$N(1.3, 0.0013)$	ψ_{slew8}	$N(5.64, 0.2)$
δ_1	$N(0, 0.5)$	ψ_{slew9}	$N(5.04, 0.2)$
δ_2	$N(0, 0.5)$	ψ_{slew10}	$N(5.75, 0.2)$
g_s	$N(1400, 14)$	ψ_{slew11}	$N(4.47, 0.1)$
κ	$U(0.6, 0.7)$	ψ_{slew12}	$N(5.53, 0.1)$
ω_{spin0}	$N(12, 1.33)$	ψ_{slew13}	$N(5.85, 0.1)$
ω_{spin1}	$N(2, 0.0667)$	ψ_{slew14}	$\Gamma(1.5, 10.5)$
ω_{spin2}	$\Gamma(11, 0.25)$	ψ_{slew15}	$\Gamma(1.5, 10.5)$
ω_{spin3}	$L(2, 0.0667)$	ψ_{slew16}	$\Gamma(1.5, 10.5)$
ω_{spin4}	$N(48, 5)$	ψ_{slew17}	$\Gamma(1.5, 10.5)$
ω_{spin5}	$N(2, 0.0667)$	ψ_{slew18}	$\Gamma(1.5, 10.5)$
ψ_{slew1}	$N(5, 0.5)$	ψ_{slew19}	$\Gamma(1.5, 10.5)$
ψ_{slew2}	$N(50.45, 5)$	q	$N(0.6, 0.06)$
ψ_{slew3}	$N(5.13, 0.5)$	$uncfuel$	$N(0, 0.05)$
ψ_{slew4}	$N(6.35, 0.6)$	F	$N(F_{table}, 7/300F_{table})$

V. Amata, G. Fasano, L. Arcaro, F. Della Croce, M. Norese, S. Palamara, R. Tadei, and F. Fragnelli. Multidisciplinary optimisation in mission analysis and design process. GSP programme ref: GSP 03/N16 contract number: 17828/03/NL/MV, European Space Agency, 2004.

M. Bandecchi, S. Melton, and F. Ongaro. *Concurrent engineering applied to space mission assessment and design*. ESA Bulletin, 1999.

V. Belton and T. J. Stewart. *Multiple criteria decision analysis: an integrated approach*. Kluwer Academic Publishers, 2002.

P. Deuflhard. *Newton Methods for Nonlinear Problems. - Affine Invariance and Adaptive Algorithms.*, volume 35 of *Series in Computational Mathematics*. Springer, 2004.

D. Dubois and H. Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. New York: Plenum Press, 1986.

D. Dubois and H. Prade. Interval-valued fuzzy sets, possibility theory and imprecise probability. In *Proceedings of International Conference in Fuzzy Logic and Technology*, 2005.

EADS. Space Propulsion Web Page, 2007.
http://cs.space.eads.net/sp/.

J. K. Erickson. Mars exploration rover: Launch, cruise, entry, descent, and landing. In *55th International Astronautical Congress of the International Astronautical Federation, the International Academy of Astronautics, and the International Institute of Space Law*, Vancouver, Canada, October 2004.

S. Ferson. What monte carlo methods cannot do. *Human and Ecological Risk Assessment*, 2:990–1007, 1996.

S. Ferson, L. Ginzburg, and R. Akcakaya. Whereof one cannot speak: When input distributions are unknown. Risk Analysis, in press, 1996.
http://www.ramas.com/whereof.pdf.

M. Fuchs, D. Girimonte, D. Izzo, and A. Neumaier. Robust and automated space system design. accepted, 2007. http://www.martin-fuchs.net/publications.php.

M. C. Grant and S. P. Boyd. Cvx: A system for disciplined convex programming. 2007. http://www.stanford.edu/∼boyd/cvx/cvx_usrguide.pdf http://www.stanford.edu/∼boyd/cvx/.

N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001. http://www.bionik.tu-berlin.de/user/niko/cmaartic.pdf.

W. Huyer and A. Neumaier. SNOBFIT - stable noisy optimization by branch and fit. submitted preprint, 2006. http://www.mat.univie.ac.at/∼neum/ms/snobfit.pdf http://www.mat.univie.ac.at/∼neum/software/snobfit/.

W. Huyer and A. Neumaier. Global optimization by multilevel coordinate search. *J. Global Optimization 14*, pages 331–355, 1999. http://www.mat.univie.ac.at/∼neum/software/mcs/.

G. Karpati, J. Martin, M. Steiner, and K. Reinhardt. The integrated mission design center (IMDC) at NASA Goddard Space Flight Center. In *IEEE Aerospace Conference*, volume 8, pages 3657–3667, 2003.

P. N. Koch, T. W. Simpson, J. K. Allen, and F. Mistree. Statistical approximations for multidisciplinary optimization: The problem of size. *Special Issue on Multidisciplinary Design Optimization of Journal of Aircraft*, 36(1):275–286, 1999.

A. Kolmogoroff. Confidence limits for an unknown distribution function. *The Annals of Mathematical Statistics*, 12(4):461–463, 1941.

V. Kreinovich. *Random Sets: Theory and Applications*, chapter Random sets unify, explain, and aid known uncertainty methods in expert systems, pages 321–345. Springer-Verlag, 1997.

W. J. Larson and J. R. Wertz. *Space Mission Analysis and Design*. Microcosm Press, third edition, 1999.

M. Lavagna and A. E. Finzi. A multi-attribute decision-making approach towards space system design automation through a fuzzy logic-based analytic hierarchical process. In *Proceedings of the 15th International Conference on Industrial and Engineering. Applications of Artificial Intelligence and Expert Systems*, 2002.

K. Lewis and F. Mistree. Modeling interactions in multidisciplinary design: A game theoretic approach. *AIAA Journal*, 35(8):1387–1392, 1997.

D. J. McCormick and J. R. Olds. A distributed framework for probabilistic analysis. In *AIAA/ISSMO Symposium On Multidisciplinary Analysis And Design Optimization*, 2002.

M. McKay, W. Conover, and R. Beckman. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 221:239–245, 1979.

MER. Mars Exploration Rover Project, 2003. http://marsrovers.nasa.gov/mission/spacecraft.html.

A. Neumaier. On the structure of clouds. Manuscript, 2003. http://www.mat.univie.ac.at/∼neum/ms/struc.pdf.

A. Neumaier. Clouds, fuzzy sets and probability intervals. *Reliable Computing 10*, pages 249–272, 2004. http://www.mat.univie.ac.at/∼neum/ms/cloud.pdf.

A. Neumaier, M. Fuchs, E. Dolejsi, T. Csendes, J. Dombi, B. Banhelyi, and Z. Gera. Application of clouds for modeling uncertainties in robust space system design. ACT Ariadna Research ACT-RPT-05-5201, European Space Agency, 2007. Available on-line at http://www.esa.int/gsp/ACT/ariadna/completed.htm.

U. Nowak and L. Weimann. A family of Newton codes for systems of highly nonlinear equations - algorithm, implementation, application. Technical report, Konrad-Zuse-Zentrum fuer Informationstechnik Berlin, 1990. http://www.zib.de/Numerik/numsoft/CodeLib/codes/nleq1_m/nleq1.m.

M. Pate-Cornell and P. Fischbeck. Probabilistic risk analysis and risk based priority scale for the tiles of the space shuttle. *Reliability Engineering and System Safety*, 40(3):221–238, 1993.

Personal communication with ESA engineers, 2007.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C*. Cambridge University Press, second edition, 1992.

Purdue School of Aeronautics and Astronautics. Satellite Propulsion Web Page, 1998. http://cobweb.ecn.purdue.edu/∼propulsi/propulsion/rockets/satellites.html.

T. J. Ross. *Fuzzy Logic with Engineering Applications*. New York, NY: McGraw-Hill, 1995.

B. Roy. *Multicriteria methodology for decision aiding*. Kluwer Academic Publishers, 1996.

G. Shafer. *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton University Press, 1976.

D. P. Thunnissen. *Propagating and Mitigating Uncertainty in the Design of Complex Multidisciplinary Systems*. PhD thesis, California Institute of Technology Pasadena, 2005.

R. C. Williamson. *Probabilistic Arithmetic*. PhD thesis, University of Queensland, 1989.

A. Zonca. Modelling and optimisation of space mission prephase A design process in a concurrent engineering environment through a decision-making software based on expert systems theory, 2004.

Validation of imprecise probability models

[1]Scott Ferson, [2]William L. Oberkampf and Lev Ginzburg
[1]*Applied Biomathematics*
[2]*Sandia National Laboratories, Stony Brook University*
email: scott@ramas.com

Abstract: Validation is the assessment of the match between a model's predictions and any empirical observations relevant to those predictions. This comparison is straightforward when the data and predictions are deterministic, but is complicated when either or both are expressed in terms of uncertain numbers (i.e., intervals, probability distributions, p-boxes, or more general imprecise probability structures). There are two obvious ways such comparisons might be conceptualized. Validation could measure the discrepancy between the *shapes* of the uncertain numbers representing prediction and data, or it could characterize the differences between *realizations* drawn from the respective uncertain numbers. When both prediction and data are represented with probability distributions, comparing shapes would seem to be the most intuitive choice because it sidesteps the issue of stochastic dependence between the prediction and the data values which would accompany a comparison between realizations. However, when prediction and observation are represented as intervals, comparing their shapes seems overly strict as a measure for validation. Intuition demands that the measure of mismatch between two intervals be zero whenever the intervals overlap at all. Thus, intervals are in perfect agreement even though they may have very different shapes. The unification between these two concepts relies on defining the validation measure between prediction and data as the shortest possible distance given the imprecision about the distributions and their dependencies.

Keywords: validation, observation, prediction, distribution, interval, p-box

1. Introduction

Validation is the comparison of the predictions of a theory or model against empirical data (AIAA 1998; ASME 2006; Oberkampf and Truncano 2002; Oberkampf et al. 2004; Oberkampf and Barone 2006; Hills 2006; Trucano et al. 2006; Romero 2007; Ferson et al. 2008). It is often contrasted with verification, which is the checking of a model's implementation against the intended specification (Oberkampf et al. 2004; Oberkampf and Trucano 2007). We also contrast validation with calibration, which is the adjustment of the model's parameters or its structure for the purpose of improving the match between its predictions and empirical reality (Kennedy and O'Hagan 2001; Trucano et al. 2006). Measures of validation might be useful in a calibration, but the processes are entirely different in their goals. Calibration seeks to correct a model, and validation seeks only to measure how correct the model is.

Several approaches to validation have recently been suggested based on simple comparisons of trends in means (e.g., Oberkampf and Barone 2006), more elaborate hypothesis testing (e.g., Hills and Trucano 2002; Hills and Leslie 2003; Rutherford and Dowding 2003; Chen et al. 2004; Dowding et al. 2004), or still more comprehensive Bayesian schemes (e.g., Hanson 1999; Kennedy and O'Hagan 2001; Hazelrigg 2003; Zhang and Mahadevan 2003; O'Hagan 2006; Chen et al. 2006; 2007). This paper concerns only the basic question of how we should summarize and measure the discrepancies between a model's predictions and relevant empirical data. Oberkampf and Truncano (2007) called this problem the 'validation assessment'. Other important issues such as how such the measure could be used to inform or quantify the predictive capability of a model or deciding whether the model is adequate for some intended use are out of our present scope.

We consider validation assessment in a context where non-negligible uncertainty is present in the prediction or the data, or both. This uncertainty can come in different forms. It may arise from natural stochasticity or randomness in the world, perhaps owing to fluctuations in processes across space or through time, heterogeneity of individuals, or variability among engineered components. This uncertainty is objective in the sense that it exists irrespective of observation by humans and it is irreducible in the sense that empirical study does not necessarily reduce it. We call it aleatory uncertainty and recognize traditional probability theory as the primary calculus for addressing it. Aleatory uncertainty is often contrasted with epistemic uncertainty which is the partial ignorance, incertitude or imprecision that arises from incomplete or imperfect scientific study and comes from small sample sizes, missing data or data censoring or other measurement uncertainties, and perhaps doubt about the proper form of a model. Epistemic uncertainty is sometimes called subjective or reducible uncertainty because it's a function of the observer rather than physical reality and because it can in principle be reduced by empirical effort. Although probability theory has often been used to address epistemic uncertainty, other approaches are also employed, notably including interval analysis.

Recently, several researchers have suggested that methods beyond traditional probability theory might be necessary for models that must distinguish aleatory and epistemic uncertainty (Shafer 1976; Walley 1991; Klir and Wierman 1999; Oberkampf et al. 2001; Nikolaidis and Haftka 2001; Ferson et al. 2003; Helton and Oberkampf 2004; inter alia). We use the phrase 'uncertain number' (Ferson et al. 2003) to denote a varying or imperfectly known quantity that is mathematically characterized by an interval, probability distribution, p-box (Ferson et al. 2003), Dempster-Shafer structure (Shafer 1976; Oberkampf et al. 2001; Oberkampf and Helton 2005), random set (Matheron 1975; Molchanov 2006), set of probability measures or 'credal set' (Levi 1980), or similar structure from the theory of imprecise probabilities (Walley 1991). In general, an uncertain number can express both aleatory uncertainty and epistemic uncertainty. One might hold that a probability distribution, as a special case of an uncertain number, expresses purely aleatory uncertainty and an interval, also a special case, expresses purely epistemic uncertainty.

The engineering value of a model's quantitative prediction is a function of both its accuracy and its precision. The precision of a prediction expressed as an uncertain number is inversely related to the

epistemic uncertainty encoded in the uncertain number. This uncertainty is sometimes called 'non-specificity' (Klir and Wierman 1999) and might be quantified as the width of an interval or the breadth between the left and right bounds of a p-box. A validation assessment lets us quantify the second essential component determining the worth of the prediction: its accuracy in the face of empirical evidence.

Section 2 considers validation for the case where both prediction and data are represented by probability distributions. Section 3 considers the more elementary problem of validation when they are both intervals. Section 4 tries to harmonize the measures developed for these two special cases. Section 5 considers some alternative solutions, and section 6 offers some conclusions.

2. Validation Metric for Comparing Probability Distributions

The difference between two probability distributions can be characterized in many ways. The comparison could be conceived in terms of differences of their realizations (i.e., real numbers) or in terms of the discrepancies between their distribution shapes. In other words, if X and Y are random numbers distributed according to their respective cumulative distribution functions F and G, then we could talk about the distribution or average of $X - Y$, or we could focus on the difference between the shapes of F and G. The characterization that seems to be most useful in the context of validation of engineering models is based on comparing the shapes of the distributions of the random variables representing the prediction and relevant observations. Random variables whose distribution functions are identical are said to be 'equal in distribution'. If the distributions are not quite identical in shape, the discrepancy can be measured with any of many possible measures that have been proposed for various purposes in fields including statistical goodness of fit (e.g., Stephens 1974; Feller 1948; Kolmogorov 1941; Smirnov 1939), probability scoring rules (Winkler 1996; Lindley et al. 1979; de Finetti 1962; Brier 1950), information theory (Song 2002; Kullback 1959; Kullback and Leibler 1951), and texture analysis (e.g., Mathiassen et al. 2002).

Ferson et al. (2008) proposed to quantify the mismatch between prediction and observation with the area between the prediction's probability distribution and the empirical distribution of observations. This area is the Minkowski L_1 metric

$$d(F, S_n) = \int_{-\infty}^{\infty} \left| F(x) - S_n(x) \right| \mathrm{d}x ,$$

where F is the cumulative distribution representing the model's prediction for the random variable and S_n is the empirical distribution function for relevant observations X_i, $i = 1,\ldots,n$, of that random variable. The empirical distribution function is

$$S_n(x) = \frac{\#\{X_i \le x\}}{n}$$

where # denotes the cardinality of the set, so $S_n(x)$ is the fraction of values in the data set that are at or below each magnitude x. The validation metric is thus computed solely from the prediction F provided by the modeler and observations X_i provided by the empiricist. A small area means there is a good match, and a large area means that prediction and data disagree.

Figure 1 illustrates an example prediction distribution for rainfall as the smooth curve drawn in gray, together with the empirical distribution functions S_n for a hypothetical data set consisting of the values 770, 790, 820, 865 in millimeters of rain. The prediction distribution is approximately normal, with mean about 810 mm and variance of about 1700. The area of the shaded region between the two functions which measures their disagreement is almost 40 mm. Note that the empirical distribution function is zero for all values smaller than the minimum of the data and one for all values larger than the maximum of the data. Likewise, beyond the range of the prediction distribution, the value of $F(x)$ is either zero or one extending to infinity in both directions. For graphical clarity, however, these flat portions at probability zero or one are not depicted when the distributions are plotted.

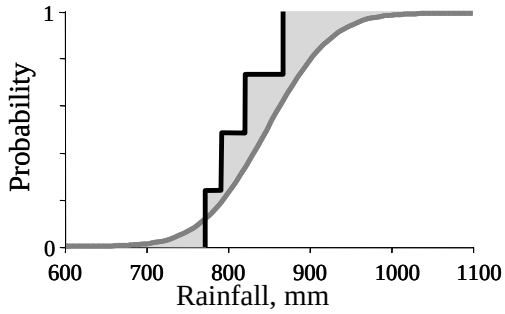


Figure 1. Area (shaded) between a prediction distribution (gray) and an empirical distribution function (black).

This metric can be computed for small data sets or even a single data value, in which case the S_n function would be the unit step function at that value. The approach can also be used even when the model is so complex and computationally expensive that it can only generate a small number of realizations for its prediction distribution. In such situations, the prediction distribution is modeled with an 'empirical' distribution formed from the sample realizations.

The area between the prediction distribution and the empirical distribution summarizing observations has several desirable properties as a formal validation measure of the mismatch between a model and evidence (Ferson et al. 2008). Most importantly, the area metric is an objective measure. Given a collection of observations and a prediction distribution, the area will be the same no matter who computes it because it does not depend on any judgments or parameters chosen by the analyst. Another important property is that the area metric generalizes deterministic comparisons between scalar values that have no uncertainty; if the prediction and the observation are both scalar point values, the area is equal to their difference. The area will tend not to be overly sensitive to minor discrepancies in the distribution tails (assuming the area is finite), but it obviously reflects the full distributions in assessing performance. In particular, it is clearly not merely a measure of the difference in the means or even the means and variances, but takes account of any differences between the prediction and observation distributions. Because probability is dimensionless, the units of the area are the same as those of the system response quantity in which the prediction and data are expressed. This property is very important in making the measure intuitively meaningful to engineers. Its units are the same as one would expect for the result of a subtraction. If it were some dimensionless index or, worse, had some complex or esoteric statistical units, its physical interpretation would be difficult. The area measure is also unbounded in the sense that, if the prediction is completely off the mark of the observations, the area characterizing this discrepancy can in principle grow to be an arbitrarily large value, which is also an intuitive feature of distances. Finally, the area measure is mathematically well behaved and well understood. So long as the area converges to a finite value, it is a true metric in the mathematical sense, which means it has the essential features of a distance function. By definition, a mathematical metric d has four properties (Fréchet 1906):

non-negativity,	$d(x, y) \geq 0$,
symmetry,	$d(x, y) = d(y, x)$,
triangle inequality,	$d(x, y) + d(y, z) \geq d(x, z)$, and
identity of indiscernibles,	$d(x, y) = 0$ if and only if $x = y$.

All of these properties suggest that the area metric will be more comprehensive and easier to interpret than any of several alternative statistical measures or some distance measure based on merely matching prediction and observation distributions in the mean or in both mean and variance.

Ferson et al. (2008) also showed how the area metric could be extended to synthesize evidence of the conformance between model and data into a single measure when observations are to be compared to *different* prediction distributions. The trick is to transform each observation X_i to $u_i = F_i(X_i)$ where F_i is the prediction distribution against which X_i is to be compared. The u_i express all the available evidence on a universal scale of probability. By the probability integral transform theorem (Angus 1994), the u_i will be uniformly distributed over the unit interval [0,1] so long as the original X_i are distributed according to their respective prediction distributions F_i, which is to say, so long as the model is predicting the observations well. Statistical tests and diagnostics are straightforward to define for this synthesis. The model's performance can be assessed directly in terms of the u_i, or the values may first be back-

transformed to a common axis that re-expresses the evidence in physical units. The back-transformation can be chosen so as to maximize the relevance of the assessment for a particular regulatory or performance question. This strategey can even be used to combine evidence about model-data conformance collected in entirely different dimensions (such as, for instance, rainfall and temperature). This synthesis abandons the interpretation of the area in original units of course, but it does allow analysts to compare the relative performance of the model for different system response quantities in a meaningful way.

2.1. WHY NOT BASE THE METRIC ON DIFFERENCES OF VALUES FROM THE TWO DISTRIBUTIONS?

One could imagine developing an alternative validation measure based on the absolute difference between a random value realized from the prediction distribution and a random value drawn from the data distribution. There would of course be a distribution of such differences. It might seem preferable to use this distribution of differences to characterize the disagreement between probability distributions (Menger 1942). A distribution could be more informative than the area metric which is a crude scalar summary that could not capture the information embodied in an entire distribution. The distribution of differences could be used itself as a characterization of the disagreement between the two distributions, or it might be summarized in various ways that might highlight aspects of the disagreement of special interest. But such a notion would need to consider the *stochastic dependence* between random values from the two distributions. Specifying an assumption about the dependence is necessary to define the distribution of differences $X - Y$ from specified distributions for X and Y. Are the values statistically independent? Do they have some correlation or a nonlinear dependence? Different assumptions can lead to starkly different distributions for the random difference.

Consider, for example, a weather model that predicts daily temperatures and, by aggregating these values, also predicts a distribution of daily temperatures over the course of a year. Suppose that relevant daily temperature observations are available. It may be the case that the predicted distribution of temperatures over the year matches the observed distribution of temperatures very well and yet the correlation between predicted and observed daily temperatures is markedly poor. For instance, if the model is out of phase with respect to seasons, it may be predicting summer temperatures during the winter and vice versa, which would lead to a correlation close to −1, even though it gets the distributions exactly right. The performance of such a model would have to be considered very poor in any sensible validation assessment. But note that this poor performance is really associated with the deterministic results from the model rather than the probabilistic ones per se. If the model had not made the deterministic predictions and confined itself to purely probabilistic forecasts, this problem would not have arisen.

Contrast the weather model with another model that does not predict individual daily temperatures, but only the summary *distribution* of daily temperatures. Essentially, this retreat changes the weather model into a climate model that does not make predictions about the temperature for any particular day, except to assert that, considered as a group over the course of many days, these temperatures will

converge in distribution to the prediction. And it is certainly not making any predictions about the dependence between values that might be drawn from the prediction distribution and observed temperature values. The model is not even saying that such temperature pairs are independent. In fact, actual temperatures have strong autocorrelation from day to day, so supposing that temperatures should be drawn independently from the predicted distribution would obviously be empirically incorrect too. It is possible, of course, to construct a probabilistic weather model of daily temperatures. Such a model might predict a probability distribution for each and every day's temperature. But these predictions would not be saying anything about dependence or even about randomness; they are asserting only that $F_i(X_i)$ are uniformly distributed, where F_i is the probability prediction for day i and X_i is the observed temperature for that day. In any case, if the model refrains from making deterministic forecasts and makes only purely probabilistic predictions about distributions without characterizing dependence, then the model would have excellent performance in a validation assessment.

If the model asserts nothing about the possible dependence between predicted and observed values of a system response quantity, then the distribution of differences between predictions and data cannot be uniquely defined. Thus, it would be seem to be impossible to base a validation metric on the distribution differences. It is possible, however, to *bound* the distribution of absolute differences even without specifying anything about the dependence between the subtrahend and the minuend. Elementary probability bounds analysis (Frank et al. 1987; Williamson and Downs 1990; Ferson 2002; Ferson et al. 2003) can be used to compute these bounds, which may be informative. Figure 2 depicts four examples of validation as characterized by the area metric and bounds on the distribution of differences. In the upper panel of graphs, prediction distributions F are depicted as gray curves, and data distributions S_n are depicted as black step functions. Under each of these four graphs, the corresponding area metric is plotted as a dotted spike. On the same graph, bounds on the distribution of absolute differences between random values from the prediction and data distributions are shown as thin lines. In each of the four comparisons, the prediction is a normal distribution with mean 2 and standard deviation 0.2, truncated at the 0.5^{th} and 99.5^{th} percentiles. In the first comparison, the data consists of a single observed value at 4, so the empirical distribution is degenerate. The validation metric in this case is 2 units, which is the area between the truncated normal and this degenerate step function. The distribution of absolute differences between random values drawn from the prediction distribution and the observed value 4 ranges between 1.5 and 2.5. The data forming the empirical distribution in the second comparison comes from 8 measurements scattered between roughly 2.2 and 3.2. The area validation metric in the second comparison is almost 0.6 units. Without specifying the stochastic dependence between the prediction and data distributions, it is impossible to define the distribution of their differences, but probability bounds analysis can bound the distribution (Ferson 2002). The thin lines in the second graph of the lower panel of Figure 2 represent the best-possible bounds on the distribution of absolute differences between predicted values and observed values. The breadth of the bounds comes from not making any assumption about the dependence between the two distributions. In the third comparison, the empirical data have a larger dispersion and the resulting area metric is somewhat larger. In the fourth comparison, the data values come much closer to the prediction distribution, so the area metric is much closer to zero. Note, however, that the distribution of differences could nevertheless include values close to one.

These few examples convey an idea for how the area metric and the bounds on the distribution of differences compare to each other. The bounds tell us how wrong we might be if dependence matters, but they do not contain the information needed to compute the area metric, so, insofar as the area metric is important or informative, the bounds on differences are incomplete as a summarization of the disagreement. Likewise, the bounds contain information not encapsulated in the area metric as well, although engineering judgment does not seem to recognize the information in the bounds as particularly relevant to the question of whether the distributions match well.

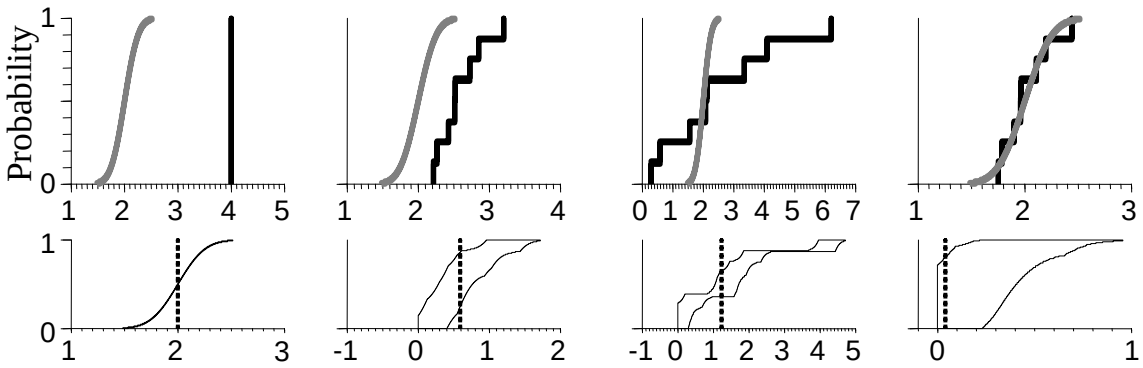


Figure 2. Predictions (gray) and data (black) yielding area metrics (dotted) and difference distributions (thin p-boxes).

3. Validation Measure for Comparing Intervals

Predictions should include epistemic uncertainty if it exists in our knowledge about the modeled physical process. Indeed, except in rare situations, precise predictions are not reasonable in real-world problems, or they only result from assumptions that modelers themselves do not unequivocally believe. Although a model may give point predictions, there is almost always an implied precision associated with each quantity. Modern notions of best practice argue that these implicit considerations be made explicit, and more and more modelers are accepting this and incorporating uncertainty analyses into their models. The simplest quantitative expression of epistemic uncertainty is an interval. Giving an interval as the representation of an estimated quantity is asserting that the value (or values) of the quantity lie somewhere within the interval. Intervals can arise in both predictions and observations. When a prediction is an interval, its width relates the modeler's inability to nail down the prediction precisely. The modeler is saying the quantity in question is within a particular range, but not saying any more than this. In particular, the modeler is not making any assertion about which possible values might be more likely than which other possible values. If there is such extra information available about a prediction, but too little to justify the selection of a particular probability distribution, the information can be expressed in a more general uncertain number such as a p-box, Dempster-Shafer structure or credal set.

Empirical observations can also contain epistemic uncertainty. Again, the simplest form of this is an interval. Uncertainty about measurements that is appropriately characterized by intervals is called incertitude, and it arises naturally in a variety of circumstances, including plus-or-minus reports, significant digits, intermittent measurement, non-detects, censoring, data binning, rounding or bit compression in data transmission, missing data and gross structural ignorance (Ferson et al. 2007; 2004). When a collection of such intervals comprise a data set, one can think of the breadths of the intervals as representing epistemic uncertainty while the scatter among the intervals represents variability or aleatory uncertainty. Recent reviews (Manski 2003; Gioia and Lauro 2005; Ferson et al. 2007) have described how interval uncertainty in data sets produces uncertain numbers containing epistemic uncertainty. When empirical observations have uncertainty of this form that is too large to simply ignore, these elementary techniques can be used to characterize it is a straightforward way.

The comparison between two fixed real numbers reduces to the scalar difference between the two. Suppose that, instead of both numbers being reals, at least one of them is an interval range representing acknowledged uncertainty. If the prediction and the observation overlap, then we should say that the prediction is *correct*, in an important sense, relative to the observation. If the prediction is an interval, this means that the model, or perhaps one would say the modeler, is being modest about what is being claimed. For example, the assertion that a regional maximum temperature will be between 20 and 40 °C is a weaker claim than saying it will be exactly 30. And it is a stronger claim than saying the temperature will be between 10 and 60. In the extreme case, a vacuous prediction, while not very useful, is certainly true, if just because it isn't claiming anything that might be false. For example, predicting that some probability will be between zero and one doesn't require any bravery, but at least it is free from contradiction. It is proper that a prediction's express uncertainty be counted toward reducing any measure of mismatch between theory and data in this way because the model(er) is admitting doubt. If it were not so, an uncertainty analysis could otherwise have no epistemological value. From the perspective of validation, when the uncertainty of prediction encompasses the actual observation, the prediction ought to be regarded as true, because *validity is distinct from precision*. Both are important in determining the usefulness of a model, but it is reasonable to distinguish them and give credit where it is due.

A reciprocal consideration applies, by the same token, if the datum is an interval to be compared against a prediction that's a real number. Validation has to give to the model whatever benefit of the doubt that arises because of the uncertainty about the datum. For instance, if the prediction is, say, 30% and the observation tells us that it was somewhere between 20% and 50%, then we would have to admit that the prediction might be perfectly correct. If on the other hand the evidence was that it was between 35% and 75%, then we would have to say that the disagreement between the prediction and the observation might be as low as 5%. We could also be interested in how bad the comparison might be, but a validation metric shouldn't penalize the model for the empiricist's imprecision. In most conceptions of the word, the 'distance' between two things is the length of the shortest path between them. The distance between England and France is the breadth of the English Channel between Dover and Calais; it doesn't matter that Newcastle and Marseilles are much further apart. Similarly, the validation measure between a point prediction and an interval datum, or vice versa, should be the shortest difference between the

characterizations of the quantities. Likewise, the validation measure between an interval prediction and an interval datum is the shortest distance between the two intervals, which will be zero if they overlap. Symbolically, the validation measure for comparing intervals A with B is

$$\inf_{\substack{X \in A \\ Y \in B}} |X - Y|.$$

where inf denotes the infimum (which just generalizes minimum for intervals that might be open or partially open). Although this choice for the validation measure shares a similar graphical intuition with the area metric discussed in section 2, this measure is quite different from it. Note, for instance, that this measure is not a mathematical metric. It violates the property of identity of indiscernibles, because a value of zero for the measure does not imply that the intervals are identical. Mathematicians call a non-negative, symmetric function that satisfies the triangle inequality but not identity of indiscernibles a 'pseudometric'. More fundamentally, this measure is not based on the shapes of the intervals like the area metric was based on the shapes of the probability distributions. Indeed, the shape of the intervals could be wildly different yet still yield a value of zero for the validation measure if they overlap at all. In fact, the formula above suggests that the measure is based instead on considering possible *realizations* of values X and Y from the respective intervals.

4. Unification of the Two Conceptualizations for General Uncertain Numbers

The key to harmonizing the shape-based comparison described in section 2 with the realization-based comparison described in section 3 is to recognize that both are essentially special cases of the Wasserstein distance (Vallender 1973; Dobrushin 1970)

$$\inf_{\substack{X \sim F \\ Y \sim Sn}} \mathrm{E}|X - Y|,$$

where the E denotes the expectation operator, and the infimum is taken over all possible random variables X and Y that are distributed according to F and S_n respectively. When the prediction F and the data distribution S_n are probability distributions, the infimum searches over all possible stochastic dependencies between the random variables X and Y (constrained by the fact that they must respect their marginal distributions F and S_n). The Wasserstein distance is a metric for any distributions for which the infimum is finite (Dobrushin 1970). When the random variables are univariate, then it equals the area metric (Vallender 1973). The infimum occurs when the X and Y are comonotonic, that is, when the

dependence between X and Y is perfect, and the correlation between them is a large as is possible given their marginal distributions. It is this fact that creates the graphical interpretation as the area between the distributions.

When the prediction and data are intervals, we interpret the tilda to mean 'is an element of' and ignore the E operator (because intervals do not have probability measures defined over them) so that the Wasserstein distance is the same as our intuitive formula for the validation measure for intervals described in section 3.

The generalization of the Wasserstein distance for uncertain numbers is now clear: it should be the infimum expectation of the absolute value of the difference between the variates, where the infimum is taken over all possible distribution with respective uncertain numbers *and* under all possible dependencies between those distributions. The computational task of identifying this infimum may be challenging for some uncertain numbers such as credal sets, but it turns out to be rather simple for p-boxes. The area measuring mismatch for general p-boxes is the integral

$$\int_{-\infty}^{\infty} \Delta([F_{\mathrm{R}}(x), F_{\mathrm{L}}(x)],[S_{n\mathrm{R}}(x), S_{n\mathrm{L}}(x)])\,\mathrm{d}x$$

where F and S_n denote the prediction and the data distributions, respectively, and the subscripts L and R denote the left and right bounds for those distributions, and

$$\Delta(A,B) = \min_{\substack{a\in A \\ b\in B}} |a-b|$$

is the shortest distance between two intervals, or zero if the intervals touch or overlap. This measure integrates the regions of non-overlap between the two sets of bounds, for every value along the probability axis.

The thin p-boxes in the lower panel of graphs in Figure 2 are bounds on all possible distributions of the difference between the two random values. Instead of all possible distributions, we want the mean of the precise distribution of differences assuming perfect dependence between the prediction F and data distribution S_n. We might therefore characterize this measure as the *mean perfect absolute difference of deviates*, but perhaps it will suffice to continue to call it the 'area measure'. It is important to keep in mind that we're *not* selecting perfect dependence as our model of how the prediction and observation distributions are expected to be related to each other. Perfect dependence would mean that locally large observations would always be associated with locally large predicted values, and small with small, in a

very strict fashion. We certainly do not believe that they would be related in this way in reality. Perfect dependence just falls out of the formula because it is the dependence that leads to the smallest possible value of the mean of the absolute differences. The smallest area is the one of interest because the distance between two things is the length of the shortest connection between them. At least for p-boxes, this also has the happy graphical interpretation as the area between the prediction and the observation.

Figure 3 depicts four more examples. As before, predictions are depicted in the upper panel in gray, and data are depicted in black, but now they are p-boxes rather than precise distributions. Under each of these four graphs, the corresponding area measure is shown as a dotted spike and bounds on the all distribution of differences between random values from the prediction and observation p-boxes are shown as thin lines. In each of the four comparisons, the prediction is a p-box of normal distributions whose means are in the interval [1.75, 2.25] with standard deviation 0.2, truncated at the 0.5^{th} and 99.5^{th} percentiles. In the first comparison, the data consists of a single interval [4,5], so the resulting area measure is 1.75. It is the area between the rightmost normal distribution inside the gray p-box and the leftmost scalar inside the black interval. It seems reasonable that the discrepancy between the prediction and data in this case is only 1.75 units even though the difference between a predicted value and an observed data value could be larger than 3.5 units. The wide breadth of the bounds on the differences comes from the epistemic uncertainty about the prediction distribution and the data distribution within their respective p-boxes and also from not making any assumption about the dependence between them. The data in the second comparison comes from 8 measurements for which measurement incertitude was ±0.25. The 8 intervals implied by this incertitude were cumulated into a p-box describing epistemic uncertainty about the empirical distribution function (Ferson et al. 2007). The area in the second comparison is about 0.34, which is the between the right edge of the graph prediction and left edge of the black data p-box. In the third comparison, the empirical data had the same sample size and the same incertitude as in the second comparison, but the values happened to have a larger dispersion. In this case, the area is the sum of the two areas where the gray prediction p-box and the black data p-box do not overlap. In the fourth comparison, the data values had the same measurement uncertainty but a smaller dispersion and central tendency so the area measure is zero because there exist distributions that lie within both the prediction and data p-boxes.

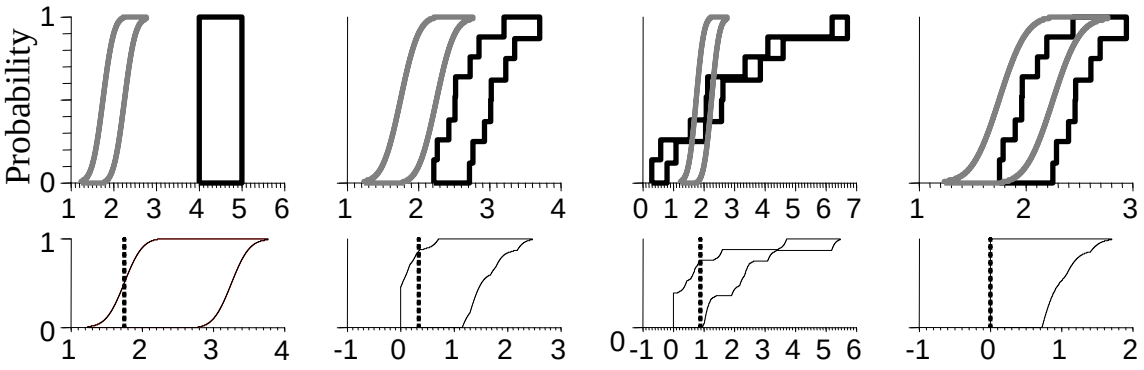


Figure 3. Predictions (gray) and data (black) yielding area measures (dotted) and difference p-boxes (thin).

5. 'Same Shape' versus 'Possibly Equal'

Although we think that using the area distance when the prediction and observations are uncertain numbers as described in the previous section is appropriate both mathematically and in practical engineering terms, we acknowledge that there are several other ways this generalization could be conceived. This section introduces three alternative generalizations of the area metric for use when uncertain numbers are used to characterize predictions or observations.

The area metric proposed in section 2 is based on the distribution functions of the predictions and the data, as distinguished from the random variables those distributions summarize. Although we chose to compare the shapes of the probability distributions when the quantities had only aleatory uncertainty, this choice does not seem satisfactory when there is epistemic uncertainty present as well. The area measure between the prediction and data in the general case as described in section 4, is no longer a mathematical metric when at least one is an interval or a more general uncertain number because the area can fall to zero without the prediction and data becoming identical (as in the rightmost graph of Figure 3). In section 4, the application of the area measure when prediction and data are characterized as uncertain numbers was based on the conventional idea that distance between two things is the length of the *shortest* line between them. There are, however, different ways to look at the question. A standard mathematical way to construct a metric between two potentially overlapping sets is to define

$$\max\left(\sup_{F \in x} \inf_{G \in y} d(F,G), \sup_{G \in y} \inf_{F \in x} d(G,F) \right)$$

where F is an element of the first set x and G is an element of the second set y and d is a metric on the space containing the sets (Pompieu 1905), which is our case would just be the area metric. The elements F and G are possible distribution functions taken from the respective prediction and data uncertain numbers x and y. This function is zero if and only if the set of distributions representing the prediction is the same as the set of distributions representing the data, that is, if their respective uncertain numbers had identical shapes. This function constitutes a much stricter view about agreement between prediction and data. It holds that perfect agreement involves not only overlapping but having exactly the same imprecision. Generalizing the area distance using this function would mean that our measure would remain a true mathematical metric, but it seems overly strict about what constitutes perfect agreement. For instance, suppose that the theoretical prediction is a simple interval and is to be compared with an observation that is also an interval and that the prediction interval is a *subset* of the observation interval. In other words, the prediction and observation agree in that they overlap, but the imprecision about the observation is wider than that of the prediction. It doesn't seem reasonable to insist that the theory and data are somehow not in perfect agreement in this situation, nor to require that the theory somehow inflate the uncertainty of its prediction simply to match the poorer precision associated with the observation.

Another way to generalize the area metric for uncertain numbers considers comparisons between distributions realized from the uncertain numbers, rather than the shapes of the uncertain numbers. For example, it might be natural to find upper and lower bounds on the areas between distributions that are consistent with the two uncertain numbers. Rather than differences between pairs of bounds, this would be bounds on differences between pairs of distributions. In this case, the measure would be the smallest and largest possible values of the underlying metric

$$\left[\inf_{F \in x, G \in y} d(F,G), \sup_{F \in x, G \in y} d(F,G) \right],$$

where F and G are distribution functions within (consistent with) the respective uncertain numbers x and y. The range would be degenerate, i.e., the infimum and supremum would be the same if the two uncertain numbers are actually particular probability distributions, neither having any epistemic uncertainty. The range being double-zero would mean that the prediction and the data distribution are identical, and that neither has any epistemic uncertainty. This generalization is not a metric because it does not have the property of identity of indiscernibles; x and y could be identical and not yield a double-zero.

Note that this scheme, like the Pompieu scheme, can be very difficult computationally because there are infinitely many distributions within the uncertain numbers to be compared. It obviously does not suffice to compare extreme distributions corresponding to the edges of the uncertain numbers. For example, consider the leftmost graph of Figure 4. It is intuitively clear that that the smallest possible value of the area between a distribution inside the prediction bounds and a distribution inside the observation bounds corresponds to the shaded area. This area corresponds to a prediction distribution that follows the

left edge of the prediction bounds (smooth gray bounds) for small probability levels and follows the right edge of the prediction bounds for large probabilities. The corresponding distribution consistent with the observation bounds (black step bounds) conversely follows the right edge of those bounds for small probabilities and the left edge for large probabilities. For intermediate probabilities, the prediction distribution and the empirical distribution are coincident monotone curves in the region where the bounds overlap. The *largest* possible area, however, is not so easy to discern from the graph. The two distributions that lead to the largest possible area are depicted on the rightmost graph of Figure 4. The distribution from within the prediction bounds is shown as a dashed line; the distribution from within the observation bounds is shown as a dotted line. The area between these two distributions is shaded in the middle graph of the figure. The non-intuitive shape of the shading gives a hint at the computational complexity of bounding the area metric. This scheme of bounding the area is not itself a mathematical metric. Firstly, it produces two numbers rather than a single scalar. Secondly, it does not satisfy the property of identity of indiscernibles. Even if the prediction uncertain number is identical to the data uncertain number, the upper bound will not be zero (unless there is no epistemic uncertainty).

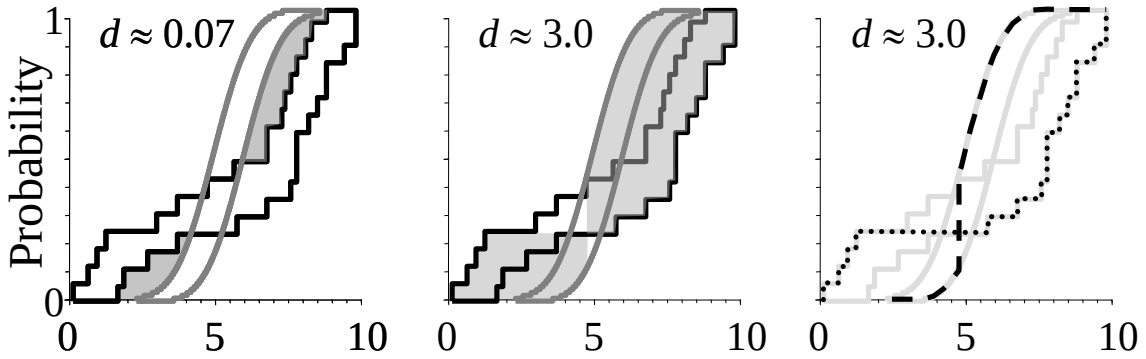


Figure 4. Smallest (left) and largest (middle and right) possible areas between a distribution inside the uncertain predictions (gray bounds) and a distribution inside the uncertain empirical observations (black step functions). The extremal distributions yielding the largest area are depicted in the right graph.

As yet another alternative, we could generalize our validation metric as the *two-dimensional* vector $Д(x,y) = (d(x_L, y_L), d(x_R, y_R))$ where the subscript L denotes the left side of a p-box and the subscript R denotes the right side, and d is our regular area metric for distributions. The left value of the pair reflects the difference between the left side of the prediction and the left side of the observations. Likewise, the right side of the distance pair reflects the difference between the right side of the prediction and the right side of the observations. This pair would constitute what we might call a double metric, $Д: \mathbf{B} \times \mathbf{B} \rightarrow \Re^+ \times \Re^+$, where $\mathbf{B}$ is the set of all p-boxes (which includes intervals, probability distributions and scalars as special cases), and $\Re^+$ is the set of all positive real numbers, satisfying the following generalizations of the four metric properties:

$$
\begin{aligned}
&Д(x, y) = (a, b) \text{ implies both } a \geq 0 \text{ and } b \geq 0 && \text{(non-negativity)},\\
&Д(x, y) = Д(y, x) && \text{(symmetry)},\\
&Д(x, y) = (0,0) \text{ if and only if } x = y && \text{(identity of indiscernibles), and}\\
&\left\{ \begin{array}{l} Д(x, y) = (a_1, b_1) \\ Д(y, z) = (a_2, b_2) \\ Д(x, z) = (a_3, b_3) \end{array} \right\} \text{imply } a_1 + a_2 \geq a_3 \text{ and } b_1 + b_2 \geq b_3 && \text{(triangle inequality).}
\end{aligned}
$$

Figure 5 shows three examples of this double metric. In the leftmost graph, a scalar prediction at $x = 7$, depicted as a gray spike, is compared to an interval observation $y = [14, 19]$ shown in black. The value 7 is compared against both sides of the interval to yield $Д(x, y) = (|14–7|, |19–7|) = (7, 12)$. In the middle graph, the comparison is between two intervals, and the two-dimensional difference is $Д([4, 9], [13, 18]) = (|13–4|, |18–9|) = (9, 9)$. In the rightmost graph of Figure 5 the black observation interval overlaps with the gray prediction interval. The double metric is $Д([3, 11], [8, 17]) = (|8–3|, |17–11|) = (5, 6)$. The value of the double metric would be (0,0) when the corresponding edges coincide exactly. Being double-zero would not mean that the uncertainty in either the evidence or prediction has gone to zero, but only that they match in both location and imprecision.

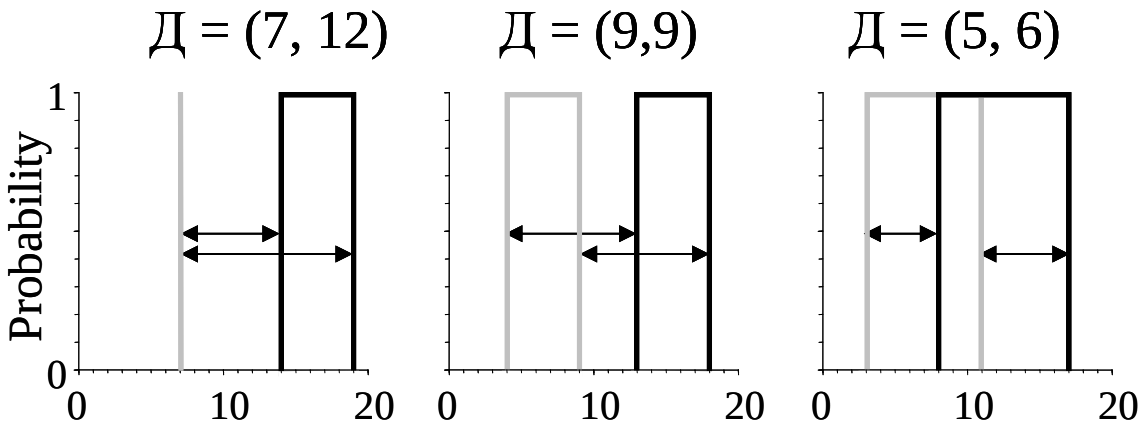


Figure 5. A generalized, two-dimensional metric between uncertain numbers (intervals).

Four possible generalizations of the area metric for epistemic uncertainty in predictions or observations have been discussed in this and the previous section. None has all the properties one might desire. Neither the shortest distance nor the range of possible areas is a true mathematical metric because they do not have the property of identity of indiscernibles. In the case of the shortest distance, the distance

being zero does not guarantee that the representation of the prediction is identical to the representation of the observations. In the case of the range of possible areas, if the prediction and observation representations are identical, the value will not generally be zero. The double metric and Pompieu's max-sup-inf both have formal metric properties (or at least generalizations of them), but they seem to be overly strict in that predictions must match observations in their uncertainties even though there's no physical or engineering reason to demand this. The double metric is the easiest to compute, followed by the shortest distance. Pompieu's max-sup-inf and the range of possible areas are hardest to compute. The shortest distance measure and the double metric are both based on the comparing the shapes of the representations of the prediction and observations, whereas the other two measures are based on comparing individual elements (i.e., distribution functions consistent with those representations). The table below summarizes these observations.

Measure	*Scheme*	*Metric*	*Compute*	*Strictness*
Shortest distance	Shape	No	Medium	Reasonable
Pompieu's max-sup-inf	Element	Yes	Hard	Too strict
Range of possible areas	Element	No	Hard	Reasonable
Double metric	Shape	Yes	Easy	Too strict

We expect that the shortest distance will be most useful in many practical applications. In some situations, the range of possible areas will be most informative.

The comparison between random numbers characterized by probability distributions could be understood in terms of their difference as real numbers that are *realizations* from those distributions or in terms of the discrepancies between the *shapes* of those distributions. When there is only aleatory uncertainty associated with the prediction and observations, it seems reasonable to use the latter comparison based on distribution shapes for the purposes of validation. The analogous comparison between uncertain numbers, i.e., characterizations of numerical quantities that express both aleatory and epistemic uncertainty, can also be considered in these two senses. But comparing the shapes of distributions does not seem completely satisfactory when there is epistemic uncertainty present as well. There are several approaches possible for handling epistemic uncertainty based on the area metric. Two of these approaches seem most promising. The first is based on comparing shapes and considers the measure of the disagreement to the smallest possible value of the area metric that would be consistent with distributions from within the express uncertainty. The second approach, based on realizations, considers the range of possible values of the area metric consistent with distributions within the uncertainty.

6. Conclusions

The comparison between random numbers that are characterized by probability distributions can be understood in terms of their difference as real numbers that are realizations from those distributions, or in

terms of the discrepancies between the shapes of their distributions. It seems reasonable to use the latter comparison based on distribution shapes for the purposes of validation for (precise) probabilistic models. The analogous comparison between uncertain numbers, i.e., characterizations of numerical quantities that simultaneously express both aleatory and epistemic uncertainty, can also be considered in these two senses. But, whereas we chose to compare the shapes of the probability distributions when the quantities had only aleatory uncertainty, this choice does not seem satisfactory when there is epistemic uncertainty present as well. In the case of comparing two simple intervals which contain only epistemic uncertainty, if the prediction interval overlaps with the datum interval, then the prediction is perfectly correct from the perspective of a validation assessment. The shapes of the two intervals could be quite different, and indeed, their overlap could be very small, yet the validation measure of their mismatch is zero if they overlap at all.

There are several ways to unify and extend these apparently disparate notions of validation for the case of general uncertain numbers that include both epistemic and aleatory uncertainty. Perhaps the most workable is the smallest area between the uncertain numbers. This is the smallest possible area between probability distributions contained in the respective uncertain numbers under any possible dependence. For many situations in which p-boxes are used to characterize the prediction and the data, the smallest area is easy to compute when the edges of the p-boxes represent admissible distributions. In these cases, the smallest area is the mean of the distribution of differences of the extremal distributions computed under the assumption of perfect dependence.

Acknowledgements

We thank Marty Pilch, Kevin Dowding and Laura Swiler at Sandia National Laboratories, Bob Nau of Duke University and Wei Chen of Northwestern University. This paper is a product of work supported by the Sandia Epistemic Uncertainty Project under contract 19094 and supported by NASA Small Business Innovation Research grants NNL06AA40P and NNL07AA06C. All opinions are those of the authors and not necessarily of the supporting agencies.

References

AIAA. 1998. Guide for the Verification and Validation of Computational Flui Dynamics Simulations. AIAA G-077-1998. American Institute of Aeronautics and Astronautics, Reston, Virginia.

Angus, J.E. 1994. The probability integral transform and related results. *SIAM Review* 36(4): 652–654.

ASME. 2006. Guide for Verification and Validation in Computational Solid Mechanics. ASME V&V 10-2006. American Society of Mechanical Engineers. http://catalog.asme.org/Codes/ PrintBook/VV_10 _2006_Guide_Verification.cfm.

Berger, J.O. 1985. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Brier, G.W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78: 1–3.

Chen, W., L. Baghdasaryan, T. Buranathiti and J. Cao. 2004. Model validation via uncertainty propagation. *AIAA Journal* 42: 1406-1415.

Chen, W., Y. Xiong, K.-L. Tsui, and S. Wang. 2006. Some metrics and a Bayesian procedure for validating predictive models in engineering design. *Proceedings of ASME 2006 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Philadelphia. American Society of Mechanical Engineers. http://ideal.mech.northwestern.edu/pdf/DAC_ validation06.pdf.

Chen, W., Y. Xiong, K.-L. Tsui, and S. Wang. 2007. A design-driven validation approach using Bayesian prediction models. *Journal of Mechanical Design* [in press].

Colyvan, M. 2004. The philosophical significance of Cox's theorem. *International Journal of Approximate Reasoning* 37(1): 71–85. http://homepage.mac.com/mcolyvan/papers/cox.pdf.

Cox, R.T. 1946. Probability, frequency and reasonable expectation. *American Journal of Physics* 14:1–13.

Devore, J.L. 2000. *Probability and Statistics for Engineers and Scientists*. Duxbury, Pacific Grove, California.

Dobrushin, R.L. 1970. Prescribing a system of random variables by conditional distributions. *Theory of Probability and its Applications* 15: 458–486.

Dowding, K.J., M. Pilch and R.G. Hills. 2008. Formulation of the thermal problem. *Computer Methods in Applied Mechanics and Engineering* [in press].

Dowding, K.J., R.G. Hills, I. Leslie, M. Pilch, B.M. Rutherford an M.L. Hobbs. 2004. *Case Study for Model Validation: Assessing a Model for Thermal Decomposition of Polyurethane Foam*. SAND2004-3632, Sandia National Laboratories, Albuquerque, NM.

Draper, D. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B* 57: 45–97.

Feller, W. 1948. On the Kolmogorov-Smirnov limit theorems for empirical distributions. *Annals of Mathematical Statistics* 19: 177–189.

Ferson, S. 2002. *RAMAS Risk Calc 4.0 Software: Risk Assessment with Uncertain Numbers*. Lewis Publishers, Boca Raton, Florida.

Ferson, S. and L.R. Ginzburg. 1996. Different methods are needed to propagate ignorance and variability. *Reliability Engineering and Systems Safety* 54:133–144.

Ferson, S., V. Kreinovich, L. Ginzburg, D.S. Myers, and K. Sentz. 2003. *Constructing Probability Boxes and Dempster-Shafer Structures.* SAND2002-4015, Sandia National Laboratories, Albuquerque, New Mexico. http://www.ramas.com/unabridged.zip.

Ferson, S., C.A. Joslyn, J.C. Helton, W.L. Oberkampf and K. Sentz. 2004. Summary from the epistemic uncertainty workshop: consensus amid diversity. *Reliability Engineering and System Safety* 85: 355–370.

Ferson, S., V. Kreinovich, J. Hajagos, W.L. Oberkampf and L. Ginzburg 2007. *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*. SAND2007-0939, Sandia National Laboratories, Albuquerque, NM. http://www.ramas.com/intstats.pdf.

Ferson, S., W.L. Oberkampf and L. Ginzburg. 2008. Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering* [in press]. Available at http://www.ramas.com/thermval.pdf.

de Finetti, B. 1962. Does it make sense to speak of "good probability appraisers"? Pages 357–363 in *The Scientist Speculates: An Anthology of Partly-Baked Ideas*, I.J. Good (ed.). Wiley, New York.

Frank, M.J., R.B. Nelsen and B. Schweizer 1987. Best-possible bounds for the distribution of a sum—a problem of Kolmogorov. *Probability Theory and Related Fields* 74:199–211.

Fréchet, M. 1906. Sur quelques points du calcul fonctionnel (Thèse). *Rendiconti Circolo Mathematico di Palermo* 22:1–74.

Gioia, F., and C.N. Lauro. 2005. Basic statistical methods for interval data. *Statistica Applicata* [Italian Journal of Applied Statistics] 17(1): 75–104.

Hansen, K.M. 1999. A framework for assessing uncertainties in simulation predictions. *Physica D* 133: 179-188.

Hazelrigg, G. 2003. Thoughts on model validation for engineering design. *Proceedings of ASME 2003 Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Chicago.

Helton, J.C., and W.L. Oberkampf (eds.) 2004. *Reliability and Engineering System Safety* 85 (issues 1–3).

Hills, R.G., M. Pilch, K.J. Dowding, I. Babuska, and R. Tempone. 2008. Model validation challenge problems: tasking document. *Computer Methods in Applied Mechanics and Engineering* [in press].

Hills, R.G. 2006. Model validation: model parameter and measurement uncertainty. *Journal of Heat Transfer* 128: 339–351.

Hills, R.G., and T.G. Truncano. 2002. *Statistical Validation of Engineering and Scientific Models: A Maximum Likelihood Based Metric*. SAND2002-1783, Sandia National Laboratories, Albuquerque, NM.

Hills, R.G., and I. Leslie. 2003. *Statistical Validation of Engineering and Scientific Models: Validation Experiments to Application*. SAND2003-0706, Sandia National Laboratories, Albuquerque, NM.

ISO [International Organization for Standardization]. 1993. *Guide to the Expression of Uncertainty in Measurement*. International Organization for Standardization, Geneva, Switzerland.

JCGM [Joint Committee for Guides in Metrology]. 2006. Evaluation of measurement data—Supplement 1 to the "Guide to the expression of uncertainty in measurement"— Propagation of distributions using a Monte Carlo method. http://www.intermet.jp/JCGM/0610news/ Supplement_to_GUM.pdf.

Kennedy, M. and A. O'Hagan. 2001. Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society, Series B* 63: 425-464.

Klir, G., and M.J. Wierman. 1999. *Uncertainy-based Information: Elements of Generalized Information Theory*. Physica-Verlag, Heidelberg.

Kolmogorov [Kolmogoroff], A. 1941. Confidence limits for an unknown distribution function. *Annals of Mathematical Statistics* 12: 461–463.

Kullback, S. 1959. *Information Theory and Statistics*. Wiley, New York.

Kullback, S., and R.A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22: 79–86.

Levi, I. 1980. *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts.

Lindley, D.V., A. Tversky and R.V. Brown. 1979. On the reconciliation of probability assessments. *Journal of the Royal Statistical Society* A 142 (Part 2): 146–180.

Manski, C.F. 2003. *Partial Identification of Probability Distributions*, Springer Series in Statistics, Springer, New York.

Matheron, G. 1975. Random Sets and Integral Geometry. J.Wiley, New York

Mathiassen, J.R., A. Skavhaug and K. Bø. 2002. Texture similarity measure using Kullback-Leibler divergence between gamma distributions. Pages 19–49 in *Computer Vision - ECCV 2002: 7th European Conference on Computer Vision, Copenhagen, Denmark, May 28–31, 2002. Proceedings, Part III*. Lecture Notes in Computer Science, volume 2352. Springer, Berlin.

Menger, K. 1942. Statistical metrics. *Proceedings of the National Academy of Science U.S.A.* 28: 535-537.

Molchanov, I. 2005. *Theory of Random Sets*. Springer, London.

Neapolitan, R.E. 1992. A survey of uncertain and approximate inference. Pages 55–82 in *Fuzzy Logic for the Management of Uncertainty*, L. Zadeh and J. Kacprzyk (eds.), John Wiley & Sons, New York.

Nikolaidis, E., and R. Haftka. 2001. Theories of uncertainty for risk assessment when data is scarce. *International Journal of Advanced Manufacturing Systems* 4(1): 49–56.

Oberkampf, W.L., and M.F. Barone. 2006. Measures of agreement between computation and experiment: validation metrics. *Journal of Computational Physics* 217: 5–36.

Oberkampf, W.L., and J.C. Helton. 2005. Evidence theory for engineering applications. Pages 10-1–10-30 in *Engineering Design Reliability Handbook*, E. Nikolaidis, D.M. Ghiocel, and S. Singhal (eds.), CRC Press, Boca Raton, Florida.

Oberkampf, W.L., and T.G. Trucano. 2007. *Verification and Validation Benchmarks*. SAND2007-0853, Sandia National Laboratories, Albuquerque, NM. To appear in *Nuclear Engineering and Design*.

Oberkampf, W.L., J.C. Helton and K. Sentz. 2001. Mathematical representation of uncertainty. American Institute of Aeronautics and Astronautics Non-Deterministic Approaches Forum, Seattle, WA, Paper No. 2001-1645, April, 2001.

Oberkampf, W.L., T.G. Trucano and C. Hirsch. 2004. Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews* 57(5): 345–384.

Pompeiu, D. 1905. *Sur la continuité des fonctions de variables complexes* (Thèse). Gauthier-Villars, Paris. *Ann. Fac. Sci. de Toulouse* 7:264-315.

Rabinovich, S. 1993. *Measurement Errors: Theory and Practice*. American Institute of Physics, New York.

Romero, V.J. 2007. Validated model? Not so fast. The need for model "conditioning" as an essential addendum to model validation. AIAA-2007-1953 in *Proceedings of the 2007 AIAA Non-Deterministic Approaches Conference*, Honolulu. American Institute of Aeronautics and Astronautics.

Rutherford, B.M., and K.J. Dowding. 2003. *An Approach to Model Validation and Model-based Prediction—Polyurethane Foam Case Study*. SAND2003-2336, Sandia National Laboratories, Albuquerque, NM.

Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey.

Song, K.-S.. 2002. Goodness-of-fit tests based on Kullback-Leibler discrimination information. *IEEE Transactions on Information Theory* 48:1103–1117

Stephens, M.A. 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* 69: 730–737.

Smirnov [Smirnoff], N. 1939. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin de l'Université de Moscou, Série internationale (Mathématiques)* 2: (fasc. 2).

Taylor, B.N. and C.E. Kuyatt. 1994. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results.* NIST Technical Note 1297, National Institute of Standards and Technology, Washington, DC. http://physics.nist.gov/Pubs/guidelines/contents.html. See also web guidance at http://physics.nist.gov/cuu/Uncertainty/index.html.

Trucano, T.G., L.P. Swiler, T. Igusa, W.L. Oberkampf and M. Pilch. 2006. Calibration, validation and sensitivity analysis: what's what. *Reliability Engineering and System Safety* 91: 1331–1357.

Williamson, R.C. and T. Downs 1990. Probabilistic arithmetic I: numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning* 4:89–158.

Winkler, R.L. 1996. Scoring rules and the evaluation of probabilities. *Test* 5: 1–60.

Yates, F. 1934. Contingency table involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society* (Supplement) 1: 217–235.

Vallender, S.S. 1973. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability and its Applications* 18: 784–786.

Walley, P. 1991. Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London.

Zhang, R., and S. Mahadevan. 2003. Bayesian methodology for reliability model acceptance. *Reliability Engineering and System Safety* 80:95-103.

Imprecise Probabilities with a Generalized Interval Form

Yan Wang
University of Central Florida, Orlando, FL 32816
email:wangyan@mail.ucf.edu

Abstract. Different representations of imprecise probabilities have been proposed, such as behavioral theory, evidence theory, possibility theory, probability bound analysis, F-probabilities, fuzzy probabilities, and clouds. These methods use interval-valued parameters to discribe probability distributions such that uncertainty is distinguished from variability. In this paper, we proposed a new form of imprecise probabilities based on generalized or modal intervals. Generalized intervals are algebraically closed under Kaucher arithmetic, which provides a concise representation and calculus structure as an extension of precise probabilities.

With the separation between proper and improper interval probabilities, *focal* and *non-focal* events are differentiated based on the modalities and logical semantics of generalized interval probabilities. Focal events have the semantics of critical, uncontrollable, specified, etc. in probabilistic analysis, whereas the corresponding non-focal events are complementary, controllable, and derived.

A generalized imprecise conditional probability is defined based on unconditional interval probabilities such that the algebraic relation between conditional and marginal interval probabilities is maintained. A Bayes' rule with generalized intervals (GIBR) is also proposed. The GIBR allows us to interpret the logic relationship between interval prior and posterior probabilities.

Keywords: imprecise probablity, conditioning, updating, interval arithmetic, generalized interval

1. Introduction

Imprecise probability differentiates uncertainty from variability both qualitatively and quantitatively, which is to complement the traditional sensitivity analaysis in probablistic reasoning. There have been several interval-based representations proposed in the past four decades and applied in various engineering domains, such as sensor data fusion (Guede and Girardi, 1997; Elouedi et al., 2004), reliability assessment (Kozine and Filimonov, 2000; Berleant and Zhang, 2004; Coolen, 2004), reliability-based design optimization (Mourelatos and Zhou, 2006; Du et al., 2006), design decision making under uncertainty (Nikolaidis et al., 2004; Aughenbaugh and Paredis, 2006). The core issue is to characterize incomplete knowledge with lower and upper probability pairs so that we can improve the robustness of decision making.

There are many representations of imprecise probabilities. For example, the Dempster-Shafer evidence theory (Dempster, 1967; Shafer, 1976) characterizes uncertainties as discrete probability masses associated with a power set of values. Belief-Plausibility pairs are used to measure likelihood. The behavioral imprecise probability theory (Walley, 1991) models behavioral uncertainties with the lower prevision (supremum acceptable buying price) and the upper prevision (infimum acceptable

selling price). A random set (Molchanov, 2005) is a multi-valued mapping from the probability space to the value space. The possibility theory (Zadeh, 1978; Dubois and Prade, 1988) provides an alternative to represent uncertainties with Necessity-Possibility pairs. Probability bound analysis (Ferson et al., 2002) captures uncertain information with p-boxes which are pairs of lower and upper probability distributions. F-probability (Weichselberger, 2000) incorporates intervals into probability values which maintains Kolmogorov properties. Fuzzy probability (Möller and Beer, 2004) considers probability distributions with fuzzy parameters. A cloud (Neumaier, 2004) is a fuzzy interval with an interval-valued membership, which is a combination of fuzzy sets, intervals, and probability distributions.

These different representations model the indeterminacy due to incomplete information very well with different forms. There are still challenges in practical issues such as assessment and computation to derive inferences and conclusions (Walley, 1996). A simple algebraic structure is important for applications in engineering and science. In this paper, we propose a new form of imprecise probabilities based on generalized intervals. Unlike traditional set-based intervals, such as the interval $[0.1, 0.2]$ which represents a set of real values between 0.1 and 0.2, generalized or modal intervals also allow the existence of the interval $[0.2, 0.1]$. With this extension, logic quantifiers ($\forall$ and $\exists$) can be integrated to provide the interpretation of intervals. Another advantage of generalized interval is that it is closed under arithmetic operations ($+, -, \times, \div$). This property simplifies the set structures.

We are interested to explore the potential of generalized interval to provide a connection between imprecise and precise probability, as well as among different representations of imprecise probability. In this paper, we study the algebraic properties of imprecise probablities with a generalized interval form and associated interpretation issues. In the remainder of the paper, Section 2 gives a brief overview of generalized intervals. Section 3 presents the interval probability with the generalized interval form. Section 4 describes the Bayes' rule based on generalized intervals.

2. Generalized Interval

Modal interval analysis (MIA) (Gardenes et al., 2001; Markov, 2001; Shary, 2002; Popova, 2001; Armengol et al., 2001) is an algebraic and semantic extension of interval analysis (IA) (Moore, 1966). Unlike the classical interval analysis which identifies an interval by a set of real numbers, MIA identifies the intervals by the set of predicates which is fulfilled by the real numbers. A generalized interval is not restricted to ordered bounds. A modal interval or generalized interval $\mathbf{x} := [\underline{x}, \overline{x}] \in \mathbb{KR}$ is called proper when $\underline{x} \leq \overline{x}$ and improper when $\underline{x} \geq \overline{x}$. The set of proper intervals is denoted by $\mathbb{IR} = \{[\underline{x}, \overline{x}] \mid \underline{x} \leq \overline{x}\}$, and the set of improper interval is $\overline{\mathbb{IR}} = \{[\underline{x}, \overline{x}] \mid \underline{x} \geq \overline{x}\}$. Operations are defined in Kaucher arithmetic (Kaucher, 1980).

Given a generalized interval $\mathbf{x} = [\underline{x}, \overline{x}] \in \mathbb{KR}$, two operators *pro* and *imp* return proper and improper values respectively, defined as

$$\text{pro}\mathbf{x} := [\min(\underline{x}, \overline{x}), \max(\underline{x}, \overline{x})] \tag{1}$$

$$\text{imp}\mathbf{x} := [\max(\underline{x}, \overline{x}), \min(\underline{x}, \overline{x})] \tag{2}$$

The relationship between proper and improper intervals is established with the operator *dual*:

Table I. The major differences between MIA and the tranditional IA

	Classical Interval Analysis	**Modal Interval Analysis**
Validity	$[3,2]$ is an invalid or empty interval	Both $[3,2]$ and $[3,2]$ are valid intervals
Semantics richness	$[2,3]+[2,4]=[4,7]$ is the only valid relation for $+$, and it only means"stack-up" and worst-case". $-,\times,\div$ are similar.	$[2,3]+[2,4]=[4,7]$, $[2,3]+[4,2]=[6,5]$, $[3,2]+[2,4]=[5,6]$, $[3,2]+[4,2]=[7,4]$ are all valid, and each has a different meaning. $-,\times,\div$ are similar.
Completeness of arithmetic	$\mathbf{a}+\mathbf{x}=\mathbf{b}$, but $\mathbf{x}\neq\mathbf{b}-\mathbf{a}$. $[2,3]+[2,4]=[4,7]$, but $[2,4]\neq[4,7]-[2,3]$ $\mathbf{a}\times\mathbf{x}=\mathbf{b}$, but $\mathbf{x}\neq\mathbf{b}\div\mathbf{a}$. $[2,3]\times[3,4]=[6,12]$, but $[3,4]\neq[6,12]\div[2,3]$ $\mathbf{x}-\mathbf{x}\neq 0$ $[2,3]-[2,3]=[-1,1]\neq 0$	$\mathbf{a}+\mathbf{x}=\mathbf{b}$, and $\mathbf{x}=\mathbf{b}-\text{dual}\mathbf{a}$. $[2,3]+[2,4]=[4,7]$, and $[2,4]=[4,7]-[3,2]$ $\mathbf{a}\times\mathbf{x}=\mathbf{b}$, and $\mathbf{x}=\mathbf{b}\div\text{dual}\mathbf{a}$. $[2,3]\times[3,4]=[6,12]$, and $[3,4]=[6,12]\div[3,2]$ $\mathbf{x}-\text{dual}\mathbf{x}=0$ $[2,3]-[3,2]=0$

$$\text{dual}\mathbf{x} := [\overline{x}, \underline{x}] \tag{3}$$

For example, $\mathbf{a}=[-1,1]$ and $\mathbf{b}=[1,-1]$ are both valid intervals. While $\mathbf{a}$ is a proper interval, $\mathbf{b}$ is an improper one. The relation between $\mathbf{a}$ and $\mathbf{b}$ can be established by $\mathbf{a}=\text{dual}\mathbf{b}$. The *inclusion* relation between generalized intervals $\mathbf{x}=[\underline{x},\overline{x}]$ and $\mathbf{y}=[\underline{y},\overline{y}]$ is defined as

$$\begin{aligned} [\underline{x},\overline{x}] \subseteq [\underline{y},\overline{y}] &\iff \underline{x}\geq\underline{y}\wedge\overline{x}\leq\overline{y} \\ [\underline{x},\overline{x}] \supseteq [\underline{y},\overline{y}] &\iff \underline{x}\leq\underline{y}\wedge\overline{x}\geq\overline{y} \end{aligned} \tag{4}$$

The *less-than-or-equal-to* and *greater-than-or-equal-to* relations are defined as

$$\begin{aligned} [\underline{x},\overline{x}] \leq [\underline{y},\overline{y}] &\iff \underline{x}\leq\underline{y}\wedge\overline{x}\leq\overline{y} \\ [\underline{x},\overline{x}] \geq [\underline{y},\overline{y}] &\iff \underline{x}\geq\underline{y}\wedge\overline{x}\geq\overline{y} \end{aligned} \tag{5}$$

Table I lists the major differences between MIA and IA. MIA offers better algebraic properties and more semantic capabilities.

For a solution set $\mathcal{S}\subset\mathbb{R}^n$ of the interval system $\mathbf{f}(\mathbf{x})=0$ where $\mathbf{x}\in\mathbb{IR}^n$, an inner estimation $\mathbf{x}^{in}$ of the solution set $\mathcal{S}$ is an interval vector that is guaranteed to be included in the solution set, and an outer estimation $\mathbf{x}^{out}$ of $\mathcal{S}$ is an interval vector that is guaranteed to include the solution set. Not only for outer range estimations, generalized intervals are also convenient for inner range estimations (Kupriyanova, 1995; Kreinovich et al., 1996; Goldsztejn, 2005).

Another uniqueness of generalized intervals is the modal semantic extension. Unlike IA which identifies an interval by a set of real numbers only, MIA identifies an interval by a set of predicates which is fulfilled by real numbers. Given a set of closed intervals of real numbers in $\mathbb{R}$, and the set of logical existential ($\exists$) and universal ($\forall$) quantifiers, each generalized interval has an associated quantifier. The semantics of $\mathbf{x}\in\mathbb{KR}$ is denoted by $(\mathrm{Q}_{\mathbf{x}}x\in\text{pro}\mathbf{x})$ where $\mathrm{Q}_{\mathbf{x}}\in\{\exists,\forall\}$. An interval

$\mathbf{x} \in \mathbb{KR}$ is called *existential* if $\mathrm{Q}_{\mathbf{x}} = \exists$. Otherwise, it is called universal if $\mathrm{Q}_{\mathbf{x}} = \forall$. If a real relation $z = f(x_1, \ldots, x_n)$ is extended to the interval relation $\mathbf{z} = \mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, the interval relation $\mathbf{z}$ is interpretable if there is a semantic relation

$$(\mathrm{Q}_{\mathbf{x}_1} x_1 \in \mathrm{pro}\mathbf{x}_1) \cdots (\mathrm{Q}_{\mathbf{x}_n} x_n \in \mathrm{pro}\mathbf{x}_n) \, (\mathrm{Q}_{\mathbf{z}} z \in \mathrm{pro}\mathbf{z}) \, (z = f(x_1, \ldots, x_n)) \tag{6}$$

In this paper, we propose an interval probability representation that incorporates the generalized interval in imprecise probability. The aim is to take the advantage of its algebraic closure so that the structure of interval probability can be simplified. At the same time, the interpretation of probablistic properties can be integrated with the logic relations in the structure.

3. Imprecise Probability based on Generalized Intervals

Given a sample space Ω and a σ-algebra $\mathcal{A}$ of random events over Ω, we define the generalized interval probability $\mathbf{p} : \mathcal{A} \mapsto [0,1] \times [0,1]$ which obeys the axioms of Kolmogorov: (1) $\mathbf{p}(\Omega) = [1,1] = 1$; (2) $0 \leq \mathbf{p}(E) \leq 1 \; (\forall E \in \mathcal{A})$; and (3) for any countable mutually disjoint events $E_i \cap E_j = \emptyset (i \neq j)$, $\mathbf{p}(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} \mathbf{p}(E_i)$. This implies $\mathbf{p}(\emptyset) = 0$. We also define

$$\mathbf{p}(E_1 \cup E_2) := \mathbf{p}(E_1) + \mathbf{p}(E_2) - \mathrm{dual}\mathbf{p}(E_1 \cap E_2) \tag{7}$$

When the probabilities of E_1 and E_2 are measurable and become precise, Eq.(7) has the same form as the traditional precise probabilities. The lower and upper probabilities in the generalized interval form do not have the traditional meanings of lower and upper envelops. Rather, they provide the algebraic closure. From Eq.(7), we have

$$\mathbf{p}(E_1 \cup E_2) + \mathbf{p}(E_1 \cap E_2) = \mathbf{p}(E_1) + \mathbf{p}(E_2) \tag{8}$$

which also indicates the generalized interval probabilities are 2-monotone (and 2-alternating) in the sense of Choquet's capacities. But the relation of Eq.(8) is stronger than 2-monotonicity.

Let $(\Omega, \mathcal{A})$ be the probability space and $\mathcal{P}$ a non-empty set of probability distribution on that space. The lower and upper probability envelopes are usually defined as

$$P_*(E) = \inf_{P \in \mathcal{P}} P(E)$$
$$P^*(E) = \sup_{P \in \mathcal{P}} P(E)$$

Not every probability envelope is 2-monotone. However, 2-monotone closed-form representations are more applicable because it may be difficult to track probability envelopes during manipulations. Therefore it is of our interest that a simple algebraic structure can provide such practical advantages for broader applications.

Furthermore, we have

$$\mathbf{p}(E_1 \cup E_2) \leq \mathbf{p}(E_1) + \mathbf{p}(E_2) \; (\forall E_1, E_2 \in \mathcal{A}) \tag{9}$$

in the new interval representation, since $\mathbf{p}(E_1 \cap E_2) \geq 0$. Note that Eq.(9) is different from the relation defined in the Dempster-Shafer structure or F-probability. Here it has the same form as the precise probability except for the newly defined inequality $(\leq, \geq)$ relations for generalized intervals.

Both lower and upper probabilities are subadditive. Similar to the precise probability, the equality of Eq.(9) occurs when $\mathbf{p}(E_1 \cap E_2) = 0$.

We also define the probability of the complement of event E as

$$\mathbf{p}(E^c) := 1 - \text{dual}\mathbf{p}(E) \tag{10}$$

which is equivalent to

$$\underline{p}(E^c) := 1 - \overline{p}(E) \tag{11}$$

$$\overline{p}(E^c) := 1 - \underline{p}(E) \tag{12}$$

The definitions in Eq.(11) and Eq.(12) are equivalent to the other forms of interval probabilities. The calculation based on generalized intervals as in Eq.(10) can be more concise.

$$\mathbf{p}(E) + \mathbf{p}(E^c) = 1 \ (\forall E \in \mathcal{A}) \tag{13}$$

In general, for a mutually disjoint event partition $\bigcup_{i=1}^{n} E_i = \Omega$, we have

$$\sum_{i=1}^{n} \mathbf{p}(E_i) = 1 \tag{14}$$

This requirement is more restrictive than the traditional coherence constraint (Walley, 1991). Suppose $\mathbf{p}(E_i) \in \mathbb{IR}$ (for $i = 1, \ldots, k$) and $\mathbf{p}(E_i) \in \overline{\mathbb{IR}}$ (for $i = k+1, \ldots, n$). If the range of an interval probability is defined as

$$\mathbf{p}'(E) := \text{pro}\mathbf{p}(E) \tag{15}$$

Eq.(14) can be interpreted as

$$\forall p_1 \in \mathbf{p}'(E_1), \ldots, \forall p_k \in \mathbf{p}'(E_k), \exists p_{k+1} \in \mathbf{p}'(E_{k+1}), \ldots, \exists p_n \in \mathbf{p}'(E_n), \sum_{i=1}^{n} p_i = 1 \tag{16}$$

based on the interpretability principles of MIA (Gardenes et al., 2001). Therefore, we call Eq.(14) the *logic coherence constraint*.

The values of interval probabilities are between 0 and 1. As a result, the interval probabilities $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{p}_3$ have the following algebraic properties:

$$\begin{aligned}
\mathbf{p}_1 \leq \mathbf{p}_2 &\Leftrightarrow \mathbf{p}_1 + \mathbf{p}_3 \leq \mathbf{p}_2 + \mathbf{p}_3 \\
\mathbf{p}_1 \subseteq \mathbf{p}_2 &\Leftrightarrow \mathbf{p}_1 + \mathbf{p}_3 \subseteq \mathbf{p}_2 + \mathbf{p}_3 \\
\mathbf{p}_1 \leq \mathbf{p}_2 &\Leftrightarrow \mathbf{p}_1\mathbf{p}_3 \leq \mathbf{p}_2\mathbf{p}_3 \\
\mathbf{p}_1 \subseteq \mathbf{p}_2 &\Leftrightarrow \mathbf{p}_1\mathbf{p}_3 \subseteq \mathbf{p}_2\mathbf{p}_3
\end{aligned}$$

3.1. Focal and Non-Focal Events

We differentiate two types of events. An event E is a *focal* event if its associated semantics is universal ($\text{Q}_{\mathbf{p}(E)} = \forall$). Otherwise it is a *non-focal* event if the semantics is existential ($\text{Q}_{\mathbf{p}(E)} = \exists$). A focal event is an event of interest in the probabilistic analysis. The uncertainties associated with focal events are critical for the analysis of a system. In contrast, the uncertainties associated with non-focal events are "complementary" and "balancing". The corresponding non-focal event is not

the focus of the assessment. The quantified uncertainties of non-focal events are derived from those of the corresponding focal events. For instance, in risk assessment, the high-consequence event of interest is the target and focus of study, such as the event of a hurricane landfall at U.S. coastline or the event of a structural failure at the half of a bridge's life expectancy, whereas the event of the hurricane landfall at Mexican coastline and the event of the structral failure when the bridge is twice as old as it was designed for may become non-focal.

In the interpretation in Eq.(16), the interval probability of a focal event E_i is proper ($\mathbf{p}(E_i) \in \mathbb{IR}$), and the interval probability of a non-focal event E_j is existential ($\mathbf{p}(E_j) \in \overline{\mathbb{IR}}$). Focal events have the semantics of critical, uncontrollable, specified in probabilistic analysis, whereas the corresponding non-focal events are complementary, controllable, and derived. The complement of a focal event is a non-focal event. For a set of mutually disjoint events, there is at least one non-focal event because of Eq.(14).

Two relations between events are defined. Event E_1 is said to be *less likely* (or *more likely*) to occur than event E_2, $E_1 \preceq E_2$ (or $E_1 \succeq E_2$), defined as

$$\begin{array}{l} E_1 \preceq E_2 \iff \mathbf{p}(E_1) \leq \mathbf{p}(E_2) \\ E_1 \succeq E_2 \iff \mathbf{p}(E_1) \geq \mathbf{p}(E_2) \end{array} \tag{17}$$

Event E_1 is said to be *less focused* (or *more focused*) than event E_2, denoted as $E_1 \sqsubseteq E_2$ (or $E_1 \sqsupseteq E_2$), defined as

$$\begin{array}{l} E_1 \sqsubseteq E_2 \iff \mathbf{p}(E_1) \subseteq \mathbf{p}(E_2) \\ E_1 \sqsupseteq E_2 \iff \mathbf{p}(E_1) \supseteq \mathbf{p}(E_2) \end{array} \tag{18}$$

LEMMA 3.1. *$E_1 \subseteq E_2 \Rightarrow E_1 \preceq E_2$.*

Proof. $E_1 \subseteq E_2 \Rightarrow \mathbf{p}(E_2) = \mathbf{p}(E_1 \cup (E_2 - E_1)) = \mathbf{p}(E_1) + \mathbf{p}(E_2 - E_1) - \text{dual}\mathbf{p}(E_1 \cap (E_2 - E_1)) \geq \mathbf{p}(E_1)$.

LEMMA 3.2. *If $E_1 \cap E_3 = \emptyset$ and $E_2 \cap E_3 = \emptyset$, $E_1 \preceq E_2 \Leftrightarrow E_1 \cup E_3 \preceq E_2 \cup E_3$, $E_1 \sqsubseteq E_2 \Leftrightarrow E_1 \cup E_3 \sqsubseteq E_2 \cup E_3$.*

Proof.
$E_1 \preceq E_2 \Leftrightarrow \mathbf{p}(E_1) \leq \mathbf{p}(E_2) \Leftrightarrow \mathbf{p}(E_1) + \mathbf{p}(E_3) \leq \mathbf{p}(E_2) + \mathbf{p}(E_3) \Leftrightarrow \mathbf{p}(E_1 \cup E_3) \leq \mathbf{p}(E_2 \cup E_3) \Leftrightarrow E_1 \cup E_3 \preceq E_2 \cup E_3$.
$E_1 \sqsubseteq E_2 \Leftrightarrow \mathbf{p}(E_1) \subseteq \mathbf{p}(E_2) \Leftrightarrow \mathbf{p}(E_1) + \mathbf{p}(E_3) \subseteq \mathbf{p}(E_2) + \mathbf{p}(E_3) \Leftrightarrow \mathbf{p}(E_1 \cup E_3) \subseteq \mathbf{p}(E_2 \cup E_3) \Leftrightarrow E_1 \cup E_3 \sqsubseteq E_2 \cup E_3$.

LEMMA 3.3. *If E_1 and E_3 are independent, and also E_2 and E_3 are independent, $E_1 \preceq E_2 \Leftrightarrow E_1 \cap E_3 \preceq E_2 \cap E_3$, $E_1 \sqsubseteq E_2 \Leftrightarrow E_1 \cap E_3 \sqsubseteq E_2 \cap E_3$.*

Proof.
$E_1 \preceq E_2 \Leftrightarrow \mathbf{p}(E_1) \leq \mathbf{p}(E_2) \Leftrightarrow \mathbf{p}(E_1)\mathbf{p}(E_3) \leq \mathbf{p}(E_2)\mathbf{p}(E_3) \Leftrightarrow \mathbf{p}(E_1 \cap E_3) \leq \mathbf{p}(E_2 \cap E_3) \Leftrightarrow E_1 \cap E_3 \preceq E_2 \cap E_3$.
$E_1 \sqsubseteq E_2 \Leftrightarrow \mathbf{p}(E_1) \subseteq \mathbf{p}(E_2) \Leftrightarrow \mathbf{p}(E_1)\mathbf{p}(E_3) \subseteq \mathbf{p}(E_2)\mathbf{p}(E_3) \Leftrightarrow \mathbf{p}(E_1 \cap E_3) \subseteq \mathbf{p}(E_2 \cap E_3) \Leftrightarrow E_1 \cap E_3 \sqsubseteq E_2 \cap E_3$.

LEMMA 3.4. *Suppose $E \cup E^c = \Omega$ and $\mathbf{p}(E) \in \mathbb{IR}$. (1) $\mathbf{p}(E) \leq \mathbf{p}(E^c)$ if $\overline{p}(E) \leq 0.5$; (2) $\mathbf{p}(E) \geq \mathbf{p}(E^c)$ if $\underline{p}(E) \geq 0.5$; (3) $\mathbf{p}(E) \supseteq \mathbf{p}(E^c)$ if $\underline{p}(E) \leq 0.5$ and $\overline{p}(E) \geq 0.5$.*

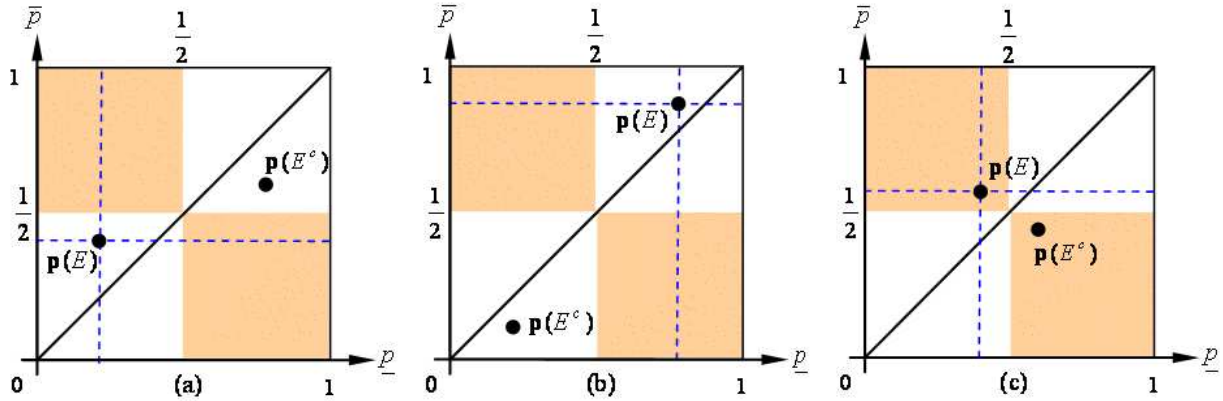


Figure 1. inf-sup diagrams for different relations between $\mathbf{p}(E)$ and $\mathbf{p}(E^c)$ when $\mathbf{p}(E) \in \mathbb{IR}$

Proof. (1) Because $\mathbf{p}(E) \in \mathbb{IR}$, $\mathbf{p}(E^c) \in \overline{\mathbb{IR}}$, and $\mathbf{p}(E) + \mathbf{p}(E^c) = 1$, it is easy to see $\underline{p}(E) \leq \underline{p}(E^c)$ and $\overline{p}(E) \leq \overline{p}(E^c)$ if $\overline{p}(E) \leq 0.5$. (2) can be verified similarly. (3) If $\underline{p}(E) \leq 0.5$ and $\overline{p}(E) \geq 0.5$, then $\underline{p}(E^c) \geq 0.5$ and $\overline{p}(E^c) \leq 0.5$. Thus $\underline{p}(E) \leq \underline{p}(E^c)$ and $\overline{p}(E) \geq \overline{p}(E^c)$.
Remark. As illustrated in Fig. 1 (a-c) respectively, a focal event E is less likely to occur than its complement if $\mathbf{p}(E) \leq 0.5$; E is more likely to occur than its complement if $\mathbf{p}(E) \geq 0.5$; otherwise, E is more focused than its complement. When E is a non-focal event, its complement E^c is a focal event. The relationships between $\mathbf{p}(E)$ and $\mathbf{p}(E^c)$ are just opposite.

For three events $E_i (i = 1, 2, 3)$,

$$\begin{aligned}\mathbf{p}(E_1 \cup E_2 \cup E_3) = \mathbf{p}(E_1) + \mathbf{p}(E_2) + \mathbf{p}(E_3) - \text{dual}\mathbf{p}(E_1 \cap E_2) \\ -\text{dual}\mathbf{p}(E_2 \cap E_3) - \text{dual}\mathbf{p}(E_1 \cap E_3) + \mathbf{p}(E_1 \cap E_2 \cap E_3)\end{aligned}$$

In general, for $A \subseteq \Omega$,

$$\mathbf{p}(A) = \sum_{S \subseteq A} (-\text{dual})^{|A|-|S|} \mathbf{p}(S) \tag{19}$$

3.2. Conditional Interval Probabilities

There have been several conditioning schemes proposed based on the Demspter-Shafer structures (Smets, 1991; Fagin and Halpern, 1991; Jaffray, 1992; Dubois and Prade, 1994; Chrisman, 1995; Kulasekere et al., 2004). Different from the coherent provision or F-probability theory, we define conditional generalized interval probabilities based on marginal probabilities. The conditional interval probability $\mathbf{p}(E|C)$ for $\forall E, C \in \mathcal{A}$ is defined as

$$\mathbf{p}(E|C) := \frac{\mathbf{p}(E \cap C)}{\text{dual}\mathbf{p}(C)} = \left[\frac{\underline{p}(E \cap C)}{\underline{p}(C)}, \frac{\overline{p}(E \cap C)}{\overline{p}(C)}\right] \tag{20}$$

when $\mathbf{p}(C) > 0$.

Not only does the definition in Eq.(20) ensure the algebraic closure of the interval probability calculus, but also it is a generalization of the canonical conditional probability in F-probabilities. Different from the Dempster's rule of conditioning or geometric conditioning, this conditional structure maintains the algebraic relation between marginal and conditional probabilities. Further,

$$\mathbf{p}(C|C) = 1$$

.

The available logic interpretations of the conditional interval probabilities are as follows.

– when $\mathbf{p}(E \cap C) \in \mathbb{IR}$, $\mathbf{p}(C) \in \overline{\mathbb{IR}}$, and $\mathbf{p}(E|C) \in \mathbb{IR}$

$$\forall p_{E\cap C} \in \mathbf{p}'(E \cap C), \forall p_C \in \mathbf{p}'(C), \exists p_{E|C} \in \mathbf{p}'(E|C), p_{E|C} = \frac{p_{E\cap C}}{p_C} \tag{21}$$

or

$$\forall p_{E|C} \in \mathbf{p}'(E|C), \exists p_{E\cap C} \in \mathbf{p}'(E \cap C), \exists p_C \in \mathbf{p}'(C), p_{E|C} = \frac{p_{E\cap C}}{p_C} \tag{22}$$

– when $\mathbf{p}(E \cap C) \in \mathbb{IR}$, $\mathbf{p}(C) \in \mathbb{IR}$, and $\mathbf{p}(E|C) \in \mathbb{IR}$

$$\forall p_{E\cap C} \in \mathbf{p}'(E \cap C), \exists p_C \in \mathbf{p}'(C), \exists p_{E|C} \in \mathbf{p}'(E|C), p_{E|C} = \frac{p_{E\cap C}}{p_C} \tag{23}$$

or

$$\forall p_{E|C} \in \mathbf{p}'(E|C), \forall p_C \in \mathbf{p}'(C), \exists p_{E\cap C} \in \mathbf{p}'(E \cap C), p_{E|C} = \frac{p_{E\cap C}}{p_C} \tag{24}$$

– when $\mathbf{p}(E \cap C) \in \mathbb{IR}$, $\mathbf{p}(C) \in \mathbb{IR}$, and $\mathbf{p}(E|C) \in \overline{\mathbb{IR}}$

$$\forall p_{E\cap C} \in \mathbf{p}'(E \cap C), \forall p_{E|C} \in \mathbf{p}'(E|C), \exists p_C \in \mathbf{p}'(C), p_{E|C} = \frac{p_{E\cap C}}{p_C} \tag{25}$$

or

$$\forall p_C \in \mathbf{p}'(C), \exists p_{E\cap C} \in \mathbf{p}'(E \cap C), \exists p_{E|C} \in \mathbf{p}'(E|C), p_{E|C} = \frac{p_{E\cap C}}{p_C} \tag{26}$$

The logic interpretations of interval conditional probabilities build the connection between point measurements and probability sets. Therefore, we may use them to check if a range estimation is a tight envelop. We use the Example 3.1 in (Weichselberger, 2000) to illustrate.

EXAMPLE 3.1. *Given the following probabilities in the sample space* $\Omega = E_1 \cup E_2 \cup E_3$,

$$\begin{array}{ll} \mathbf{p}'(E_1) = [0.10, 0.25] & \mathbf{p}'(E_2 \cup E_3) = [0.75, 0.90] \\ \mathbf{p}'(E_2) = [0.20, 0.40] & \mathbf{p}'(E_1 \cup E_3) = [0.60, 0.80] \\ \mathbf{p}'(E_3) = [0.40, 0.60] & \mathbf{p}'(E_1 \cup E_2) = [0.40, 0.60] \end{array}$$

A partition of Ω *is*

$$\mathcal{C} = \{C_1, C_2\} \quad \textit{where } C_1 = E_1 \cup E_2 \textit{ and } C_2 = E_3$$

$$\mathbf{p}(C_1) = [0.40, 0.60] \;\; \mathbf{p}(C_2) = [0.60, 0.40]$$

Suppose $\mathbf{p}(E_1) = [0.10, 0.25]$ *and* $\mathbf{p}(C_1) = [0.60, 0.40]$*, we have*

$$\mathbf{p}(E_1|C_1) = \frac{[0.10, 0.25]}{[0.40, 0.60]} = [0.1666, 0.6250]$$

The interpretation of

$$\forall p_{E_1} \in [0.10, 0.25], \forall p_{C_1} \in [0.40, 0.60], \exists p_{E_1|C_1} \in [0.1666, 0.6250], p_{E_1|C_1} = \frac{p_{E_1}}{p_{C_1}}$$

indicates that the range estimation $\mathbf{p}(E_1|C_1) = [0.1666, 0.6250]$ *is* complete *in the sense that it considers all possible occurences of* $p(E_1)$ *and* $p(C_1)$*. However, the range estimation is not necessarily a tight envelop.*

On the other hand, if $\mathbf{p}(E_1) = [0.25, 0.10]$ *and* $\mathbf{p}(C_1) = [0.40, 0.60]$*, we have*

$$\mathbf{p}(E_1|C_1) = \frac{[0.25, 0.10]}{[0.60, 0.40]} = [0.6250, 0.1666]$$

The interpretation of

$$\forall p_{E_1|C_1} \in [0.1666, 0.6250], \exists p_{E_1} \in [0.10, 0.25], \exists p_{C_1} \in [0.40, 0.60], p_{E_1|C_1} = \frac{p_{E_1}}{p_{C_1}}$$

indicates that the range estimation $[0.1666, 0.6250]$ *is also* sound *in the sense that the range estimation is a tight envelop.*

Suppose $\mathbf{p}(E_1) = [0.25, 0.10]$*,* $\mathbf{p}(E_2) = [0.20, 0.40]$*, and* $\mathbf{p}(C_1) = [0.60, 0.40]$*, we have*

$$\mathbf{p}(E_1|C_1) = \frac{[0.25, 0.10]}{[0.40, 0.60]} = [0.4166, 0.25]$$

$$\mathbf{p}(E_2|C_1) = \frac{[0.20, 0.40]}{[0.40, 0.60]} = [0.3333, 1.0]$$

The interpretations are

$$\forall p_{E_1|C_1} \in [0.25, 0.4166], \forall p_{C_1} \in [0.40, 0.60], \exists p_{E_1} \in [0.10, 0.25], p_{E_1|C_1} = \frac{p_{E_1}}{p_{C_1}}$$

$$\forall p_{E_2} \in [0.20, 0.40], \forall p_{C_1} \in [0.40, 0.60], \exists p_{E_2|C_1} \in [0.3333, 1.0], p_{E_2|C_1} = \frac{p_{E_2}}{p_{C_1}}$$

respectively. Combining the two, we can have the interpretation of

$$\forall p_{E_2} \in [0.20, 0.40], \forall p_{C_1} \in [0.40, 0.60], \forall p_{E_1|C_1} \in [0.25, 0.4166], \exists p_{E_1} \in [0.10, 0.25] \exists p_{E_2|C_1} \in [0.3333, 1.0], p_{E_1|C_1} = \frac{p_{E_1}}{p_{C_1}}, p_{E_2|C_1} = \frac{p_{E_2}}{p_{C_1}}$$

If events A and B are independent, then

$$\mathbf{p}(A|B) = \frac{\mathbf{p}(A)\mathbf{p}(B)}{\text{dual}\mathbf{p}(B)} = \mathbf{p}(A) \tag{27}$$

For a mutually disjoint event partition $\bigcup_{i=1}^{n} E_i = \Omega$, we have

$$\mathbf{p}(A) = \sum_{i=1}^{n} \mathbf{p}(A|E_i)\mathbf{p}(E_i) \tag{28}$$

LEMMA 3.5. *If* $B \cap C = \emptyset$, *(1)* $\mathbf{p}(A|C) \subseteq \mathbf{p}(A|B) \Leftrightarrow \mathbf{p}(A|B \cup C) \subseteq \mathbf{p}(A|B)$. *(2)* $\mathbf{p}(A|B \cup C) \supseteq \mathbf{p}(A|B) \Leftrightarrow \mathbf{p}(A|C) \supseteq \mathbf{p}(A|B)$.

Proof. (1) $\mathbf{p}(A|C) \subseteq \mathbf{p}(A|B) \Leftrightarrow \mathbf{p}(A \cap C)/\text{dual}\mathbf{p}(C) \subseteq \mathbf{p}(A|B) \Leftrightarrow \mathbf{p}(A \cap C) \subseteq \mathbf{p}(A|B)\mathbf{p}(C) \Leftrightarrow \mathbf{p}(A|B)\mathbf{p}(B)+\mathbf{p}(A \cap C) \subseteq \mathbf{p}(A|B)\mathbf{p}(B)+\mathbf{p}(A|B)\mathbf{p}(C) \Leftrightarrow \mathbf{p}(A \cap B)+\mathbf{p}(A \cap C) \subseteq \mathbf{p}(A|B)\mathbf{p}(B \cup C) \Leftrightarrow \mathbf{p}\left(A \cap (B \cup C)\right) \subseteq \mathbf{p}(A|B)\mathbf{p}(B \cup C) \Leftrightarrow \mathbf{p}\left(A \cap (B \cup C)\right)/\text{dual}\mathbf{p}(B \cup C) \subseteq \mathbf{p}(A|B) \Leftrightarrow \mathbf{p}(A|B \cup C) \subseteq \mathbf{p}(A|B)$. (2) can be verified similarly.

Remark. The interpretation of the relationship (1) is that if there are two pieces of evidence (B and C), and one (C) may provide more precise estimation about a focal event (A) than the other (B) may, then the new estimation of probability about the focal event (A) based on the disjunctively combined evidence can be more precise than the one based on only one of them (B), even though the two pieces of information are contradictory to each other. The other direction of the reasoning is that if the precision of the focal event estimation with the newly introduced evidence (C) is improved, the new evidence (C) must be more informative than the old one (B) although these two are controdictory.

Remark. The interpretation of the relationship (2) is that if the estimation about a focal event (A) becomes more precise if some new evidence (B) excludes some possibilities (C) from the original evidence ($B \cup C$), then the estimation of probability about the focal event (A) based on the new evidence (B) must be more precise than the one based on the excluded one (C) along. The other direction of the reasoning is that if the precision of the focal event estimation with a contradictory evidence (C) is not improved compared to the old one with another evidence (B), then the new evidence ($B \cup C$) does not improve the estimation of the focal event (A).

4. Bayes' Rule with Generalized Intervals

The Bayes' rule with generalized intervals (GIBR) is defined as

$$\mathbf{p}(E_i|A) = \frac{\mathbf{p}(A|E_i)\mathbf{p}(E_i)}{\sum_{j=1}^{n} \text{dual}\mathbf{p}(A|E_j)\text{dual}\mathbf{p}(E_j)} \tag{29}$$

where $E_i (i = 1, \ldots, n)$ are mutually disjoint event partitions of Ω and $\sum_{j=1}^{n} \mathbf{p}(E_j) = 1$. The lower and upper probabilities are calculated as

$$\left[\underline{p}(E_i|A), \overline{p}(E_i|A)\right] = \left[\frac{\underline{p}(A|E_i)\underline{p}(E_i)}{\sum_{j=1}^{n} \underline{p}(A|E_j)\underline{p}(E_j)}, \frac{\overline{p}(A|E_i)\overline{p}(E_i)}{\sum_{j=1}^{n} \overline{p}(A|E_j)\overline{p}(E_j)}\right] \tag{30}$$

We can see Eq.(29) is algebraically consistent with the conditional definition in Eq.(20), with $\sum_{j=1}^{n} \text{dual}\mathbf{p}(A|E_j)\text{dual}\mathbf{p}(E_j) = \sum_{j=1}^{n} \text{dual}\left[\mathbf{p}(A|E_j)\mathbf{p}(E_j)\right] = \text{dual}\sum_{j=1}^{n} \mathbf{p}(A \cap E_j) = \text{dual}\mathbf{p}(A)$.

When $n = 2$, $\mathbf{p}(E) + \mathbf{p}(E^c) = 1$. Let $\mathbf{p}(E^c) \in \overline{\mathbb{IR}}$. Eq.(29) becomes

$$\underline{p}(E|A) = \frac{\underline{p}(A|E)\underline{p}(E)}{\underline{p}(A|E)\underline{p}(E) + \underline{p}(A|E^c)\underline{p}(E^c)} = \frac{\underline{p}(A \cap E)}{\underline{p}(A \cap E) + \underline{p}(A \cap E^c)} \tag{31}$$

$$\overline{p}(E|A) = \frac{\overline{p}(A|E)\overline{p}(E)}{\overline{p}(A|E)\overline{p}(E) + \overline{p}(A|E^c)\overline{p}(E^c)} = \frac{\overline{p}(A \cap E)}{\overline{p}(A \cap E) + \overline{p}(A \cap E^c)} \tag{32}$$

When $\mathbf{p}(A \cap E) \in \mathbb{IR}$ and $\mathbf{p}(A \cap E^c) \in \overline{\mathbb{IR}}$, the relation is equivalent to the well-known 2-monotone tight envelop (Fagin and Halpern, 1991; de Campos et al., 1990; Wasserman and Kadan, 1990; Jaffray, 1992; Chrisman, 1995), given as:

$$P_*(E|A) = \frac{P_*(A \cap E)}{P_*(A \cap E) + P^*(A \cap E^c)} \tag{33}$$

$$P^*(E|A) = \frac{P^*(A \cap E)}{P^*(A \cap E) + P_*(A \cap E^c)} \tag{34}$$

where P_* and P^* are the lower and upper probability bounds defined in the traditional interval probabilities. Here $P^*(A \cap E^c) = \underline{p}(A \cap E^c)$ and $P_*(A \cap E^c) = \overline{p}(A \cap E^c)$ are the estimations of the lower and upper probability envelops.

LEMMA 4.1. $\mathbf{p}(A|E) \subseteq \mathbf{p}(A|E^c) \Leftrightarrow \mathbf{p}(E|A) \subseteq \mathbf{p}(E)$. $\mathbf{p}(A|E) \supseteq \mathbf{p}(A|E^c) \Leftrightarrow \mathbf{p}(E|A) \supseteq \mathbf{p}(E)$.

Proof. $\mathbf{p}(A|E) \subseteq \mathbf{p}(A|E^c) \Leftrightarrow \mathbf{p}(A \cap E)/\text{dual}\mathbf{p}(E) \subseteq \mathbf{p}(A \cap E^c)/\text{dual}\mathbf{p}(E^c) \Leftrightarrow \mathbf{p}(A \cap E)\mathbf{p}(E^c) \subseteq \mathbf{p}(A \cap E^c)\mathbf{p}(E) \Leftrightarrow \underline{p}(A \cap E)\underline{p}(E^c) \geq \underline{p}(A \cap E^c)\underline{p}(E)$ and $\overline{p}(A \cap E)\overline{p}(E^c) \leq \overline{p}(A \cap E^c)\overline{p}(E) \Leftrightarrow \underline{p}(A \cap E)\left[1 - \underline{p}(E)\right] \geq \underline{p}(A \cap E^c)\underline{p}(E)$ and $\overline{p}(A \cap E)\left[1 - \overline{p}(E)\right] \leq \overline{p}(A \cap E^c)\overline{p}(E) \Leftrightarrow \underline{p}(A \cap E) \geq \underline{p}(A \cap E)\underline{p}(E) + \underline{p}(A \cap E^c)\underline{p}(E)$ and $\overline{p}(A \cap E) \leq \overline{p}(A \cap E)\overline{p}(E) + \overline{p}(A \cap E^c)\overline{p}(E) \Leftrightarrow \mathbf{p}(A \cap E) \subseteq \mathbf{p}(A \cap E)\mathbf{p}(E) + \mathbf{p}(A \cap E^c)\mathbf{p}(E) \Leftrightarrow \mathbf{p}(A \cap E) \subseteq \left[\mathbf{p}(A \cap E) + \mathbf{p}(A \cap E^c)\right]\mathbf{p}(E) \Leftrightarrow \mathbf{p}(A \cap E)/\text{dual}\left[\mathbf{p}(A \cap E) + \mathbf{p}(A \cap E^c)\right] \subseteq \mathbf{p}(E) \Leftrightarrow \mathbf{p}(E|A) \subseteq \mathbf{p}(E)$.

The proof of $\mathbf{p}(A|E) \supseteq \mathbf{p}(A|E^c) \Leftrightarrow \mathbf{p}(E|A) \supseteq \mathbf{p}(E)$ is similar.

Remark. When the likelyhood functions $\mathbf{p}(A|E)$ and $\mathbf{p}(A|E^c)$ as well as prior and posterior probabilities are proper intervals, we can interpret the above relation as follows. If the likelyhood estimation of event A given E occurs is more accurate than that of event A given event E does not occur, then the extra information A can reduce the ambiguity of the prior estimation.

LEMMA 4.2. $\mathbf{p}(A|E) \geq \mathbf{p}(A|E^c) \Leftrightarrow \mathbf{p}(E|A) \geq \mathbf{p}(E)$. $\mathbf{p}(A|E) \leq \mathbf{p}(A|E^c) \Leftrightarrow \mathbf{p}(E|A) \leq \mathbf{p}(E)$.

Proof. The proof is similar to the previous Lemma.

Remark. If the occurance of event E increases the likelyhood estimation of event A compared to the one without the occurance of event E, then the extra information A will increase the probability of knowing that event E occurs.

LEMMA 4.3. $\mathbf{p}(A|E) = \mathbf{p}(A|E^c) \Leftrightarrow \mathbf{p}(E|A) = \mathbf{p}(E)$.

Proof. From either of the above two lemmas, $\mathbf{p}(A|E) = \mathbf{p}(A|E^c) \Leftrightarrow \mathbf{p}(A|E) \supseteq \mathbf{p}(A|E^c)$ and $\mathbf{p}(A|E) \subseteq \mathbf{p}(A|E^c) \Leftrightarrow \mathbf{p}(E|A) \supseteq \mathbf{p}(E)$ and $\mathbf{p}(E|A) \subseteq \mathbf{p}(E) \Leftrightarrow \mathbf{p}(E|A) = \mathbf{p}(E)$. Or $\mathbf{p}(A|E) =$

$\mathbf{p}(A|E^c) \Leftrightarrow \mathbf{p}(A|E) \geq \mathbf{p}(A|E^c)$ and $\mathbf{p}(A|E) \leq \mathbf{p}(A|E^c) \Leftrightarrow \mathbf{p}(E|A) \geq \mathbf{p}(E)$ and $\mathbf{p}(E|A) \leq \mathbf{p}(E) \Leftrightarrow \mathbf{p}(E|A) = \mathbf{p}(E)$.
Remark. The extra information A does not add much value to the assessment of event E if we have very similar likelyhood ratios, $\mathbf{p}(A|E)$ and $\mathbf{p}(A|E^c)$.

One of common issues associated with the Bayes' rule based on traditional set-based intervals is the loss of information during belief updating. The general bounds of posterior probabilities obtained depend on the sequence in which updates are performed (Pearl, 1990; Chrisman, 1995). That is, the posterior lower and upper bounds obtained by applying a series of evidences sequencially may disagree with the bounds obtained by conditioning the prior with all of the evidences in a single step. The belief updating based on Eq.(29) is sequence-independent because $\mathbf{p}(E|A)$ can be calculated incrementally, given as follows.

LEMMA 4.4. $\mathbf{p}(E|A \cap B) = \mathbf{p}(E \cap B|A)/\text{dual}\mathbf{p}(B|A)$ *for* $\forall A, B, E \in \mathcal{A}$.

Proof. $\mathbf{p}(E|A \cap B) = \mathbf{p}(E \cap A \cap B)/\text{dual}\mathbf{p}(A \cap B) = [\mathbf{p}(E \cap B|A)\mathbf{p}(A)] / \text{dual}[\mathbf{p}(B|A)\mathbf{p}(A)] = \mathbf{p}(E \cap B|A)/\text{dual}\mathbf{p}(B|A)$.
At the same time, $\mathbf{p}(E)$ can be calculated incrementally based on

$$\mathbf{p}(A \cap B) = \mathbf{p}(B|A)\mathbf{p}(A)$$

The above sequence-independent property is due to the algebraic closure of the conditional probability defined in Eq.(20).

4.1. Logic Interpretation

Some examples of logic interpretations for the relationships between prior and posterior interval probabilities in Eq.(29) are as follows.

- when $\mathbf{p}(A|E_i) \in \mathbb{IR}$, $\mathbf{p}(E_i) \in \mathbb{IR}$, $\mathbf{p}(A|E_j) \in \mathbb{IR}\,(j = 1, \ldots, n,\, j \neq i)$, $\mathbf{p}(E_{j_1}) \in \mathbb{IR}\,(j_1 = 1, \ldots, k,\, j_1 \neq i)$, $\mathbf{p}(E_{j_2}) \in \overline{\mathbb{IR}}\,(j_2 = k+1, \ldots, n,\, j_2 \neq i)$ and $\mathbf{p}(E_i|A) \in \overline{\mathbb{IR}}$

$$\begin{aligned} &\forall_{j\neq i} p_{A|E_j} \in \mathbf{p}'(A|E_j), \forall_{j_1 \neq i} p_{E_{j1}} \in \mathbf{p}'(E_{j_1}), \\ \exists p_{A|E_i} \in \mathbf{p}'(A|E_i), \exists p_{E_i} \in \mathbf{p}'(E_i), \exists_{j_2 \neq i} p_{E_{j2}} \in \mathbf{p}'(E_{j_2}), &\exists p_{E_i|A} \in \mathbf{p}'(E_i|A), \\ &p_{E_i|A} = \frac{p_{A|E_i} p_{E_i}}{\sum_{j=1}^{n} p_{A|E_j} p_{E_j}} \end{aligned} \tag{35}$$

- when $\mathbf{p}(A|E_i) \in \overline{\mathbb{IR}}$, $\mathbf{p}(E_i) \in \overline{\mathbb{IR}}$, $\mathbf{p}(A|E_j) \in \overline{\mathbb{IR}}\,(j = 1, \ldots, n,\, j \neq i)$, $\mathbf{p}(E_j) \in \overline{\mathbb{IR}}\,(j = 1, \ldots, n,\, j \neq i)$, and $\mathbf{p}(E_i|A) \in \overline{\mathbb{IR}}$

$$\begin{aligned} \forall_{j\neq i} p_{A|E_j} \in \mathbf{p}'(A|E_j), \forall_{j\neq i} p_{E_j} \in \mathbf{p}'(E_j), \forall p_{E_i|A} \in \mathbf{p}'(E_i|A), \\ \exists p_{A|E_i} \in \mathbf{p}'(A|E_i), \exists p_{E_i} \in \mathbf{p}'(E_i), \\ p_{E_i|A} = \frac{p_{A|E_i} p_{E_i}}{\sum_{j=1}^{n} p_{A|E_j} p_{E_j}} \end{aligned} \tag{36}$$

Notice that because both $\mathbf{p}(A|E_i)$ and $\text{dual}\mathbf{p}(A|E_i)$ occur in Eq.(29), the associated logic interpretation about $\mathbf{p}(A|E_i)$ is always existential. This indicates that the completeness of the posterior

probability $\mathbf{p}(E_i|A)$ cannot be checked by the interpretation itself. Yet the soundness of the posterior probability estimation can be checked by some interpretations such as the one in Eq.(36).

5. Concluding Remarks

In this paper, we presented a new form of imprecise probability based on generalized intervals. Generalized intervals allow the coexistence of proper and proper intervals. This enables the algebraic closure of arithmetic operations. We differentiate focal events from non-focal events by the modalities and semantics of interval probabilities. An event is focal when the semantics associated with its interval probability is universal, whereas it is non-focal when the semantics is existential. This differentiation allows us to have a simple and unified representation based on a logic coherence constraint, which is a stronger restriction than the regular 2-monotoniciy. This stronger requirement appears to be the cost we pay for the algebraic closure.

New rules of conditioning and updating are defined with generalized intervals. The new conditional probabilities ensure the algebraic relation with marginal interval probabilities. It is also shown that the new Bayes' updating rule is a generalization of the 2-monotone tight envelop updating rule under the new representation. This enables sequence-independent updating. Generalized intervals also allow us to interpret the algebraic relations among intervals in terms of the first-order logic. This helps us to understand the relationship between individual measurements and probability sets as well as to check completeness and soundness of bounds.

In summary, the algebraic closure of the new form provides some advantages for a simpler probability calculus, which is helpful in engineering and computer science practices. Future work may include the study of interpretation with the new form for assessment guidance. That is, we need to understand the algebraic conclusions better and take appropriate actions. Even though the computation is simplified, the completeness of lower and upper envelop estimations based on generalized intervals is not clear in general. We need to study how generalized intervals may underestimate envelops. We also need to investigate how much difference between the new and the traditional interval forms because of the logic coherence constraint.

Acknowledgements

The author thanks Scott Ferson for discussions.

References

Armengol, J., J. Vehi, L. Trave-Massuyes, and M. A. Sainz. Application of modal intervals to the generation of error-bounded envelopes. *Reliable Computing*, 7(2):171–185, 2001.

Aughenbaugh, J. M. and C. J. J. Paredis. The value of using imprecise probabilities in engineering design. *Journal of Mechanical Design*, 128(4):969–979, 2006.

Berleant D. and J. Zhang. Representation and problem solving with Distribution Envelope Determination (DEnv). *Reliability Engineering & System Safety*, 85(1-3):153–168, 2004.

Buede, D.M. and P. Girardi. A Target Identification Comparison of Bayesian and Dempster-Shafer Multisensor Fusion. *IEEE Transactions on Systems, Man & Cybernetics, Part A*, 27(5):569–577, 1997.

Chrisman, L. Incremental conditioning of lower and upper probabilities. *International Journal of Approximate Reasoning*, 13(1):1–25, 1995.

Coolen, F.P.A. On the use of imprecise probabilities in reliability. *Quality and Reliability Engineeering International*, 20(3):193–202, 2004.

de Campos, L. M., M. T. Lamata, and S. Moral. The concept of conditional fuzzy measure. *International Journal of Intelligent Systems*, 5(3):237–246, 1990.

Dempster, A. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38(2):325–339, 1967.

Du. L., K. K. Choi, B. D. Youn, and D. Gaorsich. Possibility-Based Design Optimization Method for Design Problems with Both Statistical and Fuzzy Input Data. *Journal of Mechanical Design*, 128(4):928–935, 2006.

Dubois, D. and H. Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*, Plenum, New York, 1988.

Dubois, D. and H. Prade. Focusing versus updating in belief function theory. In R.R. Yager, M. Fedrizzi and J. Kacprzyk, eds, *Advances in the Dempster–Shafer Theory of Evidence*, pp.71–95, John Wiley and Sons, New York, NY, 1994.

Elouedi, Z., K. Mellouli, and P. Smets. Assessing Sensor Reliability for Multisensor Data Fusion within the Transferable Belief Model. *IEEE Transactions on Systems, Man & Cybernetics, Part B*, 34(1):782–787, 2004.

Fagin, R. and J. Y. Halpern. A new approach to updating beliefs. In P.P. Bonissone, M. Henrion, L.N. Kanal, and J. Lemer, eds, *Uncertainty in Artificial Intelligence* , Vol.VI, pp.347–374, Elsevier, Amsterdam, 1991.

Ferson, S., V. Kreinovich, L. Ginzburg, D. S. Myers, and K. Sentz. Constructing probability boxes and Dempster-Shafer structures. Sandia National Laboratories Technical report SAND2002-4015, Albuquerque, NM, 2002.

Gardenes, E., M. A. Sainz, L. Jorba, R. Calm, R. Estela, H. Mielgo, and A. Trepat. Modal intervals. *Reliable Computing*, 7(2):77–111, 2001.

Goldsztejn, A. A right-preconditioning process for the formal-algebraic approach to inner and outer estimation of AE-solution sets. *Reliable Computing*. 11(6):443–478, 2005.

Jaffray, J.-Y. Bayesian Updating and Belief Functions. *IEEE Transactions on Systems, Man, & Cybernetics*, 22(5):1144–1152, 1992.

Kaucher, E. Interval analysis in the extended interval space IR. *Computing Supplement*, 2:33–49, 1980.

Kozine, I. O. and Y. V. Filimonov. Imprecise reliabilities: experiences and advances). *Reliability Engineering & System Safety*, 67(1):75–83, 2000.

Kreinovich, V., V. M. Nesterov, and N. A. Zheludeva. Interval methods that are guaranteed to underestimate (and the resulting new justification of Kaucher arithmetic). *Reliable Computing*, 2(2):119–124, 1996.

Kulasekere, E. C., K. Premaratne, D. A. Dewasurendra, M.-L. Shyu, and P. H. Bauer Conditioning and updating evidence. *International Journal of Approximate Reasoning*, 36(1), 75–108, 2004.

Kupriyanova, L. Inner estimation of the united solution set of interval algebraic system *Reliable Computing*, 1(1):15–41, 1995.

Markov, S. On the algebraic properties of intervals and some applications. *Reliable Computing*, 7(2):113–127, 2001.

Molchanov, I. *Theory of Random Sets*. Springer, London, 2005.

Möller, B. and Beer M. *Fuzzy Randomness: Uncertainty in Civil Engineering and Computational Mechanics*. Springer, Berlin, 2004.

Moore, R. E. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 1966.

Mourelatos, Z.P. and J. Zhou. A design optimization method using evidence theory. *Journal of Mechanical Design*, 128(4):901–908, 2006.

Neumaier, A. Clouds, fuzzy sets, and probability intervals. *Reliable Computing*, 10(4):249–272, 2004.

Nikolaidis, E., Q. Chen, H. Cudney, R. T. Haftka, and R. Rosca. Comparison of Probability and Possibility for Design Against Catastrophic Failure Under Uncertainty. *Journal of Mechanical Design*, 126(3):386–394, 2004.

Pearl, J. Reasoning with belief functions: An analysis of compatibility. *International Journal of Approximate Reasoning*, 4(5-6):363–389, 1990.

Popova, E. D. Multiplication distributivity of proper and improper intervals. *Reliable Computing*, 7(2):129–140, 2001.

Shafer, G. *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, NJ, 1990.

Shary, S. P. A new technique in systems analysis under interval uncertainty and ambiguity. *Reliable Computing*, 8(2):321–418, 2002.

Smets, P. About Updating. *Proc. 7th conference on Uncertainty in Artificial Intelligence*, San Mateo, CA, pp.378–385, 1991.

Walley, P. *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, London, 1991.

Walley, P. Measures of uncertainty in expert systems. *Artificial Intelligence*, 83(1):1–58, 1996.

Wasserman, L. A. and J. B. Kadane. Bayes' theorem for Choquet capacities. *The Annals of Statistics*, 18(3):1328–1339, 1990.

Weichselberger, K. The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24(2-3), 149–170, 2000.

Zadeh, L. A. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Systems*, 1(1):3–28, 1978.

Extreme probability distributions of random/fuzzy sets and p-boxes

A. Bernardini
Dpt di Costruzioni e Trasporti, Università degli Studi di Padova
email: alberto.bernardini@unipd.it

F. Tonon
Dpt of Civil Engineering, University of Texas
email:tonon@mail.utexas.edu

Abstract: Uncertain information about a system variable described by a random set or an equivalent Dempster-Shafer structure on a finite space of singletons determines an infinite convex set of probability distributions, given by the convex hull of a finite set of extreme distributions. Extreme distributions allow one to evaluate (through the Choquet integral) exact upper/lower bounds of the expectation of monotonic and non-monotonic functions of uncertain variables, for example in reliability evaluation of engineering systems. The paper considers the simple case of a single variable, and details applications to random sets with nested focal elements (consonant random sets or the equivalent fuzzy set) and to p-boxes. A simple direct procedure to derive extreme distributions from a p-box is described through simple numerical examples.

Keywords: random sets, fuzzy sets, p-boxes

1 Introduction

In Civil Engineering practice, the growing need for rationally including uncertainty in engineering modelling and calculations is witnessed by the adoption of reliability-based EuroCodes or Load and Resistance Factor Design codes (Level I). More sophisticated reliability-based approaches are used in research or special practical problems (Levels II and III). This need, however, has been accompanied by the realization of the limitations that affect probabilistic modelling of uncertainty when dealing with imprecise data (Walley. 1991).

On one hand, in the enlarged ambit of a multi-valued logic, alternative models of uncertainty have been propounded that attempt to capture qualitative or ambiguous aspects of engineering models. Particularly important models are based on the idea of fuzzy sets and relations, and positive applications have been reported in the fields of automatic controls in robotics and artificial intelligence, more generally in the field of optimal decisions and approximate reasoning. Less convincing and frequently charged with leading to unrealistic or unverifiable conclusions are the tentative applications of fuzzy

models in predicting or simulating objective phenomena, for example to evaluate the reliability of an engineering design or to assess the reliability of an existing engineering system.

On the other, new models of uncertainty have been formulated, based on a generalisation of the probabilistic paradigm, and in particular its objective interpretation as relative frequency of events. The main point is the considerations of "imprecise probabilities" of events or "imprecise previsions" of functions, based on the idea of bounded sets of probability distributions compatible with the available information or, alternatively, on the combination of a probability distribution (randomness) with imprecise events (set uncertainty). Because these models retain the semantics of probability theory, comparisons with probability theory are straightforward.

The subjectivist formulation of this approach (Theory of evidence, (Shafer. 1976)) is compatible with a different interpretation based on statistics of objective but imprecise events (Theory of random sets). When imprecise events are nested, it includes the notion of fuzzy set as a particular case.

After a quick review of the definitions and properties of imprecise probabilities and classification of the corresponding upper/lower bounds according to the order of Choquet capacities, the paper focuses on the theory of random sets, with particular emphasis on fuzzy sets (consonant random sets) and p-boxes (non consonant random sets that contain, as a particular case, the ordinary probability distributions). Both fuzzy sets and p-boxes are indexable-type random set, i.e the imprecise events can be ordered and univocally determined by an index varying from 0 to 1 (Alvarez. 2006). This property is very useful in applications involving numerical simulations.

With reference to a finite probability space for a single variable, the paper continues by discussing the properties of the infinite convex set of probability distributions, and of the finite set of extreme distributions generated by random sets, fuzzy sets, and p-boxes. The finite sets of extreme distributions are particularly useful in evaluating exact expectation bounds for a real-valued function of the considered variable, in the case of both monotonic and not monotonic functions.

A simple and direct procedure to derive extreme distributions from a p-box is described.

2 Imprecise probabilities and convex sets of probability distributions

2.1 COHERENT UPPER AND LOWER PROBABILITIES AND PREVISIONS

Le us consider a finite probability space $(\Omega, \mathcal{F}, P)$, where $\mathcal{F}$ is the σ-algebra generated by a finite partition of Ω into elementary events (or singletons) $S = \{s_1, s_2 \ldots, s_j, \ldots s_n\}$. Hence the probability space is fully specified by the probabilities $P(s_j)$, which sum up to 1 (in the following: the "probability distribution").

Imprecise probabilities arise when the available information does not allow one to uniquely determine a unique probability distribution. In this case, the information could be given by means of upper and/or lower probabilities, $\mu_{LOW}(T_i)$, $\mu_{UPP}(T_i)$, of some events $T_i \in \mathcal{F}$, or directly through a set of probability distributions, Ψ.

The foundation of a theory of imprecise probabilities is mainly due to the work of Peter Walley in the 1980s/90s on a new theory of probabilistic reasoning, statistical inference and decision, under uncertainty, partial information or ignorance ((Walley. 1991), or for a concise introduction (Walley. 2000)). In his work, the basic idea of upper/lower probabilities is enlarged to the more general concept of upper/lower *previsions* for a family of bounded and point-valued functions f_i: $S \rightarrow Y = \Re$. For a specific precise probability distribution $P(s_j)$, the prevision is equivalent to the linear expectation:

$$E_P[f_i] = \sum_{s_j \in S} f_i(s_j) P(s_j) \qquad (2.1)$$

Since the probability of an event T_i is equal to the expectation of its indicator function (equal to: 1 if $s_j \in T_i$, 0 if $s_j \notin T_i$), upper/lower previsions generalize and hold upper/lower probabilities as a particular case.

Let us now focus on the information about the space of events in S given by upper and/or lower previsions, $E_{LOW}[f_i]$ and $E_{UPP}[f_i]$, for a family of bounded and point-valued functions f_i, $\mathcal{K}$. This is accomplished by the set, Ψ^E, of probability distributions $P(s_j)$ compatible with $E_{LOW}[f_i]$ and $E_{UPP}[f_i]$:

$$\Psi^E = \left\{ P : E_{LOW}[f_i] \le E_P[f_i] \le E_{UPP}[f_i] \quad \forall f_i \in \mathcal{K} \right\} \qquad (2.2)$$

Ψ^E is convex and closed. One is interested in checking two basic conditions of the suggested bounds:

1. A preliminary, strong condition requires that set Ψ^E should be non-empty. If set Ψ^E is empty, it means there is something basically irrational in the suggested bounds. For example, the set Ψ^E is empty if $E_{LOW}[f_i] > \max_j f_i(s_j)$ or $E_{UPP}[f_i] < \min_j f_i(s_j)$ (for upper/lower probabilities: $\mu_{LOW}(T_i) > 1$ or $\mu_{UPP}(T_i) < 0$). In the behavioural interpretation adopted by Walley, the functions f_i are called *gambles*, and this basic condition is said to *avoid sure loss*.
2. A second, weaker but reasonable condition requires that the given bounds should be the same as the *naturally extended* expectation bounds that can be derived from Ψ^E (*coherence* according to Walley's nomenclature)

$$E_{LOW,c}[f_i] = \min_{P \in \Psi^E} E_P[f_i] \qquad (2.3)$$
$$E_{UPP,c}[f_i] = \max_{P \in \Psi^E} E_P[f_i]$$

In this case, one says that $E_{LOW}[f_i]$ and $E_{UPP}[f_i]$ are (lower and upper, respectively) envelopes of Ψ^E.

If the given bounds are not coherent, i.e. envelopes to Ψ^E, because they do not satisfy Eq. (2.3), the given bounds can be restricted without changing the probabilistic content of the original information, i.e. set Ψ^E. These restricted bounds, calculated by using Eq. (2.3), are called "*natural extension*" of the given bounds $E_{LOW}[f_i]$ and $E_{UPP}[f_i]$. For example if bounds are given for both function f_i and the opposite $-f_i$

coherence requires the "duality condition": $E_{UPP}[f_i] = - E_{LOW}[-f_i]$ (equivalently for upper/lower probabilities of complementary sets T_i and T_i^c : $\mu_{UPP}(T_i) = 1- \mu_{LOW}(T_i^c)$).

The applications that follow are restricted to the special case when $\mathcal{K}$ is a set of indicator functions, i.e. previsions coincide with probabilities. In this special case, there is no one-to-one correspondence between imprecise probabilities and closed convex sets of probability distributions because several closed convex sets of probability distributions could give the same imprecise probabilities. This one-to-one correspondence only holds between previsions and convex sets of probability distributions when $\mathcal{K}$ is the set of *all* functions. In other terms, imprecise probabilities are less informative than previsions.

2.2 CHOQUET CAPACITIES AND ALTERNATE CHOQUET CAPACITIES

An important criterion for classifying monotonic (with respect to inclusion) measures of sets was introduced by Choquet in his theory of *capacities* (Choquet. 1954). Given a finite set *S*, let $\mathcal{P}(S)$ be the power set (set of all subsets) of *S*. A regular monotone set function μ: $\mathcal{P}(S) \rightarrow [0, 1] \mid \mu(\varnothing) = 0$, $\mu(S)= 1$ is called 2-monotone (or a *Choquet Capacity of order k* = 2) if, given two subsets T_1 and T_2:

$$\mu(T_1 \cup T_2) \geq \mu(T_1) + \mu(T_2) - \mu(T_1 \cap T_2) \qquad (2.4)$$

The dual coherent upper probabilities ($\mu_{UPP}(T_i) = 1- \mu(T_i^c)$) are called *Alternate Choquet Capacity of order k* = 2, and satisfy the relation:

$$\mu_{UPP}(T_1 \cap T_2) \leq \mu_{UPP}(T_1) + \mu_{UPP}(T_2) - \mu_{UPP}(T_1 \cup T_2) \qquad (2.5)$$

More generally, monotone dual set functions (μ, μ_{UPP}) are *k*-monotone (*Choquet Capacity of order k*), and, respectively, *Alternate Choquet Capacity of order k*, if, given *k* subsets T_1, T_2 $\ldots T_k$:

$$\mu(T_1 \cup T_2 \ldots \cup T_k) \geq \sum_{\varnothing \subset K \subseteq \{1,2..k\}} (-1)^{|K|+1} \mu\left(\bigcap_{i \in K} T_i\right) \qquad (2.6)$$

$$\mu_{UPP}(T_1 \cap T_2 \ldots \cap T_k) \leq \sum_{\varnothing \subset K \subseteq \{1,2..k\}} (-1)^{|K|+1} \mu_{UPP}\left(\bigcup_{i \in K} T_i\right)$$

Precise probability distributions are both an Choquet Capacity and an Alternate Choquet Capacity of order $k= \infty$, that satisfy relations (2.5) and (2.6) as equalities.

Choquet and dual Alternating Choquet capacities of order $k > 1$ are coherent lower and upper probabilities respectively. Indeed, compare the above properties with the necessary conditions for coherent upper/lower probabilities:

- Monotonicity with inclusion: $T_1 \subseteq T_2 \Rightarrow \quad \mu_{LOW}(T_1) \leq \mu_{LOW}(T_2)$; $\quad \mu_{UPP}(T_1) \leq \mu_{UPP}(T_2)$

- Super-additivity of μ_{LOW} for disjoint sets ($T_1 \cap T_2 = \varnothing$): $\mu_{LOW}(T_1 \cup T_2) \geq \mu_{LOW}(T_1) + \mu_{LOW}(T_2)$
- Sub-additivity of μ_{UPP} for any pair of sets T_1, T_2: $\mu_{UPP}(T_1 \cup T_2) \geq \mu_{UPP}(T_1) + \mu_{UPP}(T_2)$.

Therefore, coherent super-additive lower probabilities are not necessarily Choquet capacities of order $k > 1$.

There is a strong connection between the order k and the *Möbius transform* of the set function $\mu(T)$:

$$^{\mu}m(A) = \sum (-1)^{|A-T|}\, \mu(T) \,|\, T \subseteq A \qquad (2.7)$$

The Möbius transform of a set function μ is a one-to-one invertible set function $^{\mu}m$: $\mathcal{P}(S) \rightarrow \Re$, and its inverse is precisely :

$$^{m}\mu(T) = \sum m(A) \,|\, A \subseteq T, \qquad \forall T \subset S; \qquad (2.8)$$

For the purposes of this study, the most interesting properties (see for example (Chateauneuf and Jaffray. 1989, Klir. 2005) are the following:

1- a set function μ is monotone if and only if:

$$m(\varnothing) = 0; \sum_{T \in \mathcal{P}(S)} {}^{\mu}m(T) = 1; \quad \forall T \in \mathcal{P}(S) : \sum_{A \subseteq T} {}^{\mu}m(A) \geq 0 \qquad (2.9)$$

and, therefore, $\forall\, j$: $^{\mu}m(\{s_j\}) \geq 0$.

2- If $\mu(T)$ is k-monotone and $|T| \leq k$ then $^{\mu}m(T) \geq 0$

3- $\mu(T)$ is ∞-monotone if and only if: $\forall\, T \in \mathcal{P}(S) : {}^{\mu}m(T) \geq 0$.

2.3 EXTREME DISTRIBUTIONS

For a given regular monotone set function μ, a permutation $\pi(j)$ of the indexes of the singletons in the set $S = \{s_1, s_2 \ldots, s_j, \ldots s_n\}$ defines the following probability distribution:

$$P\left(s_{\pi(j)=1}\right) = \mu\left(\left\{s_{\pi(j)=1}\right\}\right) \qquad (2.10)$$
$$P\left(s_{\pi(j)=k>1}\right) = \mu\left(\left\{s_{\pi(j)=1},\ldots s_k\right\}\right) - \mu\left(\left\{s_{\pi(j)=1},\ldots s_{k-1}\right\}\right)$$

The |S|! possible permutations define a finite set of probability distributions, *EXT*, together with its convex hull, Ψ^{EXT}.

If the same permutation is applied to a pair (μ_{LOW}, μ_{UPP}) of dual regular monotone set functions, a pair of dual distinct probability distributions is generated, but (μ_{LOW}, μ_{UPP}) always generate the same set *EXT*.

Now, one would wonder what the relationship is between Ψ^{EXT} and the set Ψ^{μ} calculated for (μ_{LOW}, μ_{UPP}) by using (Eq. 2.2). It turns out that the two sets could be different, and satisfy the inclusion: $\Psi^{E} \subseteq \Psi^{EXT}$. Precisely:

- For coherent monotone measures (k = 1), Eq. (2.10) could generate probability distributions in *EXT* that do not satisfy the bounds in (Eq. 2.2); hence Ψ^{μ} could be strongly included in Ψ^{EXT};

- for monotone measures with k > 1, all probability distributions in *EXT* (and in Ψ^{EXT}) satisfy the bounds in (Eq. 2.2), and thus $\Psi^{EXT} = \Psi^{\mu}$; *EXT* coincides with the set of the extreme points (or the *profile*) of the closed convex set Ψ^{μ} .

2.4 EXPECTATION BOUNDS AND CHOQUET INTEGRALS FOR REAL VALUED FUNCTIONS

When the sets Ψ^{μ} or Ψ^{EXT} are known, or when a generic set Ψ is assigned, the upper and lower expectation bounds for any real function f: $S \rightarrow Y = \Re$ could be calculated by solving the optimization problems in Eqs (2.3) by substituting Ψ^{μ}, Ψ^{EXT}, or Ψ for Ψ^{E}, respectively. However, the *Choquet integral* (a direct calculation based on the dual upper/lower probabilities) is generally suggested in the literature to solve the problem more easily.

The expectation of a point valued function f: $S \rightarrow Y=[y_L, y_R] \subset \Re$ with CDF $F(y)$ can be calculated as follows by using the Stieltjes Integral and equivalent expressions:

$$E[y = f] = \int_{y_L}^{y_R} f \cdot dF = \left[yF\right]_{y_L}^{y_R} - \int_{y_L}^{y_R} F dy = y_R - \int_{y_L}^{y_R} F dy = y_L + \int_{y_L}^{y_R} (1-F) dy =$$
$$= y_L + \int_{y_L}^{y_R} P(f > \alpha) d\alpha = y_L + \int_{y_L}^{y_R} P({}^{\alpha}T = \{s \in S \mid f(s) > \alpha\}) d\alpha \qquad (2.11)$$

The Choquet Integral is the direct extension of the last functional expression to a monotonic measure μ, for the ordered family of subsets ${}^{\alpha}T$, which depend on the selected function *f*:

$$C(f,\mu) = y_L + \int_{y_L}^{y_R} \mu({}^{\alpha}T)d\alpha \qquad (2.12)$$

Indeed, the Choquet integral gives a numerical value that coincides with the expectation of the function *f* for a particular probability distribution. The latter distribution is obtained by the permutation leading to a monotonic (decreasing) ordering of the function values.

The expectation bounds are therefore obtained through the dual probability distributions obtained by applying Eq. (2.11) to the dual upper/lower probabilities (μ_{LOW}, μ_{UPP}). The Choquet integral determines optimal bounds with respect to the set EXT (or Ψ^{EXT}) defined in Section 2.3: hence, for general monotone measures ($k = 1$), it can give larger bounds than the correct bounds calculated by using the extreme points of Ψ^{μ}; on the other hand, for $k > 1$, the Choquet integral gives exact expectation bounds.

3 Random sets

3.1 General properties of random sets

Among the different definitions of random set, we refer here to the formalism of the Theory of Evidence, but with no particular limitation to the subjectivist emphasis of this theory. The original information is described by a family of pairs of nonempty subsets A^i ("*focal elements*") and attached $m^i = m(A^i) > 0$, i=1, 2, …n ("*probabilistic assignment*"), with the condition that the sum of m^i is equal to 1. The (total) probability of any subset T of S can therefore be bounded by means of the additivity rule. Shafer suggested the words *Belief* (*Bel*) and *Plausibility* (*Pla*) for the lower and upper bounds, respectively. Formally:

$$\begin{aligned} \forall T \subset S : \mu_{UPP}(T) &= Pla(T) = \sum_i m^i \mid A^i \cap T \neq \varnothing, \\ \mu_{LOW}(T) &= Bel(T) = \sum_i m^i \mid A^i \subseteq T \end{aligned} \qquad (3.1)$$

Comparison with Eq. (2.8) demonstrates that *Bel* is the inverse Möbius transform of the non-negative set function *m*: hence *Bel* is a ∞-monotone set function, and *Pla* an Alternate Choquet capacity of order k =

∞. As explained in Section 2.3, Ψ^{Bel} (calculated with Eq. 2.2 for *Bel*) coincides with the set Ψ^{EXT}, where EXT (calculated with Eq. 2.11) is the set of extreme distributions that can be used to evaluate exact expectation bounds for a function of interest.

3.2 FUZZY SETS

The conclusions in Section 3.1 also apply in the particular case of a *consonant* random set; i.e. when focal elements are *nested*, and hence can be ordered in such a way that:

$$A^1 \subseteq A^2 \subseteq .... \subseteq A^n \tag{3.2}$$

Consonant random sets satisfy the relation:

$$Pla(T_1 \cup T_2) = \max(Pla(T_1), Pla(T_2)) \tag{3.3}$$

and hence (similar to classical Probability measures) they satisfy the following "*decomposability property*": the measure of uncertainty of the union of any pair of disjointed sets depends solely on the measures of the individual sets. Therefore, in the case of a consonant random set, the point-valued *contour function* (Shafer. 1976) μ: $S \rightarrow [0, 1]$:

$$\mu(s_j) = Pla(\{s_j\}) \tag{3.4}$$

completely defines the information on the measures of any subset $T \subset S$, exactly in the same way as the probability distribution $P(s_j)$ defines, although through a different rule (the additivity rule), the probability of every subset T in the algebra generated by the singletons. Indeed:

$$Pla(T) = \max_{s_j \in T} \mu(s_j); \qquad Bel(T) = 1 - \max_{s_j \in T^c} \mu(s_j) \tag{3.5}$$

Moreover, the Möbius inversion (2.7) of the set function *Bel* allows the (nested) family of focal elements to be determined through the set function *m*.

More directly, let us assume:

$$\alpha_1 = \max_{\mathrm{j}}\left(\mu\left(s_j\right)\right) = 1$$

.....

$$\alpha_i = \max_{\mathrm{j}|\mu\left(s_j\right)<\alpha_{i-1}}\left(\mu\left(s_j\right)\right); \qquad (3.6)$$

.....

$$\alpha_n = \max_{\mathrm{j}|\mu\left(s_j\right)<\alpha_{n-1}}\left(\mu\left(s_j\right)\right) = \min_{\mathrm{j}}\left(\mu\left(s_j\right)\right);$$

$$\alpha_{n+1} = 0$$

The family of focal elements and related probabilistic assignments (summing up to 1) are given by:

$$A^i = \left\{s_j \in S \mid \mu\left(s_j\right) \geq \alpha_i\right\} \; ; \quad m^i = \alpha_i - \alpha_{i+1} \qquad (3.7)$$

The number of focal elements, *n*, is therefore equal to the cardinality of the range of *S* through μ; of course this cardinality is less then or equal to |*S*|, because some singletons could map onto the same value of plausibility.

There is a narrow correspondence between consonant random sets and other decomposable measures of uncertainty: fuzzy sets and possibility distributions. This connection can clearly be envisaged using the dual representation of a fuzzy set through their α-cuts ${}^{\alpha}A$. They are classical subsets of *S* defined, for any selected value of membership α, by the formula:

$${}^{\alpha}A = \left\{s \in S \mid \mu\left(s\right) \geq \alpha\right\} \qquad (3.8)$$

When a fuzzy set is implicitly given through the (finite or infinite) sequence of its α-cuts ${}^{\alpha}A$, its membership function can be reconstructed through the equation:

$$\mu\left(s_j\right) = \max_{\alpha} \quad \min\left(\alpha, \chi_{{}^{\alpha}A}\left(s\right)\right) \qquad (3.9)$$

where $\chi_{{}^{\alpha}A}\left(s\right)$ is the indicator function of the classical subset ${}^{\alpha}A$.

By comparing Eq. (3.8) with Eq. (3.7), it is clear that the α-cuts ${}^{\alpha}A$ of any given normal fuzzy set are a nested sequence of subsets of set *S*, and therefore the family of focal elements of an associated consonant random set: the membership function of normal fuzzy sets gives the contour function of the corresponding random sets, and the basic probabilistic assignment (for a finite sequence of α-cuts) is given by $m(A^i = {}^{\alpha i}A) = \alpha_i - \alpha_{i+1}$.

By considering Eq. (3.5) from this point of view, the membership function of a fuzzy subset A allows measures of Plausibility and Belief to be attached to every classical subset $T \subseteq S$; this very different interpretation of a fuzzy set was recognized by Zadeh himself in 1978 (Zadeh. 1978), as the basis of a theory of *Possibilities* defined by a *possibility* distribution numerically equal to $\mu_A(s)$, and later extensively developed by other authors, in particular Dubois and Prade (Dubois and Prade. 1988).

This comparison suggests a probabilistic (objective or subjective) content of the information summarized by a fuzzy set and allows one to evaluate by means of the set *EXT* exact expectation bounds for real functions of a fuzzy variable. Although the discussion was restricted to finite discrete variables, the conclusion can be extended to continuous variables.

Example 3—1. Consider $S = \{s_1, s_2, s_3, s_4\}$, and the point-valued function $f(s_j)$ mapping to the set $Y = \{5, 20, 10, 0\}$. The fuzzy set of S is measured by the set of membership values (0, 0.1, 1, 0.1). Eqs. (3.6) give: $\alpha_1 = 1$; $\alpha_2 = 0.1$; $\alpha_3 = 0$. The associated consonant random set is defined by the set of pairs $\{ (A^1 = \{s_3\}, m^1 = 1 - 0.1 = 0.9), (A^2 = \{ s_2, s_3, s_4\}, m^2 = 0.1 - 0 = 0.1)\}$.
The permutation leading to a monotonic decreasing ordering of the function $f(s_j)$ is the following:
($\pi(s_2) = 1$, $\pi(s_3) = 2$, $\pi(s_1) = 3$, $\pi(s_4) = 4$). Table 3—1 shows the corresponding dual extreme distribution according to Eq. (2.10) and the dual set functions *Pla* and *Bel*.

Table 3—1. Dual extreme distributions for Example 3—1

T	$Pla(T)$	$P_{EXT,UPP}(s)$	T^c	$Bel(T) = 1 - Pla(T^c)$	$P_{EXT,LOW}(s)$
$T_1 = \{s_2\}$	0.1	$P(s_2) = Pla(T_1) = 0.1$	$\{s_1, s_3, s_4\}$	0	$P(s_2) = Bel(T_1) = 0$
$T_2 = \{s_2, s_3\}$	1	$P(s_3) = Pla(T_2) - Pla(T_1) = 0.9$	$\{s_1, s_4\}$	0.9	$P(s_3) = Bel(T_2) - Bel(T_1) = 0.9$
$T_3 = \{s_2, s_3, s_1\}$	1	$P(s_1) = Pla(T_3) - Pla(T_2) = 0$		0.9	$P(s_1) = Bel(T_3) - Bel(T_2) = 0$
$T_4 = S$	1	$P(s_4) = Pla(T_4) - Pla(T_3) = 0$	$\varnothing$	1	$P(s_4) = Bel(T_4) - Bel(T_3) = 0.1$

Hence:
$E_{UPP}[f] = E_{P_{EXT,UPP}}[f] = 20 \times 0.1 + 10 \times 0.9 = 11$; $E_{LOW}[f] = E_{P_{EXT,LOW}}[f] = 10 \times 0.9 + 0 \times 0.1 = 9$.
The same results can be obtained through the Choquet integral (Eq. (2.12)). For example:

$C(f, \mu_{LOW} = Bel) = 0 + Bel(\{s_2, s_3, s_1\}) \times (\Delta\alpha = f(s_1) - f(s_4)) + Bel(\{s_2, s_3\}) \times (\Delta\alpha = f(s_3) - f(s_1)) + Bel(\{s_2\}) \times (\Delta\alpha = f(s_2) - f(s_3)) = 0 + 0.9 \times (5-0) + 0.9 \times (10 - 5) + 0 \times (20 - 10) = 9$

$C(f, \mu_{UPP} = Pla) = 0 + Pla(\{s_2, s_3, s_1\}) \times (\Delta\alpha = f(s_1) - f(s_4)) + Pla(\{s_2, s_3\}) \times (\Delta\alpha = f(s_3) - f(s_1)) + Pla(\{s_2\}) \times (\Delta\alpha = f(s_2) - f(s_3)) = 0 + 1 \times (5-0) + 1 \times (10 - 5) + 0.1 \times (20 - 10) = 11$

3.3 P-BOXES

Given a finite space S, a set Ψ^F of probability distributions is implicitly defined by lower and upper bounds, $F_{LOW}(s_j)$ and $F_{UPP}(s_j)$, of the cumulative distribution functions $F(s_j)$:

$$\Psi^F = \left\{ P : F_{LOW}(s_j) \le F(s_j) = P\left(\left\{s_1,...,s_j\right\}\right) \le F_{UPP}(s_j),\ j = 1 \text{ to } |S| \right\} \quad (3.10)$$

The set Ψ^F is non-empty if $F_{LOW}(s_k) \le F_{UPP}(s_j)$ for any $k \le j$.

However, coherence clearly requires stronger conditions: the bounds $F_{LOW}(s_j)$ and $F_{UPP}(s_j)$ should be non-negative, non-decreasing in j, and both must be equal to 1 for $j = |S|$ (Walley. 1991, § 4.6.6).

Explicit evaluation of set Ψ^F can be obtained by solving the constraints (3.10) for the probabilities of the singletons $P(s_j)$:

$$\begin{aligned}
&F_{LOW}(s_1) \le P(s_1) \le F_{UPP}(s_1); && P(s_1) \ge 0 \\
&F_{LOW}(s_2) \le P(s_1) + P(s_2) \le F_{UPP}(s_2); && P(s_2) \ge 0 \\
&\text{.....} \\
&F_{LOW}(s_j) \le P(s_j) + \sum_{i=1}^{j-1} P(s_i) \le F_{UPP}(s_j); && P(s_j) \ge 0 \\
&\text{.....} \\
&P\left(s_{j=|S|}\right) + \sum_{i=1}^{|S|-1} P(s_i) = 1; && P\left(s_{j=|S|}\right) \ge 0
\end{aligned} \quad (3.11)$$

A simple iterative procedure can be used. For example, the explicit solution of the first two constraints is shown in Figure 3—1: observe that the p-box defines 4 (case a)) or 5 (case b)) extreme points of the projection of set Ψ^F on the two-dimensional space ($P(s_1)$, $P(s_2)$).

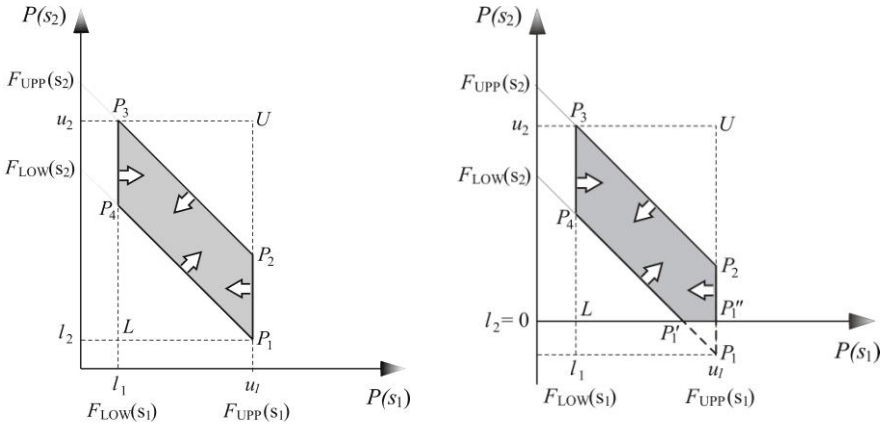


Figure 3—1. Explicit solution of the first 2 constraints in Eq. (3.11): case a): $F_{LOW}(s_2) - F_{UPP}(s_1) > 0$; case b: $F_{LOW}(s_2) - F_{UPP}(s_1) < 0$. Projection of set Ψ^F is shown hatched.

The interval bounds for the probability of the singletons are given by the intervals:

$$[l_1, u_1] = [F_{LOW}(s_1), F_{UPP}(s_1)],$$
$$[l_2, u_2] = [\max(0, F_{LOW}(s_2) - F_{UPP}(s_1)), F_{UPP}(s_2) - F_{LOW}(s_1)]$$

However, the set Ψ^{F*} generated by the same interval probabilities thought of as being non-interactive could be much larger. Indeed, provided that the last constraint in (3.11) is satisfied, the extreme points $U=(u_1, u_2)$ and $L=(l_1, l_2)$ could be in Ψ^{F*} together with the entire Cartesian product $[l_1, u_1] \times [l_2, u_2]$.

More generally, the interval probabilities for singleton $\{s_j\}$ are given by the intervals:

$$[l_j, u_j] = \left[\max\left(0, F_{LOW}(s_j) - F_{UPP}(s_{j-1})\right), F_{UPP}(s_j) - F_{LOW}(s_{j-1}) \right] \qquad (3.12)$$

The extreme points of the projection of set Ψ^F on the j-dimensional space $(P(s_1), \ldots, P(s_j))$ can be derived from each extreme point on the $j-1$-dimensional space, by considering that the sum $P(s_1) + \ldots + P(s_j)$ must be bounded by $F_{LOW}(s_j)$ and $F_{UPP}(s_j)$.

A constructive algorithm to evaluate the extreme distributions compatible with the information given by a p-box can be obtained by selecting the set, *EXT*, corresponding to the cumulative (non-decreasing) distribution functions jumping from $F_{LOW}(s_j)$ to $F_{UPP}(s_j)$ at some points s_j and from $F_{UPP}(s_k)$ to $F_{LOW}(s_k)$ at other points s_k (or at least non-decreasing values of F, case b) in Figure 3—1). Of course, the set *EXT* contains the distribution functions corresponding to the bounds of the p-box: $P_{EXT,LOW}(s_j) = F_{LOW}(s_j) - F_{LOW}(s_{j-1})$; $P_{EXT,UPP}(s_j) = F_{UPP}(s_j) - F_{UPP}(s_{j-1})$.

The same set *EXT* (and therefore the same set $\Psi^R=\Psi^F$ of probability distributions) can be given by an equivalent random set, *R*, with focal elements and probabilistic assignment derived from the p-box by using a rule quite similar to the algorithm for deriving an equivalent random set from a normal fuzzy set (when the membership function is meant as a possibility distribution; see § 3.2).

Define :

$$s_j^- : F_{LOW}\left(s_j^-\right) = \lim_{\varepsilon\to 0^+} F_{LOW}\left(s_j-\varepsilon\right) = F_{LOW}\left(s_j\right) - P_{LOW}\left(s_j\right)$$

Let:

$$\alpha_1 = 1 = \max_j\left(F_{LOW}\left(s_j\right)\right) = \max_j\left(F_{UPP}\left(s_j\right)\right);$$

$$\alpha_2 = \max\left(\max_{j|F_{LOW}(s_j^-)<\alpha_1}\left(F_{LOW}\left(s_j\right)\right), \max_{j|F_{UPP}(s_j)<\alpha_1}\left(F_{UPP}\left(s_j\right)\right)\right);$$

.....

$$\alpha_i = \max\left(\max_{j|F_{LOW}(s_j^-)<\alpha_{i-1}}\left(F_{LOW}\left(s_j\right)\right), \max_{j|F_{UPP}(s_j)<\alpha_{i-1}}\left(F_{UPP}\left(s_j\right)\right)\right) \qquad (3.13)$$

.....

$$\alpha_n = \max\left(\max_{j|F_{LOW}(s_j^-)<\alpha_{n-1}}\left(F_{LOW}\left(s_j\right)\right), \max_{j|F_{UPP}(s_j)<\alpha_{n-1}}\left(F_{UPP}\left(s_j\right)\right)\right) =$$

$$= \min\left(\min_j\left(F_{LOW}\left(s_j\right)\right), \min_j\left(F_{UPP}\left(s_j\right)\right)\right); \qquad \alpha_{n+1} = 0$$

and define:

$$A^i = \left\{s_j \in S \mid F_{UPP}\left(s_j\right) \geq \alpha_i; F_{LOW}\left(s_j^-\right) < \alpha_i\right\}; \quad m(A^i) = \alpha_i - \alpha_{i+1} \qquad (3.14)$$

Consequently:

- the lower/upper probabilities for subsets $T \subseteq S$ are Choquet capacities and Alternate Choquet capacities of order ∞ respectively (or Belief and Plausibility set functions respectively);

- the probabilistic assignment of the equivalent random set can alternatively be derived from the Belief function through the Möbius transform;

- the upper bounds u_j of the singletons (Eq. (3.12)) give the contour function of the equivalent random set *R*.

In (Alvarez 2006) the procedure is extended to p-boxes on infinite spaces, thus deriving equivalent random sets with infinite focal elements given by the α-cuts of the upper/lower CDFs.

Example 3—2. Let us consider $S = \{s_1, s_2, s_3, s_4\}$ and the p-box defined in the first three columns of Table 3—2. The table also displays the bounds for the singletons. The upper bounds give the contour function of the associated non-consonant random set, R. The five extreme points in the two-dimensional space ($P(s_1)$, $P(s_2)$) (case b)) determine the 10 extreme points shown in Figure 3—2a for the projection in the three-dimensional space ($P(s_1)$, $P(s_2)$, $P(s_3)$). Of course in the four-dimensional space ($P(s_1)$, $P(s_2)$, $P(s_3)$, $P(s_4)$) 10 extreme distributions are obtained when $P(s_4)=1- P(s_1) -P(s_2) -P(s_3)$. The extreme points $P_{EXT,1}$ and $P_{EXT,2}$ correspond to the cumulative distribution functions $F_{LOW}(s_j)$ and $F_{UPP}(s_j)$ respectively.

Table 3—3 presents the lower probabilities for all of the subsets in S together with their Möbius transform m, which confirms the rules given by Eqs. (3.13) and (3.14). The resulting focal elements and probabilistic assignments for R are calculated in Table 3—3 and displayed in Figure 3—2b. The random set is completely described by a stack of rectangular boxes: the width of each box identifies its (in this particular case convex, but more generally non convex) focal element along the S axis, and the height of each box is equal to its probabilistic assignment. Hence, the total height of the stack is equal to 1. The focal elements are here ordered in such a way as to obtain a stack enclosed by the cumulative upper and lower bounds of the p-box.

Table 3—2. Bounds and lower/upper CDF in Example 3-2.

s_j	$F_{LOW}(s_j)$	$F_{UPP}(s_j)$	$F_{LOW}(s^-_j)$	$l = Bel(\{s_j\})$	$u = Pla(\{s_j\}) = \mu(s_j)$
s_1	0	0.2	0	0	0.2
s_2	0.1	0.3	0	$\max(0, 0.1 - 0.2) = 0$	$0.3 - 0 = 0.3$
s_3	0.7	1.0	0.1	$\max(0, 0.7 - 0.3) = 0.4$	$1.0 - 0.1 = 0.9$
s_4	1.0	1.0	0.7	$\max(0, 1 - 1) = 0$	$1.0 - 0.7 = 0.3$

Table 3—3. Set functions in Example 3-2.

i	$\chi_i(s_1)$	$\chi_i(s_2)$	$\chi_i(s_3)$	$\chi_i(s_4)$	$\mu_{LOW}(A^i)$	$m^i = m(A^i)$
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	0	0	1	0	0.4	0.4
4	0	0	0	1	0	0
5	1	1	0	0	0.1	0.1
6	0	1	1	0	0.5	0.5-0.4=0.1
7	0	0	1	1	0.7	0.7-0.4=0.3
8	1	0	1	0	0.4	0.4-0.4=0
9	0	1	0	1	0	0
10	1	0	0	1	0	0
11	1	1	1	0	0.7	0.7-1+0.4=0.1
12	0	1	1	1	0.8	0.8-1.2+0.4=0
13	1	0	1	1	0.7	0.7-1.1+0.4=0
14	1	1	0	1	0.1	0.1-0.1+0=0
15	1	1	1	1	1.0	1-2.3+1.7-0.4=0

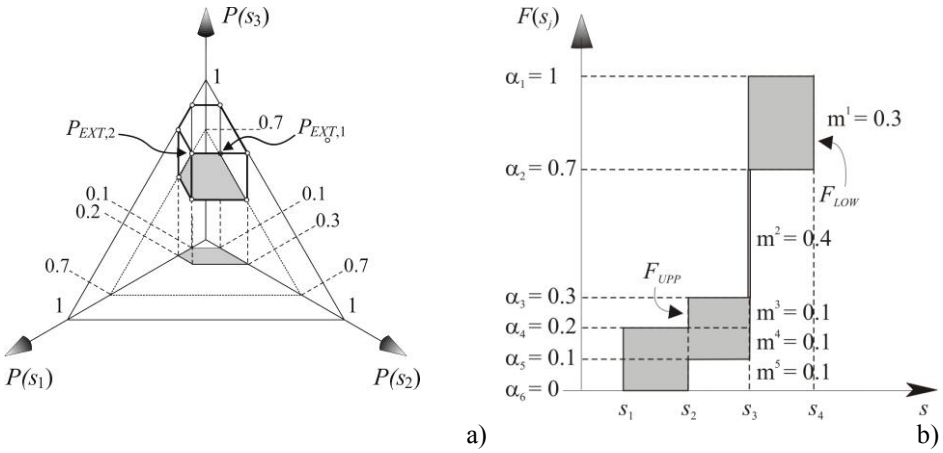


Figure 3—2.: Example 3-2 a) extreme points in the 3-dimensional space ($P(s_1)$, $P(s_2)$, $P(s_3)$); b) equivalent random set R.

Table 3—4. Set functions in Example 3-2.

i	α_i	A^i	$m^i = m(A^i)$
1	1	$\{s_3, s_4\}$	1 – 0.7 = 0.3
2	max(max(0, 0.1, 0.7), max (0.2, 0.3)) = 0.7	$\{s_3\}$	0.7 – 0.3 = 0.4
3	max(max(0, 0.1), max (0.2, 0.3)) = 0.3	$\{s_2, s_3\}$	0.3 – 0.2 = 0.1
4	max(max(0, 0.1, 0.7), max (0.2)) = 0.2	$\{s_1, s_2, s_3\}$	0.2 – 0.1 = 0.1
5	max(max(0, 0.1)) = 0.1	$\{s_1, s_2\}$	0.1 – 0 = 0.1
6	max(max(0)) = 0		

Table 3—5. Dual extreme distributions for Example 3-2.

T	$Pla(T)$	$P_{EXT,UPP}(s)$	T^c	$Bel(T) = 1 - Pla(T^c)$	$P_{EXT,LOW}(s)$
$T_1 = \{s_2\}$	0.3	$P(s_2) = 0.3$	$\{s_1, s_3, s_4\}$	0	$P(s_2) = 0$
$T_2 = \{s_2, s_3\}$	1	$P(s_3) = 0.7$		0.5	$P(s_3) = 0.5$
$T_3 = \{s_2, s_3, s_1\}$	1	$P(s_1) = 0$		0.7	$P(s_1) = 0.2$
$T_4 = S$	1	$P(s_4) = 0$	$\varnothing$	1	$P(s_4) = 0.3$

Now, let us evaluate the expectation bounds for the same function considered in Example 3—1, i.e. the point-valued function $f(s_j)$ mapping onto the set $Y = \{5, 20, 10, 0\}$. The extreme distributions are identified in Table 3—5: events T are the same as events T as in Table 3—1. $Pla(T)$ and $Bel(T)$ are calculated by using m from Table 3—3. The expectation bounds are:
$E_{UPP}[f] = 20\text{x } 0.3 + 10 \text{ x } 0.7 = 13$; $E_{LOW}[f] = 10\text{x } 0.5 + 5\text{x}0.2 + 0 \text{ x } 0.3 = 6$

It is easy to show that the random set R determined by Eqs. (3.13) and (3.14) is not the only random set compatible with the p-box. However it must be considered as the natural extension of the information given by the p-box because the set Ψ^R determined by Eqs. (3.13) and (3.14) includes all probability distributions compatible with the p-box, and also the set Ψ^{R*} of any other random set $R*$ compatible with the p-box.

For example, when the maximum of the contour function defined by the p-box (Eq. (3.12) with $\mu(s_j)=u_j$) is equal to 1, the algorithm (3.5)-(3.6) can be used to derive a consonant random set compatible with the p-box: the focal elements are now the α-cuts of the contour function and the probabilistic assignment is again defined by the increments of α. In other words: the information given by the p-box together with additional information suggesting that the structure of the underlying random set should be consonant determine a consonant random set R' and a corresponding set $\Psi^{R'}$ of probability distributions, and of course $\Psi^{R'} \subseteq \Psi^R$.

Example 3—3. Table 3—6 presents a slightly modified p-box (with respect to the p-box discussed in Example 3-2). The 8 extreme points EXT^F of set Ψ^F and the underlying non-consonant random set are shown in Figure 3—3 a) and b), respectively. The projection of Ψ^F onto the two-dimensional space ($P(s_1)$, $P(s_2)$) now contains 4 extreme points because $F_{LOW}(s_1) = F_{LOW}(s_2)$. Table 3—7 shows that the extreme distributions giving the expectation bounds are the same as in Example 3-2 (compare with Table 3—5): hence $E[f] = [6, 13]$.

Table 3—6. Reachable bounds and lower/upper CDF in Example 3—3.

s_j	$F_{LOW}(s_j)$	$F_{UPP}(s_j)$	$l = Bel(\{s_j\})$	$u = Pla(\{s_j\}) = \mu(s_j)$
s_1	0	0.2	0	0.2
s_2	0	0.3	max(0, 0 – 0.2) = 0	0.3 - 0 = 0.3
s_3	0.7	1.0	max(0, 0.7 – 0.3) = 0.4	1.0 – 0 = 1
s_4	1.0	1.0	max(0, 1 – 1) = 0	1.0 – 0.7 = 0.3

Table 3—7. Dual extreme distributions for Example 3—3.

T	$Pla(T)$	$P_{EXT,UPP}(s)$	T^c	$Bel(T) = 1 - Pla(T^c)$	$P_{EXT,LOW}(s)$
$T_1 = \{s_2\}$	0.3	$P(s_2) = 0.3$	$\{s_1, s_3, s_4\}$	0	$P(s_2) = 0$
$T_2 = \{s_2, s_3\}$	1	$P(s_3) = 0.7$		0.5	$P(s_3) = 0.5$
$T_3 = \{s_2, s_3, s_1\}$	1	$P(s_1) = 0$		0.7	$P(s_1) = 0.2$
$T_4 = S$	1	$P(s_4) = 0$	$\varnothing$	1	$P(s_4) = 0.3$

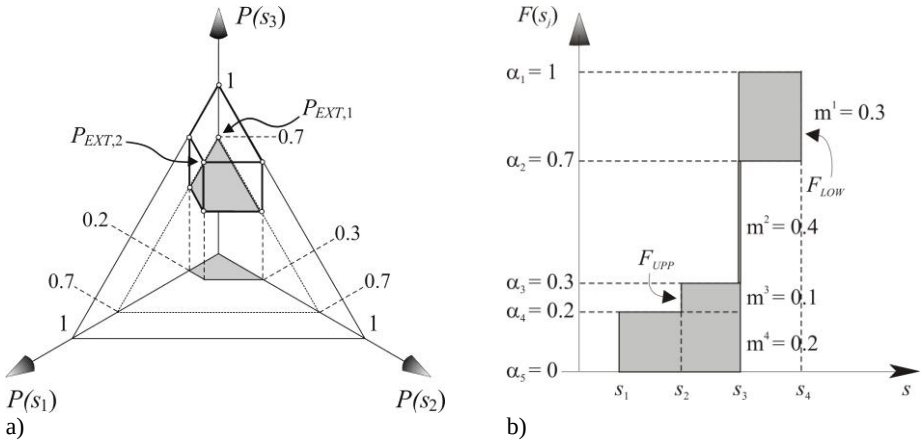


Figure 3—3. Example 3—3: a) extreme points in the 3-dimensional space ($P(s_1)$, $P(s_2)$, $P(s_3)$); b) equivalent random set.

Since now $\mu(s_3) = 1$, the contour function can be assumed to be a possibility distribution that defines a consonant random set *R'*, and corresponding set $EXT^{R'}$ of extreme distributions shown in Figure 3—4. The set $EXT^{R'}$ contains only 5 of the 8 extremes in set EXT^{F}. These 5 extreme points are the vertices of a pyramid with vertex in $P_{EXT,1}$ and quadrangular base on the equilateral triangle $P(s_4)= 1 - P(s_1)- P(s_2) - P(s_3) = 0$. Both $EXT^{R'}$ and EXT^{F} contain the extreme points $P_{EXT,1}$ and $P_{EXT,2}$, which correspond to the cumulative distribution functions $F_{LOW}(s_j)$ and $F_{UPP}(s_j)$ respectively.

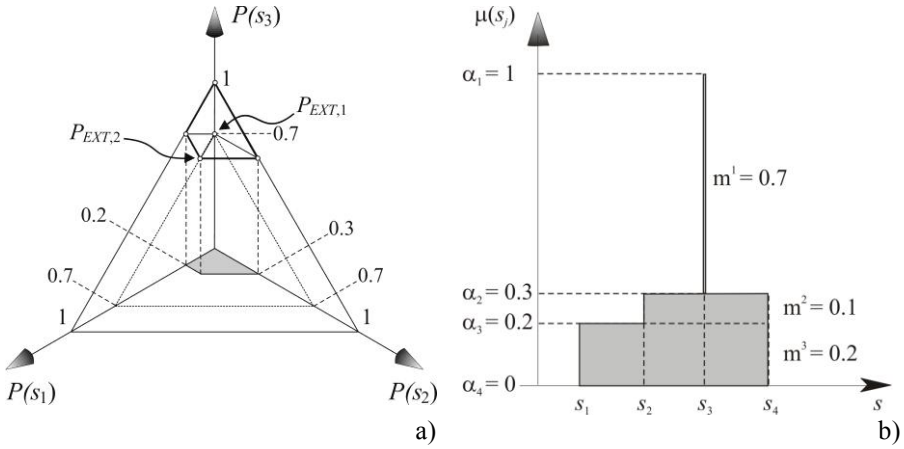


Figure 3—4. Consonant random set in Example 3—3: a) extreme points in the 3-dimensional space

Table 3—8. Dual extreme distributions for the consonant random set in Example 3—3

T	$Pla(T)$	$P_{EXT,UPP}(s)$	T^c	$Bel(T) = 1- Pla(T^c)$	$P_{EXT,LOW}(s)$
$T_1 =\{s_2\}$	0.3	$P(s_2)= 0.3$	$\{s_1, s_3, s_4\}$	0	$P(s_2)= 0$
$T_2 =\{s_2, s_3\}$	1	$P(s_3)= 0.7$		0.7	$P(s_3)= 0.7$
$T_3 =\{s_2, s_3, s_1\}$	1	$P(s_1)= 0$		0.7	$P(s_1)= 0$
$T_4 =S$	1	$P(s_4)= 0$	$\varnothing$	1	$P(s_4)= 0.3$

Hence: $E_{UPP}[f] = 13$; $E_{LOW}[f] = 10\text{x } 0.7 + 0 \text{ x } 0.3 = 7$

The same procedure (to get a consonant random set) cannot be applied to the p-box discussed in Exampe 3-2 because the contour function maximum value is equal to 0.9; however, it is easy to derive a second random set compatible with the p-box in Example 3-2 that has a nearly consonant structure: it is enough to modify the third focal element displayed in Figure 3—4 b) by taking $m^3 = 0.1$ and introducing a fourth focal element $A^4 = =\{s_1, s_2\}$, with $m^4 = 0.1$.

4 Conclusions

Random sets, which combine aleatory and set uncertainty, appear to be a powerful generalization of the classical probability theory. On the other hand, they are particular cases of a more general theory of monotone non-additive measures, Choquet capacities of different orders, coherent upper/lower probabilities, and previsions. More precisely, belief functions are coherent lower probabilities and Choquet capacities of infinite order.

The set of probability distributions compatible with the information given by a random set coincides with the natural extension of the belief/plausibility set functions, and also with the convex hull of a set of extreme distributions.

Therefore, exact bounds of the expectation of any real-valued function can be derived through the Choquet integral or equivalently by a couple of dual extreme distributions. This property seems to be very useful in engineering applications, optimal design and decision making under strong uncertainty conditions.

Fuzzy sets and p-boxes can be considered as particular indexable-type random sets, whose set of focal elements are ordered and uniquely determined by a single real number. In both the cases, simple rules can be given to derive the corresponding family of focal elements, the probabilistic assignment, and the extreme distributions of the associated random set.

Finally, the possibility of considering a hierarchy of random sets ordered by the inclusions of the corresponding sets of probability distributions has been highlighted. For example, conditions have been given to derive an included consonant random set (a fuzzy set) from the contour function of the random set corresponding to a p-box.

References

Alvarez D. A. On the calculation of the bounds of probability of events using infinite random sets. *International Journal of Approximate Reasoning*, 43(3):241-267, 2006.

Chateauneuf A.and J. Y. Jaffray. Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. *Mathematical Social Sciences*, 17:263-283, 1989.

Choquet G. Theory of capacities. *Annales de L'Institut Fourier*, 5:131-295, 1954.

Dubois D.and H. Prade. *Possibility theory: an approach to computerized processing of uncertainty.* Plenum Press, 1988.

Klir G. J. *Uncertainty and Information. Foundations of generalized Information Theory.* Wiley & Sons, Inc., 2005.

Shafer G. *A mathematical theory of evidence.* Princeton University Press, 1976.

Walley P. Towards a Unified Theory of Imprecise probabilities. *International Journal of Approximate Reasoning*, 24:125-148; 125, 2000.

Walley P. *Statistical reasoning with Imprecise Probabilities.* Chapman and Hall, 1991.

Zadeh L. A. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1):3-28, 1978.

On using global optimization method for approximating interval hull solution of parametric linear systems

Iwona Skalna
Department of Applied Computer Science
AGH University of Science and Technology
Krakow, Poland

Andrzej Pownuk
Department of Mathematical Sciences
University of Texas at El Paso
Texas, El Paso, TX 79968

Abstract. Systems of parametric linear interval equations are encountered in many practical applications. Parametric linear interval system is a family of real linear systems. Parametric solution set is a set of all solutions of real systems from the family. In general case the parametric solution set is not an interval vector. Hence, instead of the parametric solution set itself, interval vector containing the solution set (outer interval solution) is calculated. The tightest outer interval solution is called an interval hull solution. To calculate the interval hull solution $2n$ constrained optimization problems are solved using the global optimization method with some accelerating techniques. The monotonicity test is performed using a direct method for solving parametric linear interval systems. Some other techniques like special ordering of subdivided boxes is also used. A bisection and multisection techniques are compared. Various subdivision direction selections rules are tested.

Keywords: parametric linear systems, hull solution, global optimization

1. Introduction

This paper focuses on solving parametric linear systems of structure mechanics with interval parameters. Parametric interval methods allow the engineering practice to account for uncertainty connected either to external factors, such as boundary conditions or applied loads, or to internal factors, such as mechanical or geometric characteristics (Aughenbaugh, 2006; Lallemand, 2000; Muhanna, 2006; Muhanna, 2006; Zalewski et. al., 2006), and to calculate the very sharp bounds on the system response for all possible scenarios in a single analysis (Mullen, 2002).

In general case the parametric solution set is not an interval vector (Neumaier, 1990). Hence, instead of the parametric solution set itself, interval vector containing the parametric solution set (*outer interval solution*) is calculated. The tightest outer interval solution is called an interval hull solution. The problem of computing the hull solution is NP-hard (Rohn and Kreinovich, 1995). However, when the parametric solution is monotone with respect to all interval parameters, interval hull can be calculated by solving at most $2n$ real linear systems.

The problem of calculating hull solution can be written as a problem of solving $2n$ constrained optimization problems. In (Skalna, 2006) an evolutionary optimization methods for approximating (from below) the hull solution has been proposed. One may argue that the underestimation is unknown. However, numerical experiments and the comparison with other methods for solving parametric systems show that the method performs very well.

In this paper global optimization method (GOM for short) with some accelerating techniques is used to calculate the interval hull solution. The monotonicity test is performed using a Direct Method for solving parametric linear interval systems. Some other techniques like special ordering of subdivided boxes are also exploited. A bisection and multisection techniques are compared. Various subdivision direction selections rules are tested.

The paper is organized as follows. The second section contains preliminaries on solving parametric interval linear systems with two disjoint sets of parameters. In the third section, the optimization problem is outlined. This is followed by a description of global optimization algorithm and selected accelerating techniques. Next, some illustrative examples of truss structures and the results of computational experiments are presented. The paper ends with summary conclusions.

2. Preliminaries

Italic faces will be used for real quantities, while bold italic faces will denote their interval counterparts. Let $\mathbb{IR}$ denote a set of real compact intervals $\boldsymbol{x} = [\underline{x}, \overline{x}] = \{x \in \mathbb{R} \mid \underline{x} \leqslant x \leqslant \overline{x}\}$. For two intervals $a, b \in \mathbb{IR}$, $a \geqslant b$, $a \leqslant b$ and $a = b$ will mean that, resp., $\underline{a} \geqslant \overline{b}$, $\overline{a} \geqslant \underline{b}$, and $\underline{a} = \underline{b}$ z $\overline{a} = \overline{b}$. $\mathbb{IR}^n$ will denote interval vectors, $\mathbb{IR}^{n\times n}$ square interval matrices (Neumaier, 1990). The midpoint $\check{x} = \mathrm{m}(\boldsymbol{x}) = (\underline{x} + \overline{x})/2$, the radius $\mathrm{r}(\boldsymbol{x}) = (\overline{x} - \underline{x})/2$, and the width $\mathrm{w}(\boldsymbol{x}) = \overline{x} - \underline{x}$ are applied to interval vectors and matrices componentwise.

Consider linear algebraic system

$$A(p)x(p,\, q) = b(q) \ , \tag{1}$$

with linear dependencies

$$a_{ij}(p) = \alpha_{ij0} + \alpha_{ij}^{\mathrm{T}} \cdot p \,, \qquad b_j(q) = \beta_{j0} + \beta_j^{\mathrm{T}} \cdot q \ , \tag{2}$$

where $\alpha_{ij0},\, \beta_{j0} \in \mathbb{R}$, $\alpha_{ij} = \{\alpha_{ij\nu}\} \in \mathbb{R}^k$, $\beta_j = \{\beta_{j\nu}\} \in \mathbb{R}^l$, $i,\, j = 1, \ldots, n$.

Now assume that some model parameters are unknown. The real vectors p and q are replaced by interval vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ (the real elements are represented by point intervals). This gives a family of the systems

$$A(p)x(p,\, q) = b(q), \quad p \in \boldsymbol{p},\, q \in \boldsymbol{q} \ , \tag{3}$$

which is usually written in a symbolic compact form

$$A(\boldsymbol{p})x(\boldsymbol{p},\, \boldsymbol{q}) = b(\boldsymbol{q}) \ , \tag{4}$$

and is called the *parametric interval linear system*. Parametric (*united*) solution set of the system (4) is defined (Jansson, 1991; Kolev, 2004; Rump, 1994) as

$$S(\boldsymbol{p},\, \boldsymbol{q}) = \{\, x \,|\, \exists p \in \boldsymbol{p},\, \exists q \in \boldsymbol{q},\ \ A(p)x(p,\, q) = b(q)\} \ . \tag{5}$$

If the solution set $\boldsymbol{S} = S(\boldsymbol{p}, \boldsymbol{q})$ is bounded, then its interval hull exists and is defined as

$$\square \boldsymbol{S} = [\inf \boldsymbol{S}, \sup \boldsymbol{S}] = \bigcap \{\boldsymbol{y} \in \mathbb{IR}^n \mid \boldsymbol{S} \subseteq \boldsymbol{y}\} \quad .$$

$\square \boldsymbol{S}$ is called an *interval hull solution*. In order to guarantee that the solution set is bounded, the matrix $A(\boldsymbol{p})$ must be regular, i.e. $A(p)$ must be non-singular for all parameters $p \in \boldsymbol{p}$.

3. Optimization problem

The problem of computing the interval hull solution of the parametric linear system (3) can be written as a problem of solving $2n$ constrained optimization problems

$$\min_{\substack{p \in \boldsymbol{p} \\ q \in \boldsymbol{q}}} x_i(p, q), \ i = 1, \ldots, n \tag{6}$$

and

$$\max_{\substack{p \in \boldsymbol{p} \\ q \in \boldsymbol{q}}} x_i(p, q), \ i = 1, \ldots, n \tag{7}$$

where $x_i(p, q) = \left\{A(p)^{-1} b(q)\right\}_i$ is the i-th coordinate of the solution of the parametric linear system (1), $\boldsymbol{p} \in \mathbb{IR}^k$ and $\boldsymbol{q} \in \mathbb{IR}^l$ are vectors of interval parameters.

Theorem 1. Let $A(\boldsymbol{p})$ be regular, $p \in \mathbb{IR}^k$, and $x^i_{\min}$, $x^i_{\max}$ denote the global solutions of the i-th minimization (6), resp. maximization (7) problems. Then the interval vector

$$\boldsymbol{x} = [x_{\min}, x_{\max}] = \left(\left[x^i_{\min}, x^i_{\max}\right]\right)_{i=1}^{n} = \square S(\boldsymbol{p}, \boldsymbol{q}). \tag{8}$$

The optimization problems (6) and (7) will be solved using a global optimization approach. As a result of the minimization (maximization) problem approximation of the solution set hull, possibly the solution hull itself, will be gained.

4. Global optimization

Global optimization refers to finding the extreme value of a given nonconvex function in a certain feasible region. Solving global optimization problems has made great gain from the interest in the interface between computer science and operations research.

It is assumed in what follows that the inclusion functions have the isotonicity property; i.e., $\boldsymbol{x} \subseteq \boldsymbol{y}$ implies $F(\boldsymbol{x}) \subseteq F(\boldsymbol{y})$ and that for all the inclusion functions holds

$$\mathrm{w}(F(\boldsymbol{x}^i)) \longrightarrow 0 \text{ as } \mathrm{w}(\boldsymbol{x}^i) \longrightarrow 0. \tag{9}$$

4.1. Algorithm

Consider $x(\boldsymbol{p},\ \boldsymbol{q})$, and define $\boldsymbol{r} \in \mathbb{IR}^{k+l}$ with $\boldsymbol{r}_i = \boldsymbol{p}_i$ for $i = 1, \ldots, k$, $\boldsymbol{r}_i = \boldsymbol{q}_i$ for $i = k+1, \ldots, k+l$. Now $x(\boldsymbol{p},\ \boldsymbol{q})$ can be written in shorter form as $x(\boldsymbol{r})$ keeping in mind that x has two vector arguments. Inclusion function is calculated using the Direct Method (Skalna, 2007). It can be easily shown that the method preserves isotonicity property.

The model algorithm is as follows:

Step 0 Set $\boldsymbol{y} = \boldsymbol{r}$ and $f = \min x(\boldsymbol{y})$. Initialize the list $L = \{(f,\ \boldsymbol{y})\}$ and the cutoff level $z = \max x(\boldsymbol{y})$.
Step 1 Choose a coordinate direction using one of the rules: $\nu \in \{1, 2, ..., k+l\}$.
Step 2 Bisect (multisect) $\boldsymbol{y}$ in direction ν: $\boldsymbol{y}_1 \cup \boldsymbol{y}_2 \left(\bigcup_{i=1}^{s} \boldsymbol{y}_i,\ \mathrm{int}(\boldsymbol{y}_i) \cap \mathrm{int}(\boldsymbol{y}_j) = \emptyset,\ i \neq j \right)$, int denotes the interior.
Step 3 Calculate $x(\boldsymbol{y}_1)$, $x(\boldsymbol{y}_2)$, and set $f_i = \min x(\boldsymbol{y}_i)$ for $i = 1,\ 2$ and $z = \min\{z,\ \max x(\boldsymbol{y}_1),\ \max x(\boldsymbol{y}_2)\}$.
Step 4 Remove $(f,\ \boldsymbol{y})$ from the list L.
Step 5 Cutoff test: discard the pair $(f_i,\ \boldsymbol{y}_i)$ if $f_i > z$ (where $i \in \{1, 2\}$).
Step 6 Monotonicity test: discard or reduce any remaining pair $(f_i,\ \boldsymbol{y}_i)$ if $0 \notin x_j(\boldsymbol{y}_i)$ for any $j \in \{1, 2, ..., n\}$ and $i = 1, 2$.
Step 7 Add any remaining pairs to the list L. If the list becomes empty, then **STOP**.
Step 8 Denote the pair with the smallest first element by $(f^*, \boldsymbol{y}^*)$.
Step 9 If the width of $x(\boldsymbol{y}^*)$ is less than ε, then print $x(\boldsymbol{y}^*)$ and $\boldsymbol{y}^*$, **STOP**.
Step 10 Go to **Step 1**.

4.2. Midpoint test

The midpoint test is used to reduce the number of intervals in the list L. The pair $(\tilde{f},\ \tilde{\boldsymbol{y}})$ which satisfies $\tilde{f} < f$ for all pairs $(f,\ \boldsymbol{y})$ of the list L is chosen out of L. Then, $\check{f} = \sup F(c)$ is computed, with $c = \mathrm{mid}(\tilde{\boldsymbol{y}})$. Now, all pairs $(f',\ \boldsymbol{y}')$ satisfying $\check{f} < f'$ can be discarded from the list L. Also, a new pair $(f'',\ \boldsymbol{y}'')$ must only be entered in the list L if $\tilde{f} \geqslant f''$ is satisfied.

4.3. Monotonicity test

The monotonicity test is used to figure out whether the function f is strictly monotone in a whole subbox $\boldsymbol{y} \subseteq \boldsymbol{x}$ Then, $\boldsymbol{y}$ cannot contain a global minimizer in its interior. Therefore, if f satisfies

$$\frac{\partial f}{\partial x_i}(\boldsymbol{y}) < 0 \quad \vee \quad \frac{\partial f}{\partial x_i}(\boldsymbol{y}) < 0 \tag{10}$$

then the subbox $\boldsymbol{y}$ can be reduced to one of its edges.

Monotonicity test is pefromed using the Method for Checking the Monotonicity (MCM for short) proposed in (Skalna, 2007). The MCM method is based on a Direct Method (Skalna, 2007) for solving parametric linear systems. Let $f = x(p, q)$. Briefly speaking, the approximations of

$\frac{\partial x}{\partial p_m}(\boldsymbol{p},\, \boldsymbol{q})$, $\frac{\partial x}{\partial q_r}(\boldsymbol{p},\, \boldsymbol{q})$ are obtained by solving the following $k+l$ parametric linear systems

$$A(\boldsymbol{p})\frac{\partial x}{\partial p_m} = b'^m(\boldsymbol{x}^*),\, m = 1,\, \ldots,\, k; \quad A(\boldsymbol{p})\frac{\partial x}{\partial q_r} = b''^r, r = 1,\, \ldots,\, l \ , \tag{11}$$

where $b_j'^m(x^*) = -\alpha_{ijm}x_j^*$, $b_j''^r = \beta_{jr}$, $j = 1,\, \ldots,\, n$, $x^* \in \boldsymbol{x}^*$. Detailed description of the MCM method can be found in (Skalna, 2007).

4.4. Subdivision direction selection

Following Ratz and Csendes the interval subdivision direction selection rules has the following merit function:

$$k := \min\{j \mid j \in \{1,\, \ldots,\, n\} \text{ and } D(j) = \max_i D(i)\} \tag{12}$$

where $D(i)$ is determined by a given rule.

Rule A. The first rule to be applied was the interval-width-oriented rule (Hansen, 1980), it can also be aaplied to non-differentiable function. This rule chooses the coordinate direction with

$$D(i) = w(\boldsymbol{y}). \tag{13}$$

and was justified by the idea that if the original interval is subdivided in a uniform way then the width of the actual subintervals goes to zero most rapidly.

Rule B. Define the indicator

$$p(f_k, f) = \frac{f_k - \underline{f}}{\overline{f} - \underline{f}} \tag{14}$$

that gives which interval is to be selected for subdivision. Here f_k is the approximation of the global minimum value in the iteration k (Casado, 200)

$$f_k = \min\{f_l \mid (f_l,\, y_l) \in L\}\,. \tag{15}$$

Rule B selects the coordinate direction for which (12) holds with

$$D(i) = p(f_k,\, f_i). \tag{16}$$

Rule C. Hansen described another rule (initiated by G.W. Walster) (Hansen, 1980). Rule C selects the coordinate direction for which (12) holds with

$$D(i) = \mathrm{w}\left(F_i'(\boldsymbol{y})\mathrm{w}(\boldsymbol{y})\right). \tag{17}$$

4.5. Multisection

Global optimization is based on successive subdivision of the set of feasible solutions. The main idea of multisection technique is to subdivide the problem (in a single step) into many (> 2) smaller problems in contrast to traditional bisection, where to new subintervals are always produced.

5. Examples

To check the performance of the method some illustrative examples of structural mechanical systems are provided. The results of the Global Optimization Method are compared with the results of the Evolutionary Optimization Method (EOM for short) (Skalna, 2006) .

Example 1. (*21-bar plane truss structure*)
For the plane truss structure shown in Fig. 1 the displacements of the nodes are computed. The truss is subjected to downward forces $P_1 = P_2 = P_3 = 30$[kN] as depicted in the figure; Young's modulus $Y = 7.0 \times 10^{10}$[Pa], cross-section area $C = 0.003$[m^2], and length $L = 2$[m]. Assume the stiffness of all bars is uncertain by $\pm 5\%$. This gives 21 interval parameters.

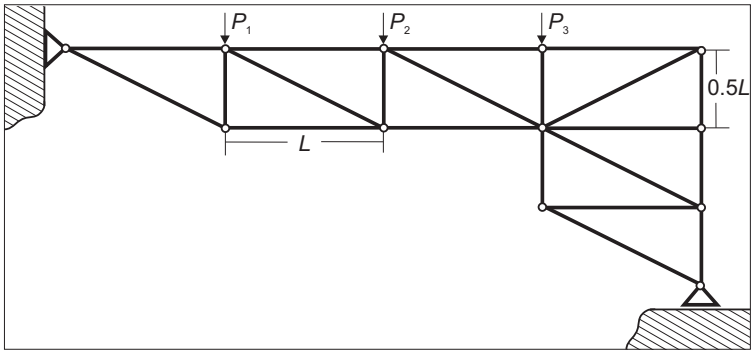


Figure 1. Example 1: 21 planar truss structure

The results produced by the GOM and the EOM methods (Table I) coincide.

Table I. Example 1: results of the GOM and the EOM methods

n	$\underline{x}$ [$\times 10^{-5}$]	$\overline{x}$ [$\times 10^{-5}$]	n	$\underline{x}$ [$\times 10^{-5}$]	$\overline{x}$ [$\times 10^{-5}$]
1	-32.53	-29.27	12	3.90	4.67
2	-1.61	-1.45	13	-16.23	-14.67
3	-26.45	-23.93	14	3.18	3.87
4	-2.41	-2.17	15	-3.63	-2.96
5	-15.78	-14.27	16	3.18	3.87
6	-1.69	-1.37	17	-0.05	0.05
7	-4.08	-3.37	18	2.35	3.02
8	-0.96	-0.57	19	-0.46	-0.40
9	0.36	0.50	20	0.85	1.47
10	3.90	4.67	21	-2.78	-2.09
11	-26.45	-23.93			

Example 2. (Baltimore bridge built in 1870)
Consider the plane truss structure shown in Figure 2 subjected to downward forces of $P_1 = 80[kN]$ at node 11, $P_2 = 120[kN]$ at node 12 and P_1 at node 15; Young's modulus Y$= 2.1 \times 10^{11}$ [Pa], cross-section area C $= 0.004[m^2]$, and length $L = 1$[m]. Assume that the stiffness of 16 bars is uncertain by $\pm 5\%$. This gives 16 interval parameters.

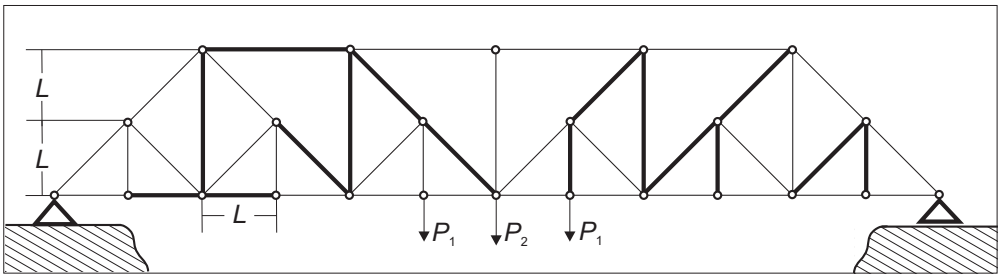


Figure 2. Example 2: Baltimore bridge (built in 1870)

Once again the results of the GOM and the EOM methods coincide. The average relative error produced by both methods equals 2.51%. Maximal relative error equals 34%. For 23 coordinates relative error equals 1%, for another 10 coordinates equals 2%.

6. Conclusions

The problem of solving parametric linear systems has been considered in Section 2. In Section 3 the global optimization method GOM for approximating the solution set hull of parametric linear systems has been described. Computations performed in Section 5 show that the GOM is a powerful tool for solving such systems. The results of the GOM method have been compared with the results of the evolutionary optimization method EOM. Both methods produced the same result which proves that both approaches are powerful tools for solving parametric linear systems. It turns out from the experiments that the monotonicity test and the cutoff test significantly speeds up the convergence of the GOM method, while different rules of subdivision direction selection have no impact on the convergence. Multisection technique is not useful for the problem of solving parametric linear systems since the computation of implicitly given inclusion function is very expensive.

References

J. Aughenbaugh and C. Paredis. Why are intervals and imprecisions important in engineering design? In R.L.Muhannah, editor, *Proceedings of the NSF Workshop on Reliable Engineering Computing (REC)*, pages 319–340, Svannah, Georgia USA, Feb. 22–24 2006.

L.G. Casado, I. Garcia, and T. Csendes. A new multisection technique in interval methods for global optimization. *Computing*, 65:263–269, 2000.

D.E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, Reading, Massachusetts, 1989.

E. Hansen. Global optimization using interval analysis - the multidimensional case. *Numer. Math.*, 43:247–270, 1980.

C. Jansson. Interval linear systems with symmetric matrices, skew-symmetric matrices and dependencies in the right hand side. *Computing*, 46(3):265–274, 1991.

L.V. Kolev. A method for outer interval solution of linear parametric systems. *Reliable Computing*, 10:227–239, 2004.

L.V. Kolev. Solving linear systems whose elements are non-linear functions of intervals. *Numerical Algorithms*, 37:213–224, 2004.

B. Lallemand, G. Plessis, T. Tison, and P. Level. Modal Behaviour of Structures Defined by Imprecise Geometric Parameters #125. In *Proc. SPIE Vol. 4062, Proceedings of IMAC-XVIII: A Conference on Structural Dynamics., p.1422*, volume 4062 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pages 1422–+, January 2000.

Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs.* Springer-Verlag, Berlin, Germany, 1996.

R.L. Muhanna and A. Erdolen. Geometric uncertainty in truss systems: an interval approach. In R.L.Muhannah, editor, *Proceedings of the NSF Workshop on Reliable Engineering Computing (REC): Modeling Errors and Uncertainty in Engineering Computations*, pages 239–247, Svannah, Georgia USA, Feb. 22–24 2006.

R.L. Muhanna, V. Kreinovich., P. Solin, J. Cheesa, R. Araiza, and G. Xiang. Interval finite element method: New directions. In R.L. Muhannah, editor, *Proceedings of the NSF Workshop on Reliable Engineering Computing (REC)*, pages 229–244, Svannah, Georgia USA, Feb. 22–24 2006.

R. Mullen and R.L. Muhanna. Efficient interval methods for finite element solutions. In *HPCS '02: Proceedings of the 16th Annual International Symposium on High Performance Computing Systems and Applications*, page 161, Washington, DC, USA, 2002. IEEE Computer Society.

A. Neumaier. *Interval Methods for Systems of Equations.* Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, UK, 1990.

A. Osyczka. *Evolutionary Algorithm for Single and Multicriteria Design Optimization.* Studies in Fuzziness and Soft Computing Physica-Verlag, Heidelberg New York, 2002.

E. Popova, R. Iankov, and Z. Bonev. Bounding the response of mechanical structures with uncertainties in all the parameters. In R.L.Mullen R.L.Muhannah, editor, *Proceedings of the NSF Workshop on Reliable Engineering Computing (REC)*, pages 245–265, Svannah, Georgia USA, Feb. 22-24 2006.

J. Rohn. A method for handling dependent data in interval linear systems. Technical report 911, Academy of Sciences of the Czech Republic, Czech Republic, 2004.

S.M. Rump. Verification methods for dense and sparse systems of equations. In Jürgen Herzberger, editor, *Topics in validated computations: proceedings of IMACS-GAMM International Workshop on Validated Computation, Oldenburg, Germany, 30 August–3 September 1993*, volume 5 of *Studies in Computational Mathematics*, pages 63–135, Amsterdam, The Netherlands, 1994. Elsevier.

I. Skalna. A method for outer interval solution of parametrized systems of linear interval equations. *Reliable Computing*, 12(2):107–120, 2006.

I. Skalna. On checking the monotonicity of parametric interval solution of linear structural systems. In R. Wyrzykowski, editor, *Proceedings of Seventh International Conference on Parallel Processing and Applied Mathematics*, Czestochowa, Poland, 18-22 September 2007.

B. Zalewski and R.L. Muhanna R. Mullen. Bounding the response of mechanical structures with uncertainties in all the parameters. In R.L.Muhannah, editor, *Proceedings of the NSF Workshop on Reliable Engineering Computing (REC)*, pages 439–456, Svannah, Georgia USA, Feb. 22–24 2006.

V. Kreinovich. Optimal solution of interval linear systems is intractable (NP-hard). *Interval Computations*, 1:6–14, 1993.

J. Rohn and V. Kreinovich. Computing exact componentwise bounds on solutions of linear systems with interval data is NP-hard. *SIAM Journal on Matrix Analysis and Applications (SIMAX)*, 16:415–420, 1995.

I. Skalna. Evolutionary optimization method for approximating the solution set hull of parametric linear systems. *LNCS: Numerical Method and Applications* 4310:361-368, 2007.

E. Hansen. Global optimization using interval analysis. *Marcel Dekker*, New York.

Propagating Uncertainties in Modeling Nonlinear Dynamic Systems

Joshua A. Enszer,[1] Youdong Lin,[1*] Scott Ferson,[2] George F. Corliss[3], Mark A. Stadtherr[1†]

[1] *Department of Chemical and Biomolecular Engineering*
University of Notre Dame, Notre Dame, IN 46556, USA

[2] *Applied Biomathematics*
Setauket, NY 11733, USA

[3] *Department of Electrical and Computer Engineering*
Marquette University, Milwaukee, WI 53201, USA

Abstract. Engineering analysis and design problems, either static or dynamic, frequently involve uncertain parameters and inputs. Propagating these uncertainties through a complex model to determine their effect on system states and outputs can be a challenging problem, especially for dynamic models. In this work, we demonstrate the use of Taylor model methods for propagating uncertainties through nonlinear ODE models. We concentrate on uncertainties whose distribution is not known precisely, but can be represented by a probability box (p-box), and show how to use p-boxes in the context of Taylor models. This allows us to obtain p-box representations of the uncertainties in the state variables and outputs of a nonlinear ODE model. Examples are used to demonstrate the potential of this approach for studying the effect of uncertainties with imprecise probability distributions.

Keywords: Ordinary differential equations, Interval analysis, Probability bounds analysis

1. Introduction

Ordinary differential equations (ODEs) are the basis for many mathematical models in the sciences and engineering. Often a system of ODEs is formulated as an initial value problem (IVP), in which the model is integrated through time beginning with specified initial values of the state variables. Especially in cases where no analytical solution exists, the numerical integration of these systems is necessary to obtain the trajectories of ODE systems.

Of particular interest here is the verified, or mathematically guaranteed, solution of systems of ODEs, especially such systems that involve uncertainty in initial conditions or model parameters. Traditional numerical methods, such as Euler's method or the Runge-Kutta schemes, only approx-

* Current address: LINDO Systems, Inc., 1415 North Dayton St., Chicago, IL USA 60622

† Author to whom all correspondence should be addressed. Phone: (574) 631-9318; Fax: (574) 631-8366; E-mail: markst@nd.edu

imate the trajectory of an ODE system since truncation errors from both function approximation and machine arithmetic are present. Furthermore, if any particular parameter or initial state is uncertain, normal use of these methods will not fully guarantee that all possible trajectories are found.

In response to the need for guaranteed results, both with and without uncertainty, interval methods have been proposed. Computations with intervals, as opposed to real numbers, can provide mathematical and computational guarantees, and further, intervals are a logical way to deal with uncertainty in any parameters or initial conditions. An excellent review of interval methods for IVPs has been given by Nedialkov et al. (1999), and more recent work has been reviewed by Neher et al. (2007). Much work has been done for the case in which the initial values are given by intervals, and there are several available software packages that deal with this case, including AWA (Lohner, 1992), VNODE (Nedialkov et al., 2001), and COSY VI (Berz and Makino, 1998). These methods can also deal with interval-valued parameters, by treating them as additional state variables with derivative of zero. In the work described here, we will use a new validated IVP solver for parametric ODEs (Lin and Stadtherr, 2007b) called VSPODE (Validating Solver for Parametric ODEs), which is used to produce guaranteed bounds on the solutions of nonlinear dynamic systems with interval-valued initial states and parameters. VSPODE treats interval-valued parameters directly without the need to increase the number of state variables. Both COSY VI and VSPODE use Taylor models (Makino and Berz, 1996; Makino and Berz, 1999; Makino and Berz, 2003), but in different ways, to deal with the uncertain quantities (parameters and initial values).

Other methods exist to solve ODE systems with uncertainty, but they do not provide a mathematical guarantee that all possible trajectories are enclosed. These methods are often a combination of a Monte Carlo process with a standard integration scheme, such as Runge-Kutta. While such methods cannot guarantee that all solutions are enclosed, they can propagate uncertainty in ways that standard interval methods cannot. Interval methods do not use knowledge about the distribution of uncertainty in a variable or parameter, while such knowledge can be put to use in Monte Carlo methods to discern the most probable trajectory of an ODE system.

When the concepts of intervals and probability distributions are combined, the result has been called a probability distribution variable (PDV), and theorems and computations with this data type have been presented by Li and Hyman (2004). Intervals and "probability boxes" (p-boxes) can be viewed as specific enclosures of the more broadly defined PDV. If there are only upper and lower bounds on the uncertainties but no known probability distribution, then this can be represented by an interval. If there is some knowledge of the probability distribution, but it is uncertain, then this can be represented by a probability box (p-box). For computations with p-boxes, we use here the risk analysis software RAMAS Risk Calc (Ferson, 2002).

In this paper, we demonstrate the use of Taylor model methods for propagating uncertainties through nonlinear ODE models. We concentrate here on uncertainties represented by p-boxes, and show how to use p-boxes in the context of Taylor models. This allows us to obtain p-box representations of the uncertainties in the state variables. Examples are used to demonstrate the potential of this approach for studying the effect of uncertainties with imprecise probability distributions.

This paper is divided as follows. The next section will provide a general statement of the problem to be addressed. Section 3 gives background on interval analysis, Taylor models, and p-boxes. In

Section 4 we outline the specific method that is used, and in Section 5 we show the results of applying this method to some specific examples.

2. Problem Statement

Here we introduce the notation used in the paper as we describe the problems to be solved. We will consider the verified solution of the parametric autonomous IVP

$$y'(t) = f(y, \theta), \qquad y(t_0) = y_0 \in Y_0, \qquad \theta \in \Theta, \tag{1}$$

where $t \in [t_0, t_m]$ for some $t_m > t_0$. Here y is the n-dimensional vector of state variables with initial value y_0, and θ is a p-dimensional vector of *time-invariant* parameters. The vectors Y_0 and Θ are intervals that enclose uncertainties in the initial states and parameters, respectively. Additional information about these uncertainties is available in the form of p-boxes, as described in the next section, for at least one component of Y_0 or Θ. We assume that f maps the variable and parameter space back to the variable space and that f is $(k-1)$ times continuously differentiable with respect to y and $(q+1)$ times continuously differentiable with respect to θ. Here k is the order of the truncation error in the interval Taylor series (ITS) method used by VSPODE, and q is the order of the Taylor model in VSPODE used to represent dependence on parameters and initial values. We also assume that f can be represented by a finite number of standard functions. Our goal is to obtain a guaranteed enclosure of the state variables y at time t_m and a probability distribution, in the form of a p-box, for the values of y within the enclosure.

3. Background

3.1. Interval Analysis

A real interval X is the set of real numbers between and inclusive of its lower bound (denoted $\underline{X}$) and upper bound (denoted $\overline{X}$). The width of an interval, denoted $w(X)$, is equal to $\overline{X} - \underline{X}$, while the midpoint $m(X)$ is $(\overline{X} + \underline{X})/2$. A real interval vector $X = (X_1, X_2, ..., X_n)^{\mathrm{T}}$ has n real interval components and can be interpreted as an n-dimensional rectangle or box. Interval matrices are similarly defined.

Basic arithmetic operations are defined on intervals according to

$$X \text{ op } Y = \{x \text{ op } y \mid x \in X, y \in Y\}, \;\; \text{op} \in \{+, -, \times, \div\}. \tag{2}$$

Division in the case of Y containing zero is only allowed in extensions of interval arithmetic (Hansen and Walster, 2004). Addition and multiplication are commutative and associative but only subdistributive. Interval versions of the elementary functions can also be defined.

For a real function $f(x)$, the interval extension $F(X)$ encloses the range of $f(x)$ for $x \in X$. When $f(x)$ can be written as a series of arithmetic operations and elementary functions, the natural interval extension is obtained by substituting the given interval X into $f(x)$ and evaluating using

interval arithmetic. Computing the interval extension in this way often results in overestimation of the function range due to the "dependency" problem. While a variable may take on any value within its interval, it must take on the *same* value each time it occurs in an expression. However, this type of dependency is not recognized when the natural interval extension is computed. In effect, when the natural interval extension is used, the range computed for the function is the range that would occur if each instance of a particular variable was allowed to take on a different value in its interval range.

Another source of overestimation that may arise in the use of interval methods is the "wrapping" effect. This occurs when an interval is used to enclose (wrap) a set of results that is not an interval. If this overestimation is propagated from step to step in an integration procedure for ODEs, it can quickly lead to the loss of a meaningful enclosure.

Several good introductions to interval analysis, as well as interval arithmetic and other aspects of computing with intervals, are available (Hansen and Walster, 2004; Jaulin et al., 2001; Kearfott, 1996; Neumaier, 1990). Implementations of interval arithmetic and elementary functions are also readily available, and recent compilers from Sun Microsystems directly support interval arithmetic and an interval data type.

3.2. Taylor Models

Makino and Berz (1996) have described a remainder differential algebra (RDA) approach for bounding function ranges and control of the dependency problem of interval arithmetic (Makino and Berz, 1999). In this method, a function is represented using a model consisting of a Taylor polynomial and an interval remainder bound. Such a model is called a Taylor model.

One way of forming a Taylor model of a function is by using the Taylor theorem. Consider a real function $f(x)$ that is $(q+1)$ times partially differentiable on X and let $x_0 \in X$. The Taylor theorem states that for each $x \in X$, there exists a real ζ with $0 < \zeta < 1$ such that

$$f(x) = p_f(x - x_0) + r_f(x - x_0, \zeta), \tag{3}$$

where p_f is a q-th order polynomial (truncated Taylor series) in $(x - x_0)$, and r_f is a remainder, which can be quantitatively bounded over $0 < \zeta < 1$ and $x \in X$ using interval arithmetic or other methods to obtain an interval remainder bound R_f. A q-th order Taylor model $T_f = p_f + R_f$ for $f(x)$ over X then consists of the polynomial p_f and the interval remainder bound R_f and is denoted by $T_f = (p_f, R_f)$. Note that $f \in T_f$ for $x \in X$, and thus T_f encloses the range of f over X.

In practice, it is more useful to compute Taylor models of functions by performing Taylor model operations. Arithmetic operations with Taylor models can be done using the RDA operations described by Makino and Berz (1996; 1999; 2003), which include addition, multiplication, reciprocal, and intrinsic functions. Using these, it is possible to start with simple functions such as the constant function $f(x) = k$, for which $T_f = (k, [0, 0])$, and the identity function $f(x_i) = x_i$, for which $T_f = (x_{i0} + (x_i - x_{i0}), [0, 0])$, and then to compute Taylor models for very complicated functions. Therefore, it is possible to compute a Taylor model for any function representable in a computer environment by simple operator overloading through RDA operations. It has been shown that, compared to other rigorous bounding methods, the Taylor model often yields sharper bounds for modest to

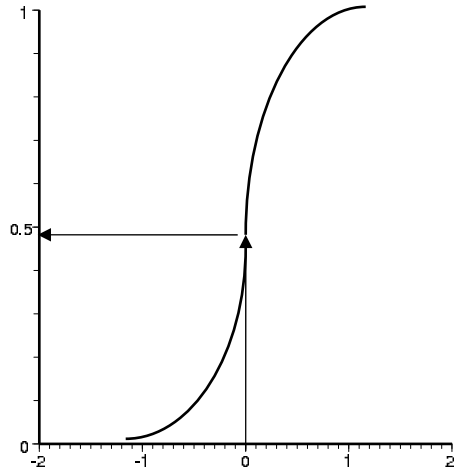


Figure 1. A cumulative probability density function is a one-to-one function where the value of a quantity x is on the abscissa and the corresponding cumulative probability is given by the ordinate. Here, $P(x \leq 0) = 0.48$.

complicated functional dependencies (Makino and Berz, 1996; Makino and Berz, 1999; Neumaier, 2003). A discussion of the uses and limitations of Taylor models has been given by Neumaier (2003).

3.3. P-Boxes

For some quantity (variable or parameter) x, the cumulative distribution function (CDF) $F(z)$ gives the probability that $x \leq z$. A sample CDF is shown in Figure 1. Here, for example, the probability that $x \leq 0$ is 0.48. Probability boxes, or p-boxes, are similar, but provide an interval of cumulative probability values represented by a pair of CDFs. A sample p-box is shown in Figure 2. This indicates, for example, that the probability of $x \leq 0$ is $[0.40, 0.59]$. The slightly stepped appearance of the p-box curves, here and below, is due to the discretized representation of a p-box used by Risk Calc. This representation is used in the implementation of p-box arithmetic.

A probability box, then, is essentially a hybrid of an interval and a probability distribution. As an interval bounds a range of real numbers, a p-box bounds a range of probability distributions. Also, as a probability distribution gives a real-valued probability for the value of a real parameter, a p-box provides an interval-valued probability for the value of a real parameter (Ferson, 2002). Read in another way, for a real-valued probability, a p-box provides the interval of values corresponding to that probability. Formally, a p-box is a pair of functions (F, G) such that the true probability distribution H of a number satisfies $F \geq H \geq G$. The function F is called the left bound of the p-box, while the function G is the right bound, which should be apparent from the definition. For a given real-valued probability, the left bound provides the lower bound of the parameter, and the right bound corresponds to the upper bound. For a given value of a parameter, the left bound corresponds to the upper bound of its probability and the right bound to its lower bound.

When the probabilities of parameters are independent, computations with p-boxes are analogous to those with intervals. They are defined beginning with arithmetic and standard functions, again using Eq. (2). In this case, p-boxes encounter the same dependency issues that intervals do. However,

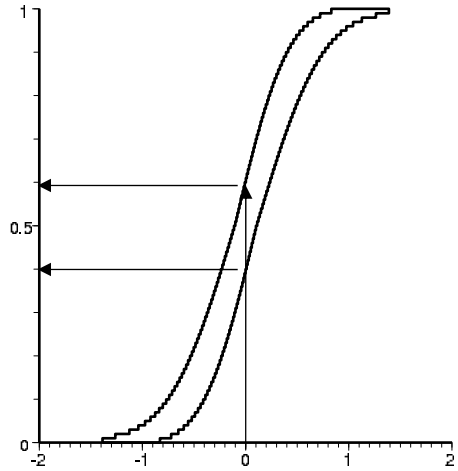


Figure 2. A p-box encloses the cumulative probability distribution function for a quantity x. Now the corresponding cumulative probability is an interval. Here, $P(x \leq 0) = [0.4, 0.59]$.

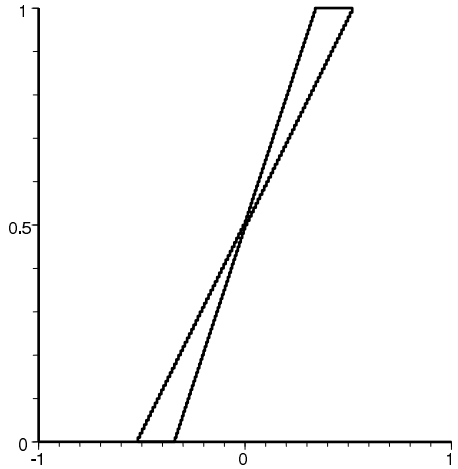


Figure 3. Sample p-box with bounds obtained from a uniform distribution with a mean of 0 and standard deviation in $[0.2, 0.3]$.

p-box arithmetic can vary depending on assumptions about the parameters; much more detail is provided by Ferson et. al. (2004).

There are three types of p-boxes employed in the examples used in Section 5. The first, as illustrated in Figure 3, is a p-box with bounds obtained from a uniform distribution with a fixed mean and an interval-valued standard deviation. Note that such a p-box encloses both uniform (straight line) and nonuniform CDFs. The second is a p-box with bounds obtained from a normal distribution, again with fixed mean and an interval-valued standard deviation, as shown in Figure 4. The third type of p-box used is shown in Figure 5 and corresponds to the case of an uncertain distribution with a specified minimum and maximum and fixed mean and standard deviation. This is referred to as the mmms distribution.

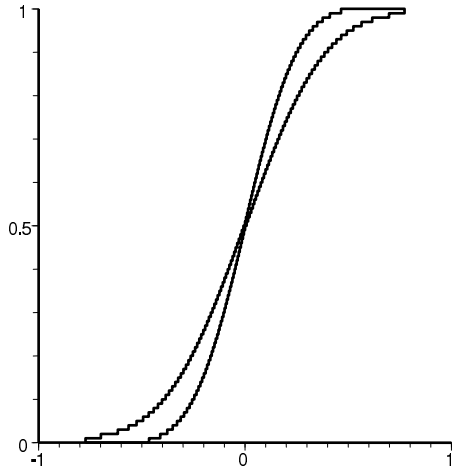


Figure 4. Sample p-box with bounds obtained from a normal distribution with a mean of 0 and standard deviation in $[0.2, 0.3]$.

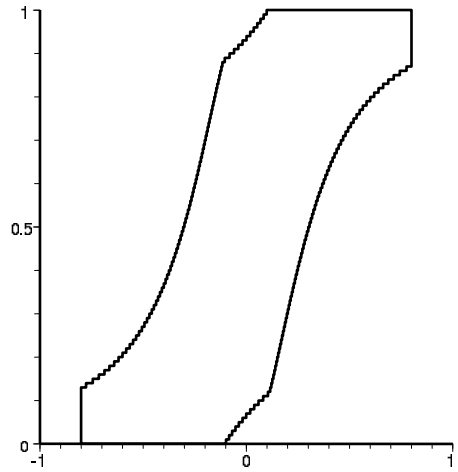


Figure 5. Sample p-box for mmms distribution. This p-box has a minimum of -0.8, maximum of 0.8, mean of 0, and standard deviation of 0.3.

4. Solution Procedure

In this section, we outline the method used for solving the problem described in Section 2. This involves using the VSPODE program for the verified integration of the IVP given by Eq. (1) and the RAMAS Risk Calc software for calculations with p-boxes. More detailed descriptions of the VSPODE program and RAMAS Risk Calc software can be found elsewhere (Lin and Stadtherr, 2007b; Ferson, 2002).

As a first step, VSPODE is used to integrate the IVP. This provides a guaranteed enclosure of the state variables at each time step in the integration and a Taylor model representation of the

final state $y_m = y(t_m)$. This Taylor model, $T_{y_m} = T_{y_m}(y_0, \theta)$, is a polynomial in y_0 and θ with an interval remainder bound, and is valid for all $y_0 \in Y_0$ and $\theta \in \Theta$.

Then, in the second step, information about the distribution of y_0 and θ values is substituted into the Taylor model T_{y_m}, and the distribution of final state values is computed using Risk Calc. Each initial state and parameter is given by either an interval (no distribution known) or by a p-box, and Risk Calc can do the necessary arithmetic with either. To reduce the occurrence of overestimation in these calculations, Risk Calc can employ a subinterval reconstitution (SIR) procedure. These methods are described in more detail by Ferson and Hajagos (2004).

5. Examples

The following examples illustrate the solution procedure when applied to a model from population ecology and to three reactor modeling problems from chemical engineering. On all of the example problems, the order of the interval Taylor series used in VSPODE was $k = 17$, and the order of the Taylor models used was $q = 5$. Unless specified otherwise, a constant step size of 0.2 was used in VSPODE, though this step size may be automatically reduced during the integration procedure if needed.

5.1. Lotka-Volterra Model

One of the most basic population ecology models is the Lotka-Volterra model of a predator-prey system. The model equations, with parameter uncertainties, can be written as

$$\frac{dx_1}{dt} = \theta_1 x_1(1 - x_2), \qquad x_1(0) = 1.2, \qquad \theta_1 \in [2.99, 3.01] \tag{4}$$

$$\frac{dx_2}{dt} = \theta_2 x_2(x_1 - 1), \qquad x_2(0) = 1.1, \qquad \theta_2 \in [0.99, 1.01]. \tag{5}$$

This example has served as a test problem for comparing interval-based ODE solvers (Lin and Stadtherr, 2007b), in which uncertainty is represented as an interval. Figure 6 reproduces the interval trajectories computed by VSPODE for $t = [0, 10]$. The Taylor model describing the solution at $t = 10$ can be combined with probability bound analysis when more specific information regarding the distribution of uncertainty is known. Figures 7 and 8 show the p-box solutions if both parameters are described by a p-box with bounds obtained from a uniform distribution with standard deviation of [0.0050, 0.0057] and mean at the interval midpoint. If the parameters were simply intervals (no distribution known), then only an interval enclosure of the states would be obtained, and these upper and lower bounds would match the upper and lower bounds of the p-boxes shown in Figures 7 and 8. If no probability bounds analysis was done, we could only say that the probability that $x_1 \leq 1.14$ is in $[0, 1]$. However, using p-boxes it can be seen from Figure 7 that the probability that $x_1 \leq 1.14$ is in the interval $[0.05, 0.5]$. We can run subinterval reconstitution to make the p-boxes tighter. Figures 9 and 10 show that the areas of the p-boxes can be drastically reduced with this technique. Now, the probability that $x_1 \leq 1.14$ is shown to be about [0.13, 0.36]. The choice of a uniform

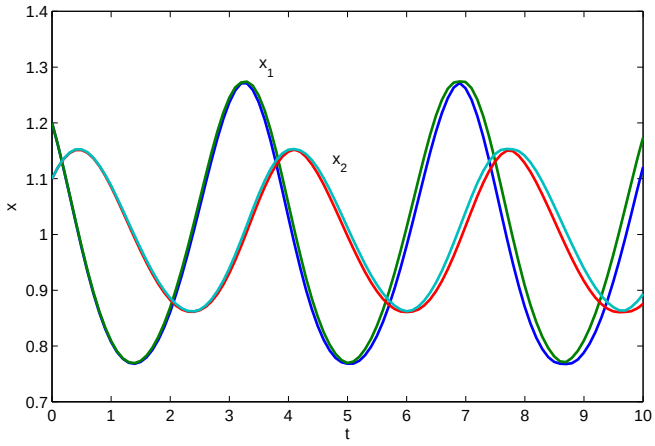


Figure 6. The interval bounds on the state trajectories of Lotka-Volterra model as computed by VSPODE.

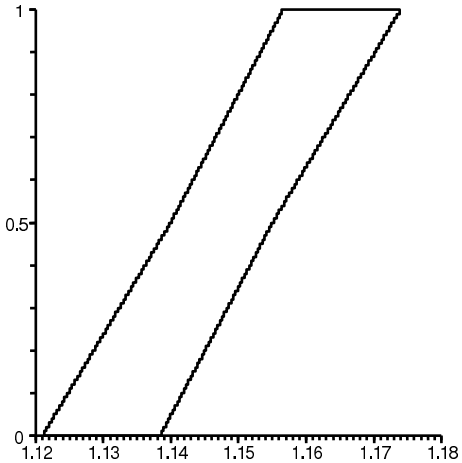


Figure 7. The p-box enclosure of x_1 at time $t = 10$ in the Lotka-Volterra model as computed by Risk Calc.

distribution for the p-box bounds in this problem was an arbitrary one. Other types of p-boxes could also be used.

5.2. Microbial Growth Model with Monod Kinetics

The system of equations for a simple bioreactor model (Lin and Stadtherr, 2007a) is

$$\frac{dX}{dt} = (\mu - \alpha D)X \tag{6}$$

$$\frac{dS}{dt} = D(S_f - S) - k\mu X, \tag{7}$$

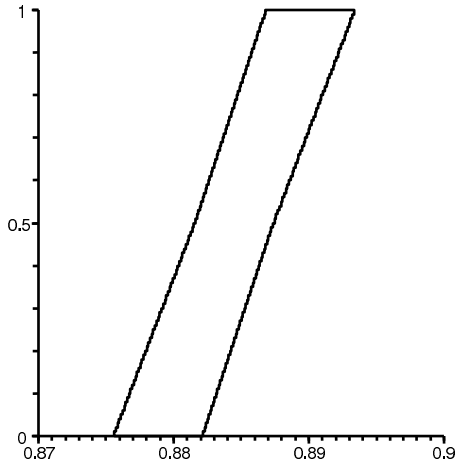


Figure 8. The p-box enclosure of x_2 at time $t = 10$ in the Lotka-Volterra model as computed by Risk Calc.

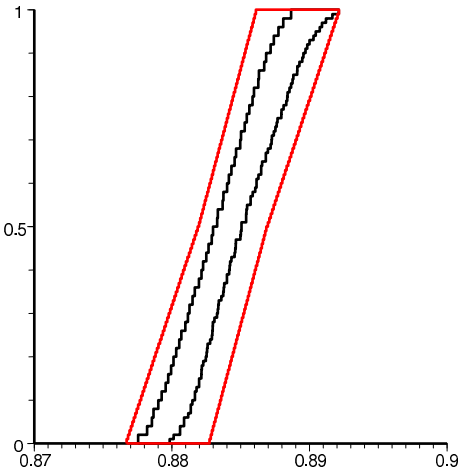


Figure 9. The inner curves show the p-box enclosure of x_1 at time $t = 10$ in the Lotka-Volterra model as computed by Risk Calc, now using the subinterval reconstitution technique. The outer box corresponds to the solution shown in Figure 7 for comparison.

where X represents the concentration of cells in the system, and S represents the concentration of substrate. The parameters α, D, S_f, and k represent the heterogeneity parameter, the dilution rate of substrate, the feed concentration of substrate, and the yield coefficient, respectively. The growth rate of cells, μ, is dependent on the concentration of substrate, S. This term may take a variety of forms. For a simple initial example, we consider Monod kinetics (Bastin and Douchain, 1990; Bequette, 2003), where

$$\mu = \frac{\mu_{max} S}{K_S + S}. \tag{8}$$

In the above expression, μ_{max} is the maximum growth rate, and K_S is the saturation parameter.

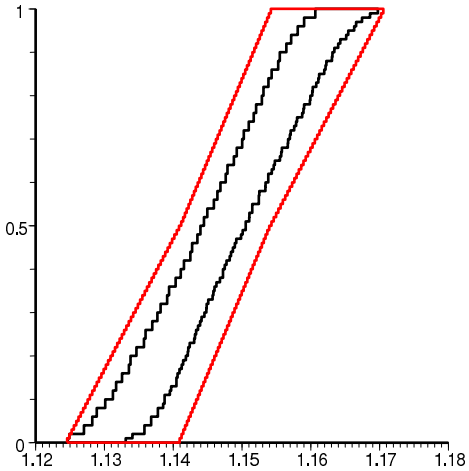


Figure 10. The inner curves show the p-box enclosure of x_2 at time $t = 10$ in the Lotka-Volterra model as computed by Risk Calc, now using the subinterval reconstitution technique. The outer box corresponds to the solution shown in Figure 8 for comparison.

We explore a subset of the uncertain conditions and parameters used by Lin and Stadtherr (2007a): $X_0 \in [0.794, 0.864]$ g/L, $\mu_{max} \in [1.15, 1.25]$ day^{-1}, and $K_S \in [6.8, 7.2]$ g/L. We assume an mmms distribution for these parameters, with the mean being the midpoint of the interval, and the standard deviation being one tenth of the width of the interval (these are arbitrary choices). Other initial conditions and parameters are expressed as real numbers: $S_0 = 0.8$ g/L, $\alpha = 0.5$, $D = 0.36$ day^{-1}, $S_f = 5.7$ g/L, and $k = 10.53$ g substrate/ g cells. We employ VSPODE to integrate the equation from $t = 0$ to $t = 10$ days. The biomass trajectory produced by VSPODE is shown in Figure 11. The resulting Taylor model that describes X and S at $t = 10$ is used with Risk Calc, and the p-box calculations give bounds on the probability distributions for the state variables as shown in Figures 12 and 13. This shows, for example, that the probability that the biomass of cells is less than or equal to 0.85 g is in the interval $[0.9, 1.0]$.

5.3. Microbial Growth Model with Haldane Kinetics

The same bioreactor model described by Eqs. (6)-(7) can be solved using the slightly more complicated Haldane kinetics (Bastin and Douchain, 1990; Lin and Stadtherr, 2007a), also called substrate inhibition kinetics (Bequette, 2003). Here we replace the growth rate equation previously given as Eq. (8) with

$$\mu = \frac{\mu_{max} S}{K_S + S + K_I S^2}. \tag{9}$$

The new parameter, K_I, is called the inhibition parameter. Following Lin and Stadtherr (2007a), we will treat this new parameter as uncertain, with its value lying in the interval $[0.0025, 0.01]$ and its uncertainty described again using the mmms p-box as discussed above.

Integrating this equation the same way as before, we obtain the transient biomass trajectory shown in Figure 14 and a Taylor model describing the variables at time $t = 10$. The p-box enclosures

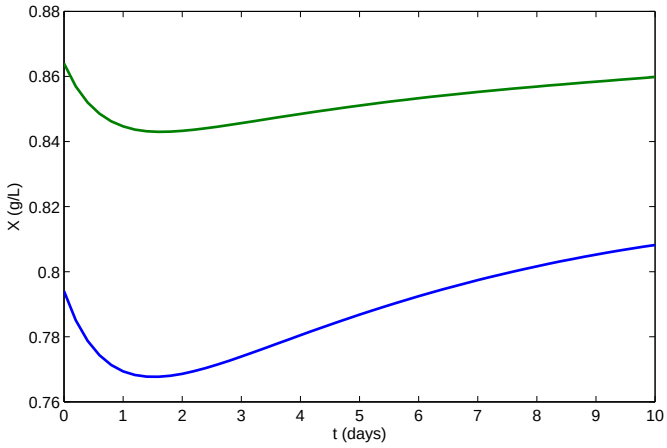


Figure 11. Interval bounds on trajectory of cell biomass X under Monod kinetics as computed by VSPODE.

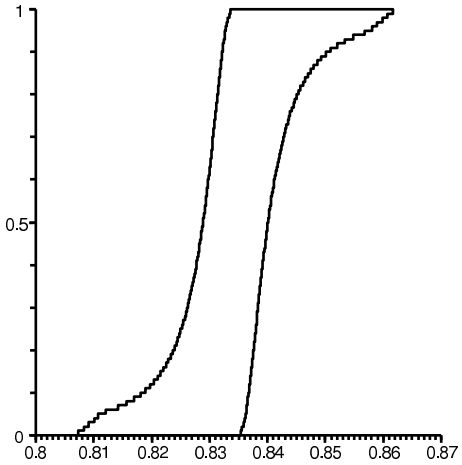


Figure 12. P-box enclosure of cell biomass X at $t = 10$ for Monod kinetics as computed by Risk Calc.

computed by Risk Calc for the state variables are shown in Figures 15 and 16. These enclosures are larger than the enclosures determined in the previous example, which is expected because there is an additional uncertain parameter. Now the probability that the biomass of cells is less than or equal to 0.85 g is in the interval $[0.86, 1.0]$.

5.4. Three-State Bioreactor Model

A second bioreactor model explored by Lin and Stadtherr (2007a) is a three-state biochemical reactor. Here, we model the growth of cells x_1 that consume substrate x_2, but which also form a product x_3. The model is

$$\frac{dx_1}{dt} = (\mu - D)x_1 \tag{10}$$

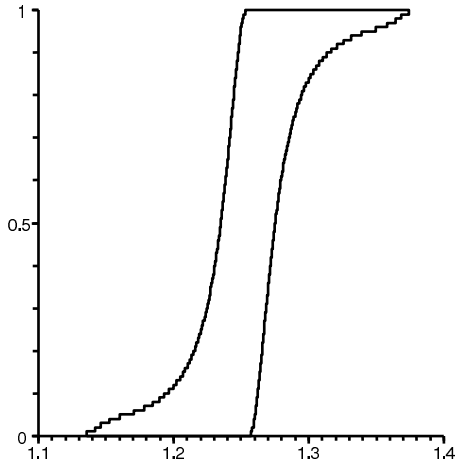


Figure 13. P-box enclosure of substrate S at $t = 10$ for Monod kinetics as computed by Risk Calc.

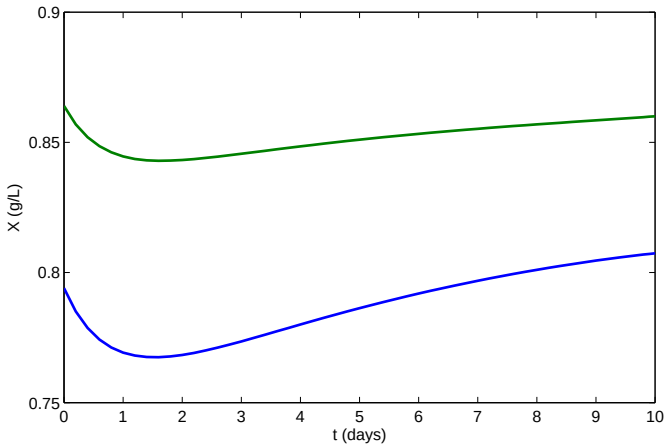


Figure 14. Interval bounds on trajectory of biomass X under Haldane kinetics as computed by VSPODE.

$$\frac{dx_2}{dt} = D(x_{2f} - x_2) - \frac{\mu x_1}{Y} \tag{11}$$

$$\frac{dx_3}{dt} = -Dx_3 + (\alpha\mu + \beta)x_1, \tag{12}$$

with the growth rate as a function of both substrate and product concentrations,

$$\mu = \frac{\mu_{max}\left[1 - (x_3/x_{3m})\right] x_2}{k_s + x_2}. \tag{13}$$

In the above equations, the initial concentration of cells is unknown but within [6.4549, 6.5676], and is represented by a p-box with bounds obtained from a uniform distribution with standard deviation of [0.028170, 0.032533]. Two parameters are uncertain; the maximum growth rate $\mu_{max} \in$

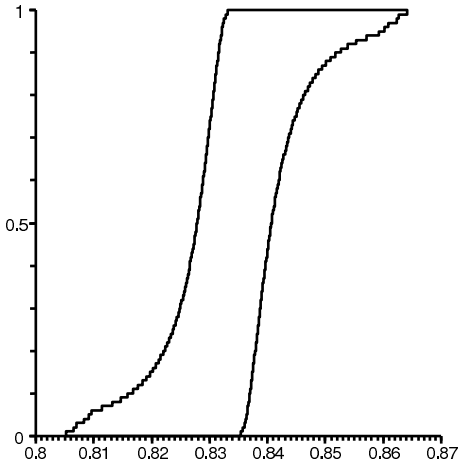


Figure 15. P-box enclosure of cell biomass X at $t = 10$ for Haldane kinetics as computed by Risk Calc.

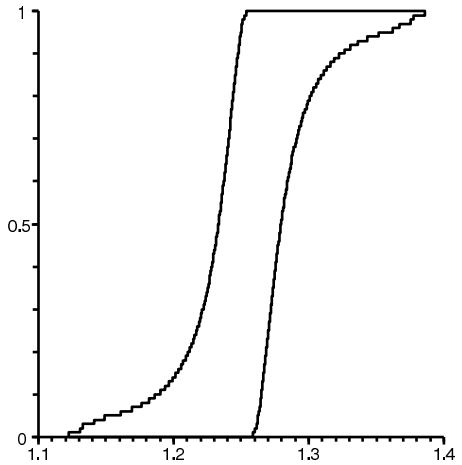


Figure 16. P-box enclosure of substrate S at $t = 10$ for Haldane kinetics as computed by Risk Calc.

$[0.46, 0.47]$ is represented by a p-box with bounds obtained from a normal distribution with standard deviation of $[0.0282, 0.0325]$, and the saturation parameter $k_s \in [1.03, 1.1]$ is represented by a p-box corresponding to the mmms distribution with standard deviation equal to 0.007. All p-box means are at the interval midpoint. All other initial conditions and parameters are known exactly: $x_{20} = 5$ g/L, $x_{30} = 15$ g/L, $Y = 0.4$ g/g, $\beta = 0.2$ hour^{-1}, $D = 0.202$ hour^{-1}, $\alpha = 2.2$ g/g, $x_{3m} = 50$ g/L, and $x_{2f} = 20$ g/L.

VSPODE provides the biomass trajectory shown in Figure 17 and the Taylor model used in Risk Calc to create Figures 18, 19, and 20, which give the probability distribution for the state variables as p-boxes at time $t = 10$. One purpose for this example is to show the ability to have uncertain conditions under a variety of probability distributions. Such an ability is essential in complicated biological models where a variety of distributions is likely.

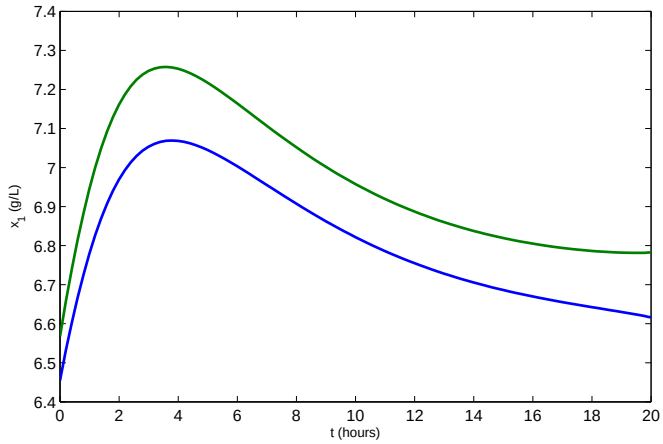


Figure 17. Interval bounds on trajectory of biomass x_1 in three-state reactor as computed by VSPODE.

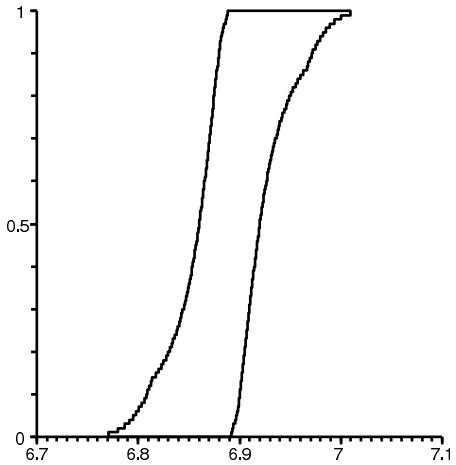


Figure 18. P-box enclosure of biomass x_1 at $t = 10$ in three-state reactor as computed by Risk Calc.

6. Concluding Remarks

The verified ODE solver VSPODE (Lin and Stadtherr, 2007b) provides a powerful tool for bounding the solutions of parametric nonlinear ODEs. Because it provides output in the form of Taylor models, VSPODE is also useful in situations in which uncertainties in parameters and initial states are represented by p-boxes. In this case, the Taylor models from VSPODE can be combined with the p-box uncertainties in initial states and parameters using RAMAS Risk Calc, resulting in a propagation of these uncertainties into the final values of the state variables. In this way, probability distributions (p-boxes) for the final state values can be obtained, as demonstrated in several example problems.

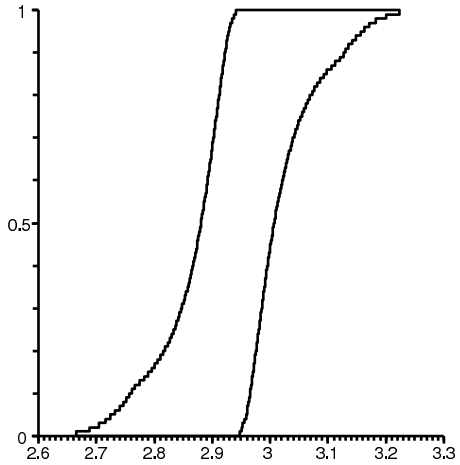


Figure 19. P-box enclosure of substrate mass x_2 at $t = 10$ in three-state reactor as computed by Risk Calc.

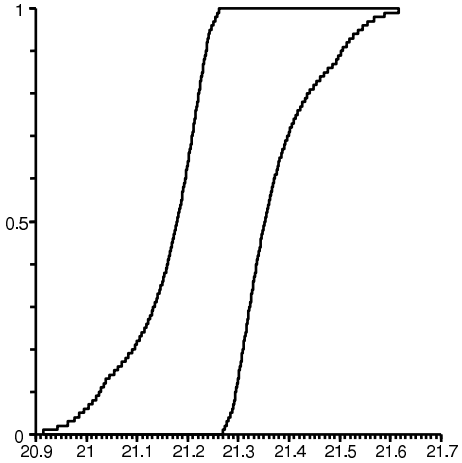


Figure 20. P-box enclosure of product mass x_3 at $t = 10$ in three-state reactor as computed by Risk Calc.

Acknowledgements

This work was supported in part by the U. S. Department of Energy under Grant DE-FG02-05CH11294 and the U. S. National Oceanic and Atmospheric Administration under Grant NA050AR4601153.

References

Bastin, G. and D. Douchain. *On-line Estimation and Adaptive Control of Bioreactors.* Elsevier, 1990.

Bequette, B. W. *Process Control: Modeling, Design, and Simulation.* Prentice-Hall, 2003.

Berz, M. and K. Makino. Verified integration of ODEs and flows using differential algebraic methods on high-order Taylor models. *Reliable Computing* 4: 361–369, 1998.

Ferson, S. *RAMAS Risk Calc 4.0: Risk Assessment with Uncertain Numbers*. Lewis Press, 2002

Ferson, S. and J. G. Hajagos. Arithmetic with uncertain numbers: Rigorous and (often) best possible answers. *Reliability Engineering and System Safety*, 85: 135–152, 2004.

Ferson, S., Nelson, R. B., Hajagos, J., Berleant, D. J., Zhang, J., Tucker, W. T., Ginzburg, L. R. and W. L. Oberkampf. Dependence in probabilistic modeling, Dempster-Shafer theory, and probability bounds analysis. Technical report, Sandia National Laboratories, 2004.

Hansen, E. R. and G. W. Walster, G. W. *Global Optimization Using Interval Analysis*. Marcel Dekker, New York, 2004.

Jaulin, L., Kieffer, M., Didrit, O. and É Walter. *Applied Interval Analysis*. Springer-Verlag, London, 2001.

Kearfott, R. B. *Rigorous Global Search: Continuous Problems*. Kluwer, Dordrecht, The Netherlands, 1996.

Li, W. and J. M. Hyman. Computer arithmetic for probability distribution variables. *Reliability Engineering and System Safety*, 85: 191–209, 2004.

Lin, Y. and M. A. Stadtherr. Guaranteed state and parameter estimation for nonlinear continuous-time systems with bounded-error measurements. *Industrial & Engineering Chemistry Research*, 46: 7198–7207, 2007a.

Lin, Y. and M. A. Stadtherr. Validated solutions of initial value problems for parametric ODEs. Applied Numerical Mathematics, 57: 1145–1162, 2007b.

Lohner, R. J. Computations of guaranteed enclosures for the solutions of ordinary initial and boundary value problems. In: Cash, J., Gladwell, I. (Eds.), *Computational Ordinary Differential Equations*. Clarendon Press, Oxford, UK, pp. 425–435, 1992.

Makino, K. and M. Berz. Remainder differential algebras and their applications. In: Berz, M., Bishof, C., Corliss, G., Griewank, A. (Eds.), *Computational Differentiation: Techniques, Applications, and Tools*. SIAM, Philadelphia, pp. 63–74, 1996.

Makino, K. and M. Berz. Efficient control of the dependency problem based on Taylor model methods. *Reliable Computing*, 5: 3–12, 1999.

Makino, K. and M. Berz. Taylor models and other validated functional inclusion methods. *International Journal of Pure and Applied Mathematics*, 4: 379–456, 2003.

Nedialkov, N. S., Jackson, K. R., and G. F. Corliss. Validated solutions of initial value problems for ordinary differential equations. *Applied Mathematics and Computation*, 105: 21–68, 1999.

Nedialkov, N. S., Jackson, K. R., and J. D. Pryce. An effective high-order interval method for validating existence and uniqueness of the solution of an IVP for an ODE. *Reliable Computing*, 7: 449–465, 2001.

Neher, M., Jackson, K. R., and N. S. Nedialkov. On Taylor model based integration of ODEs. *SIAM Journal on Numerical Analysis*, 45: 236–262, 2007.

Neumaier, A. *Interval Methods for Systems of Equations*. Cambridge University Press, Cambridge, UK, 1990.

Neumaier, A. Taylor forms – Use and limits. *Reliable Computing*, 9: 43–79, 2003.

A Comparison of Information Management Using Imprecise Probabilities and Precise Bayesian Updating of Reliability Estimates

J. M. Aughenbaugh[1] and J. W. Herrmann[2]
[1]Applied Research Laboratories
University of Texas at Austin
email: jason@arlut.utexas.edu

[2]Department of Mechanical Engineering
University of Maryland College Park
email: jwh2@umd.edu

Abstract: Assessing the reliability of components and systems is an important problem in engineering design. Estimates of the reliability of a design can play a significant role in final design decisions. Data for making these estimates is often scarce during the design process. However, designers also frequently have the option to acquire more information by expending resources. Designers thus face the dual questions of how to update their estimates and whether it is valuable to collect additional information. Various statistical updating methods exist and can be used in reliability estimation, including precise Bayesian updating and methods based on imprecise probabilities. In this paper, we explore the management of information collection using these two approaches. These ideas combine elements from sensitivity analysis, value of information calculations, and uncertainty measures. Rather than dealing with abstract measures of total of uncertainty for a particular distribution or set of distributions, we explore the relationships between variance-based sensitivity analysis of the prior and posterior estimates of the mean and variance over all possible results of a particular test. The goal is to gain insight into the many tradeoffs that occur when comparing different information collection actions, especially when the exact outcome of the action is uncertain. These tradeoffs are explored using the example reliability modeling of a simple parallel-series system with three components.

Keywords: reliability assessment, imprecise probabilities, information management

1. Introduction

Modeling uncertainty is an increasingly important activity in engineering applications. As engineers progress from deterministic approaches to nondeterministic approaches, the question of how to model the uncertainty in the nondeterministic approaches must be answered. Researchers have proposed various methods for modeling and propagating uncertainty. A great deal of literature has been devoted to developing and applying individual methods, and others are devoted to philosophical debates of the appropriateness of different measures. More recently, there is a growing interest in practical comparisons

of the methods (Oberkampf et al., 2001; Nikolaidis et al., 2004; Soundappan et al., 2004; Aughenbaugh and Paredis, 2006; Hall, 2006; Kokkolaras et al., 2006; Aughenbaugh and Herrmann, 2007).

Most of this work has focused on what we will call the *problem solution* stage of engineering decisions. In this stage, the engineer makes a decision about a product's design. For example, the engineer determines the dimensions of a component or chooses a particular architecture for the system. This stage follows and is distinct from the *problem formulation* phase, which includes tasks such as identifying design alternatives, eliciting stakeholder preferences, and modeling the state of the world. One step of this formulation phase is *information management*. In this step, the engineers make decisions about what information to collect, how to collect it, and how to process it. For example, the design of experiments falls into this stage. The focus of this paper is on how to model uncertainty in order to best support information management decisions. In particular, we consider the problem of system reliability assessment, as discussed in Section 2.

Managing information collection activities during engineering design is clearly related to the concept of the value of information. This concept has been used in the context of engineering design for incorporating the cost of decision making (Gupta, 1992), for model selection (Radhakrishnan and McAdams, 2005), and for catalog design (Bradley and Agogino, 1994). Some recent work to improve engineering design processes has considered this problem from a frequentist updating perspective (Ling et al., 2006) and developed a method for managing multiple sources of information in engineering design using imprecise probabilities (Schlosser and Paredis, 2007). This work used the principles of *information economics* (Howard, 1966; Matheson, 1968; Marschak, 1974; Lawrence, 1999). At a basic level, these principles state that one should explicitly consider the *expected net value of information*, which is the expected benefit of the information minus the cost of acquiring that information.

In this paper, we focus on a Bayesian updating problem and consider problem-independent measures of the value of the information. Ideally, the value of information would be measured in terms of the value of the final product and the cost of the design process. However, such value and cost models are not always available, particularly early in the design process when the design is only very vaguely defined (Malak et al., 2007). It is thus important to have some statistical metrics for guiding information collection that are independent of the value context of the problem, while still adequately accounting for the information state and the known structure of the system being designed.

Section 2 presents the example problem. Section 3 reviews the precise and imprecise Bayesian statistical models. Section 4 presents the uncertainty metrics that we will use. Experimental results are presented and discussed in Section 5. Section 6 gives a general discussion, and Section 7 concludes the paper.

2. Example problem description

We consider the case where a designer is considering additional testing of some key components in a system in order to get better estimates of the system reliability. From a reliability perspective, the example system can be modeled as a parallel-series system, as shown in Figure 1. We assume that the failures of each component are independent events. The designer has some prior information about the reliability of each component and thus can create an estimate of the system reliability. However, the engineer hopes that additional testing will refine the estimate.

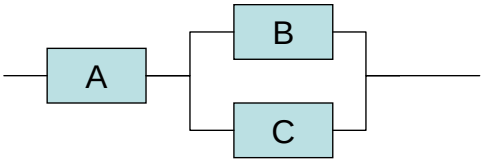


Figure 1. Reliability Block Diagram for the System.

It will be convenient to frame things in terms of failure probability instead of reliability. If component *i* has a reliability of R_i, then the corresponding failure probability of component *i* is $P_i = 1 - R_i$. Let θ be the failure probability of the system, which is the parameter of interest. Ideally, in order to determine this, the designer would have enough data to make a precise assessment of P_A, P_B, and P_C, such as "$P_A = 0.05$." However, there are practical reasons why the designer cannot or is unwilling to make a precise assessment despite holding some initial beliefs about the failure probability (Malak et al., 2007).

We consider the case in which only component testing is feasible. No system-level tests are possible. The designer will use the results of additional testing to update his beliefs about the components' failure probabilities, which will yield an updated estimate of the system failure probability. In particular, the engineer is interested in knowing which component should be tested. Since testing requires resources, it is not reasonable to test every component a large number of times. In this work we consider the number of tests as a surrogate for the cost of testing, which is reasonable if all tests require roughly the same amount of resources.

The failure probabilities P_A, P_B, and P_C are modeled as independent random variables with a beta distribution. If $P_A \sim beta(\alpha_A, \beta_A)$, then

$$E[P_A] = \alpha_A / (\alpha_A + \beta_A);$$

$$V\left[P_A\right] = \frac{\alpha_A \beta_A}{\left(\alpha_A + \beta_A\right)^2 \left(\alpha_A + \beta_A + 1\right)};$$

$$E\left[P_A^2\right] = V\left[P_A\right] + \left(E\left[P_A\right]\right)^2 = \left(\frac{\alpha_A}{\alpha_A + \beta_A}\right)\left(\frac{\alpha_A + 1}{\alpha_A + \beta_A + 1}\right).$$

The mathematical model for the reliability of the system shown in Figure 1 follows.

$$R_{sys} = R_A\left(1-(1-R_B)(1-R_C)\right) \tag{1}$$

$$\theta = P_{sys} = P_A + P_B P_C - P_A P_B P_C \tag{2}$$

$$E[\theta] = E[P_A] + E[P_B]E[P_C] - E[P_A]E[P_B]E[P_C] \tag{3}$$

$$\begin{aligned} E[\theta^2] = {} & E[P_A{}^2] + 2E[P_A]E[P_B]E[P_C] - 2E[P_A{}^2]E[P_B]E[P_C] + \\ & + E[P_B{}^2]E[P_C{}^2] - 2E[P_A]E[P_B{}^2]E[P_C{}^2] + E[P_A{}^2]E[P_B{}^2]E[P_C{}^2] \end{aligned} \tag{4}$$

$$V[\theta] = E[\theta^2] - (E[\theta])^2 \tag{5}$$

3. Formalisms for modeling uncertainty

In this paper we will compare two different approaches for updating reliability assessments: the precise Bayesian and the imprecise beta model, which is useful for both the robust Bayesian approach and the imprecise probability approach. Some introductory material is provided here. For a more complete discussion, see the cited references and the discussion in Aughenbaugh and Herrmann (2007).

3.1. PRECISE BAYESIAN

The Bayesian approach (e.g. Box and Tiao, 1973; Berger, 1985) provides a way to combine existing knowledge and new knowledge into a single estimate by using Bayes's Theorem. One of the requirements of Bayesian analysis is a prior distribution that will be updated. The objective selection of a prior distribution in the absence of relevant prior information is a topic of extensive research and debate. The approaches proposed include the use of non-informative priors (Jeffreys, 1961; Zellner, 1977; Berger and Bernardo, 1992), maximum-entropy priors (Fougere, 1990), and data-dependent empirical Bayes approaches (Maritz and Lewin, 1989). Still, whether a single prior distribution, especially the uniform prior, can reflect all of the uncertainty is an open question to some observers. However, in many engineering problems, designers do have some prior information, such as data and experience from similar systems, and the Bayesian approach allows this information to be included in the analysis.

To support analytical solutions, the form of the prior is often restricted to conjugate distributions with respect to the measurement model, in which case the posterior distribution that results from the update has the same type as the prior. For the problem considered in this paper, in which the number of failures in a given number of tests is a binomial random variable, it is convenient to model the prior distribution of a component's failure probability as a beta distribution with parameters α and β. If the prior distribution is $Beta(\alpha_0, \beta_0)$ and one observes m failures out of n trials, then the posterior distribution is $Beta(\alpha_0 + m, \beta_0 + n - m)$. Consequently, the update involves simple addition and subtraction, an enormous improvement in efficiency over the general case.

3.2. ROBUST BAYESIAN AND IMPRECISE PROBABILITIES APPROACHES

Two alternatives to a precise Bayesian approach are the robust Bayesian and imprecise probabilities approaches. The two approaches are mathematically similar, but differ in motivation.

The robust Bayesian approach addresses the problem of lack of confidence in the prior (Berger, 1984; Berger, 1985; Berger, 1993; Insua and Ruggeri, 2000). The core idea of the approach is to perform a "what-if" analysis by changing the prior. The analyst considers several reasonable prior distributions and performs the update on each to get a set of posterior distributions. After additional data is collected, each candidate prior is updated, resulting in a set of posterior distributions. This set of posterior distributions yields a range of point estimates and a set of credible intervals. If there is no significant change in the conclusion across this set of posteriors, then the conclusion is *robust* to the selection of the prior.

This analysis is not possible with a single prior. For example, if the designer is unsure about the failure probability, one precise Bayesian approach for dealing with this lack of confidence in the estimate is to increase the variance of the prior model, thus reflecting more uncertainty of some kind. Taken to the extreme, a complete lack of information generally leads to a uniform distribution. Unfortunately, the use of a uniform distribution confounds two cases: first, that nothing is known; second, that all failure probabilities between 0 and 1 are equally likely, which is actually substantial information.

In the context of a large engineering project in which there are many individuals, this is an important distinction. For example, one engineer's complete lack of knowledge about some aspect of the system may be offset by another engineer's expertise or by additional experimentation. However, if substantial analysis has already led to the conclusion that certain outcomes are equally likely, then it would be inefficient to expend additional resources examining those probabilities. A precise approach confounds these two scenarios, therefore adding confusion to information management decisions. The robust Bayesian approach allows one to consider the different scenarios independently rather than aggregating them together. This affords the design team the opportunity to make different, more appropriate decisions under the two scenarios.

The theory of imprecise probabilities, formalized by Walley (1991), has previously been considered in design decisions and set-based design (Aughenbaugh and Paredis, 2006; Rekuc et al., 2006) and in reliability analysis (Coolen, 1994; Coolen, 2004; Utkin, 2004a; Utkin, 2004b). The theory of imprecise probabilities uses the same fundamental notion of rationality as de Finetti's work (1974). However, the theory allows a range of indeterminacy—prices at which a decision-maker will not enter a gamble as either a buyer or a seller. These in turn correspond to ranges of probabilities. For the problem of updating beliefs, imprecise probability theory essentially allows prior and posterior beliefs to be expressed as sets of density functions.

The imprecise model captures two aspects of uncertainty: the imprecision in the prior beliefs (whether inherent or due to incomplete elicitation) and the probabilistic uncertainty in the parameter value. The distinction between these two types of uncertainty is not always obvious, but the distinction can be valuable in practice (Winkler, 1996).

The consideration of imprecision is the primary difference between the motivation for imprecise probabilities and the motivation for a robust Bayesian approach. Whereas the imprecise probability view is that the analyst's beliefs can be imprecise, the robust Bayesian view is that there exists a single prior that captures the analyst's beliefs perfectly, although it may be hard to identify this distribution in practice. Either motivation leads to the consideration of sets of priors and posteriors.

For the robust updating approach, it is convenient to use the imprecise beta model and to re-parameterize the beta so that the density of $beta(s,t)$ is as given in Equation (6) (Walley, 1991; Walley et al., 1996).

$$\pi_{s,t}(\theta) \propto \theta^{st-1}(1-\theta)^{s(1-t)-1} \tag{6}$$

Compared to the standard parameterization of $beta(\alpha,\beta)$, this means that $\alpha = s \cdot t$ and $\beta = s \cdot (1-t)$ or equivalently that $s = \alpha + \beta$ and $t = \alpha/(\alpha+\beta)$. The convenience of this parameterization is that t is the mean of the distribution, which has an easily grasped meaning for both the prior assessment and the posterior analysis. The model is updated as follows: if the prior parameters are s_0 and t_0, then, after n trials with m failures, the posterior parameters are $s_n = s_0 + n$ and $t_n = (s_0 t_0 + m)/(s_0 + n)$. Since $s_n = s_0 + n$, s_0 can be interpreted to be a virtual sample size of the prior information; it captures how much weight to place on the prior compared to the observed data. Selecting this parameter therefore depends on the available information. Following Walley (1991), the parameters *t* and *s* can be imprecise and expressed as intervals $\left[\underline{t}_0, \overline{t}_0\right]$ and $\left[\underline{s}_0, \overline{s}_0\right]$. That is, the priors are the set of beta distributions with $\alpha_0 = s_0 t_0$ and $\beta_0 = s_0(1-t_0)$ such that $\underline{t}_0 \le t_0 \le \overline{t}_0$ and $\underline{s}_0 \le s_0 \le \overline{s}_0$. When the test results are collected, each prior in the set is updated as described above.

4. Metrics of uncertainty

When considering measures of uncertainty in context of planning additional tests to reduce uncertainty, it is important to keep the following points in mind.

First, if one is modeling the system performance (e.g., system failure probability) as a precise probability distribution, then the mean, variance, and other statistics about that distribution are specific numbers for the prior distribution. If *n* identical tests are conducted, and each test result is a pass or fail, then there are $n+1$ possible test results. Thus, there are $n+1$ possible posterior distributions and $n+1$ possible means, variances, and other statistics. Characterizing the quality a test plan (which has uncertain outcomes) is a classic problem in decision-making. Decision-makers have different attitudes towards such situations. For example, some decision-makers will want to know the complete distribution of outcomes, some will want the worst-case, and others will want the "average" value of a statistic.

Modeling the system performance as an imprecise probability distribution introduces an additional complexity: the mean, variance, and other statistics about that distribution are imprecise; that is, there is a set of means for the prior distribution. If *n* identical tests are conducted, and each test result is a pass or fail, then there are $n+1$ possible imprecise posterior distributions, with $n+1$ possible sets of means, variances, and other statistics. Characterizing any specific result requires some way to describe the set, either by selecting a subset of the distributions or finding the range of values for that result. After that, one still has the problem of uncertain outcomes, as discussed above.

This paper presents and demonstrates approaches for evaluating information gathering plans using metrics of uncertainty. In addition, we will compare the types of results that these different approaches give. These results can be used as the input to existing approaches for decision-making under uncertainty, including those for determining the economic value of information. The integration with such approaches we leave for future work. Therefore, we will focus on the required methods and demonstrating them with metrics that display the range of results.

For the precise Bayesian approach, we will use a variance-based sensitivity analysis and the dispersion of the mean and variance of the posterior distributions of system failure probability. For the imprecise beta model, we will consider an imprecise variance-based sensitivity analysis (Hall, 2006), the imprecision in the mean before and after additional tests are conducted, and the range of the mean and variance of the prior and posterior distributions of system failure probability.

4.1. METRICS FOR PRECISE DISTRIBUTIONS

For both precise and imprecise priors, we will consider two different strategies. The first is a variance-based sensitivity analysis of the prior distribution, which allows one to ignore the possible test results. The second strategy considers the possible outcomes of a test plan.

4.1.1. Variance-based sensitivity analysis

One can avoid the problem of considering a large number of possible test results by ignoring them entirely and focusing on the current state of information. One such approach is variance-based sensitivity analysis, which calculates the total variance of the system performance and determines how each input variable contributes to this (Sobol, 1993; Chan et al., 2000). The sensitivity of the system performance to an input variable X_i is described by the sensitivity index SV_i. The sensitivity index is the ratio of the variance of the conditional expectation to the total variance.

For test planning, a large sensitivity index indicates that reducing the variance of that variable can reduce the system performance variance. This suggests that, in order to reduce system performance variance, a test plan should focus on reducing that input variable's variance. A small sensitivity index for an input variable suggests that reducing that variable's variance should be a low priority for testing.

In the case considered here, the failure probabilities of the three components (P_A, P_B, and P_C) are the input random variables, and the failure probability of the system is the system performance (or output random variable). In particular, we can calculate the sensitivity indices for our example system as follows:

$$\begin{aligned} SV_A &= \frac{1}{V(\theta)}\left(1-E[P_B]E[P_C]\right)^2 V[P_A] \\ SV_B &= \frac{1}{V(\theta)}\left(E[P_C]\right)^2\left(1-E[P_A]\right)^2 V[P_B] \\ SV_C &= \frac{1}{V(\theta)}\left(E[P_B]\right)^2\left(1-E[P_A]\right)^2 V[P_C] \end{aligned} \tag{7}$$

4.1.2. Observing Mean and Variance for different results

The variance of the probability distribution is a measure of uncertainty about the parameter that the probability distribution models. In general, a distribution with smaller variance means that there is less uncertainty about the parameter. For the problem of test planning, we may hope to conduct tests that will yield a posterior distribution with a variance that is smaller than some threshold.

Pham-Gia and Turkann (1992) derived lower bounds on the number of additional samples needed to satisfy an upper bound on the posterior variance for a random probability modeled with a beta distribution. Unfortunately this result is not directly applicable to reducing the variance of the system failure probability by testing only the components.

In the case considered here, a test plan conducts n_A tests of component A, n_B tests of component B, and n_C tests of component C. The test plan can be summarized as $T = \{n_A, n_B, n_C\}$. If there x_A failures of component A, x_B failures of component B, and x_C failures of component C, then the posterior distributions of the component failure probabilities are as follows: $P_A \sim beta(\alpha_A + x_A, \beta_A + n_A - x_A)$, $P_B \sim beta(\alpha_B + x_B, \beta_B + n_B - x_B)$, and $P_C \sim beta(\alpha_C + x_C, \beta_C + n_C - x_C)$. From these posterior distributions, one can calculate the mean and variance of the system failure probability as discussed in Section 2. Of course, this must be repeated for each of the $(n_A+1)\times(n_B+1)\times(n_C+1)$ possible test results.

4.2. METRICS OF UNCERTAINTY FOR IMPRECISE DISTRIBUTIONS

As we did with the precise priors, we will consider two different strategies. The first is a variance-based sensitivity analysis of the prior distribution, which allows one to ignore the possible test results. The second strategy considers the possible outcomes of a test plan.

One of the motivations for using imprecise probabilities instead of precise probabilities is that they allow the total uncertainty to be captured more adequately, by separating imprecision and probability. If the variance generally captures the variability, the natural question follows: *how can imprecision be measured*? Or more generally, how can we measure the total uncertainty?

This issue has been pursued by various authors (see (Klir and Smith, 2001) for an overview). In short, the search for a single, useful measure of total uncertainty has been largely unsuccessful. We begin our examination of the problem by considering the extension of precise measures to the imprecise case.

4.2.1. Imprecise variance-based sensitivity analysis

Hall (2006) presented an approach to extend variance-based sensitivity analysis to imprecise probability distributions. The generalization from the precise case is to consider the minimum and maximum sensitivity indices across the set of input distributions. Let F be the set of input distributions (jointly across all inputs). Let $SV_{i,p}$ be the sensitivity to input i given the input distribution p. Then the bounds are given by the following:

$$\begin{aligned} \underline{SV}_i &= \min_{p \in F} \left(SV_{i,p} \right) \\ \overline{SV}_i &= \max_{p \in F} \left(SV_{i,p} \right) \end{aligned} \tag{8}$$

The difficulty in calculating these is the need to optimize these indices over the set *F*. In the case considered here, each input distribution p is a joint distribution over the component failure probabilities. Each marginal distribution comes from the imprecise prior distribution for that component.

We will use a numerical approach that selects distributions from the set F in the following way. First, we select a parameter N_e that determines the number of intermediate values for each parameter. For parameter s_A, we calculate the following set of values:

$$\left\{\underline{s}_{0A}, \underline{s}_{0A}+\frac{1}{N_e+1}\left(\overline{s}_{0A}-\underline{s}_{0A}\right), \underline{s}_{0A}+\frac{2}{N_e+1}\left(\overline{s}_{0A}-\underline{s}_{0A}\right), \ldots, \underline{s}_{0A}+\frac{N_e}{N_e+1}\left(\overline{s}_{0A}-\underline{s}_{0A}\right), \overline{s}_{0A}\right\}$$

This yields N_e+2 values for this parameter. We repeat for the other five parameters. We then take all of the combinations, which yields $\left(N_e+2\right)^6$ joint prior distributions. We choose $N_e=3$, which was determined to be adequate for this problem. More complex systems will require a more complex parameter sampling scheme.

4.2.2. Dispersion of mean and variance

In Section 4.1.2, the dispersion of the mean and variance were considered for a precise prior. Given an imprecise prior, a specific test result will yield an imprecise posterior distribution, which has a range of means and a range of variances, as discussed above. The dispersion of the mean and variance (over the possible test results) is no longer a sequence of points, as in the precise case; it is instead a sequence of sets of mean-variance pairs.

In the case considered here, given imprecise priors for the failure probabilities of the three components, we can compare different test plans (e.g. test just Component A or test just Component B) and determine how they affect the dispersion of the mean and variance.

As before, let *F* be the entire set of prior joint distributions for the component failure probabilities, and consider a test plan that conducts n_A tests of component A, n_B tests of component B, and n_C tests of component C. If there x_A failures of component A, x_B failures of component B, and x_C failures of component C, then this result yields a set $F'\left(x_A, x_B, x_C, n_A, n_B, n_C\right)$ of posterior distributions. There is a different F' for every test result. Each posterior distribution $p' \in F'\left(x_A, x_B, x_C, n_A, n_B, n_C\right)$ is determined by updating a prior distribution $p \in F$ as described in Section 4.1.2. From the posterior distribution, one can calculate the mean and variance of the system failure probability as discussed in Section 2. We will select distributions from *F* using the procedure described in Section 4.2.1.

4.2.3. Imprecision in the mean

A fundamental measure of imprecision is the range of the mean value across the set of probability distributions. For the imprecise beta model, this measure is simply $\overline{t}-\underline{t}$. We can measure this range for the prior distribution and for each imprecise posterior distribution that results from a specific test result. Each result has a particular posterior imprecision (of the mean) associated with it. Ideally, the analyst

would like this posterior imprecision to be as small as possible over all results, so we consider the maximum imprecision that results across all results.

In the case considered here, given the range of means for the failure probabilities of the three components, it is easy to see that the minimal failure probability of the system is determined by the components' minimal failure probabilities. Likewise, the maximal failure probability of the system is determined by the components' maximal failure probabilities. Therefore, the prior imprecision in the system failure probability can be calculated as follows:

$$\begin{aligned}\Delta_0(\theta) &= \max_{p\in F} E\left[\theta \middle| p\right] - \min_{p\in F} E\left[\theta \middle| p\right] \\ &= \left(\overline{t}_{0A} + \overline{t}_{0B}\overline{t}_{0C} - \overline{t}_{0A}\overline{t}_{0B}\overline{t}_{0C}\right) - \left(\underline{t}_{0A} + \underline{t}_{0B}\underline{t}_{0C} - \underline{t}_{0A}\underline{t}_{0B}\underline{t}_{0C}\right)\end{aligned} \tag{9}$$

Each possible result of a test plan that conducts a total of *n* tests will yield a set F' of posterior distributions. The posterior imprecision in the system failure probability (given this result) can be determined as follows:

$$\begin{aligned}\Delta_{n,F'}(\theta) &= \max_{p'\in F'} E\left[\theta \middle| p'\right] - \min_{p'\in F'} E\left[\theta \middle| p'\right] \\ &= \left(\overline{t}_{nA} + \overline{t}_{nB}\overline{t}_{nC} - \overline{t}_{nA}\overline{t}_{nB}\overline{t}_{nC}\right) - \left(\underline{t}_{nA} + \underline{t}_{nB}\underline{t}_{nC} - \underline{t}_{nA}\underline{t}_{nB}\underline{t}_{nC}\right)\end{aligned} \tag{10}$$

The maximum posterior imprecision over all possible F' (that is, over all possible results for this test plan, $0 \le x_A \le n_A$ $0 \le x_B \le n_B$ $0 \le x_C \le n_C$) can be denoted as follows:

$$\Delta_{\max}(\theta) = \max_{F'}\left\{\Delta_{n,F'}(\theta)\right\} \tag{11}$$

One can also consider the average mean over the results for a given prior. For each prior and possible test result, that prior is used to determine the probability of that test result and the posterior mean given that result. Here, let $\mu\left(x, n, p_0(\cdot)\right)$ be the posterior mean, which depends upon the prior $p_0(\cdot)$ and the test result. These posterior means and prior probabilities of each result are then used to calculate an average mean for that prior: $\tilde{\mu}_{p_0} = E_{p(x)}\left[\mu(x, n, p_0)\right]$. Across a set of priors $p_0 \in F$, one can find the minimum and maximum of $\tilde{\mu}_{p_0}$.

4.2.4. Imprecision in the variance

Just as the mean of the posterior depends on both the priors and the experimental results, so does the variance. The variance is a traditional measure of uncertainty in precise formulations of probability. Even in an imprecise approach, the analyst would like the variance to be as small as possible. However, the variance is no longer a precise measure, but rather an interval for each possible result. Strictly speaking, if

the analyst requires a posterior variance below some threshold, then he must consider the maximum variance across all combinations of prior distributions and possible experimental results.

As we did with the mean, the analyst could calculate the expected posterior variance across all results given the prior. When the prior is precise, this yields a single number. When the prior is imprecise, this also results in an interval. The motivation for such as approach is that although the experimental results may not match the prior mean estimate, the analyst believes (by definition) that the priors are a reasonably accurate model of the results. For example, assume the prior mean is the range $[0.05, 0.10]$. Then if one performs 20 experiments, it is highly unlikely that one will observe 20 failures. Thus, this result is discounted by its improbability, unlike the maximum variance approach that would consider this extreme case on an equal footing with all others.

One can also consider measuring the imprecision using the range of the variance across the set of probability distributions. The imprecision in the variance reflects how well the variance is known. Ideally, an analyst would pick a test design that will result in a posterior variance that tends to be low and well known. The prior imprecision in the variance, the posterior imprecision in the variance given a particular result, and the maximum imprecision over all results are shown in Equations (12)–(14) respectively.

$$\Delta_0(V) = \max_{p \in F} V\left[\theta \middle| p\right] - \min_{p \in F} V\left[\theta \middle| p\right] \tag{12}$$

$$\Delta_{n,F'}(V) = \max_{p' \in F'} V\left[\theta \middle| p'\right] - \min_{p' \in F'} V\left[\theta \middle| p'\right] \tag{13}$$

$$\Delta_{\max}(V) = \max_{F'} \left\{ \Delta_{n,F'}(V) \right\} \tag{14}$$

To estimate these measures, we will select distributions for each F' by updating the distributions from F that are generated using the procedure described in Section 4.2.1.

5. Results

As mentioned above, in this paper we are primarily concerned with determining which component should receive more tests. To illustrate the approaches to evaluating test plans, we will consider the example of Section 2 using both precise priors and imprecise priors about the failure probabilities of the three components. We consider two scenarios for each approach.

5.1. SCENARIO 1

In the first scenario, the priors for the failure probability distributions are precise beta distributions. The parameters are shown in Table 1. The high prior mean for Component C is chosen for illustrative

purposes; it is unlikely that any real system would have a component with such a high mean estimate of probability of failure. For the distribution of the system failure probability, the prior mean equals 0.2201, and the prior variance equals 0.0203.

Table 1. Priors for Scenario 1

Component	A	B	C
Beta parameters	$t_0 = 0.15$ $s_0 = 10$	$t_0 = 0.15$ $s_0 = 2$	$t_0 = 0.55$ $s_0 = 10$

5.1.1. Scenario 1: Variance-based Sensitivity Analysis

The variance-based sensitivity analysis gives the following values:

$$SV_A = 0.4814$$
$$SV_B = 0.4583$$
$$SV_C = 0.0181$$

These values indicate that the output variance is similarly dependent on the variance of the failure probabilities for Components A and B. It is highly insensitive to component C. This suggests that testing be focused on Components A and B, but it does not suggest the appropriate allocation between them.

5.1.2. Scenario 1: Observing Mean and Variance for different results

Figure 1 shows the dispersion of the mean and variance for seven test plans: (1) 12 tests of Component A, (2) 12 tests of Component B, (3) 12 tests of Component C, (4) 4 tests of each component, (5) 6 tests of Components A and B, (6) 6 tests of Components A and C, and (7) 6 tests of Components B and C. The multiple points for each test plan correspond to the set of possible outcomes of the test.

Figure 1 reveals that the test plan makes a significant difference in the mean and variance of the possible posterior distributions. Because $E[P_C]$ is near 0.5 and the $s_0 = 10$, test plan 3 can change $E[P_C]$ very little for any test result. When Component C fails, Component B becomes serially connected to Component A and its influence on the system failure is greatly increased.

Test plans 1 and 2, on the other hand, can change the mean a great deal, from a low near 0.15 to a max near 0.65, and can substantially reduce $V(\theta)$. Test plans 4, 5, 6, and 7 likewise have a large range of possible posterior means. Test plans 4, 6, and 7 have generally larger posterior variance than test plan 5, which tests only the two components with the largest sensitivity indices. In this scenario, it seems that testing the components with the largest sensitivity indices is a worthwhile plan because this testing can reduce the variance of the corresponding component failure probability distributions, which has a large impact on the variance of the system failure probability distribution.

5.1.3. Posterior variance

Table 2 shows the minimum and maximum variance of $V(\theta)$, the posterior system failure probability distribution, for each of the seven test plans (taken over the possible results for each plan). Test plans 1 and 2 test the components with the largest sensitivity indices and reduce variance significantly. Test plan 5 can significantly reduce variance, as it has a significantly lower minimum variance. Its maximum variance is moderate, as poor test results for both Components A and B would increase the $V(P_A)$ and $V(P_B)$, increasing $V(\theta)$. Test plan 4 has similar performance. Test plan 3 has the largest minimum and

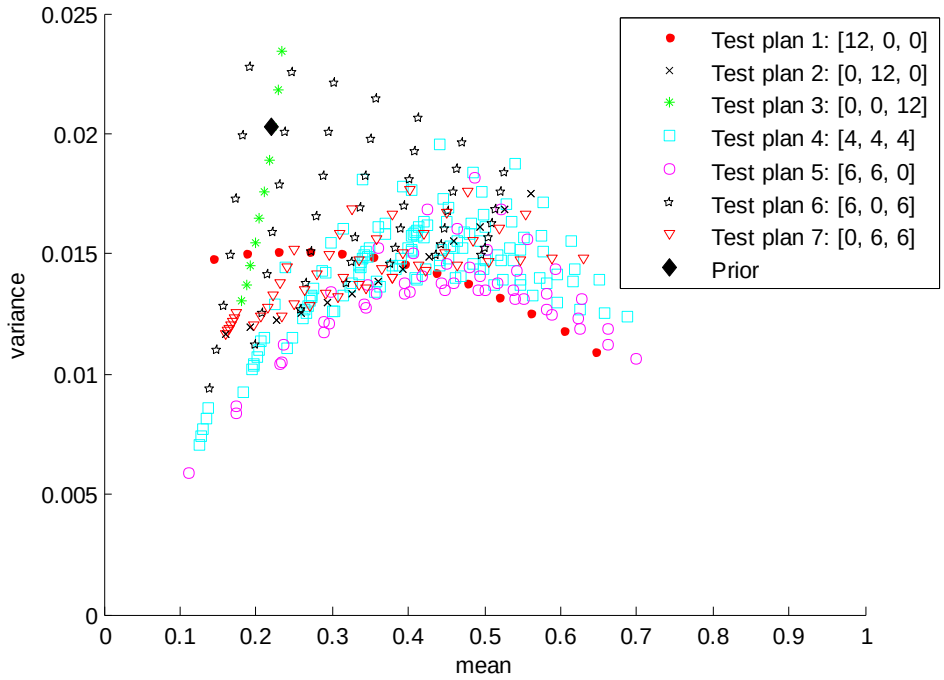


Figure 1. Dispersion of the mean and variance of the system failure probability for different test plans for Scenario 1.

maximum posterior variance. Reducing the variance of Component C's failure probability distribution cannot reduce $V(\theta)$ much, but poor test results could increase $E[P_C]$ greatly, which in this case increases $V(\theta)$ by making it more sensitive to the highly uncertain performance of Component B as in Equation (7).

Table 2. Posterior variance for scenario 1

Test Plan #: $\{n_A, n_B, n_C\}$	Posterior Variance Across Test Results	
	Min	Max
1: {12,0,0}	0.0110	0.0151
2: {0,12,0}	0.0117	0.0175
3: {0,0,12}	0.0131	0.0291
4: {4,4,4}	0.0071	0.0195
5: {6,6,0}	0.0059	0.0181
6: {6,0,6}	0.0094	0.0228
7: {0,6,6}	0.0117	0.0177

5.2. SCENARIO 2

In the second scenario, as in the first, the priors for the failure probability distributions are precise beta distributions. The parameters are shown in Table 3. The difference from Scenario 1 is that Component C now has the same distribution as Component A. For the distribution of the system failure probability, the mean equals 0.1691, and the variance equals 0.0116.

Table 3. Priors for Scenario 2

Component	A	B	C
Beta parameters	$t_0 = 0.15$ $s_0 = 10$	$t_0 = 0.15$ $s_0 = 2$	$t_0 = 0.15$ $s_0 = 10$

5.2.1. Scenario 2: Variance-based Sensitivity Analysis

The variance-based sensitivity analysis gives the following values:

$$SV_A = 0.8982$$
$$SV_B = 0.0560$$
$$SV_C = 0.0153$$

The important position of Component A yields a large sensitivity index. Compared to Scenario 1, $E[P_C]$ is now much smaller, which reduces SV_B, as suggested by Equation (7). These values suggest that reducing $V(P_A)$ by testing Component A should reduce $V(\theta)$ significantly.

5.2.2. Scenario 2: Observing Mean and Variance for different results

Figure 2 shows the dispersion of the mean and variance of the posterior mean and variance of the system failure probability distribution for seven different test plans (the same test plans used in Scenario 1). Although the prior distributions for the failure probabilities for Components A and C are the same, testing Component A (which is essential for system operation and has a much greater sensitivity index) makes a bigger change in $E[\theta]$ and $V(\theta)$. In this scenario, the mean-variance dispersion confirms the suggestion made by the variance-based sensitivity analysis: testing Component A appears to be the best strategy.

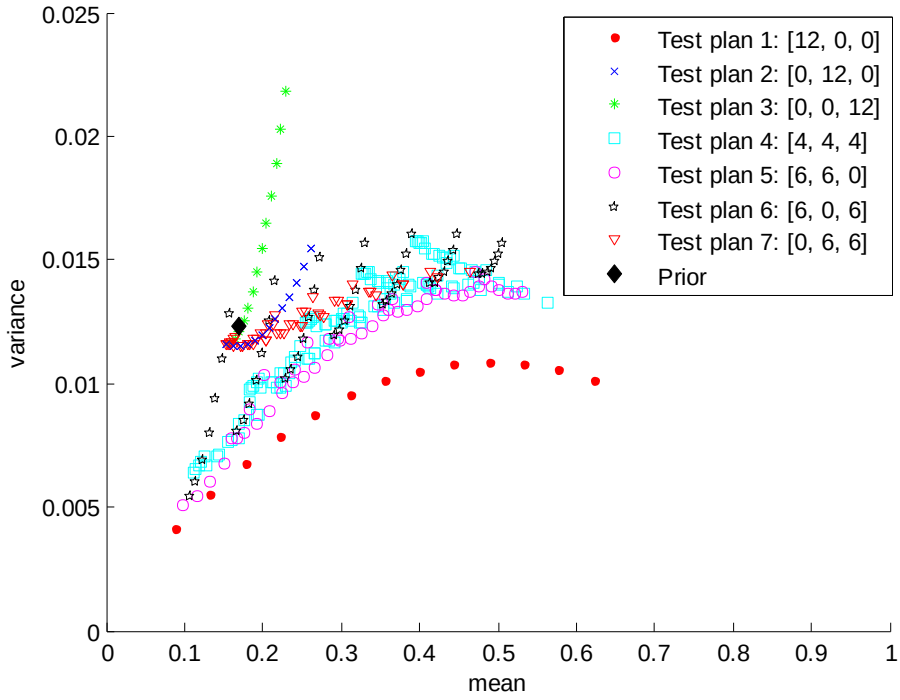


Figure 2. Dispersion of the mean and variance of the system failure probability for different test plans for Scenario 2.

Components B and C have the same position in the system, but Component B has a smaller *s* parameter and a higher variance. Therefore, testing Component B makes a bigger change in $E[P_B]$ and $V(P_B)$ than the same number of tests of Component C would make in $E[P_C]$ and $V(P_C)$. Moreover, the fact that $SV_B > SV_C$ suggests that this impact will be larger on $V(\theta)$ than on $E[\theta]$, which we see by comparing test plans 2 and 3 and then comparing test plans 5 and 6. Testing Component B reduces $V(\theta)$ more than testing Component C.

If we look at the possible results of test plan 2, which tests Component B, we see an interesting "hook" pattern that occurs because the test results with a small number of failures tend to confirm the prior, which reduces $V(P_B)$. However, results with more failures increase both $E[P_B]$ and $V(P_B)$, which increase $E[\theta]$ and $V(\theta)$.

5.2.3. Posterior variance

Table 4 shows the minimum and maximum variance of $V(\theta)$, the posterior system failure probability distribution, for each of the seven test plans (taken over the possible results for each plan). Test plans 1, 4, 5, and 6 all have low minimum $V(\theta)$ because all test Component A and can lower $V(P_A)$, which make a large impact, as we know because Component A has the largest sensitivity index. Test plan 1 also has a low maximum $V(\theta)$, which makes this test plan particularly desirable. As in Scenario 1, test plan 3 has the largest minimum and maximum posterior variance. Reducing the variance of Component C's failure probability distribution cannot reduce $V(\theta)$, but poor test results could increase $E[P_C]$ greatly, which in this case increases $V(\theta)$ because the large variance $V(P_B)$ becomes more important, as in Equation (7).

Table 4. Posterior variance for scenario 2

Test Plan #: $\{n_A, n_B, n_C\}$	Posterior Variance Across Test Results	
	Min	Max
1: {12,0,0}	0.0042	0.0109
2: {0,12,0}	0.0115	0.0155
3: {0,0,12}	0.0116	0.0218
4: {4,4,4}	0.0064	0.0158
5: {6,6,0}	0.0051	0.0145
6: {6,0,6}	0.0054	0.0160
7: {0,6,6}	0.0115	0.0145

5.3. SCENARIO 3

In the third scenario, prior distributions are imprecise (we use the imprecise beta model parameterized by *t* and *s*). Table 5 lists the parameters for each component's failure probability distribution. We assume that

less is known about Component B than the other components, as indicated by the small values for s and the large range for t. The estimated probability of failure of Component C is assumed to be quite large; while these values may not make sense in a real system, they are illustrative of interesting information management behavior. Note that the precise priors given for the first scenario (Table 1) are included in these sets. For selecting priors for the numerical results, as discussed in Section 4.2.1, we use $N_e = 3$.

Table 5. Imprecise priors for Scenario 3

Component	A	B	C
Imprecise beta parameters	$\underline{t}_0 = 0.15$ $\overline{t}_0 = 0.20$ $\underline{s}_0 = 10$ $\overline{s}_0 = 12$	$\underline{t}_0 = 0.15$ $\overline{t}_0 = 0.55$ $\underline{s}_0 = 2$ $\overline{s}_0 = 5$	$\underline{t}_0 = 0.55$ $\overline{t}_0 = 0.60$ $\underline{s}_0 = 10$ $\overline{s}_0 = 12$

The mean of the system failure probability distribution ranges from 0.2201 to 0.4640, which is an imprecision of 0.2439. The variance ranges from 0.0136 to 0.0332, which has an imprecision of 0.0196.

5.3.1. Scenario 3: Variance-based sensitivity analysis

The imprecise variance-based sensitivity analysis yields the results shown in Table 2. These results suggest that it is similarly important to test A and B (using the upper bounds), and it is much less important to test component C. The maximum for component C is about the same as the minimum for Component B, so the values strongly suggest that testing B is more valuable than testing C.

Table 6. Imprecise variance-based sensitivity analysis Scenario 3

Component i	A	B	C
$\min\{SV_{ij}\}$	0.1363	0.2406	0.0116
$\max\{SV_{ij}\}$	0.7204	0.6960	0.2512

5.3.2. Scenario 3: Dispersion of mean and variance

We will consider the same seven test plans used in Scenarios 1 and 2. Based on the sensitivity indices, it appears that test plans 1 (12, 0, 0), 2 (0, 12, 0), and 5 (6, 6, 0) should have the most potential to reduce $V(\theta)$. We begin by examining the dispersion of the mean and variance estimates across all possible experimental results for these three test plans, as shown in Figure 3. In general, a figure showing all of these points gets very difficult to display and view due to overlap. Consequently, we generally will display just the convex hull of each result of each test plan, as shown in Figure 4, which are clearer when

viewed in color. While these sets of points are not always convex, this approximation is reasonable for the qualitative analysis performed with them.

All three test plans (1, 2, and 5) can significantly change $E[\theta]$. The impact of test plan 2 (0, 12, 0) is mitigated by the system structure, in which Component B is parallel to Component C. The maximum $V(\theta)$ of test plan 1 (12, 0, 0) is much greater than the other two test plans. The significant imprecision in the priors, especially when combined, leads to large imprecision for any test result, especially in test plan 1 (12, 0, 0). Because the *s* parameters for Component B are smaller than those for Component A, testing Component B reduces $V(P_B)$ more than testing Component A reduces $V(P_A)$. Of course, testing both components (as in test plan 5) can reduce both component variances, which is quite effective at reducing $V(\theta)$ while still being responsive to the mean.

Figure 5 shows the convex hulls for the results of test plans 3 (0, 0, 12), 6 (6, 0, 6), and 7 (0, 6, 6). Test plan 6 leads to results that have a wide range of means and variances. Test plan 3 also has results with large variance, though not as large a range as test plan 6. The results for test plan 7 are similar to those of test plan 5 (shown in Figure 4), but the variance is not as small. As suggested by the sensitivity indices, testing A has more impact than testing C.

Figure 6 includes the convex hulls for the results of test plan 4 (4, 4, 4), as well as the promising plans of 2, 5, and 7. Plan 4 yields results that are quite close to those of test plan 5. Because it tests all three components and can change all three mean values, the max $E[\theta]$ is larger in the results of test plan 4. The extreme results of test plan 5 were limited by no change in $E[P_C]$.

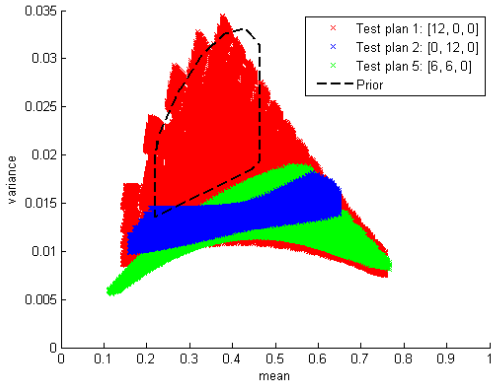


Figure 3. Sample results for test plans 1, 2, and 5 for Scenario 3.

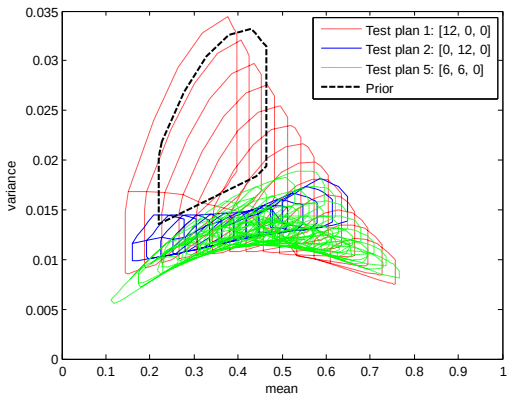


Figure 4. Convex hull for each result of test plans 1, 2, and 5 for Scenario 3.

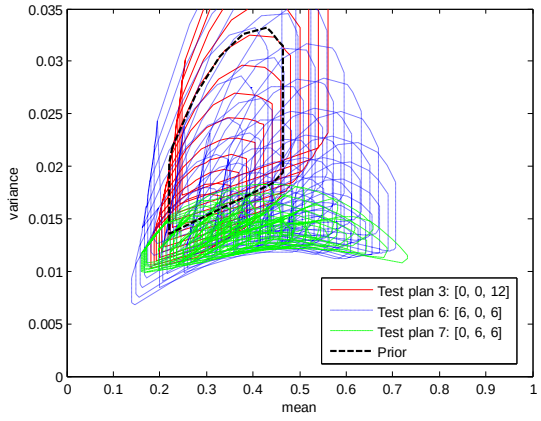


Figure 5. Convex hull for each result of test plans 3, 6, and 7 for Scenario 3.

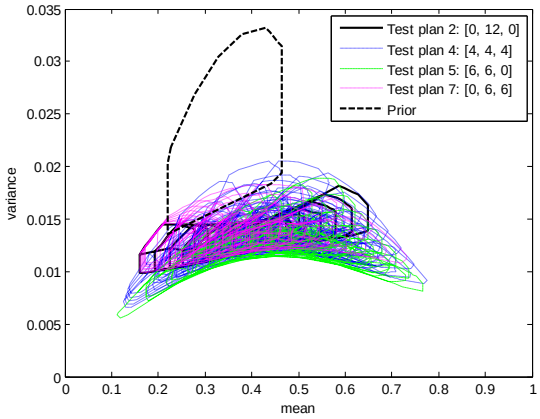


Figure 6. Convex hull for each result of test plans 2, 4, 5, and 7 for Scenario 3.

5.3.3. Scenario 3: Imprecision in the mean

Table 7 describes the imprecision of the $E[\theta]$ that can result from the various test plans. In each row, the first column is the test plan. The second column ("Minimum minimum") is the minimum possible mean over all possible distributions and test results. The third column ("Maximum maximum") is the maximum possible mean over all possible distributions and test results. The fourth column ("Minimum average") is the minimum average mean (see Section 4.2.3). The fifth column ("Maximum average") is the maximum average mean over all the priors. The sixth and seventh columns are different. Here, the imprecision in $E[\theta]$ is calculated for each possible test result using Equation (10), and the minimum and maximum are taken over the possible test results.

In these results, test plan 4 (4, 4, 4) stands out for its low minimum minimum, low minimum average, and low maximum average. This occurs because this test plan is more likely to have zero failures (than other test plans that run more tests of a component) and it includes the possibility of zero failures of all three components. Either result would significantly reduce the components' means and thus $E[\theta]$. Such a result would also leave little imprecision in $E[\theta]$, as indicated in its very low minimum imprecision. Test plans 2, 4, 5, and 7 have a maximum imprecision that is less than the imprecision in the prior. All of these plans test Component B and reduce the large imprecision in $E[P_B]$, which reduces the imprecision in $E[\theta]$. Note that testing Component A (as in test plans 1 or 6) does not reduce the imprecision significantly, suggesting that the sensitivity indices are not good predictors of which tests will do well on this measure. Similarly, the suggestion to not test Component C based on the sensitivity indices is contradicted.

Table 7. Posterior mean analysis for scenario 3

Test Design #: $\{n_A, n_B, n_C\}$	$E[\theta]$				Imprecision in $E[\theta]$	
	Minimum minimum	Maximum maximum	Minimum average	Maximum average	Minimum	Maximum
Prior	0.2201	0.4640	n.a.	n.a.	0.2439	
1: {12,0,0}	0.1451	0.7564	0.2143	0.4680	0.1463	0.2519
2: {0,12,0}	0.1600	0.6491	0.2113	0.4766	0.1085	0.1486
3: {0,0,12}	0.1819	0.5600	0.2192	0.4678	0.1501	0.3112
4: {4,4,4}	0.1119	0.6367	0.1644	0.2916	0.0709	0.1817
5: {6,6,0}	0.1124	0.7662	0.2116	0.4778	0.1172	0.1952
6: {6,0,6}	0.1405	0.7063	0.2153	0.4690	0.1474	0.3019
7: {0,6,6}	0.1610	0.7325	0.2151	0.4773	0.1126	0.2174

5.3.4. Scenario 3: Imprecision in the variance

Table 8 describes the imprecision of $V[\theta]$ that can result from the various test plans. The table structure and results shown are similar to those of Table 7. In these results, test plans 4 and 5 (which test both Components A and B) are notable for their low values on all of the measures. Both plans reduce the variance associated with these components' failure probability distributions, which can significantly reduce $V[\theta]$, as the sensitivity indices indicate.

As mentioned above, because the *s* parameters for Component B are smaller than those for Component A, testing Component B reduces $V(P_B)$ more than testing Component A reduces $V(P_A)$.

Table 8. Posterior variance analysis for scenario 3

Test Design #: $\{n_A, n_B, n_C\}$	$V[\theta]$				Imprecision in $V[\theta]$	
	Minimum minimum	Maximum maximum	Minimum average	Maximum average	Minimum	Maximum
Prior	0.0136	0.0332	n.a.	n.a.	0.0196	
1: {12,0,0}	0.0075	0.0344	0.0094	0.0304	0.0046	0.0259
2: {0,12,0}	0.0099	0.0181	0.0103	0.0153	0.0035	0.0051
3: {0,0,12}	0.0103	0.0465	0.0134	0.0310	0.0070	0.0293
4: {4,4,4}	0.0059	0.0162	0.0075	0.0118	0.0020	0.0054
5: {6,6,0}	0.0056	0.0189	0.0083	0.0150	0.0022	0.0063
6: {6,0,6}	0.0068	0.0458	0.0107	0.0295	0.0041	0.0309
7: {0,6,6}	0.0100	0.0183	0.0109	0.0183	0.0026	0.0060

Consequently, when comparing plans that test Component B to those that test Component A, we see that test plan 2 reduces $V[\theta]$ and the imprecision in $V[\theta]$ more than test plan 1, and test plan 7 reduces these measures more than test plan 6. The exceptions are the minimum-minimum and minimum average because test plans 1 and 6 include the possibility of dramatically reducing $V(P_A)$ and $V[\theta]$ if no failures are observed.

These results are more consistent with the sensitivity indices. Testing just Component C leads to the worst performance (according to most metrics). However, test 4, in which all three components are tested, performs very well, even though it includes testing C. This is because testing A and B change the actual sensitivities. This is related to the difference between batch testing and sequential (i.e. one-at-a-time) testing.

5.4. SCENARIO 4

For the fourth scenario, consider the imprecise prior distributions given in Table 9. The difference from Scenario 3 is only in component C: we now assume the probability of failure is believed to be much lower and more realistic. Note that the precise priors given for the first scenario are included in these sets. For selecting priors for the numerical results, as discussed in Section 4.2.1, we use $N_e = 3$.

Table 9. Imprecise priors for Scenario 4

Component	A	B	C
Imprecise beta parameters	$\underline{t}_0 = 0.15$ $\overline{t}_0 = 0.20$ $\underline{s}_0 = 10$ $\overline{s}_0 = 12$	$\underline{t}_0 = 0.15$ $\overline{t}_0 = 0.55$ $\underline{s}_0 = 2$ $\overline{s}_0 = 5$	$\underline{t}_0 = 0.15$ $\overline{t}_0 = 0.20$ $\underline{s}_0 = 10$ $\overline{s}_0 = 12$

The mean of the system failure probability distribution ranges from 0.1691to 0.2880, which is an imprecision of 0.1189. The variance ranges from 0.0100 to 0.0173, which is an imprecision of 0.0073.

5.4.1. Scenario 4: Variance-based sensitivity analysis

The imprecise variance-based sensitivity analysis yields the results shown in Table 10. SV_A and SV_C have remained roughly the same. Compared to Scenario 3, SV_B has dropped due to the drop in $E[P_C]$. These results suggest that testing Component A and reducing its variance will have the most impact on reducing $V(\theta)$.

Table 10. Imprecise variance-based sensitivity analysis Scenario 4

Component *i*	A	B	C
$\min\{SV_{ij}\}$	0.5438	0.0210	0.0095
$\max\{SV_{ij}\}$	0.9590	0.1819	0.2515

5.4.2. Scenario 4: Dispersion of mean and variance

We will consider the same seven test plans used in the previous scenarios. Based on the sensitivity indices, it appears that test plan 1 (12, 0, 0) should have the most potential to reduce $V(\theta)$. Because testing Component B can reduce the large imprecision in $E[P_B]$, we expect that test plans that include Component B will reduce the imprecision in $E[\theta]$. Figure 7–Figure 9 show the convex hull of each result of each test plan.

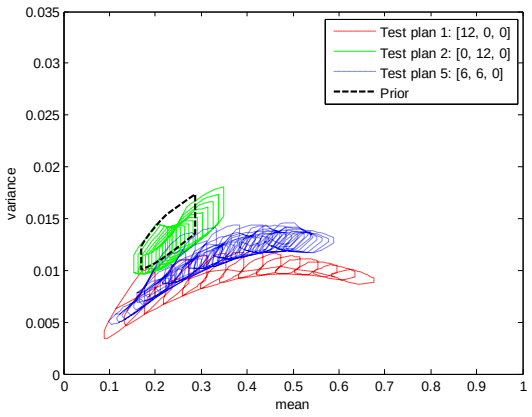


Figure 7. Convex hull of each result for Scenario 4, test plans 1, 2, and 5.

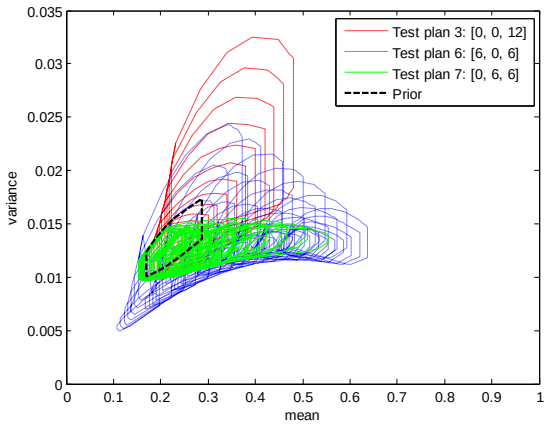


Figure 8. Convex hull of each result for Scenario 4, test plans 3, 6, and 7.

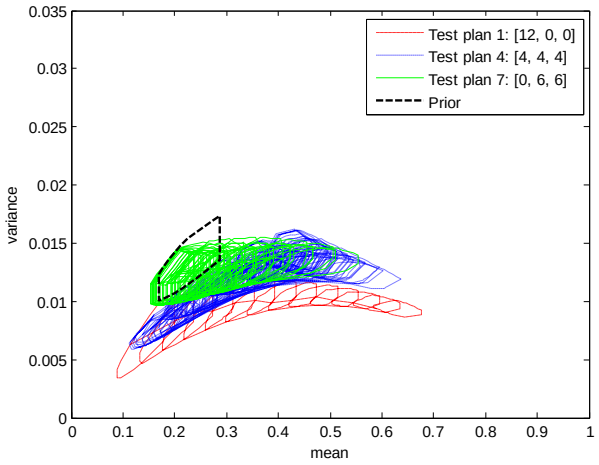


Figure 9. Convex hull of each result for Scenario 4, Tests 1, 4 and 7

Test plan 1 has the greatest range of $E[\theta]$, which reflects the critical location of Component A in the system. Moreover, this plan reduces $V(\theta)$ significantly, as the sensitivity index suggested. Test plan 5 has a slightly smaller range of $E[\theta]$ and does not reduce $V(\theta)$ as much, though it does more than test plan 2. In the results of test plan 2, we see again the behavior noted in Scenario 2 (the "hook" in Figure 2), but now multiplied for a number of priors. The entire convex hull follows this trajectory. For a given prior, when the test results confirm the prior, testing Component B reduces $V(P_B)$. However, poor test results increase both $E[P_B]$ and $V(P_B)$, which increase $E[\theta]$ and $V(\theta)$.

Testing Component C (test plan 3) is not helpful. Test results that confirm the prior tend to shrink the range of $E[\theta]$ compared to the prior. However, poor test results increase both $E[P_C]$ and $V(P_C)$, which increase $E[\theta]$ and $V(\theta)$. Test plan 6 can also give high-variance results because it does not reduce $V(P_B)$, which is relatively large, and poor results can increase both $V(P_A)$ and $V(P_C)$. Test plan 7 can reduce both $V(P_B)$ and $V(P_C)$, but, as the sensivity indices suggest, this cannot reduce $V(\theta)$ as much as reducing $V(P_A)$. Test plan 4 can reduce all three component-level variances, but the limited number of test results means that the $V(P_A)$ is not reduced as much as it is in test plan 1, which limits the reduction of $V(\theta)$.

5.4.3. Scenario 4: Imprecision in the mean

Table 11 describes the imprecision of $E[\theta]$ that can result from the various test plans. The table structure and the types of results shown are identical to those of Table 7. Test plan 1 yields the most extreme values of minimum-minimum and maximum- maximum because no failures (or all failures) significantly affects $E[P_A]$, which has a large impact on $E[\theta]$ due Component A's position in the system.

Most of the test plans have the same minimum average and maximum average, which are close to the minimum and maximum prior $E[\theta]$. This is not surprising since extreme test results (and large changes from a prior to its posterior) such as observing all failures are unlikely when the number of tests is large enough.

As in Scenario 3, the test plans that include Component B reduce the large imprecision in $E[P_B]$, which reduces the imprecision in $E[\theta]$. Test plans 3 and 6, which don't include Component B, not only fail to reduce the large imprecision in $E[P_B]$ but also add imprecision when a large number of failures for Component C add imprecision to $E[P_C]$. Similarly, though not to the same degree, test plan 4 can add imprecision. As noted in the results of Scenario 3, testing just Component A (as in test plans 1) does not significantly reduce the imprecision, suggesting that the sensitivity index is not a good predictor of which tests will do well on this measure. The greatest potential reduction in imprecision can occur when A and B are tested equally in test plan 5.

Table 11. Posterior mean analysis for scenario 4

Test Design #: $\{n_A, n_B, n_C\}$	$E[\theta]$				Imprecision in $E[\theta]$	
	Minimum minimum	Maximum maximum	Minimum average	Maximum average	Minimum	Maximum
Prior	0.1691	0.2880	n.a.	n.a.	0.1189	
1: {12,0,0}	0.0891	0.6764	0.1643	0.2934	0.0918	0.1099
2: {0,12,0}	0.1527	0.3497	0. 1671	0. 2916	0.0732	0.1041
3: {0,0,12}	0.1587	0.4800	0.1682	0.2912	0.0853	0.2567
4: {4,4,4}	0.1120	0.6367	0.1656	0.2918	0.0709	0.1817
5: {6,6,0}	0.0988	0.5888	0.1647	0.2955	0.0685	0.1100
6: {6,0,6}	0.1065	0.6375	0.1644	0.2926	0.0809	0.2190
7: {0,6,6}	0.1530	0.5550	0.1667	0.2936	0.0737	0.1790

5.4.4. Scenario 4: Imprecision in the variance

Table 12 describes the imprecision of $V[\theta]$ that can result from the various test plans. The table structure and types of results shown are identical to those of Table 8. In these results, test plan 1 is notable for its low values on almost all of the measures (the only exception being the maximum imprecision). This plan can substantially reduce $V(P_A)$, which reduces $V[\theta]$, as the sensitivity indices indicate. As we saw in Scenario 2, poor test results for Component C can greatly increase $V[\theta]$, and we see that here in the maximum maximum for test plan 3.

Unlike the results for the mean in Table 11, here we see that testing Component A, as test plans 1, 5, and 6 do, can reduce the minimum average and maximum average (compared to the prior) because they substantially reduce $V(P_A)$, which reduces $V[\theta]$, as the sensitivity indices indicate. The other test plans have less impact because the sensitivity indices of the other components are smaller. All of the test plans except test plans 3 and 6 (which can greatly increase $V[\theta]$) reduce the imprecision in $V[\theta]$.

Table 12. Posterior variance analysis for scenario 4

Test Design #: $\{n_A, n_B, n_C\}$	$V[\theta]$				Imprecision in $V[\theta]$	
	Minimum minimum	Maximum maximum	Minimum average	Maximum average	Minimum	Maximum
Prior	0.0100	0.0173	n.a.	n.a.	0.0073	
1: {12,0,0}	0.0034	0.0117	0.0053	0.0110	0.0015	0.0068
2: {0,12,0}	0.0097	0.0181	0.0098	0.0155	0.0045	0.0061
3: {0,0,12}	0.0098	0.0325	0.0100	0.0160	0.0048	0.0189
4: {4,4,4}	0.0059	0.0162	0.0074	0.0117	0.0020	0.0054
5: {6,6,0}	0.0048	0.0146	0.0067	0.0116	0.0019	0.0062
6: {6,0,6}	0.0049	0.0243	0.0068	0.0119	0.0024	0.0165
7: {0,6,6}	0.0097	0.0155	0.0098	0.0145	0.0028	0.0050

6. Discussion

The above results, though for specific scenarios and a specific system design, demonstrate some principles that we believe are generally applicable to problems of this type.

First, examining the dispersion of the mean and variance is a useful way to determine the possible outcomes of a test plan. Comparing different dispersions can identify which plans are most likely to reduce system-level variance and have a large impact on system-level mean.

Next, the variance-based sensitivity analysis is not a substitute for looking at the dispersion of the mean and variance, especially in the imprecise scenarios. It does give some prediction into which components should be tested. Because it is computationally less expensive to calculate the sensitivity indices than the potential posteriors across all results, this is important. In particular, testing a component with a high sensitivity index can reduce system-level variance substantially if the number of tests is large enough relative to the *s* parameter (a small number of tests won't change the component-level variance enough if the *s* parameter is large). However, testing a component with a small sensitivity index may greatly increase system-level variance; only examining the dispersion of the mean and variance can reveal that.

However, the sensitivity indices do not give adequate insight into joint testing—that is, testing multiple components. In Scenario 3, the sensitivity indices clearly suggested that testing Component C was much less important than testing A or B. However, the smallest maximum-maximum and second smallest minimum-minimum posterior variances actually occur with test plan 4, which tests all three components equally (see Table 8). This test plan also yields the smallest maximum imprecision in the variance, which means that its worst case result leads to the most information about the variance than any other test's worst case. This is ideal in that not only does the variance have the smallest maximum, but it will be known accurately, whatever the actual result. It should be noted that one could also consider joint sensitivity indices, an analysis that was not performed in this study and should be considered in future work.

A sensitivity index does not give much insight into how testing that component will affect the imprecision of the system-level mean. While expected, since they deal with the variance and not the mean, the results confirm this result. The adjustment from the precise sensitivity indices to the imprecise ones is necessary when using imprecise probabilities, but it does not sufficiently capture all important aspects of the imprecision. For example, in Scenario 4, the sensitivity indices clearly suggest that testing Component A is most important, and from a variance perspective, it is. However, the best reduction in the imprecision of the mean actually occurs in test plan 5, when both A and B are tested. Similarly, in Scenario 3, the best reduction in imprecision in the mean goes from either testing just A or testing all three equally (Table 7), although the sensitivity indices clearly suggested that testing C was unimportant, and were relatively inconclusive between A and B.

Testing a component with large imprecision in its mean failure probability is useful because it reduces the component-level imprecision, which reduces the system-level imprecision. However, if the component-level imprecision is low, testing that component may increase imprecision of the system-level mean and variance if the results contradict the prior information. Again, the dispersion plot will show this potential.

The minimum and maximum average measures (for system-level mean and variance) are not very useful. In Scenario 4, they change very little from the values for the prior. In Scenario 3, they can change significantly, but the dispersion plot will show this as well. Additionally, the minimum-minimum and maximum-maximum metrics yield similar rankings to those from the minimum-average and maximum-average respectively. Theoretically, the average metrics give a more accurate insight into the actual posterior means and variances that would result from test plans, but as far as choosing a test plan, it is only the ranking that matters. Additionally, the average values are computationally more expensive to compute.

In this example, many posterior statistics were analytically computable, as shown in Equations (1)-(5) and Equation (7). In general, the posterior system distribution would need to be calculating using a

double-loop Monte Carlo simulation, or a more advanced method (for a summary, see Bruns and Paredis, 2006). This greatly increases the computational costs over this example. However, having an estimate of the posterior distribution allows one to use other uncertainty metrics, such as the entropy, the Aggregate Uncertainty (Klir and Smith, 2001), or imprecise posterior breadth measures (Ferson and Tucker, 2006). Consideration of these metrics is left for future work.

7. Summary

This paper has presented and compared different strategies for measuring the uncertainty of precise and imprecise distributions for use in making test planning decisions. In this paper we have not considered specific approaches for making decisions in the presence of uncertainty or estimating the economic value of the information, since these depend on the problem context and the preferences of the decision-maker.

Instead, we considered the variance and imprecision of the posterior distributions more directly. In some cases, this will be sufficient to make a decision. Future work will need to consider how to integrate the approaches presented here with approaches in information economics, decision analysis, and optimization to help one select the best test plan.

Acknowledgements

This work was sponsored in part by Applied Research Laboratories at the University of Texas at Austin IR&D grant 07-09.

References

Aughenbaugh, J. M. and J. W. Herrmann. Updating Uncertainty Assessments: A Comparison of Statistical Approaches. *2007 ASME International Design Engineering Technical Conferences*. Las Vegas, NV, 2007.

Aughenbaugh, J. M. and C. J. J. Paredis. The Value of Using Imprecise Probabilities in Engineering Design. *Journal of Mechanical Design* **128**(4): 969-979, 2006.

Berger, J. O. The Robust Bayesian Viewpoint. *Robustness of Bayesian Analysis*. J. B. Kadane, ed. New York, North-Holland**:** 63-124, 1984.

---. *Statistical Decision Theory and Bayesian Analysis* (2nd edn.), Springer, 1985.

---. *An Overview of Robust Bayesian Analysis*. Report num. Technical Report #93-53c. West Lafayette, IN, Purdue University, 1993.

Berger, J. O. and J. M. Bernardo. On the Development of the Reference Prior Method. *Bayesian Statistics 4*. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds. Oxford, Oxford University Press, 1992.

Box, G. E. P. and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Reading, Massachusetts, Addison-Wesley, 1973.

Bradley, S. R. and A. M. Agogino. An Intelligent Real Time Design Methodology for Component Selection: An Approach to Managing Uncertainty. *Journal of Mechanical Design* **116**(4): 980-988, 1994.

Bruns, M. and C. J. J. Paredis. Numerical Methods for Propagating Imprecise Uncertainty. *2006 ASME International Design Engineering Technical Conferences and Computers in Information Engineering Conference*, Philadelphia, PA, DETC2006-99237, 2006.

Chan, K., S. Tarantola, A. Saltelli and I. M. Sobol. Variance-Based Methods. *Sensitivity Analysis*. A. Saltelli, K. Chan and E. M. Scott, eds. New York, Wiley, 2000.

Coolen, F. P. A. On Bernoulli Experiments with Imprecise Prior Probabilities. *The Statistician* **43**(1): 155-167, 1994.

---. On the Use of Imprecise Probabilities in Reliability. *Quality and Reliability in Engineering International* **20**: 193-202, 2004.

de Finetti, B. *Theory of Probability Volume 1: A Critical Introductory Treatment*. New York, Wiley, 1974.

Ferson, S. and W. T. Tucker. *Sensitivity in Risk Analyses with Uncertain Numbers*. Albuquerque, NM, Sandia National Laboratories, 2006.

Fougere, P., ed. *Maximum Entropy and Bayesian Methods*. Dordrecht, Klower Academic Publishers, 1990.

Gupta, M. M. Intelligence, Uncertainty and Information. *Analysis and Management of Uncertainty: Theory and Applications*. B. M. Ayyub, M. M. Gupta and L. N. Kanal, eds. New York, North-Holland**:** 3-11, 1992.

Hall, J. W. Uncertainty-Based Sensitivity Indices for Imprecise Probability Distributions. *Reliability Engineering & System Safety* **91**(10-11): 1443-1451, 2006.

Howard, R. A. Information Value Theory. *IEEE Transactions on Systems Science and Cybernetics* **SEC-2**(1): 22, 1966.

Insua, D. R. and F. Ruggeri. *Robust Bayesian Analysis*. New York, Springer, 2000.

Jeffreys, H. *Theory of Probability* (3rd edn.). London, Oxford University Press, 1961.

Klir, G. J. and R. M. Smith. On Measuring Uncertainty and Uncertainty-Based Information: Recent Developments. *Annals of Mathematics and Artificial Intelligence* **32**(1-4): 5-33, 2001.

Kokkolaras, M., Z. P. Mourelatos and P. Y. Papalambros. Impact of Uncertainty Quantification on Design: An Engire Optimisation Case Study. *International Journal of Reliability and Safety* **1**(1/2): 225-237, 2006.

Lawrence, D. B. *The Economic Value of Information*. New York, Springer-Verlag, 1999.

Ling, J. M., J. M. Aughenbaugh and C. J. J. Paredis. Managing the Collection of Information under Uncertainty Using Information Economics. *Journal of Mechanical Design* **128**(4): 980-990, 2006.

Malak, R. J., Jr., J. M. Aughenbaugh and C. J. J. Paredis. Multi-Attribute Utility Analysis in Set-Based Conceptual Design. *Journal of Computer Aided Design* **(in press)**, 2007.

Maritz, J. S. and T. Lewin. *Empirical Bayes Methods* (2nd edn.). London, Chapman and Hall, 1989.

Marschak, J. *Economic Information, Decision, and Prediction: Selected Essays*. Boston, D. Reidel Publishing Company, 1974.

Matheson, J. E. The Economic Value of Analysis and Computation. *IEEE Transactions on Systems Science and Cybernetics* **SSC-4**(3): 325-332, 1968.

Nikolaidis, E., S. Chen, H. Cudney, R. T. Haftka and R. Rosca. Comparison of Probability and Possibility for Design against Catastrophic Failure under Uncertainty. *Journal of Mechanical Design* **126**(3): 386-394, 2004.

Oberkampf, W. L., J. C. Helton, C. A. Joslyn, S. F. Wojtkiewicz and S. Ferson. *Challenge Problems: Uncertainty in System Response Given Uncertain Parameters*, 2001.

Pham-Gia, T. and N. Turkkan. Sample Size Determination in Bayesian Analysis. *The Statistician* **41**(4): 389-397, 1992.

Radhakrishnan, R. and D. A. McAdams. A Methodology for Model Selection in Engineering Design. *Journal of Mechanical Design* **127**(May): 378-387, 2005.

Rekuc, S. J., J. M. Aughenbaugh, M. Bruns and C. J. J. Paredis. Eliminating Design Alternatives Based on Imprecise Information. *Society of Automotive Engineering World Congress* Detroit, MI, 2006-01-0272, 2006.

Schlosser, J. and C. J. J. Paredis. Managing Multiple Sources of Epistemic Uncertainty in Engineering Decision Making. *SAE World Congress*. Detroit, MI, 2007.

Sobol, I. M. Sensitivity Analysis for Non-Linear Mathematical Models. *Mathematical modeling and computational experiment* **1**(1): 407-414, 1993.

Soundappan, P., E. Nikolaidis, R. T. Haftka, R. Grandhi and R. Canfield. Comparison of Evidence Theory and Bayesian Theory for Uncertainty Modeling. *Reliability Engineering & System Safety* **85**(1-3): 295-311, 2004.

Utkin, L. V. Interval Reliability of Typical Systems with Partially Known Probabilities. *European Journal of Operational Research* **153**(3 SPEC ISS): 790-802, 2004a.

---. Reliability Models of M-out-of-N Systems under Incomplete Information. *Computers and Operations Research* **31**(10): 1681-1702, 2004b.

Walley, P. *Statistical Reasoning with Imprecise Probabilities*. New York, Chapman and Hall, 1991.

Walley, P., L. Gurrin and P. Burton. Analysis of Clinical Data Using Imprecise Prior Probabilities. *The Statistician* **45**(4): 457-485, 1996.

Winkler, R. L. Uncertainty in Probabilistic Risk Assessment. *Reliability Engineering & System Safety* **54**(2-3): 127-132, 1996.

Zellner, A. Maximal Data Information Prior Distributions. *New Methods in the Applications of Bayesian Methods*. A. Aykac and C. Brumat, eds. Amsterdam, North-Holland, 1977.

The probability of type I and type II errors in imprecise hypothesis testing

Ingo Neumann and Hansjörg Kutterer
Geodetic Institute, Leibniz University of Hannover, Nienburger Straße 1, D-30419 Hannover
e-mail: [neumann, kutterer]@gih.uni-hannover.de

Abstract: In many engineering disciplines the interesting model parameters are estimated from a large number of heterogeneous and redundant observations by a least-squares adjustment. The significance of the model parameters and the model selection itself are checked with statistical hypothesis tests. After formulating a null hypothesis, the test decision is based on the comparison of a test value with a quantile value. The acceptance and the rejection of the null hypothesis are strongly related with two types of errors. A type I error occurs if the null hypothesis is rejected, although it is true. A type II error occurs if the null hypothesis is accepted, although it is false. This procedure is well known in case of only random errors for the observations.

If the uncertainty budget of the observations is assumed to comprise both random variability (probabilistic errors) and imprecision (interval errors), the classical test strategies have to be extended accordingly. In this study we focus on the relation of imprecision and the probability of type I and type II errors. These steps are based on newly developed one- and multidimensional hypothesis tests in case of imprecise data. The applied procedure is outlined in detail showing both theory and one numerical example for the parameterization of a geodetic monitoring network. Its main benefit is an improved interpretation of the influence of imprecision in model selection and significance tests. In addition the well known sensitivity analysis in parameter estimation can now generally be treated in terms of imprecise data.

Keywords: hypothesis testing, imprecision, probability, type I/II error

1. Introduction

Hypothesis tests are of wide interest for many applications in engineering and mathematical science. Different approaches to hypothesis testing exist, which are due to different methods for the description of the occurring uncertainties, e. g., in the performed measurements and the prior knowledge about the model formulation (for further data processing) and in model selection. The probably most popular approaches are statistical tests in parameter estimation, where interesting model parameters are estimated from a large number of heterogeneous and redundant observations by a least-squares adjustment. The uncertainties are assessed in a stochastic framework: measurement and system errors are modeled using random variables and probability distributions. However, the quantification of the uncertainty budget of

empirical measurements is often too optimistic due to, e.g., the ignorance of non-stochastic errors in the analysis process (Ferson et al., 2007). For this reason in this paper a more general formulation is presented which may be closer to the situation in real-world applications.

The paper is organized as follows: first, the main steps in uncertainty modeling with respect to non-stochastic measurement errors are briefly reviewed, see, e. g. (Kutterer, 2004; Neumann et al., 2006). Second, two linear hypotheses are introduced as a general approach to imprecise hypothesis testing. The main part of the paper deals with the relation of imprecision and the probability of type I and type II errors in imprecise hypothesis testing. The applied procedure is outlined in detail showing both theory and numerical examples for the parameterization of a geodetic monitoring network.

2. Hypothesis testing in parameter estimation under interval-/fuzzy-uncertainty

2.1. Modeling of Uncertainty

In this paper *uncertainty* is treated in terms of fuzzy-intervals (e. g., Bandemer and Näther 1992), see Fig. 1. With a fuzzy-interval $\tilde{A}$ it is possible to describe uncertain quantities by their membership function $m_{\tilde{A}}(x)$ over the set $\mathbb{R}$ of real numbers with a membership degree between 0 and 1:

$$\tilde{A} := \left\{ (x, m_{\tilde{A}}(x)) \middle| x \in \mathbb{R} \right\} \quad \text{with} \quad m_{\tilde{A}} : \mathbb{R} \rightarrow [0,1]. \tag{1}$$

The membership function of a fuzzy interval can be described by its left (*L)* and right (*R*) reference functions (see also Fig. 1)

$$m_{\tilde{A}}(x) = \begin{cases} L\left(\dfrac{x_m - x - r}{c_l} \right), & x < x_m - r \\ 1, & x_m - r \le x \le x_m + r \\ R\left(\dfrac{x - x_m - r}{c_r} \right), & x > x_m + r \end{cases} \tag{2}$$

with x_m denoting the midpoint, r the radius, and c_l, c_r the spread parameters of the monotonously decreasing reference functions (convex fuzzy intervals).

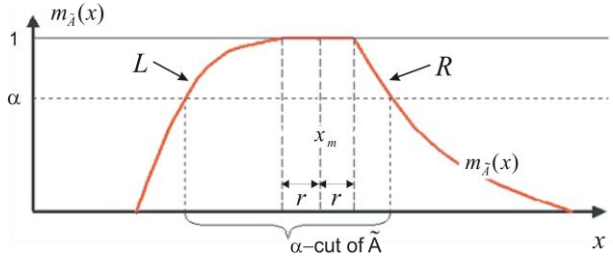


Figure 1. Fuzzy interval and its α -cut

The α-*cut* of a fuzzy-interval $\tilde{A}$ is defined by:

$$\tilde{A}_\alpha := \left\{ x \in X \,\middle|\, m_{\tilde{A}}(x) \geq \alpha \right\}, \tag{3}$$

with $\alpha \in [0,1]$. Each α-*cut* represents in case of monotonously decreasing reference functions a classical interval. The lower bound $\tilde{A}_{\alpha,\min}$ and upper bound $\tilde{A}_{\alpha,\max}$ of an α-*cut* are obtained as:

$$\tilde{A}_{\alpha,\min} = min\left(\tilde{A}_\alpha\right), \tag{4}$$

$$\tilde{A}_{\alpha,\max} = max\left(\tilde{A}_\alpha\right). \tag{5}$$

Throughout the paper we assume symmetric fuzzy intervals. Hence, an equivalent representation of symmetric α-*cuts* can be found by the midpoint A_m and radius $\tilde{A}_{\alpha,r}$ representation:

$$\tilde{A}_{\alpha,\min} = A_m - \tilde{A}_{\alpha,r}, \tag{6}$$

$$\tilde{A}_{\alpha,\max} = A_m + \tilde{A}_{\alpha,r}. \tag{7}$$

The integral over all α-*cuts* equals the membership function:

$$m_{\tilde{A}}(x) = \int_0^1 m_{\tilde{A}_\alpha}(x) d\alpha. \tag{8}$$

Furthermore, basic operations on fuzzy intervals are the *intersection* and the *complement*; they are defined through the following membership functions:

$$\textit{Intersection:} \quad \tilde{C} = \tilde{A} \cap \tilde{B} \Leftrightarrow m_{\tilde{A}\cap\tilde{B}}(x) = \min\left(m_{\tilde{A}}(x), m_{\tilde{B}}(x)\right) \quad \forall \, x \in \square \tag{9\_a}$$

$$\textit{Complement:} \quad \tilde{C} = \tilde{A}^C \Leftrightarrow m_{\tilde{A}^C}(x) = 1 - m_{\tilde{A}}(x) \quad \forall \, x \in \square \tag{9\_b}$$

Fuzzy intervals serve as basic quantities: *Random variability* is introduced through the fuzzy-interval midpoint which is modeled as a random variable and hence treated by methods of stochastics. Here random variability is superposed by *imprecision* which is due to non-stochastic errors of the measurements and the physical model with respect to reality. The standard deviation σ_x is the carrier of the stochastic uncertainty, and the spread of the fuzzy-intervals describes the range of *imprecision*.

For the modeling of imprecision it is important to know that the original measurement results are typically preprocessed before they are used in the further calculations. These preprocessing steps comprise several factors **p** influencing the observations (see also Fig. 2):

- Physical parameters (model constants) for the reduction and correction steps from the original to the reduced measurements
- Sensor parameters (e. g., remaining error sources that cannot be modeled)
- Additional information (e. g., temperature and pressure measurements for the reduction steps of a distance measurement)

Most of these influence factors are uncertain realisations of random variables; their imprecision is meaningful by many reasons:

- The model constants are only partially representative for the given situation (e. g., the model constants for the refraction index for distance measurements).
- The number of additional information (measurements) may be too small to estimate reliable distributions.
- Displayed measurement results are affected by rounding errors.
- Other non-stochastic errors of the reduced observations occur due to neglected correction and reduction steps and for effects that cannot be modeled.

Figure 2 shows the interaction between the observation and analysis model and their influence factors. While correction and reduction steps are systematic, the imprecision of the influence parameters is directly transferred to the measurements, which are now carrier of random variability and imprecision.

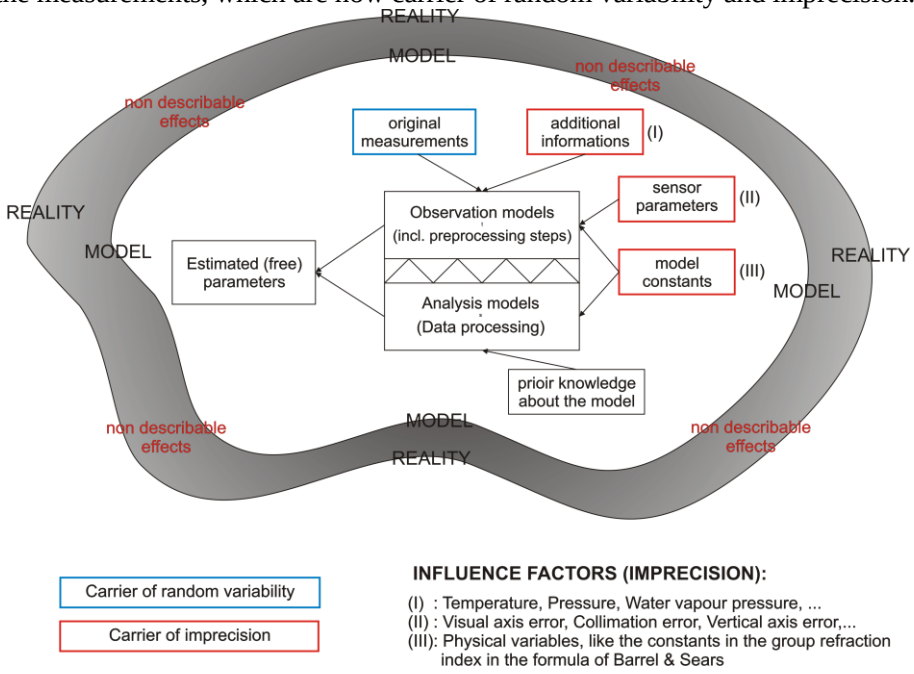


Figure 2. Interaction between the observation/analysis model and their influence factors

The non-stochastic part of the influence factors is described by fuzzy-intervals. This step is based on expert knowledge and on error models concerning the deterministic behavior of these parameters. The propagation of uncertainty is then separated into two parts. The stochastic part is treated with the law of variance-covariance propagation. Based on the assumption that imprecision is small in comparison with the measured values, we derive the data imprecision by means of a sensitivity analysis of the mostly sophisticated observation models (Neumann et al., 2006).

2.2. General Form of a Linear Hypothesis in Imprecise Hypothesis Testing

2.2.1. *The pure stochastic case*

In this subsection a general approach to imprecise hypothesis testing in parameter estimation is presented. We focus on the standard case where the vector $\mathbf{y}$ is assumed to be normal distributed with expectation vector

$$\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}_{\mathrm{y}}, \tag{10_a}$$

and (positive definite) variance-covariance matrix $\boldsymbol{\Sigma}_{\mathrm{yy}}$ (vcm)

$$\mathrm{D}(\mathbf{y}) = \boldsymbol{\Sigma}_{\mathrm{yy}} = \sigma_0^2 \mathbf{Q}_{\mathrm{yy}}, \tag{10_b}$$

where σ_0^2 denotes the variance of the unit weight and $\mathbf{Q}_{\mathrm{yy}}$ the associated cofactor matrix. Such a random vector may either be an observable quantity or a derivable quantity such as the parameters estimated by means of a least-squares (LS) adjustment. The next steps of these well-known test procedure leads to a quadratic form, which may be given by

$$\mathbf{y}^T \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{y} \sim \chi^2(f, \lambda). \tag{11}$$

In general, the quadratic form follows a non-central chi-square distribution with $\mathrm{f} = \mathrm{rank}(\boldsymbol{\Sigma}_{\mathbf{yy}})$ degrees of freedom and the non-centrality parameter λ. In the following, the vector $\mathbf{y}$ is assumed as the vector of reduced observations $\mathbf{y} = \mathbf{l} - \mathbf{a_0}$ within a least-squares adjustment, with the random vector of observations $\mathbf{l}$ and the deterministic vector of approximate observations $\mathbf{a_0}$. Then the estimated parameters $\hat{\mathbf{x}}$ of a least-squares adjustment (Gauß-Markov model) are given by the following equation:

$$\hat{\mathbf{x}} = \boldsymbol{f}(\mathbf{l}, \mathbf{x}_0) = \mathbf{x}_0 + (\mathbf{A}^{\mathbf{T}}\mathbf{P}\mathbf{A})^{+} \mathbf{A}^{\mathbf{T}}\mathbf{P}(\mathbf{l} - \mathbf{a}_0), \tag{12}$$

with the $n \times u$ column regular design matrix $\mathbf{A}$, the $n \times 1$ vector of approximate values $\mathbf{x}_0$ of the parameters $\mathbf{x}$, the $n \times n$ regular weight matrix $\mathbf{P} = \mathbf{Q}_{yy}^{-1}$. The number of observations is n and the number of parameters is u. In geodetic networks the normal equations matrix $\mathbf{A}^{\mathbf{T}}\mathbf{P}\mathbf{A}$ can be rank-deficient due to an incomplete definition of the coordinate frame through the configuration. If for example such a network is composed of distance observations only, it is not possible to estimate coordinates which are required in practice. This problem can be overcome if the pseudo-inverse matrix $(\mathbf{A}^{\mathbf{T}}\mathbf{P}\mathbf{A})^{+}$ is used; see, e. g., (Koch, 1999) which is a standard reference in geodetic literature on parameter estimation (and hypotheses tests). Finally, the imprecise vector of estimated parameters $\tilde{\hat{\mathbf{x}}}$ is constructed, based on a sufficient number of α-*cuts*:

$$\tilde{\hat{\mathbf{x}}}_{\alpha,\min} = \mathbf{x}_0 + (\mathbf{A}^{\mathbf{T}}\mathbf{P}\mathbf{A})^{+} \mathbf{A}^{\mathbf{T}}\mathbf{P}\mathbf{y} - |\mathbf{F}|\left(\tilde{\mathbf{p}}_{\alpha,r}\right), \tag{13_a}$$

$$\tilde{\hat{\mathbf{x}}}_{\alpha,\max} = \mathbf{x}_0 + (\mathbf{A}^{\mathbf{T}}\mathbf{P}\mathbf{A})^{+} \mathbf{A}^{\mathbf{T}}\mathbf{P}\mathbf{y} + |\mathbf{F}|\left(\tilde{\mathbf{p}}_{\alpha,r}\right), \tag{13_b}$$

$$m_{\tilde{\tilde{\mathbf{x}}}}(x) = \int_0^1 m_{\tilde{\tilde{\mathbf{x}}}_\alpha}(x)\,d\alpha \qquad \text{and} \qquad m_{\tilde{\tilde{\mathbf{x}}}_\alpha} = \left[\tilde{\tilde{\mathbf{x}}}_{\alpha,\min}, \tilde{\tilde{\mathbf{x}}}_{\alpha,\max}\right], \tag{13\_c}$$

with the matrix of partial derivatives $\mathbf{F} = \dfrac{\partial \mathbf{x}}{\partial \mathbf{p}}$ and $|\square|$ denoting the element-by-element absolute value of the matrix.

2.2.2. *A linear hypothesis for the standard model in parameter estimation*

The standard model in parameter estimation is given by

$$\mathrm{E}(\mathbf{y}) = \mathbf{A}\mathbf{x}, \tag{14}$$

where the expected value of the reduced observations $\mathrm{E}(\mathbf{y})$ equals $\mathbf{A}\mathbf{x}$. The null hypothesis of a linear hypothesis is then introduced as:

$$H_0: \quad \mathbf{C}\mathbf{x} = \mathbf{w}, \tag{15\_a}$$

provided that $\mathbf{C}\mathbf{x}$ must be a testable hypothesis, cf. (Koch, 1999) for details concerning the matrix $\mathbf{C}$ and the vector $\mathbf{w}$. The null hypothesis must be compared with the alternative hypothesis

$$H_A: \quad \mathbf{C}\mathbf{x} = \bar{\mathbf{w}} \neq \mathbf{w}. \tag{15\_b}$$

This leads after a few calculation steps to a quadratic form:

$$T = (\mathbf{C}\hat{\mathbf{x}} - \mathbf{w})^T \left[\mathbf{C}\left(\mathbf{A}^\mathbf{T}\mathbf{P}\mathbf{A}\right)^+ \mathbf{C}^\mathbf{T}\right]^+ (\mathbf{C}\hat{\mathbf{x}} - \mathbf{w}) \;\square\; \chi^2(h,0) \text{ under } \mathrm{H}_0, \tag{16}$$

that follows under the null hypothesis a central chi-square distribution ($\lambda = 0$) with $h = \operatorname{rank}\left[\mathbf{C}\left(\mathbf{A}^\mathbf{T}\mathbf{P}\mathbf{A}\right)^+ \mathbf{C}^\mathbf{T}\right]$ degrees of freedom. In order to avoid overestimation in imprecise hypothesis testing, the general form of a linear hypothesis has to be converted to a quadratic form of imprecise influence parameters $\tilde{\mathbf{p}}$; it is obtained as:

$$\tilde{T}_{\alpha,\min} = \min\left(\begin{bmatrix}\boldsymbol{\Delta}\mathbf{p}\\ \mathbf{y}_m\\ \mathbf{w}\end{bmatrix}^T \begin{bmatrix}\mathbf{F}^\mathbf{T}\mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{K}\mathbf{F} & \mathbf{F}^\mathbf{T}\mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{K} & -\mathbf{F}^\mathbf{T}\mathbf{K}^\mathbf{T}\mathbf{D}\\ \mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{K}\mathbf{F} & \mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{K} & -\mathbf{K}^\mathbf{T}\mathbf{D}\\ -\mathbf{D}\mathbf{K}\mathbf{F} & -\mathbf{D}\mathbf{K} & \mathbf{D}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Delta}\mathbf{p}\\ \mathbf{y}_m\\ \mathbf{w}\end{bmatrix}\right), \tag{17\_a}$$

$$\tilde{T}_{\alpha,\max} = \max\left(\begin{bmatrix}\boldsymbol{\Delta}\mathbf{p}\\ \mathbf{y}_m\\ \mathbf{w}\end{bmatrix}^T \begin{bmatrix}\mathbf{F}^\mathbf{T}\mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{K}\mathbf{F} & \mathbf{F}^\mathbf{T}\mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{K} & -\mathbf{F}^\mathbf{T}\mathbf{K}^\mathbf{T}\mathbf{D}\\ \mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{K}\mathbf{F} & \mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{K} & -\mathbf{K}^\mathbf{T}\mathbf{D}\\ -\mathbf{D}\mathbf{K}\mathbf{F} & -\mathbf{D}\mathbf{K} & \mathbf{D}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Delta}\mathbf{p}\\ \mathbf{y}_m\\ \mathbf{w}\end{bmatrix}\right), \tag{17\_b}$$

$$m_{\tilde{\tilde{T}}}(x) = \int_0^1 m_{\tilde{\tilde{T}}_\alpha}(x)\,d\alpha \qquad \text{and} \qquad m_{\tilde{\tilde{T}}_\alpha} = \left[\tilde{T}_{\alpha,\min}, \tilde{T}_{\alpha,\max}\right]. \tag{17\_c}$$

with $\Delta\mathbf{p} \in \tilde{\mathbf{p}}_\alpha = \left[\tilde{\mathbf{p}}_{\alpha,\min} - \mathbf{p}_\mathrm{m} \;,\; \tilde{\mathbf{p}}_{\alpha,\max} - \mathbf{p}_\mathrm{m}\right]$, $\mathbf{K} = \mathbf{C}^\mathbf{T}\left(\mathbf{A}^\mathbf{T}\mathbf{P}\mathbf{A}\right)^+ \mathbf{A}^\mathbf{T}\mathbf{P}$, $\mathbf{D} = \left[\mathbf{C}\left(\mathbf{A}^\mathbf{T}\mathbf{P}\mathbf{A}\right)^+ \mathbf{C}^\mathbf{T}\right]^+$ and $\mathbf{y}_m$ the midpoint of the reduced observations.

2.2.3. *A linear hypothesis for an extended model in parameter estimation*

The presented strategy from Section 2.2.2 has some shortcomings concerning the computational complexity. If additional parameters $\mathbf{z}$, e. g., in model selection and outlier detection shall be tested in the given enviroment, the model from Equation (15) has to be reformulated and must be fully analyzed (including the inversion of the normal equations). This problem can be overcome by an extended model in parameter estimation:

$$\mathrm{E}(\mathbf{y}) = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}\,. \tag{18}$$

The linear hypothesis may then be given by:

$$H_0: \quad \mathbf{C}\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix} = \mathbf{w} \qquad \text{versus} \qquad H_A: \quad \mathbf{C}\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix} = \bar{\mathbf{w}} \neq \mathbf{w}\,. \tag{19}$$

Starting with the extended normal equations (Koch, 1999)

$$\begin{bmatrix}\mathbf{A}^\mathbf{T}\mathbf{P}\mathbf{A} & \mathbf{A}^\mathbf{T}\mathbf{P}\mathbf{B}\\ \mathbf{B}^\mathbf{T}\mathbf{P}\mathbf{A} & \mathbf{B}^\mathbf{T}\mathbf{P}\mathbf{B}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{x}}\\ \hat{\mathbf{z}}\end{bmatrix} = \begin{bmatrix}\mathbf{A}^\mathbf{T}\mathbf{P}(\mathbf{l} - \mathbf{a}_0)\\ \mathbf{B}^\mathbf{T}\mathbf{P}(\mathbf{l} - \mathbf{a}_0)\end{bmatrix}, \tag{20}$$

this procedure leads after a few calculation steps to a modified quadratic form:

$$T = (\mathbf{C}\begin{bmatrix}\hat{\mathbf{x}}\\ \hat{\mathbf{z}}\end{bmatrix} - \mathbf{w})^T \left[\mathbf{C}\begin{bmatrix}\mathbf{A}^\mathbf{T}\mathbf{P}\mathbf{A} & \mathbf{A}^\mathbf{T}\mathbf{P}\mathbf{B}\\ \mathbf{B}^\mathbf{T}\mathbf{P}\mathbf{A} & \mathbf{B}^\mathbf{T}\mathbf{P}\mathbf{B}\end{bmatrix}^+ \mathbf{C}^\mathbf{T}\right]^+ (\mathbf{C}\begin{bmatrix}\hat{\mathbf{x}}\\ \hat{\mathbf{z}}\end{bmatrix} - \mathbf{w}) \sim \chi^2(j, 0) \text{ under } \mathrm{H}_0\,. \tag{21}$$

This quadratic form follows under the null hypothesis a central chi-square distribution ($\lambda = 0$) with $j = \mathrm{rank}\left[\mathbf{C}\begin{bmatrix}\mathbf{A}^\mathbf{T}\mathbf{P}\mathbf{A} & \mathbf{A}^\mathbf{T}\mathbf{P}\mathbf{B}\\ \mathbf{B}^\mathbf{T}\mathbf{P}\mathbf{A} & \mathbf{B}^\mathbf{T}\mathbf{P}\mathbf{B}\end{bmatrix}^+ \mathbf{C}^\mathbf{T}\right]$ degrees of freedom. If only the additional parameters $\mathbf{z}$ have to be tested, the null hypothesis H_0 can be reformulated as follows:

$$H_0: \quad \mathbf{C}\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix} = \begin{bmatrix}\mathbf{C}_1 & \mathbf{C}_2\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix} = \begin{bmatrix}\mathbf{0} & \mathbf{C}_2\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix} = \mathbf{w} = \begin{bmatrix}\mathbf{0}\\ \mathbf{w}_2\end{bmatrix}, \tag{22}$$

and the quadratic form is now easy to handle (Koch, 1999):

$$T = (\mathbf{C}_2\hat{\mathbf{z}} - \mathbf{w}_2)^\mathbf{T}\left[\mathbf{C}_2\left(\mathbf{B}^\mathbf{T}\left(\mathbf{P} - \mathbf{P}\mathbf{A}(\mathbf{A}^\mathbf{T}\mathbf{P}\mathbf{A})^+\mathbf{A}^\mathbf{T}\mathbf{P}\right)\mathbf{B}\right)^{-1}\mathbf{C}_2^\mathbf{T}\right]^+ (\mathbf{C}_2\hat{\mathbf{z}} - \mathbf{w}_2) \sim \chi^2(j, 0) \text{ under } \mathrm{H}_0\,. \tag{23}$$

According to Section 2.2.2 this quadratic form from Equation (23) has to be converted to a quadratic form of imprecise influence parameters $\tilde{\mathbf{p}}$. With $\hat{\mathbf{z}} = \left(\mathbf{B}^{\mathrm{T}}\mathbf{P}\mathbf{Q}_{\hat{v}\hat{v}}\mathbf{P}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}\mathbf{Q}_{\hat{v}\hat{v}}\mathbf{P}(\mathbf{l}-\mathbf{a}_0)$ and $\mathbf{Q}_{\hat{v}\hat{v}} = \left(\mathbf{P}^{-1} - \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{P}\mathbf{A})^{+}\mathbf{A}^{\mathrm{T}}\right)$, $\mathbf{J} = \left(\mathbf{B}^{\mathrm{T}}\mathbf{P}\mathbf{Q}_{\hat{v}\hat{v}}\mathbf{P}\mathbf{B}\right)^{-1}\mathbf{C}_2\mathbf{B}^{\mathrm{T}}\mathbf{P}\mathbf{Q}_{\hat{v}\hat{v}}\mathbf{P}$ and $\mathbf{M} = \left[\mathbf{C}_2\left(\mathbf{B}^{\mathrm{T}}\mathbf{P}\mathbf{Q}_{\hat{v}\hat{v}}\mathbf{P}\mathbf{B}\right)^{-1}\mathbf{C}_2^{\mathrm{T}}\right]^{+}$ we obtain:

$$\tilde{T}_{\alpha,\min} = \min\left(\begin{bmatrix}\Delta\mathbf{p}\\ \mathbf{y}_m\\ \mathbf{w}_2\end{bmatrix}^T \begin{bmatrix}\mathbf{F}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{M}\mathbf{J}\mathbf{F} & \mathbf{F}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{M}\mathbf{J} & -\mathbf{F}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{M}\\ \mathbf{J}^{\mathrm{T}}\mathbf{M}\mathbf{J}\mathbf{F} & \mathbf{J}^{\mathrm{T}}\mathbf{M}\mathbf{J} & -\mathbf{J}^{\mathrm{T}}\mathbf{M}\\ -\mathbf{M}\mathbf{J}\mathbf{F} & -\mathbf{M}\mathbf{J} & \mathbf{M}\end{bmatrix}\begin{bmatrix}\Delta\mathbf{p}\\ \mathbf{y}_m\\ \mathbf{w}_2\end{bmatrix}\right), \qquad (24\_a)$$

$$\tilde{T}_{\alpha,\max} = \max\left(\begin{bmatrix}\Delta\mathbf{p}\\ \mathbf{y}_m\\ \mathbf{w}_2\end{bmatrix}^T \begin{bmatrix}\mathbf{F}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{M}\mathbf{J}\mathbf{F} & \mathbf{F}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{M}\mathbf{J} & -\mathbf{F}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{M}\\ \mathbf{J}^{\mathrm{T}}\mathbf{M}\mathbf{J}\mathbf{F} & \mathbf{J}^{\mathrm{T}}\mathbf{M}\mathbf{J} & -\mathbf{J}^{\mathrm{T}}\mathbf{M}\\ -\mathbf{M}\mathbf{J}\mathbf{F} & -\mathbf{M}\mathbf{J} & \mathbf{M}\end{bmatrix}\begin{bmatrix}\Delta\mathbf{p}\\ \mathbf{y}_m\\ \mathbf{w}_2\end{bmatrix}\right), \qquad (24\_b)$$

$$m_{\tilde{T}}(x) = \int_0^1 m_{\tilde{T}_\alpha}(x)\,d\alpha \quad \text{and} \quad m_{\tilde{T}_\alpha} = \left[\tilde{T}_{\alpha,\min}, \tilde{T}_{\alpha,\max}\right]. \qquad (24\_c)$$

The quadratic form from the Equations (23) and (24) is computable from the residuals without a new parameter estimation. Therefore the computational complexity is significant reduced.

2.2.4. *Final Test decision based on the card criterion*

The fuzzy evaluation of the quadratic forms from the Equations (17) and (24) is based on Zadeh's extension principle (Zadeh 1965), which can be equivalently replaced by the min-max operator of an optimization algorithm, cf. (Dubois and Prade, 1980, p. 37) for the theoretical concept and (Möller and Beer, 2004) for applications in civil engineering. The optimization problem can be solved, e. g., with a standard Newton algorithm, cf. (Coleman and Li, 1996). Figure 3 shows a constructed test value $\tilde{T}$ and the comparison of the imprecise test value with the imprecise regions of acceptance $\tilde{A}$ and rejection $\tilde{R}$ (Neumann et al., 2006).

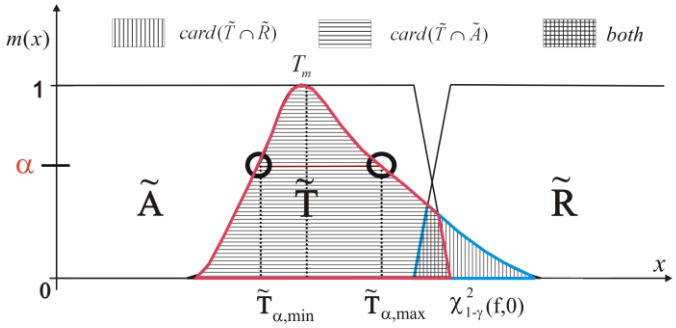


Figure 3. Comparison of the constructed test value $\tilde{T}$ with the regions of acceptance $\tilde{A}$ and rejection $\tilde{R}$

Whereas the influence of imprecision on the test decision for a smaller number of observations is unimportant, it gets more important for a larger number of observations. This is in full accordance to the theoretical concept, because the goodness of fit for the stochastic uncertainty of the parameters increases with the number of observations.

The final test decision is based on the set-theoretical comparison of the imprecise test value (constructed using an α-*cut* optimization algorithm) with the region of acceptance $\tilde{A}$ and the region of rejection $\tilde{R}$ (see Fig. 3), cf. (Kutterer 2004) and (Neumann et al. 2006) for detailed explanations. The hypotheses are defined by

$$T_m \sim \chi^2(k,\lambda)\ ; \quad \lambda \begin{cases} = 0 & \mid \mathrm{H}_0 \text{ the null hypothesis,} \\ \neq 0 & \mid \mathrm{H}_\mathrm{A} \text{ the alternative hypothesis,} \end{cases} \tag{25}$$

with the non-centrality parameter λ. The midpoint of the test value follows under the null hypothesis a central chi-square distribution with $k \in \{h\ , j\}$ degrees of freedom. The regions of acceptance $\tilde{A}$ and rejection $\tilde{R} = \tilde{A}^C$ are defined as fuzzy intervals. The degree of the rejectability $\rho_R(\tilde{T})$ of the null hypothesis H_0 under the condition of $\tilde{T}$ is computed based on the degree of agreement of the test value with the region of rejection $\gamma_{\tilde{R}}(\tilde{T})$ and the degree of disagreement of the test value with the region of acceptance $\delta_{\tilde{A}}(\tilde{T})$. We use the card criterion, because it allows a more suitable description of the degree of agreement between two fuzzy intervals. This leads to the equations given below (see also Fig. 3):

$$\gamma_{\tilde{R}}(\tilde{T}) = \frac{card\left(\tilde{T} \cap \tilde{R}\right)}{card\left(\tilde{T}\right)} \quad \text{and} \quad \delta_{\tilde{A}}(\tilde{T}) = 1 - \frac{card\left(\tilde{T} \cap \tilde{A}\right)}{card\left(\tilde{T}\right)} \tag{26\_a}$$

$$\rho_{\tilde{R}}(\tilde{T}) = \min\left(\gamma_{\tilde{R}}(\tilde{T}), \delta_{\tilde{A}}(\tilde{T})\right) \tag{26\_b}$$

For the final test decision, the degree of rejectability $\rho_{\tilde{R}}(\tilde{T})$ of the null hypothesis has to be compared with a suitable critical value $\rho_{crit} \in [0,1]$:

$$\rho_{\tilde{R}}\left(\tilde{T}\right) \begin{Bmatrix} \leq \\ > \end{Bmatrix} \rho_{\mathrm{crit}} \in [0,1] \quad \Rightarrow \quad \begin{cases} \text{Do not reject } \mathrm{H}_0 \\ \text{Reject } \mathrm{H}_0 \end{cases} \tag{27}$$

The test is only rejected, if the test value agrees with the region of rejection and disagrees with the region of acceptance. This is in full accordance with the theoretical expectations, where observation imprecision is an additive term of uncertainty during the measurement process. The choice of ρ_{crit} depends on the particular application and must be based on expert knowledge. For outlier detection we propose to choose $\rho_{crit} \to 1$ and for safety-relevant measures $\rho_{crit} \to 0$.

3. Probability of type I errors in imprecise hypothesis testing

In this subsection we focus on the relation of imprecision and the probability of a type I error. The probability γ_{impr} of a type I error in the imprecise case is defined by:

$$\gamma_{impr} = P\left(\rho_{\tilde{R}}\left(\tilde{T}\right) > \rho_{crit} \,|H_0\right). \tag{28}$$

The index „impr“ denotes the case of imprecision. Equation (28) can be reformulated as follows

$$\gamma_{impr} = P\left(\mathrm{f}\,(T_m) > \rho_{crit} \,|H_0\right), \tag{29}$$

with the degree of rejectability $\rho_{\tilde{R}}(\tilde{T})$ of the null hypothesis under the condition of $\tilde{T}$ as a function f of the midpoint T_m of the imprecise test value $\tilde{T}$.

$$\gamma_{\text{impr}} = \mathrm{P}\left(\mathrm{f}^{-1}(\rho_{\text{crit}}) > \mathrm{T_m} \,|\mathrm{H}_0\right). \tag{30}$$

This leads with respect to Equation (30) after a few calculation steps to the quantile value $\chi^2_{1-\gamma_{\text{impr}}}$ of the chi-square distribution (k degrees of freedom) in the imprecise case

$$\chi^2_{1-\gamma_{impr}}(k,0) = \mathrm{f}^{-1}(\rho_{crit}), \tag{31}$$

with f^{-1} denoting the inverse function of f. In order to illustrate the theoretical concept, an example will be shown in Section 5. See (Kutterer, 2004) for a close mathematical formulation in case of classical regions of acceptance and rejection in the one-dimensional case. Based on the quadratic form from Equation (24), the influence of imprecision on the tests decision is analyzed for different positions for the midpoint T_m of the test value (see figure 4).

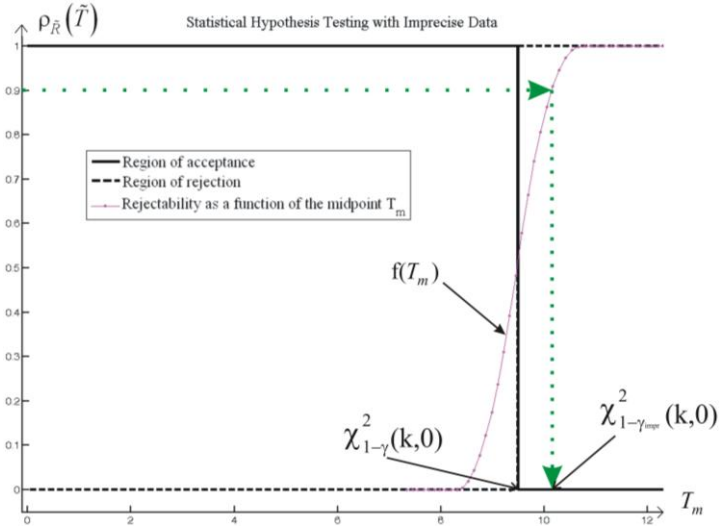


Figure 4. Calculation of the probability of a type I error in the imprecise case (for $\rho_{crit} = 0.9$)

The calculation of the probability of a type I error is then easy to handle and can be solved by the followings steps:

Step 1: Choose an adequate value for ρ_{crit} (see Section 2.2.4).

Step 2: Compute $f^{-1}(\rho_{crit}) \rightarrow \chi^2_{1-\gamma_{impr}}(k,0)$.

Step 3: Find γ_{impr} in such a way that Equation (31) is fulfilled within a negligible threshold.

4. Probability of type II errors in imprecise hypothesis testing

The probability β_{impr} of a type II error in the imprecise case can be derived by

$$\beta_{impr} = P\left(\rho_{\tilde{R}}\left(\tilde{T}\right) \leq \rho_{crit} \middle| H_A\right). \tag{32}$$

According to the probability of a type I error, Equation (32) can be reformulated as follows

$$\beta_{impr} = P\left(f(T_m) \leq \rho_{crit} \middle| H_A\right), \tag{33}$$

with the degree of rejectability $\rho_{\tilde{R}}(\tilde{T})$ of the null hypothesis under the condition of $\tilde{T}$ as a function f of the midpoint T_m of the imprecise test value $\tilde{T}$. In order to analyze Equation (33), either the non-centrality parameter λ_{impr} in the imprecise case or the probability β_{impr} of a type II error in the imprecise case must be set in advance. This leads after a few calculation steps to the comparison of two chi-square distributions (with k degrees of freedom). The first central chi-square distribution is related to the probability of a type I error in the imprecise case and the second one (with the non-centrality parameter λ_{impr} in the imprecise case) is related to the probability of a type II error.

$$\chi^2_{1-\gamma_{impr}}(k,0) = \chi^2_{\beta_{impr}}(k,\lambda_{impr}). \tag{34}$$

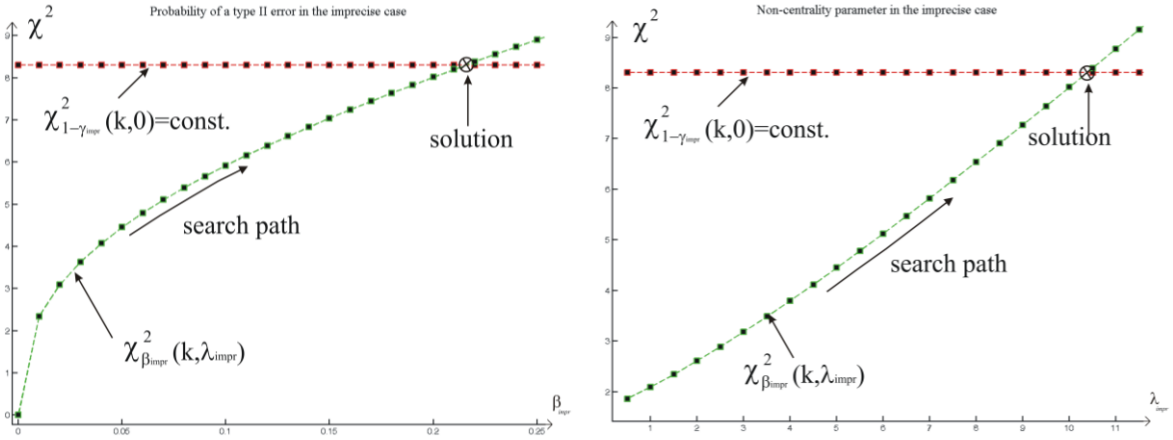
The calculation of the probability of a type II error and of the non-centrality parameter in the imprecise case can be seen as the following search problem (see figure 5a and 5b):

1. Calculation of the type II error in imprecise hypothesis testing:

   Step 1: Compute the probability of a type I error in the imprecise case (see Section 3).
   Step 2: Choose an adequate value for λ_{impr}.
   Step 3: Find β_{impr} in such a way that Equation (34) is fulfilled within a negligible threshold.

2. Calculation of the non-centrality parameter in imprecise hypothesis testing:

   Step 1: Compute the probability of a type I error in the imprecise case (see Section 3).
   Step 2: Choose an adequate value for β_{impr}.
   Step 3: Find λ_{impr} in such a way that Equation (34) is fulfilled within a negligible threshold.



a. Calculation of a type II error ($\lambda_{impr} = const$)

b. Calculation of the non-centrality parameter ($\beta_{impr} = const$)

Figure 5. Calculation of the probability of a type II error (a) and the non-centrality parameter (b) in the imprecise case

5. Example for the parameterization of a geodetic monitoring network

In order to illustrate the theoretical concept, three exemplary applications in the parameterization of a geodetic monitoring network are presented. The aim of the geodetic monitoring network is to detect significant changes of a lock due to changing water levels inside the lock. Figure 6 shows the lock and the geodetic monitoring network, which consist of four object points on top of the lock (101-104) and eight control points around the lock; see (Neumann et al., 2006) for a detailed description about the geodetic monitoring network.

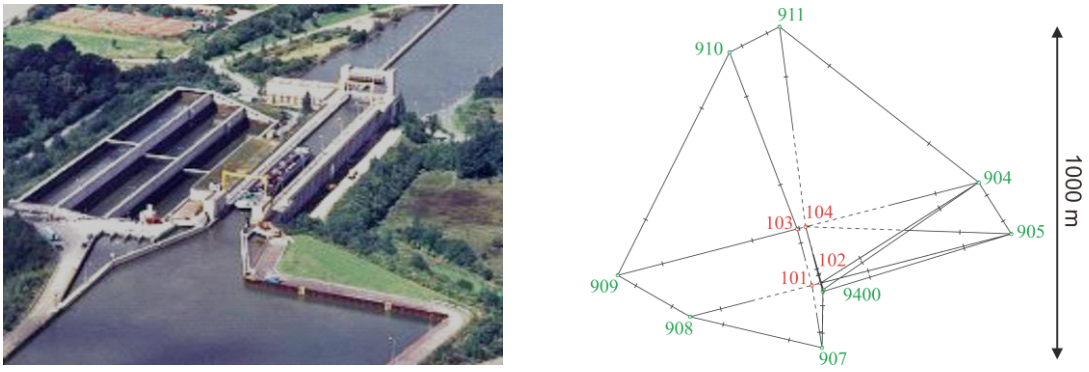


Figure 6 – The lock and the geodetic monitoring network

The coordinates of the object points are estimated within a least-squares adjustment. Therefore special geodetic measurements like horizontal directions (a), zenith angles (b) and distances (c) were carried out

between the object and control points. The measurements are affected by different types of uncertainty (see Table 1 and Section 2.1). The non-stochastic uncertainties are analyzed within a sensitivity analysis (see Table 1).

Influence factors **p**	*Interval radii* ($\alpha = 0$) *(imprecision)*	*Affected measurements*
Temperature	1.0 °C	(c)
Pressure	1.0 hPa	(c)
Visual axis error	0.1 mgon	(a)
Collimation error	0.1 mgon	(a)
Vertical axis error	0.2 mgon	(a) and (b)

a. Main influence factors for the observations

Observations	*Interval radii* ($\alpha = 0$) *(imprecision)*	*Standard deviation*
Horizontal direction	0.1 mgon	0.5 mgon
Zenith angle	0.5 mgon	1.5 mgon
Distance	0.5 mm	3 mm

b. Uncertainties of the observations

Table 1. Influence factors and uncertainties of the observations

First we focus on a single and multiple outlier test. Then a congruence test is evaluated in terms of imprecision. For a straightforward comparison to the pure stochastic case, the region of acceptance is given by a classical interval with a significance level of $\gamma = 5\%$. All computations are based on 11 different $\alpha - cuts$.

5.1. EXAMPLES IN OUTLIER DETECTION

5.1.1. *Testing procedure for a single measurement*

The first example shows an outlier test for a distance measurement. The construction of the test value $\tilde{T}$ is based on the imprecise evaluation of the quadratic form in Equation (24) with an $\alpha - cut$ optimization method.

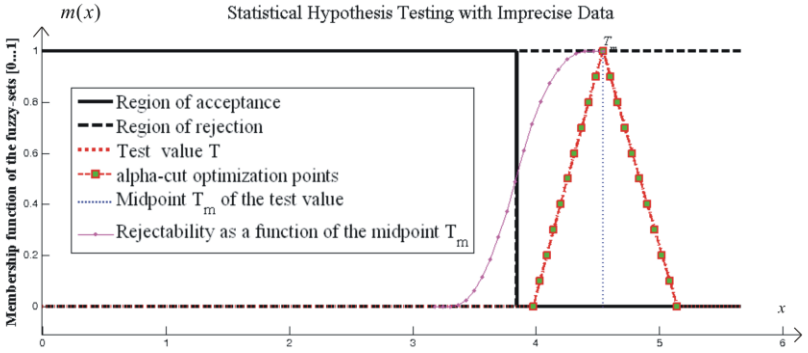


Figure 7 – The degree of rejectability $\rho_{\tilde{R}}(\tilde{T})$ of a single outlier test as a function of the midpoint T_m of the test value $\tilde{T}$

Figure 7 shows the degree of rejectability $\rho_{\tilde{R}}(\tilde{T})$ of the null hypothesis H_0 under the condition of $\tilde{T}$ as a function f of the midpoint T_m of the imprecise test value $\tilde{T}$. Obviously, in this example the observation imprecision is small in comparison to the stochastic uncertainty. For this reason, the test value is tight and close to symmetric.

The probability of a type I error in the imprecise case γ_{impr} is strongly related to the choice of the critical value ρ_{crit} for the test decision, see Equation (31). Figure 8 shows the probability of a type I error in relationship to the choice of ρ_{crit}.

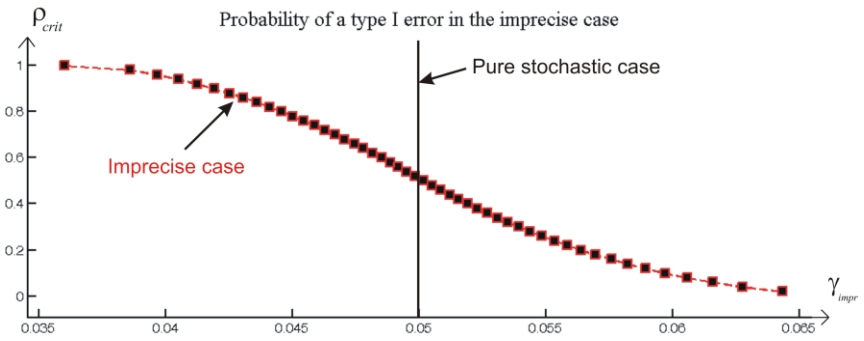


Figure 8 – Probability of a type I error in the imprecise case for a single outlier test (depending on the choice of ρ_{crit})

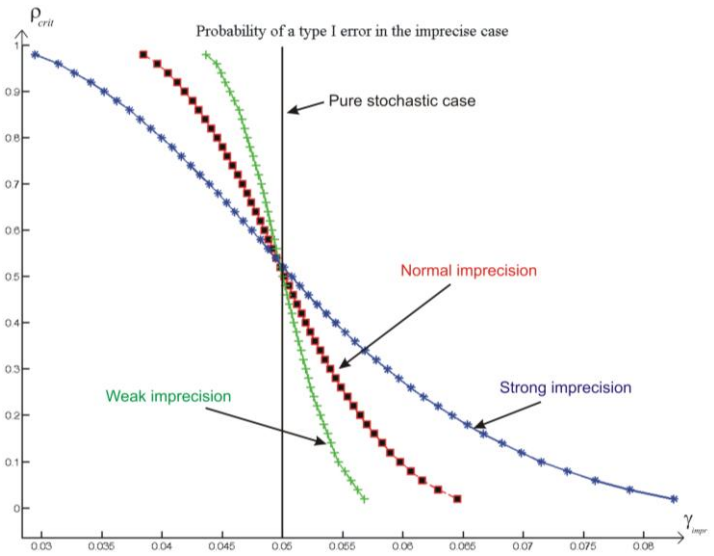


Figure 9 – Variation of the probability of a type I error in the imprecise case (depending on the choice of ρ_{crit} and the order of magnitude of imprecision)

The choice of ρ_{crit} depends on the particular application and must be based on expert knowledge. For outlier detection we propose to choose $\rho_{crit} \to 1$ and for safety-relevant measures $\rho_{crit} \to 0$. The variation of a type I error in the imprecise case γ_{impr} depends also on the order of magnitude of imprecision. If imprecision is more important in comparison to the stochastic uncertainty, the variation of a type I error in the imprecise case increases. Figure 9 shows an example with strong imprecision (twice of the imprecision of Table 1), normal imprecision and small imprecision (half of the imprecision of Table 1):

5.1.2. *Testing procedure for multiple measurements*

The second example shows a multiple outlier test due to an assumed centering error of the instrument, while measuring a set of distances at station 102. The construction of the test value $\tilde{T}$ is based on the imprecise evaluation of the quadratic form in Equation (24) with the $\alpha - cut$ optimization method. In this example, the number of tested observations is four ($j = 4$). Figure 10 shows the probability of a type I error in the imprecise case γ_{impr} .

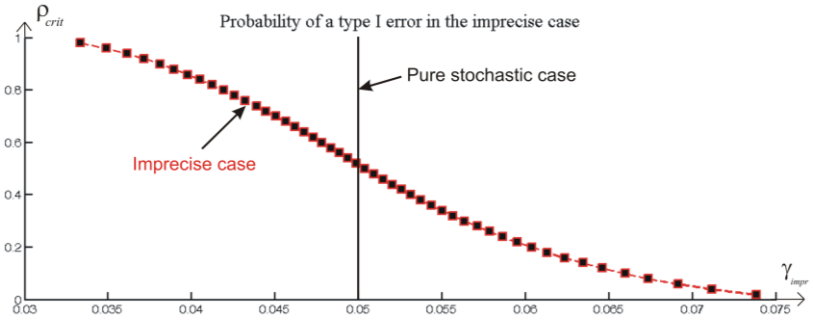


Figure 10 – Probability of a type I error in the imprecise case for a multiple outlier test (depending on the choice of ρ_{crit})

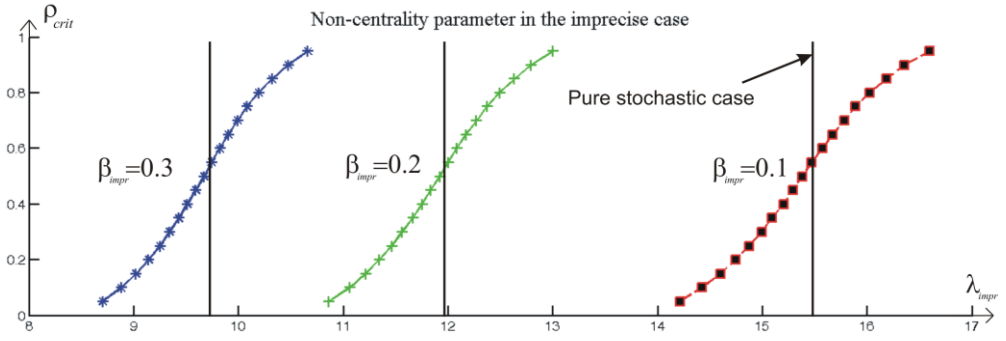


Figure 11 – The non-centrality parameter in the imprecise case (depending on the choice of ρ_{crit} and β_{impr})

To study the influence of imprecision on the probability β_{impr} of a type II error and the non-centrality parameter λ_{impr} in the imprecise case, it is more meaningful to hold the probability of a type II error constant. We focus on three standard cases with $\beta_{impr} = 0.1$, $\beta_{impr} = 0.2$ and $\beta_{impr} = 0.3$. The non-centrality parameter is obtained by the search problem described in Section 4 (see Figure 11). For $\rho_{crit} > 0.5$ the non-centrality parameter in the imprecise case is greater than in the precise case. This leads to a reduced sensitivity regarding the rejection of the null hypothesis.

5.2. EXAMPLE FOR A CONGRUENCE TEST (EPOCH COMPARISON)

The third example demonstrates an epoch comparison between the years 1999 and 2004. Both epochs are estimated within a partially constrained trace minimization with respect to the same six network points. The construction of the test value $\tilde{T}$ is based on the imprecise evaluation of the quadratic form from Equation (17). Figure 12 shows the numerical test situation with the probability of a type I error in the imprecise case and Table 2 some specifications about the two epochs and the geodetic monitoring network. Please note that the configurations in both epochs are different from each other.

Specification	Epoch 1999	Epoch 2004
Observations n	317	144
Parameters u	60	39

Table 2. Specifications about the geodetic monitoring network in the epochs 1999 and 2004

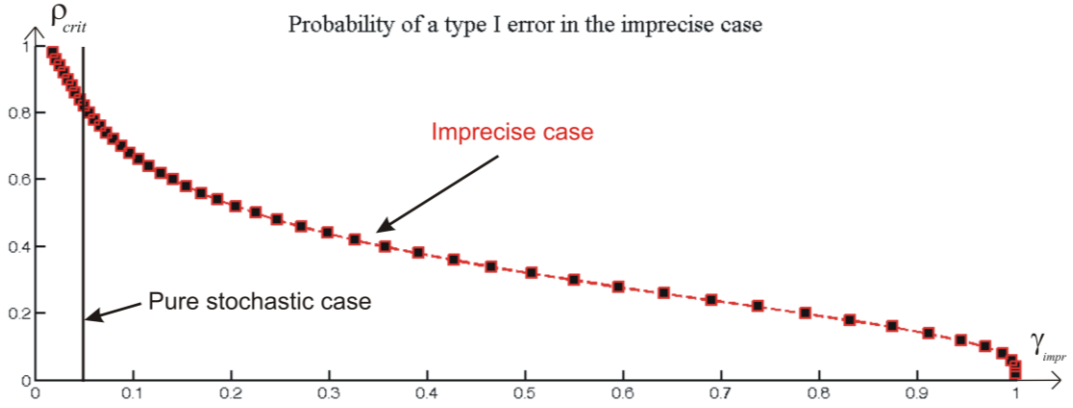


Figure 12 – Probability of a type I error in the imprecise case for a congruence test (depending on the choice of ρ_{crit})

The significant imprecision of the test value in this example is caused by the strong effects of remaining systematics in epoch comparison of a geodetic monitoring network. The influence of imprecision on the given test situation depends also on the geometric configuration of the geodetic monitoring network. Whereas a weak configuration leads to a wider expansion of the test value, a strong configuration

decreases the influence of imprecision in the test situation. The strong imprecision leads to a wide variation of the probability of a type I error in the imprecise case. In case of $\rho_{crit} \to 0$ the null hypothesis will be rejected in any rate. For this reason, the probability of a type I error in the imprecise case (for $\rho_{crit} \to 0$) is equal to one.

6. Conclusions

In this paper, we show a joint treatment of stochastic and interval/fuzzy uncertainty (*imprecision*) in hypothesis testing in parameter estimation. Imprecision is considered as an additive term of uncertainty what leads to a more reluctant rejection of the null hypothesis in case of outlier detection and to an earlier rejection of the null hypothesis in case of safety-relevant applications. If imprecision is absent, the results of the pure stochastic case are obtained. We focus on the probability of a type I and type II error and the non-centrality parameter in the imprecise case. In case of outlier detection the probability of a type I error in the imprecise case is lower than in the pure stochastic case and the non-centrality parameter in the imprecise case is greater than in the pure stochastic case. In order to detect the same changes than in the pure stochastic case, e. g., in a risk analysis, more precise measurements have to be carried out.

However, the quantification of the uncertainty budget of empirical measurements is often too optimistic due to, e.g., the ignorance of non-stochastic errors in the analysis process (Ferson et al., 2007). For this reason the above mentioned results in this paper are in our opinion closer to the situation in real-world applications. In addition, the well known sensitivity analysis in parameter estimation can now generally be treated in terms of imprecise data to decide about a suitable model for the collected data.

Further work has to deal with a significant reduction of the computational complexity of the numerical solutions. In addition, it seems to be very promising, that for special types of reference functions analytic solutions for type I and type II errors in the imprecise case can be found.

Acknowledgements

The presented paper shows results and new ideas developed during the research project KU 1250/4-1 "Geodätische Deformationsanalysen unter Berücksichtigung von Beobachtungs-impräzision und Objektunschärfe", which is funded by the German Research Foundation (DFG). This is gratefully acknowledged by the authors.

References

Bandemer, H. and Näther, W.: *Fuzzy Data Analysis*. Kluwer, 1992.

Dubois, D.J. and Prade, H.M.: *Fuzzy Sets and Systems: theory and applications*, Academic Press, New York, 1980.

Ferson, S.; Kreinovich, V.; Hajagos, J.; Oberkampf, W. and Ginzburg, L.: *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*. Sandia National Laboratories, SAND2007-0939, 2007.

Koch, K. R.: *Parameter Estimation and Hypothesis Testing in Linear Models*. Springer, Berlin New York, 1999.

Kutterer, H.: *Statistical hypothesis tests in case of imprecise data*. Proc. of the 5. Hotine-Marussi-Symposium, International Association of Geodesy Symposia, Springer, Berlin and New York, 2004, pp. 49-56.

Möller, B. and Beer, M.: *Fuzzy Randomness -Uncertainty in Civil Engineering and Computational Mechanics*-. Springer, Berlin and New York, 2004.

Neumann, I.; Kutterer, H. and Schön, S.: *Outlier detection in geodetic applications with respect to observation imprecision*. Proceedings of the NSF Workshop on Reliable Engineering Computing - Modeling Errors and Uncertainty in Engineering Computations -, Savannah, Georgia, USA, 2006, pp. 75-90.

Zadeh, L.A.: *Fuzzy sets.* In: Information Control, Vol. 8, 1965, pp. 338-353.

Uncertain Processes and Numerical Monitoring of Structures

Wolfgang Graf, Bernd Möller, and Matthias Bartzsch

Institute for Structural Analysis
Technische Universität Dresden
D-01062 Dresden
Germany
email: wolfgang.graf@tu-dresden.de

Abstract. A structure is subjected to numerous alterations and modifications during its lifetime. The entirety of the modifications of structures constitutes the process of modifications. Numerical monitoring of a structure during its lifetime close to reality requires considering the complete load and modification processes simultaneously. Both processes run discontinuously. They cause time dependent, discontinuous result values. The parameters of the load and modification process are usually uncertain parameters. Due to their predominantly informal and lexical uncertainty, they are described as fuzzy processes, respectively fuzzy functions. Taking account of this uncertainty in the nume-rical simulation of the load and modification process requires a fuzzy structural analysis in the time domain. The fuzzy variables and the fuzzy functions are mapped on the fuzzy result variables with the aid of a crisp or uncertain analysis algorithm. The numerical simulation is based on an optimization procedure. This procedure searches for special points in the input space of the fuzzy variables. Each point of the input space represents a deter-ministic parameter data set, which is introduced in a deterministic fundamental solution. In this paper the geometrically and physically nonlinear analysis of plane rein-forced concrete, prestressed concrete, textile concrete, and steel bar structures is chosen as deterministic fundamental solution. The algorithms are demonstrated by way of examples.

Keywords: uncertainty modeling, numerical monitoring, nonlinear numerical analysis

1. Numerical Monitoring of Structures – Conceptual Idea

Numerical monitoring of structures is the numerical simulation of the behaviour of structures during the lifetime. A structure is subject to numerous alterations during its lifetime. These modifications may result from:

- Sequence of different states during construction
- Changes in material, e.g., the change of material behavior due to physical or chemical processes
- Structural alteration resulting from, e.g., refurbishing, bonding of prestressing elements, strengthening
- Changes in load, described by a loading process

For structural alterations and the sequence of different states during construction the term "system modification" is adopted. The system modification comprises cross section modifi-cation, modification of structural members, and modification of support conditions [Bartzsch, Graf, Möller & Sickert 2004]. The change of prestressing forces may also be understood as system modification. The entirety of the system modifications constitutes the modification process. Analyzing a structure during the lifetime close to reality requires considering the complete load and modification processes simultaneously. Both processes run discontinuous-ly. These processes must be described by means of suitable mechanical models. They cause time dependent, discontinuous result values $\underline{z}(\underline{t})$:

$$\underline{z}(\underline{t}) = f(g(\underline{t}), p(\underline{t}), \underline{F}_p(\underline{t}), \underline{T}(\underline{t}), \underline{A}(\underline{t}), \underline{I}(\underline{t}), \underline{E}(\underline{t})) \tag{1}$$

with

$\underline{z}$	vector of structural responses (e.g., displacements and internal forces)
$\underline{g}(\underline{t})$	dead load
$\underline{p}(\underline{t})$	statically and dynamic external loads
$\underline{F}_p(\underline{t})$	prestressing forces (internal and external prestressing)
$\underline{T}(\underline{t})$	parameters of temperature
$\underline{A}(\underline{t})$, $\underline{I}(\underline{t})$	parameters of geometry representing time dependent values in the modification process (e.g., cross sections, dimensions of the system, location of the reinforcement, and the prestressing elements)
$\underline{E}(\underline{t})$	material parameters
$\underline{t} = (\underline{\theta}, \tau, \underline{\varphi})$	spatial coordinates $\underline{\theta} = \theta_1, \theta_2, \theta_3$, time τ, further parameters $\underline{\varphi}$, e.g. temperature

The parameters of the load and modification process are usually uncertain parameters. The following mathematical models are available to describe uncertainty (see also Figure 1):

- Randomness
- Fuzziness
- Fuzzy randomness

whereas fuzziness and randomness are considered as special cases of the general model fuzzy randomness [Möller & Beer 2004]. The choice of the model depends on the available data.

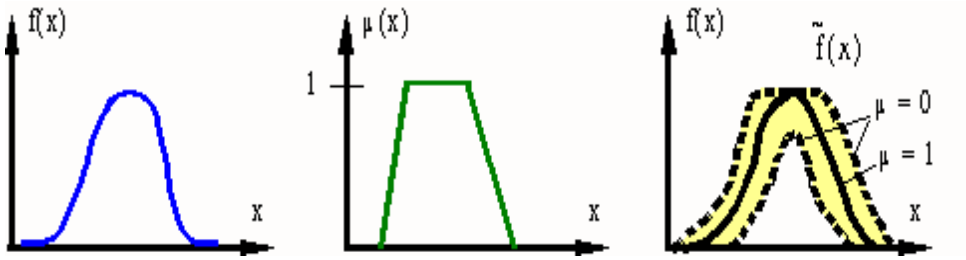


Figure 1. Mathematical models of uncertainty

If sufficient statistical data exist for a parameter the parameter may be described stochastically. Thereby the choice of the type of the probability distribution function affects the result considerably. Often statistically not ensured samples exist for a parameter. Then the description by the uncertainty model fuzziness is recommended. The model comprehends both objective and subjective information. The uncertain parameters are characterized by aid of a membership function μ(x), see eq. (1). The membership function assesses the gradual membership of elements to a set [Möller & Beer 2004].

$$\tilde{x}=\{(x;\mu_x(x)) \mid x \in X\}; \quad \mu_x(x) \geq 0 \ \ \forall \ \ x \in X \tag{2}$$

The uncertainty model fuzzy randomness is a superordinate model that both stochastic and non-stochastic properties of parameters enclose. Fuzzy random variables are used if, e.g., reproduction conditions vary during the period of observation, or if expert knowledge complements the statistical material. A fuzzy random variable is the fuzzy set of their originals, see eq. (3). The originals are probability functions of random variables.

$$\tilde{f}(x)=\{(f(x);\mu_f(f(x))) \mid f \in f\};$$
$$\mu_f(f(x)) \geq 0 \ \ \forall \ \ f \in f \tag{3}$$

Due to the predominantly informal and lexical fuzziness of the parameters of the load and modification process the uncertain parameters are described by the mathematical model fuzziness. As the parameters are time dependent they are considered as fuzzy functions $\tilde{x}(\underline{\tilde{t}}) = \tilde{x}(\underline{\tilde{\theta}}, \tilde{\tau}, \underline{\tilde{\varphi}})$ or fuzzy processes $\tilde{x}(\tilde{\tau})$.

2. Formal Description of Uncertain Discontinuous Processes

A fuzzy vector $\tilde{\underline{x}}$ describes uncertain parameters at discrete points. A fuzzy function $\tilde{x}(\tilde{\underline{t}})$ enables the formal description of at least piecewise continuous uncertain parameters in $|^1$, $|^2$, or $|^3$. The following definition of fuzzy functions is introduced. Given are

- the fundamental sets **T** ϕ | and **X** ϕ |
- the set **F(T)** of all fuzzy variables $\tilde{\underline{t}}$ on the fundamental set **T**
- the set **F(X)** of all fuzzy variables $\tilde{x}$ on the fundamental set **X**.

An uncertain mapping of **F(T)** to **F(X)** that assigns exactly one $\tilde{x}$ 0 **F(X)** to each $\tilde{\underline{t}}$ 0 **F(T)**, respectively, is referred to as a fuzzy function denoted by

$$\tilde{x}(\tilde{\underline{t}}): \quad \mathbf{F(T)} \xrightarrow{\sim} \mathbf{F(X)} \tag{4}$$

$$\tilde{x}(\tilde{\underline{t}}) = \{(\tilde{x}_f = x(\tilde{\underline{t}}) \quad \forall \quad \tilde{\underline{t}} \mid \tilde{\underline{t}} \in \mathbf{F(T)}\} \tag{5}$$

In system modification the fundamental set **T** may contain both the uncertain time coordinate $\tilde{\tau}$ and the crisp spatial coordinate $\underline{\theta}$. In this case the assigned fuzzy function is denoted by $\tilde{x}(\tilde{\underline{t}}) = \tilde{x}(\underline{\theta}, \tilde{\tau})$ with $\tilde{\underline{t}} = (\underline{\theta}, \tilde{\tau})$. The fuzzy function $\tilde{x}(\underline{\theta}, \tilde{\tau})$ enables the modeling of processes with uncertain time points. This is of interest if the system is modified at non-precise known points in time. If the time points are crisp, the special case

$$\tilde{x}(\underline{\theta}, \tau) = \tilde{x}(\underline{t}) = \{(\tilde{x}_t = \tilde{x}(\underline{t})) \quad \forall \quad \underline{t} \mid \underline{t} \in \mathbf{T})\} \tag{6}$$

is obtained [Möller & Beer 2004]. Figure 2 shows a fuzzy process $\tilde{x}(\underline{\theta}_j, \tau)$ for a specific point with the coordinate $\underline{\theta}_j$.

For the numerical simulation of system modifications the bunch parameter representation of a fuzzy function is applied.

$$x(\tilde{\underline{s}}, \underline{t}) = \{(\tilde{x}_t = x(\tilde{\underline{s}}, \underline{t})) \quad \forall \quad \underline{t} \mid \underline{t} \in \mathbf{F(T)}\} \tag{7}$$

For each crisp bunch parameter vector $\underline{s} \in \tilde{\underline{s}}$ with the assigned membership value $\mu(\underline{s})$ a crisp function $x(\underline{t}) = (x(\underline{s}, \underline{t})) \in \tilde{x}(\underline{t})$ with $\mu(x(\underline{t})) = \mu(\underline{s})$ is obtained. The fuzzy function $\tilde{x}(\tilde{\underline{t}})$ may thus be represented by the fuzzy set of all real valued functions $x(\underline{t}) \in \tilde{x}(\underline{t})$ with $\mu(x(\underline{t})) = \mu(x(\underline{s}, \underline{t})) = \mu(\underline{s})$

$$x(\underline{\tilde{s}}, \underline{t}) = \{(x(\underline{t}), \mu(x(\underline{t}))) \mid x(\underline{t}) = x(\underline{s}, \underline{t})\}; \quad (8)$$
$$\mu(x(\underline{t})) = \mu(\underline{s}) \;\; \forall \; \underline{s} \mid \underline{s} \in \underline{\tilde{s}}$$

which may be generated from all possible real vectors $\underline{s} \in \underline{\tilde{s}}$. For every $\underline{t}$ 0 $\underline{\mathbf{T}}$ takes values which are simultaneously contained in the associated fuzzy functional values $\tilde{x}(\underline{t})$. The real functions x($\underline{t}$) of $\tilde{x}(\underline{t})$ are defined for all $\underline{t}$ 0 $\underline{\mathbf{T}}$. These are referred to as trajectories.

Numerical processing of fuzzy functions $\tilde{x}(\underline{t}) = (x(\underline{\tilde{s}}, \underline{t}))$ demands the discretization of their arguments $\underline{t}$ in space and time.

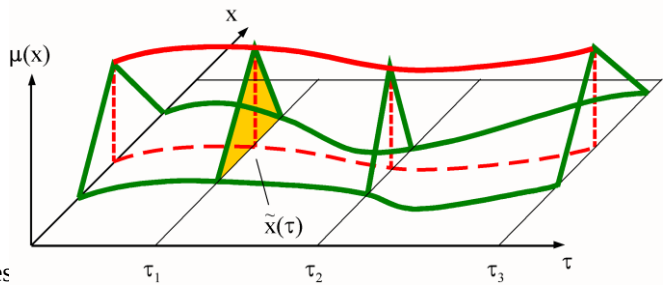


Figure 2. Fuzzy proces

3 Numerical Processing of Uncertain Discontinuous Processes

In deterministic structural analysis crisp structural input vectors $\underline{x}$ containing parameters, for example, for loads, geometrical and material properties are mapped with the aid of a computational model to structural responses such as stresses, internal forces, and displace-ments. This mapping may be denoted as

$$\underline{x} \quad \rightarrow \quad \underline{z} \quad (9)$$

in which the arrow indicates the computational model as the mapping model. This deterministic computational model is subsequently referred to as deterministic fundamental solution within the framework of an uncertain analysis.

If the structural parameters possess uncertainty in the form of fuzziness, eq. (9) may be rewritten as

$$\underline{\tilde{x}} \quad \rightarrow \quad \underline{\tilde{z}} \quad (10)$$

representing a fuzzy structural analysis. The input vectors $\underline{\tilde{x}}$ are then formed by fuzzy structural parameters $\tilde{x}_i$; and the fuzzy structural response vectors $\underline{\tilde{z}} = (..., \tilde{z}_j, ...)$ are determined on the basis of fuzzy set operations. For processing fuzzy quantities through structural computations in a general and

numerically efficient manner a global optimization scheme referred to as α-level optimization has been developed [Möller, Graf & Beer 2000]. This includes a modified evolution strategy as the kernel solution technique.

The concept of α-discretization is applied to numerically represent the fuzzy structural parameters $\tilde{x}_i$ as a set of α-level sets for a sufficiently high umber of α-levels. All fuzzy input parameters are discretized using the same number of α-levels α_k, k = 1 … r. With the aid of the deterministic fundamental solution (mapping model) crisp elements from the fuzzy input vectors, $\underline{x} \in \underline{\tilde{x}}$, are processed to obtain crisp elements of the fuzzy structural response vectors, $\underline{z} \in \underline{\tilde{z}}$. In terms of α-level optimization this means the mapping of $\underline{x} \in \underline{X}_{\alpha k}$ to $\underline{z} \in \underline{Z}_{\alpha k}$, in which $\underline{X}_{\alpha k}$ and $\underline{Z}_{\alpha k}$ are crisp input and result subspaces, respectively, for each α-level. The mapping of all elements of $\underline{X}_{\alpha k}$ yields the crisp subspace $\underline{Z}_{\alpha k}$. Once the largest element $z_{j,\alpha k_r}$ and the smallest element $z_{j,\alpha k_l}$ of the dimension j of the crisp subspace $\underline{Z}_{\alpha k}$ have been found, two points of the membership function $\mu(z_j)$ of the fuzzy result z_j are known. The search for these extreme elements $z_{j,\alpha k_r}$ and $z_{j,\alpha k_l}$ on each α-level represents an optimization problem and is referred to as α-level optimization, see Figure 3. For the detection of $z_{j,\alpha k_r}$ and $z_{j,\alpha k_l}$ with a high probability in general cases with no restrictions regarding the properties of the mapping model, which represents the objective function in the optimization procedure, the modified evolution strategy according to [Möller, Graf & Beer 2000] is employed. This procedure possesses a simple structure, exhibits a reasonable robustness with regard to numerical noise in the mapping model, and can be applied very flexibly in dependence on the problem by adjusting several effective control parameters. The computational costs of the modified evolution strategy increases approximately linearly with the number of dimensions of the problem. For a further improvement of the performance of the procedure a post-computation is carried out after the completion of all optimizations for all α-levels. This includes a recheck of all $z_{j,\alpha k_r}$ and $z_{j,\alpha k_l}$ with the aid all information gathered during all individual optimizations and a re-optimization of those results, which are identified as being not yet optimum. The features robustness, numerical efficiency, and general applicability of the modified evolution strategy enable an application of α-level optimization in combination with arbitrary nonlinear algorithms as mapping models for structural analysis.

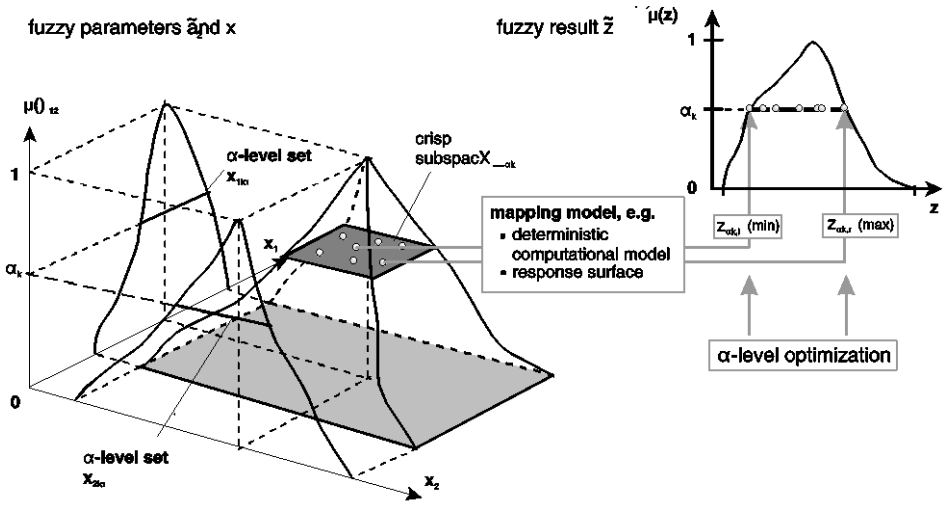


Figure 3a. α-level optimization

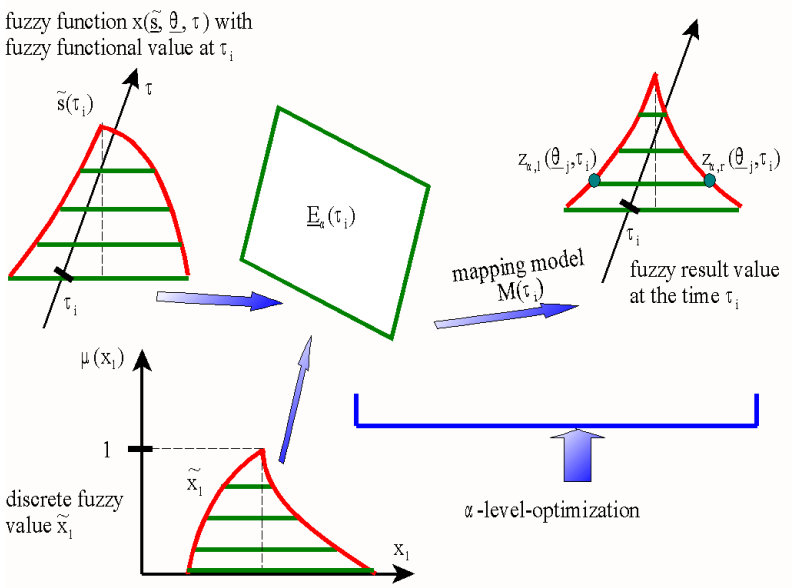


Figure 3b. α-level optimization

The deterministic fundamental solution represents the respective analysis algorithm and is selectable. In this paper the geometrically and physically nonlinear analysis of plane reinforced concrete, prestressed concrete, and steel bar structures [Bartzsch 2006] is chosen as deterministic fundamental solution. The bars are subdivided into integration sections, the cross sections are subdivided into layers. On this basis an incrementally formulated system of second order differential equations for the straight or imperfectly straight bar is obtained. The slip at the bond joint is regarded as an additional degree of freedom s.

$$\left[\frac{d\Delta z(\theta_1)}{d\theta_1}\right]_{(n)}^{[k]} = \underline{A}(\theta_1, \underline{z})_{(n-1)} \cdot \Delta \underline{z}(\theta_1)_{(n)}^{[k]} + \Delta \underline{b}(\theta_1, \underline{z})_{(n)}^{[k-1]} + \\ \ldots + \underline{d}(\theta_1, \underline{z})_{(n-1)} \cdot \Delta \dot{\underline{z}}_1(\theta_1)_{(n)}^{[k]} + \underline{m}(\theta_1, \underline{z})_{(n-1)} \cdot \Delta \ddot{\underline{z}}_1(\theta_1)_{(n)}^{[k]} \tag{11}$$

with

[k]	counter of iteration steps
(n)	counter of increments
θ_1	bar coordinate
Δ	increment
$\underline{z}$	vector of structural response, $\underline{z} = \{\underline{z}_1; \underline{z}_2\} = \{u\ w\ \nu\ s;\ N\ Q\ M\ N_s\}$
$\underline{A}$	matrix of coefficients (constant within the increment)
$\underline{b}$	"right hand side" of the system of differential equations with loads and varying parts resulting from geometrically nonlinearities, with physically nonlinear correction forces, as well as with forces from unbonded prestressing
$\underline{d}$	damping matrix
$\underline{m}$	mass matrix

The implicit nonlinear system of differential equations for the differential bar sections is linearized by increments. All geometrically and physically nonlinear components in the $\Delta\underline{b}$-vector are recalculated after every iteration step, and the $\underline{A}$-, $\underline{d}$-, and $\underline{m}$-matrix are recalculated after the completion of the iteration within the increment. The solution of the system of differential equations by a Runge-Kutta integration results in the system of differential equations of the unknown incremental displacements $\Delta\underline{v}$, velocities $\Delta\dot{\underline{v}}$, and accelerations $\Delta\ddot{\underline{v}}$ of the nodes.

$$\underline{K}_{T(n-1)} \cdot \Delta \underline{v}_{(n)}^{[k]} + \underline{D}_{(n-1)} \cdot \Delta \dot{\underline{v}}_{(n)}^{[k]} + \underline{M} \cdot \Delta \ddot{\underline{v}}_{(n)}^{[k]} = \Delta \underline{P}_{(n)} - \Delta \overset{o}{\underline{F}}_{(n)}^{[k]} + \Delta\Delta \underline{F}_{(n-1)} \tag{12}$$

Due to the system modification components of the systems of differential equations (11) and (12) is changes. A special modification increment is adopted for the numerical processing of these changes. Layers of cross sections or structural members which are added to the system within a system modification are inserted stress-free and strain-free into the system. This is numerically processed by

modifications of the corresponding components of eqs. (11) and (12). If additionally layers of cross sections or structural members are removed from the structure, the stresses of those components are transferred to the residual system.

4. Examples

4.1 STEEL CONCRETE STRUCTURE

For the steel-concrete-composite beam that is displayed in Figure 4, the process of manufacturing and loading is analyzed numerically.

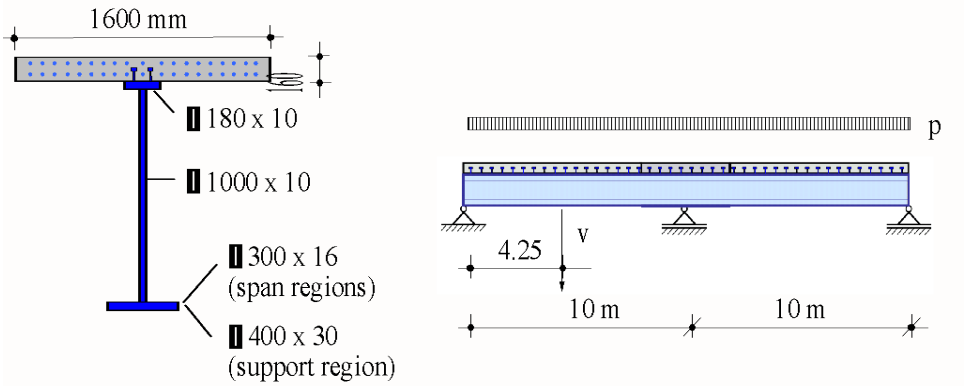


Figure 4. Cross section, system

In the states of manufacturing first the span region of the composite beam are concreted, after it the support region. According this in the numerical analysis first the fresh concrete load is considered and afterwards the respectively concrete layers are taken into consideration within a specific system modification increment. Finally the traffic load of p = 400 kN/m (about 60% of the ultimate load) is applied.

concrete C35/45 f_{ctm}= 3,2 N/mm²
$f_{cm,cyl}$ = 43 N/mm²
construction steel S355 f_y = 360 N/mm²
f_u = 510 N/mm²
reinforcement steel f_y = 500 N/mm²
f_u = 550 N/mm²

Between concrete and steel a nonlinear shear stress slip dependency is regarded, see continuous lines in Figure 5. It is considered as fuzzy function with the likewise in Figure 5 displayed bunch parameter. In comparison the structure is analyzed additionally with a linear shear stress slip dependency with the same initial stiffness (dashed lines) and with a rigid bond (dotted line). The linear shear stress slip dependency is also considered as fuzzy function with the bunch parameter in Figure 5.

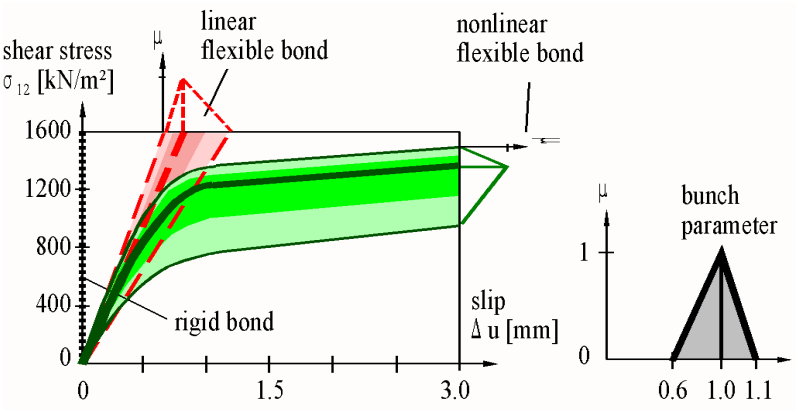


Figure 5. Fuzzy functions of shear stress slip dependency, bunch parameter s

The alteration of the vertical displacement of the girder in the span region at the longitudinal bar coordinate 4.25 m is a selected fuzzy result. The fuzzy displacement is shown in Figure 6 for the three cases of shear stress slip dependencies.

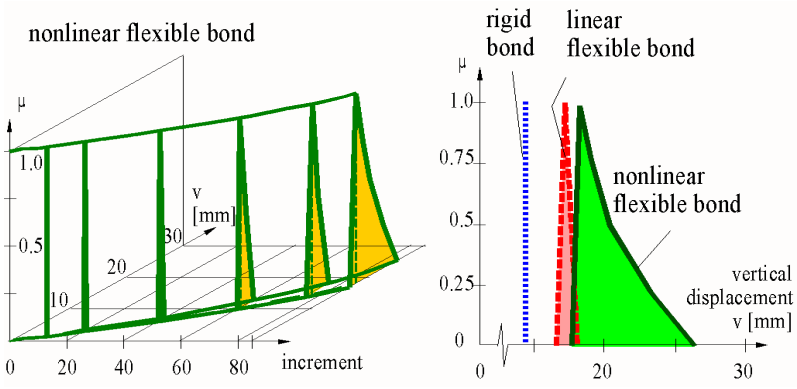


Figure 6. Fuzzy vertical displacements

4.2 NATURAL STONE ARCH BRIDGE

The second example regards the Syratal bridge in Plauen (Germany) built 1903, world wide the widest span natural stone arch bridge at that time. The span is ninety meters, see Figure 7.

Seven years ago (in 2000) the bridge was reconstructed and the masonry was grouted. The main parts of the bridge are the arch and the lateral masonry.

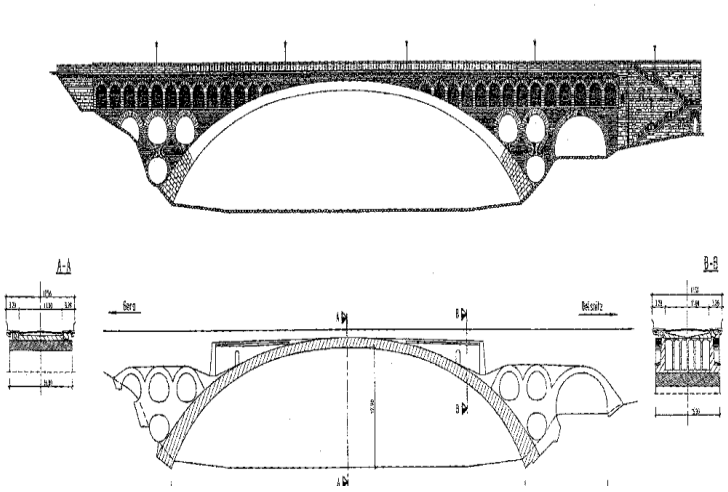


Figure 7. System, see [Schmiedel & Setzpfand 1999]

The system takes into consideration the interaction between the arch and the masonry on the right and left side of the arch. The horizontal displacements of the arch activate the stiffness of the lateral masonry. This effect is modeled by nonlinear node springs, see Figure 8.

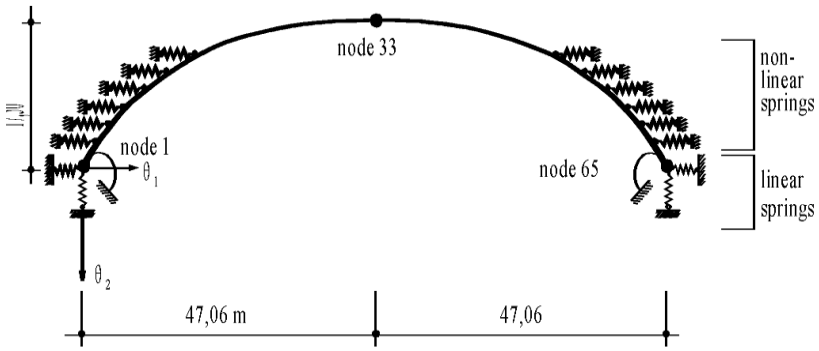


Figure 8. Computational model

In Figure 9 is shown the nonlinear force displacement dependency for the nodes springs and the fuzzy stiffness factor f_{KF}.

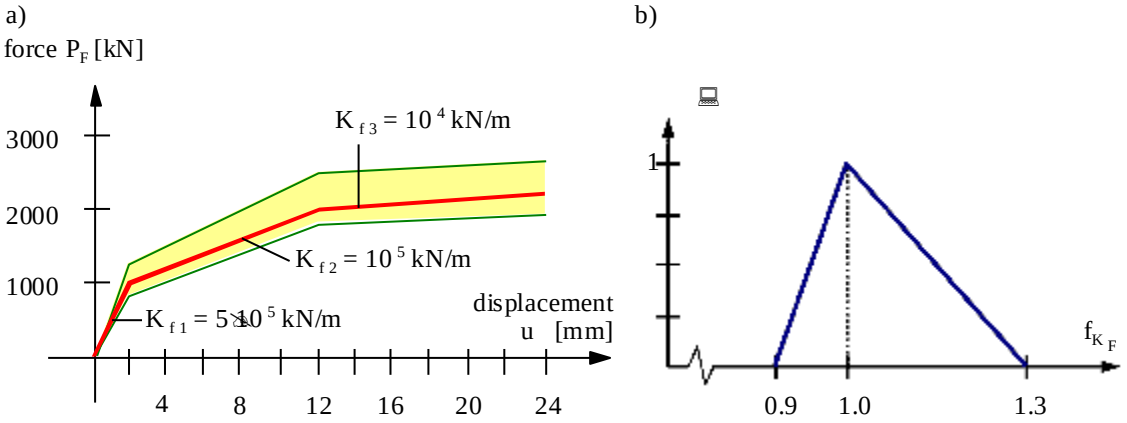


Figure 9. Uncertain force displacement dependency as fuzzy function

The system modification is caused by grouting of masonry. The modification process has a discontinuity as consequence of the rehabilitation. A representative load process is shown in Figure 10.

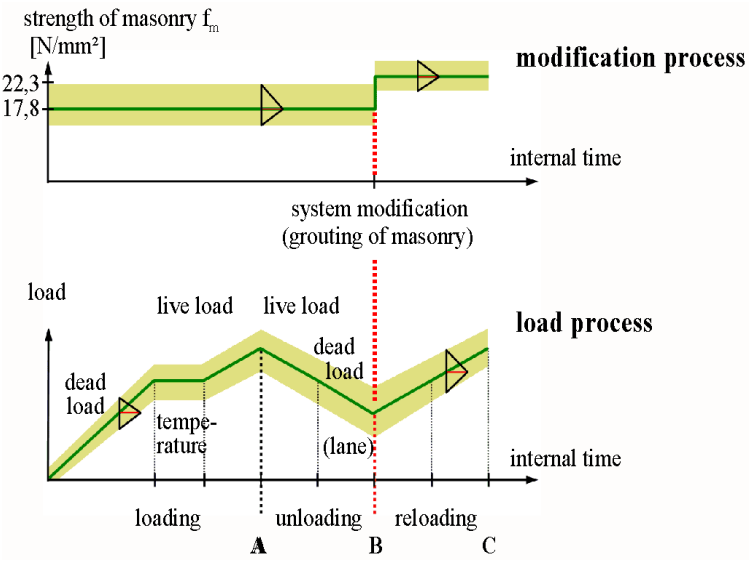


Figure10. Load and modification process

The masonry rehabilitation causes a modification of the constitutive relationship. The curve I in Figure 11 stands for the original constitutive law. The curve III shows the modified constitutive law, and the curve II is a specific sigma-epsilon-path for the modification.

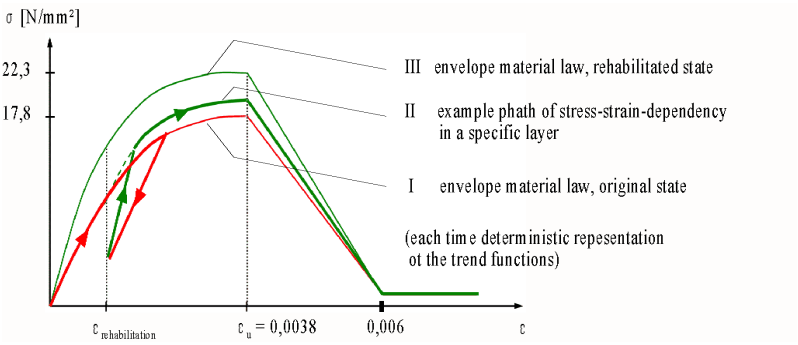


Figure 11. Trend functions of constitutive laws

The failure load factor is computed. That characterizes the ultimate traffic load, and leads to system failure. The ultimate traffic load is equal given live load multiplied by failure load factor η. In Figure 12 is given the fuzzy failure load factor η. Case I investigates the arch without system modification, case II with the unrehabilitated and rehabilitated masonry strength for the system modification process. Case III leads to overestimation of the load bearing capacity.

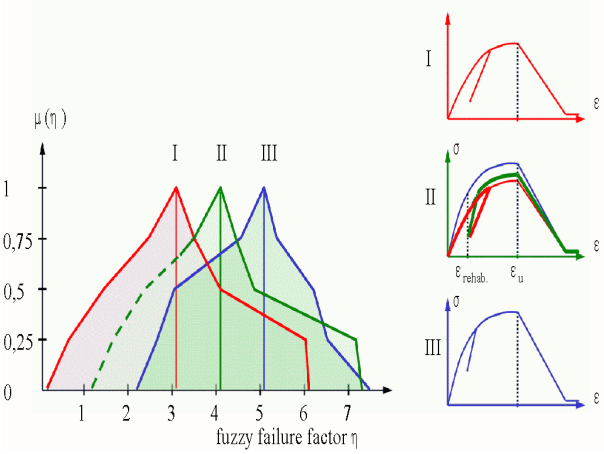


Figure 12. Fuzzy failure factor

In Figure 13 are results of the numerical monitoring, the fuzzy results for the vertical displacement of the crown of the arch (node 33) at the internal time points A, B, and C.

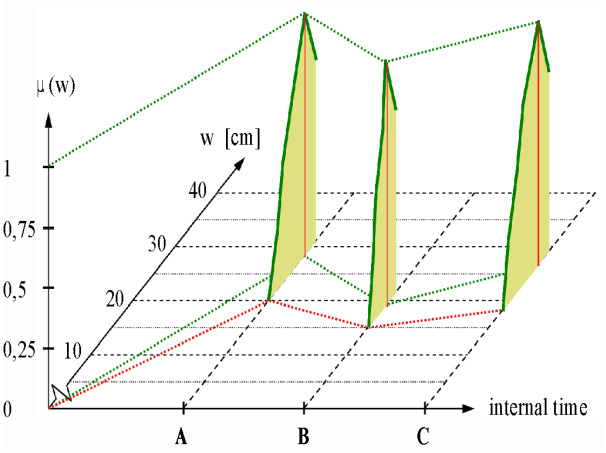


Figure 13. Fuzzy vertical displacements

Conclusions

Analyzing a structure close to reality requires consider the complete load and modification process. The parameters of the load and modification process are generally uncertain. They may be described by fuzzy processes for a numerical monitoring.

Acknowledgement

The authors gratefully acknowledge the support of the German Research Foundation (DFG).

References

Bartzsch, M. 2006. *Tragwerksmodifikation als unstetiger und unscharfer Prozeß.* Technische Universität Dresden, Institut für Statik und Dynamik der Tragwerke, H. 12

Bartzsch, M., Graf, W., Möller, B. & Sickert, J.-U. 2004. Modification of structures with uncertain parameters. In *ECCOMAS*, Jyväskylä, CD-ROM

Möller, B. & Beer, M. 2004. *Fuzzy Randomness - Uncertainty in Civil Engineering and Computational Mechanics.* Springer, Berlin, Heidelberg

Möller, B. & Graf, W. 2005. Tragwerksprozesse in der Baustatik. In *Baustatik-Baupraxis 9*, TU Dresden, Bericht, S. 381-393

Möller, B., Graf, W. & Beer, M. 2000. Fuzzy structural analysis using α-level optimization. *Computational Mechanics*, 26(2000), pp. 547-565

Möller, B., Graf, W. & Beer, M. 2003. Safety assessment of structures in view of Fuzzy randomness. *Computers & Structures*, 81(2003), pp. 1567-1582

Möller, B., Graf, W., Liebscher, M., Pannier, S. & Sickert, J.-U. 2007. An inverse solution of the lifetime-oriented design problem. In *3th ICLODC*, Ruhr-Universität Bochum

Möller, B., Graf, W. & Nguyen S. H. 2004. Modeling the life-cycle of a structure using fuzzy processes. *International Journal of Computer-Aided Civil and Infrastructure Engineering* 19(2004), pp. 157-169, Blackwell Publ., Malden, Cambridge, Oxford

Schmiedel, J. & Setzpfandt, G. 1999. Syratalbrücke Plauen (Friedensbrücke). In *Bundesministerium für Verkehr, Bau- und Wohnungswesen (Hrsg.),* Steinbrücken in Deutschland, Teil 2: Berlin, Brandenburg, Mecklenburg-Vorpommern, Sachsen-Anhalt, Sachsen, Thüringen. Verl. Bau + Technik, Düsseldorf, S. 335-338

Design under Uncertainty using a Combination of Evidence Theory and a Bayesian Approach

Jun Zhou and Zissimos P. Mourelatos
Mechanical Engineering Department
Oakland University, Rochester, MI 48309
email: mourelat@oakland.edu

Abstract: Early in the engineering design cycle, it is difficult to quantify product reliability due to insufficient data or information to model uncertainties. Probability theory can not be therefore, used. Design decisions are usually based on fuzzy information which is imprecise and incomplete. Various design methods such as Possibility-Based Design Optimization (PBDO) and Evidence-Based Design Optimization (EBDO) have been developed to systematically treat design with non-probabilistic uncertainties. In practical engineering applications, information regarding the uncertain variables and parameters may exist in the form of sample points, and uncertainties with sufficient and insufficient information may exist simultaneously. Most of the existing optimal design methods under uncertainty can not handle this form of incomplete information. They have to either discard some valuable information or postulate the existence of additional information. In this paper, a design optimization method is proposed based on evidence theory, which can handle a mixture of epistemic and random uncertainties. Instead of using "expert" opinions to form the basic probability assignment, a Bayesian approach is used with a limited number of sample points. A pressure vessel example demonstrates the merit of the proposed design optimization method. The results are compared with those from existing design methodologies under uncertainty.

1. INTRODUCTION

Engineering design under uncertainty has recently gained a lot of attention. Uncertainties are usually modeled using probability theory. In Reliability-Based Design Optimization (RBDO), variations are represented by standard deviations which are typically assumed constant, and a mean performance is optimized subject to probabilistic constraints [1-5]. In general, probability theory is very effective when sufficient data is available to quantify uncertainty using probability distributions. However, when sufficient data is not available or there is lack of information due to ignorance, the classical probability methodology may not be appropriate. For example, during the early stages of product development, quantification of the product's reliability or compliance to performance targets is practically very difficult due to insufficient data for modeling the uncertainties. A similar problem exists when the reliability of a complex system is assessed in the presence of incomplete information on the variability of certain design variables, parameters, operating conditions, boundary conditions etc.

Uncertainties can be classified in two general types; aleatory (stochastic or random) and epistemic (subjective) [6-10]. Aleatory or irreducible uncertainty is related to inherent variability and is efficiently modeled using probability theory. However, when data is scarce or there is lack of information, the probability theory is not useful because the needed probability distributions cannot be accurately constructed. In this case, epistemic uncertainty, which describes subjectivity, ignorance or lack of information, can be used. Epistemic uncertainty is also called reducible because it can be reduced with increased state of knowledge or collection of more data.

Formal theories to handle uncertainty have been proposed in the literature including evidence theory (or Dempster – Shafer theory) [9, 10], possibility theory [11] and interval analysis [12]. Two large classes of fuzzy measures, called belief and plausibility measures, respectively, characterize the mathematical theory of evidence. They are mutually dual in the sense that one of them can be uniquely determined from the other. Evidence theory uses plausibility and belief (upper and lower bounds of probability) to measure the likelihood of events. When the plausibility and belief measures are equal, the general evidence theory reduces to the classical probability theory. Therefore, the classical probability theory is a special case of evidence theory.

Possibility theory handles epistemic uncertainty if there is no conflicting evidence among experts [9]. It uses a special subclass of dual plausibility and belief measures, called possibility and necessity measures, respectively. In possibility theory, a fuzzy set approach is common, where membership functions characterize the input uncertainty [13]. Even if a probability distribution is not available due to limited information, lower and upper bounds (intervals) on uncertain design variables are usually known. In this case, interval analysis [12, 14, 15] and fuzzy set theory [13] have been extensively used to characterize and propagate input uncertainty in order to calculate the interval of the uncertain output. An efficient method for reliability estimation with a combination of random and interval variables is presented in [16]. However, it is not implemented in a design optimization framework. A few design optimization studies have been also reported, where some or all of the uncertain design variables are in interval form [17-19].

Optimization with input ranges has also been studied under the term anti-optimization [20, 21]. Anti-optimization is used to describe the task of finding the "worst-case" scenario for a given problem. It solves a two-level (usually nested) optimization problem. The outer level performs the design optimization while the inner level performs the anti-optimization. The latter seeks the worst condition under the interval uncertainty [21]. A decoupled approach is suggested in [21] where the design optimization alternates with the anti-optimization rather than nesting the two. It was mentioned that this method takes longer to converge and may not even converge at all if there is strong coupling between the interval design variables and the rest of the design variables. A "worst-case" scenario approach using interval variables has also been considered in multidisciplinary systems design [19, 22].

Very recently, possibility-based design algorithms have been proposed [23-25] where a mean performance is optimized subject to possibilistic constraints. It was shown that more conservative results are obtained compared with the probability-based RBDO. A comprehensive comparison of probability and possibility theories is given in [26] for design under uncertainty.

Evidence theory is more general than probability and possibility theories, even though the methodologies of uncertainty propagation are completely different [27, 28]. It can be used in design under

uncertainty if limited, and even conflicting, information is provided from experts. Furthermore, the basic axioms of evidence theory allow to combine aleatory (random) and epistemic uncertainty in a straightforward way without any assumptions [28]. Evidence theory however, has been barely explored in engineering design. One of the reasons may be its high computational cost due mainly to the discontinuous nature of uncertainty quantification. Evidence-based methods have been only recently used to propagate epistemic uncertainty [28, 29] in large-scale engineering systems. Although a computationally efficient method is proposed in [28, 29], the design issue is not addressed. We are aware of only one study which propagates epistemic uncertainty using evidence theory and also performs a design optimization [30]. The optimum design is calculated for multidisciplinary systems under uncertainty using a trust region sequential approximate optimization method with surrogate models representing the uncertain measures as continuous functions.

In engineering design, information regarding the uncertain quantities is usually available in the form of a set of finite samples, either from historical data or from actual measurements. These samples are not enough to infer a probability distribution. However, if we collapse them into intervals, we discard valuable information. Collecting more samples is often not possible due to the cost or time limitations. So RBDO, PBDO (Possibility-Based Design Optimization) [23, 24], and EBDO (Evidence-Based Design Optimization) [31] may not satisfactorily address the presence of incomplete information. We must utilize Bayesian inference to estimate design reliability with incomplete information.

Bayesian inference is an approach to statistics in which all forms of uncertainty are expressed in terms of probability. A Bayesian approach starts with the formulation of a model to describe the situation of interest. A prior distribution is formulated over the unknown parameters of the model, which is meant to capture the belief about the situation before seeing the data. Using available data, we apply Bayesian's rule to obtain a posterior distribution for these unknowns, which accounts for both the prior and the new data.

In this paper, a Bayesian approach is used to account for uncertainty in the design when limited information is provided by a limited number of sample points. A Bayesian approach is proposed using the extreme value distribution of the smallest value. The approach can handle both a mixture of epistemic and random uncertainties or pure epistemic uncertainties. The accuracy of predictions improves with the use of more sample points. Previous research, such as in [32-34] illustrate how to use a Bayesian approach in design utilizing the confidence percentile concept. In this paper, the available methodologies are improved by using the extreme value distribution of the smallest value instead of the conventional beta distribution. The extreme value distribution approach is necessary because we have only a small set of sample points which are different at each experiment. A Bayesian approach to design optimization (BADO) using the extreme value distribution is proposed. We show that the optimal design is conservative.

The proposed BADO approach can handle epistemic uncertainties or a mixture of aleatory and epistemic uncertainties. Also if only the number of sample points within a certain range is known instead of the exact distribution of the sample points, we propose a design methodology which combines the evidence theory and the Bayesian approach.

The difference between Possibility-Based Design Optimization (PBDO) and Bayesian-Based Design Optimization (BADO) is in the format of uncertain variables. Possibilistic variables are in the form of intervals and Bayesian uncertain variables are in the form of sample points. The latter provide more information compared with the possibilistic variables. Both of them are based on the confidence percentile concept. In PBDO, a membership function is constructed for each possibilistic variables. The PBDO approach provides a worst-case design because there is a minimal amount of information in the form of intervals. However, more information is available for the Bayesian uncertain variables in the form of sample points. For this reason, we will show that the BADO design is less conservative than the PBDO design.

The paper is organized as follows. Section 2 gives an introduction to the fundamentals of evidence theory. Section 3 presents an overview of an Evidence-Based Design Optimization (EBDO) algorithm. Section 4 presents the proposed Bayesian-Based Design Optimization (BADO) procedure and a methodology to estimate the BPA structure from limited available data using Bayesian statistics. The concepts in section 4 are demonstrated with a pressure vessel example. Comparisons among RBDO, EBDO, PBDO and BADO are also provided in order to demonstrate the value of added information in design. Finally, a summary and conclusions are given in section 5.

2. FUNDAMENTALS OF EVIDENCE THEORY

This section gives the fundamentals of evidence theory, how it can be used in design optimization and an introduction to fuzzy measures. Detailed information is provided in [8, 9, 11, 31, 35]. The role of fuzzy measures and the axiomatic definition of evidence theory are explained.

Evidence theory is based on the belief (*Bel)* and Plausibility (*Pl*) fuzzy measures. Fuzzy measures provide the foundation of fuzzy set theory. Before we introduce the basics of fuzzy measures, it is helpful to review the used notation on set representation. A universe X represents the entire collection of elements having the same characteristics. The individual elements in the universe X are denoted by x, and are usually called singletons. A set A is a collection of some elements of X. All possible sets of X constitute a special set called the power set $\wp(X)$.

A fuzzy measure is defined by a function g: $\wp(X) \rightarrow [0,1]$ which assigns to each crisp subset of X a number in the unit interval [0,1]. The assigned number in the unit interval for a subset $A \in \wp(X)$, denoted by g(A), represents the degree of available evidence or belief that a given element of X belongs to the subset A.

In order to qualify as a fuzzy measure, the function g must have certain properties. These properties are defined by axioms that are *weaker* than the probability theory axioms [8, 9]. Every fuzzy measure obeys the following three axioms:

Axiom 1 (boundary conditions): $g(\varnothing)=0$ and $g(X)=1$.

Axiom 2 (monotonicity): For every $A, B \in \wp(X)$, if $A \subseteq B$, then $g(A) \leq g(B)$.

Axiom 3 (continuity): For every sequence ($A_i \in \wp(X)$, *i=1,2,…*) of subsets of $\wp(X)$, if either $A_1 \subseteq A_i \subseteq \ldots$ or $A_1 \supseteq A_2 \supseteq \ldots$ (i.e., the sequence is monotonic), then $\lim_{i\to\infty} g(A_i) = g(\lim_{i\to\infty} A_i)$.

A belief measure is a function *Bel*: $\wp(X) \to [0,1]$ which satisfies the three axioms of fuzzy measures and the following additional axiom [9]:

$$Bel(A_1 \cup A_2) \geq Bel(A_1) + Bel(A_2) - Bel(A_1 \cap A_2) \quad . \tag{1}$$

The axiom (1) can be expanded for more than two sets. For $A \in \wp(X)$, *Bel*(*A*) is interpreted as the degree of belief, based on available evidence, that a given element of *X* belongs to the set *A*.
A plausibility measure is a function

$$Pl : \wp(X) \Rightarrow [0,1] \tag{2}$$

which satisfies the three axioms of fuzzy measures and the following additional axiom [9]

$$Pl(A_1 \cap A_2) \leq Pl(A_1) + Pl(A_2) - Pl(A_1 \cup A_2) \tag{3}$$

Every belief measure and its dual plausibility measure can be expressed with respect to the non-negative function

$$m : \wp(X) \Rightarrow [0,1] \tag{4}$$

such that $m(\varnothing) = 0$ and

$$\sum_{A\in\wp(X)} m(A) = 1. \tag{5}$$

The function *m* is called Basic Probability Assignment (BPA) due to the resemblance of Eq. (5) with a similar equation for probability distributions. The basic probability assignment *m*(*A*) is interpreted either as the degree of evidence supporting the claim that a specific element of *X* belongs to the set *A* or as the degree to which we believe that such a claim is warranted. Every set $A \in \wp(X)$ for which *m*(*A*)>0 is called a focal element of *m*. Focal elements are subsets of *X* on which the available evidence focuses; i.e. available evidence exists.

Given a BPA *m*, a belief measure and a plausibility measure are uniquely determined by

$$Bel(A) = \sum_{B\subseteq A} m(B) \tag{6}$$

and

$$Pl(A) = \sum_{B\cap A\neq 0} m(B). \tag{7}$$

which are applicable for all $A \in \wp(X)$.

In Eq. (6), *Bel*(*A*) represents the total evidence or belief that the element belongs to *A* as well as to various subsets of *A*. The *Pl*(*A*) in Eq. (7) represents not only the total evidence or belief that the element

in question belongs to set *A* or to any of its subsets but also the additional evidence or belief associated with sets that overlap with *A*. Therefore,

$$Pl(A) \geq Bel(A). \tag{8}$$

It should be noted that belief and plausibility are complementary in the sense that one of them can be uniquely derived from the other.

Probability theory is a subcase of evidence theory. When the additional axiom of belief measures (see Eq. (1)) is replaced with the stronger axiom

$$Bel(A \cup B) = Bel(A) + Bel(B) \text{ where } A \cap B = \varnothing, \tag{9}$$

we obtain a special type of belief measures which are the classical probability measures. In this case, the right hand sides of Eq. (6) and (7) become equal and therefore,

$$Bel(A) = Pl(A) = \sum_{x \in A} m(x) = \sum_{x \in A} p(x) \tag{10}$$

for all $A \in \wp(X)$, where *p(x)* is the probability distribution function (PDF). Note that the BPA *m(x)* is equal to *p(x)*. Therefore with evidence theory, we can simultaneously handle a mixture of input parameters. Some of the inputs can be described probabilistically (random uncertainty) and some can be described through expert opinions (epistemic uncertainty with incomplete data). In the first case, the range of each input parameter will be discretized using a finite number of intervals. The BPA value for each interval must be equal to the PDF area within the interval.

Evidence obtained from independent sources or experts must be combined. If the BPA's m_1 and m_2 express evidence from two experts, the combined evidence *m* can be calculated by the following Dempster's rule of combining [36]

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - K} \quad \text{for} \quad A \neq 0 \tag{11}$$

where

$$K = \sum_{B \cap C = 0} m_1(B) m_2(C) \tag{12}$$

represents the *conflict* between the two independent experts. Dempster's rule filters out any conflict, or contradiction among the provided evidence, by normalizing with the complementary degree of conflict. It is usually appropriate for relatively small amounts of conflict where there is some consistency or sufficient agreement among the opinions of the experts. Yager [10] has proposed an alternative rule of combination where all degrees of contradiction are attributed to total ignorance. Other rules of combining can be found in [36].

2.1. ASSESSING BELIEF AND PLAUSIBILITY WITH DEMPSTER-SHAFER THEORY

The previous section described a methodology to quantify epistemic uncertainty, even when the experts provide conflicting evidence. This section shows how to propagate epistemic uncertainty through a given model (transfer function). We will illustrate that, using the following simple transfer function

$$y = f(a,b) \tag{13}$$

where $a \in A, b \in B$ are two independent input parameters and *y* is the output. The combined BPA's for both *a* and *b* are obtained from Dempster's rule of combining of Eq. (11) if multiple experts have provided evidence for either *a* or *b*. With combined information for each input parameter, we define a vector $c = \lfloor a_{ci}, b_{cj} \rfloor$, needed to calculate the output *y* as

$$C = A \times B = \left\{ c = \lfloor a_{ci}, b_{cj} \rfloor, a_{ci} \in A, b_{cj} \in B \right\} \tag{14}$$

where subscript *c* stands for "combined" and *i,j* indicate focal elements.

Taking advantage of assumed parameter independency, the BPA for *c* is

$$m_c(h_{ij}) = m(a_{ci}) m(b_{cj}) \tag{15}$$

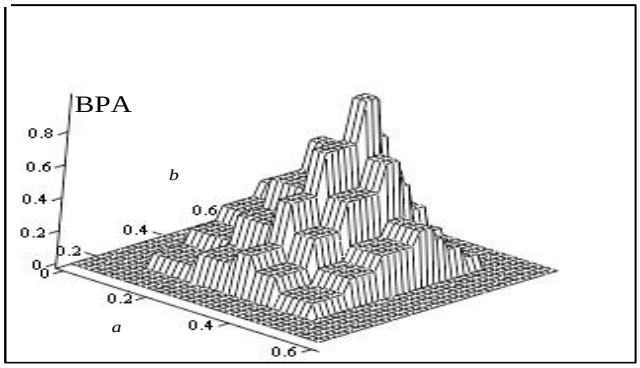


Figure 1. Representative BPA structure for two parameters *a* and *b*.

where $h_{ij} = \lfloor a_{ci}, b_{cj} \rfloor$ and a_{ci}, b_{cj} denote intervals such that $a \in a_{ci}$ and $b \in b_{cj}$. Eq. (15) can be used to calculate the combined BPA structure for the entire domain *C*. For every $(a,b) \in c \mid c \in C$, needed to evaluate the output *y*, the combined BPA m_c is used. A representative combined BPA structure is shown in Figure 1.

The Cartesian product *C* of Eq. (14) is also called frame of discernment (FD) in the literature. It consists of all focal elements (rectangles in Figure 1 with nonzero combined BPA) and can be viewed as the finite sample space in probability theory.

If a domain *F* is defined as

$$F = \{g : g = f(a,b) - y_0 > 0, (a,b) \in c, c = [a_c, b_c] \subset C\} \quad (16)$$

where y_0 is a specified value, *Bel(F)* and *Pl(F)* can be calculated from Eqs (6) and (7) where set *F* replaces set *A*. According to evidence theory, *Bel(F)* and *Pl(F)* bracket the true probability $p_f = P(g > 0)$ [9,27]; i.e.

$$Bel(F) \le p_f \le Pl(F). \quad (17)$$

The *Bel (F)* and *Pl (F)* are calculated using Eqs (6) and (7) where set *A* is equal to set *F* of Eq. (16) and *B* is a rectangular domain (focal element) such that $B \subseteq A$ for Eq. (6) and $B \cap A \neq 0$ for Eq. (7). In other words, $B \subseteq A$ means that the focal element must be entirely within the domain *g>0* and $B \cap A \neq 0$ means that the focal element must be entirely or partially within the domain *g>0* (see Fig. 2). In general, in order to identify if a focal element *B* satisfies $B \subseteq A$ or $B \cap A \neq 0$, the following minimum and maximum values of *g* must be calculated

$$[g_{\min}, g_{\max}] = \left[\min_{\mathbf{X}} g(\mathbf{X}), \max_{\mathbf{X}} g(\mathbf{X})\right] \quad (18)$$

for $\mathbf{X}^L \le \mathbf{X} \le \mathbf{X}^U$ where $(\mathbf{X}^L, \mathbf{X}^U)$ defines the focal element domain. For monotonic functions, the vertex method [37] can be used to calculate the minimum and maximum values in Eq. (18) by simply identifying the minimum and maximum values among all vertices of the focal element domain. If for a focal element, $g_{\min}$ and $g_{\max}$ are both positive, the focal element will contribute to the calculation of belief and plausibility according to Eqs (6) and (7). On the other hand, if $g_{\min}$ and $g_{\max}$ are both negative, the focal element will not contribute to the calculation of belief or plausibility. If however, $g_{\min}$ is negative and $g_{\max}$ is positive, the focal element will not contribute to the belief but it will contribute to the plausibility calculation. This is shown schematically in Figure 2.

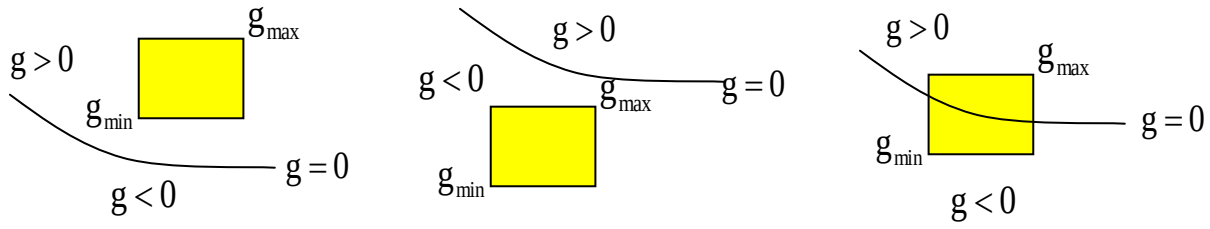


Figure 2. Schematic illustration of focal element contribution to belief and plausibility measures.

In summary the following tasks are performed in order to calculate the belief and plausibility of the failure region:

1) For each input parameter, combine the evidence from the experts by combining the individual BPA's from each expert using Dempster's rule of combining (Eq. (11)).
2) Construct the BPA structure for the *m*-dimensional frame of discernment, where *m* is the number of input parameters. Assuming independent input parameters, Eq. (15) is used.
3) Identify the failure region space (set *F* of Eq. (16)).
4) Use Eqs (6) and (7) to calculate the belief and plausibility measures of the failure region. The failure region must be identified only within the frame of discernment. The true probability of failure is bracketed according to Eq. (17).

3. EVIDENCE-BASED DESIGN OPTIMIZATION (EBDO)

In deterministic design optimization, an objective function is minimized subject to satisfying each constraint. In Reliability-Based Design Optimization (RBDO), where all design variables are characterized probabilistically, an objective function is usually minimized subject to the probability of satisfying each constraint, being greater than a specified high reliability level. In this section, a methodology is presented on how to use evidence theory in design. We will show that the evidence theory-based design is conservative compared with all RBDO designs obtained with different probability distributions.

If feasibility of a constraint g is expressed with the non-negative null form $g \geq 0$, we have shown in the previous section that $P(g \geq 0)$ is bracketed by the belief $Bel(g \geq 0)$ and plausibility $Pl(g \geq 0)$; i.e. $Bel(g \geq 0) \leq P(g \geq 0) \leq Pl(g \geq 0)$. Therefore,

$$P(g < 0) \leq p_f \text{ is satisfied if } Pl(g < 0) \leq p_f \tag{19}$$

where p_f is the probability of failure which is usually a small prescribed value. The above statement is equivalent to

$$P(g \geq 0) \geq R \text{ is satisfied if } Bel(g \geq 0) \geq R \tag{20}$$

where $R = 1 - p_f$ is the corresponding reliability level.

An evidence theory-based design optimization (EBDO) problem can be formulated as

$$\begin{aligned} &\min_{\mathbf{d},\mathbf{X}^N} f\left(\mathbf{d}, \mathbf{X}^N, \mathbf{P}^N\right) \\ \text{s.t.} \quad &Pl\left(g_i(\mathbf{d}, \mathbf{X}, \mathbf{P}) < 0\right) \leq p_{f_i},\ i = 1,...,n \\ &\mathbf{d}_L \leq \mathbf{d} \leq \mathbf{d}_U \\ &\mathbf{X}_L^N \leq \mathbf{X}^N \leq \mathbf{X}_U^N \end{aligned} \tag{21}$$

where $\mathbf{d} \in R^k$ is the vector of deterministic design variables, $\mathbf{X} \in R^m$ is the vector of uncertain design variables, $\mathbf{P} \in R^q$ is the vector of uncertain design parameters, $f(\ )$ is the objective function and *n*, *k*, *m* and *q* are the number of constraints, deterministic design variables, uncertain design variables and uncertain design parameters, respectively. According to the used notation, a bold letter indicates a vector, an upper case letter indicates an uncertain variable or parameter and a lower case letter indicates a realization of the uncertain variable. The superscript "N" in Problem (21) indicates nominal value of each uncertain design variable or design parameter. The uncertainty is provided by expert opinions.

It should be noted that the plausibility measure is used instead of the equivalent belief measure, in Problem (21). The reason is that at the optimum, the failure domain for each active constraint is usually much smaller than the safe domain over the frame of discernment (FD). As a result, the computation of the plausibility of failure is much more efficient than the computation of the belief of safe region.

3.1. IMPLEMENTATION OF THE EBDO ALGORITHM

This section describes a computationally efficient solution of Problem (21). As a geometrical interpretation of Problem (21), we can view the design point (**d**,**X**) moving within the feasible domain so that the objective *f* is minimized. If the entire FD is in the feasible domain, the constraints are satisfied and are inactive. A constraint becomes active if part of the FD is in the "failure" region so that the plausibility of constraint violation is equal to p_f. In general, Problem (21) represents movement of a hyper-cube (FD) within the feasible domain.

In order to save computational effort, the bulk of the FD movement, from the initial design point to the *vicinity* of the optimal point (point B of Figure 3), can be achieved by *moving a hyper-ellipse which contains the FD*. The center of the hyper-ellipse is the "approximate" design point and each axis is arbitrarily taken equal to three times the standard deviation of a hypothetical normal distribution. This assumes that each dimension of the FD hyper-cube is equal to six times the standard deviation of the hypothetical normal distribution. The hyper-ellipse can be easily moved in the design space by solving a Reliability-Based Design Optimization (RBDO) problem. The RBDO optimum (point B of Figure 3) is in the vicinity of the

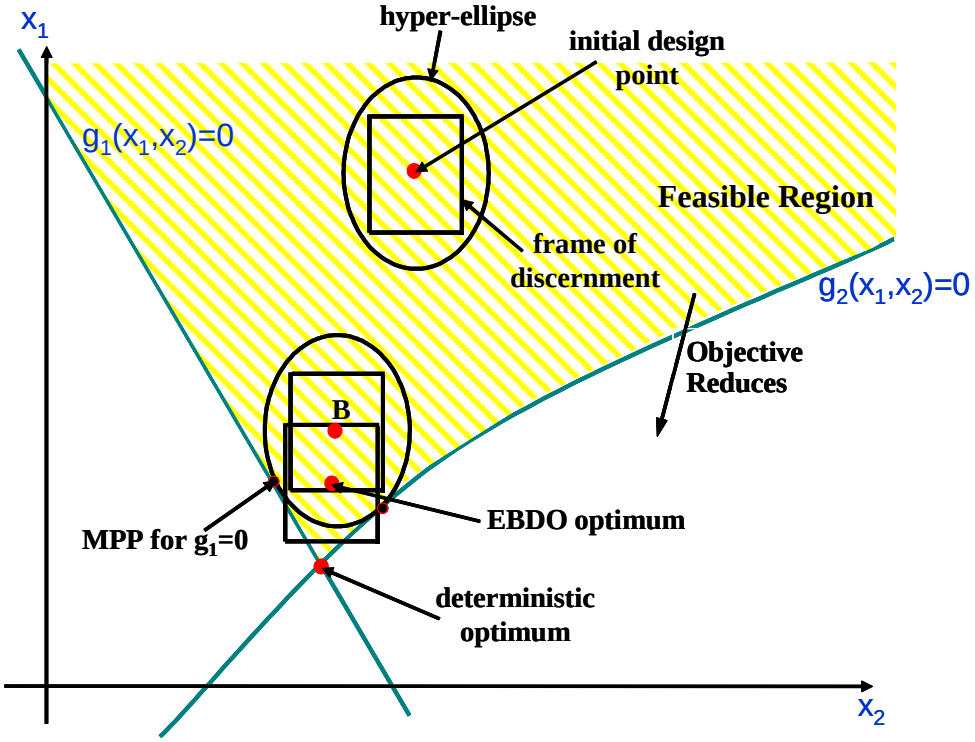


Figure 3. Geometrical interpretation of the EBDO algorithm.

solution of Problem (21) (EBDO optimum). The RBDO solution also identifies all active constraints and their corresponding most probable points (MPP's). The maximal possibility search algorithm [38] can also be used to move the FD hyper-cube in the feasible domain. It should be noted that the 3-sigma axes hyper-ellipse is arbitrary. The size of the hyper-ellipse is not however, crucial to the user because it is only used to calculate the initial point (point B of Figure 3) of the EBDO algorithm. The latter calculates the true EBDO optimum accurately. From our experience, 3 to 4-sigma size works fine.

At this point, we generate a *local* response surface of each active constraint around its MPP. In this work, the Cross-Validated Moving Least Squares (CVMLS) [39] method is used based on an Optimum Symmetric Latin Hypercube (OSLH) [40] "space-filling" sampling.

A derivative-free optimizer calculates the EBDO optimum. It uses as initial point the previously calculated RBDO optimum which is close to the EBDO optimum. Problem (21) is solved, considering only the identified active constraints. For the calculation of the plausibility of failure $Pl\left(g<0\right)$ of each active constraint, the algorithm of next section is used. The algorithm identifies all focal elements which contribute to the plausibility of failure. The computational effort is significantly reduced because accurate local response surfaces are used for the active constraints. The cost can be much higher if the optimization

algorithm evaluates the actual active constraints instead of their efficient surrogates (response surfaces). It should be noted that a derivative-free optimizer is needed due to the discontinuous nature of the combined BPA structure. The DIRECT (DIvisions of RECTangles) derivative-free, global optimizer is used in this work. DIRECT is a modification of the standard Lipschitzian approach that eliminates the need to specify a Lipschitz constant [41].

3.1. CALCULATION OF PLAUSIBILITY OF FAILURE

In Problem (21), the plausibility of failure or equivalently the plausibility of constraint violation, $Pl(g<0)$, must be calculated every time the optimizer evaluates a constraint. The algorithm is given below.

Step 1. Initialize sets C = {FD} and F = {0} and counter m = 1

Step 2. Consider all sets $E_k : E_k \subset C$ or $E_k \subseteq C$

Initialize counter n = 0
Empty set C; i.e. C = {0}
For k = 1 to m
Partition E_k into E_k^1 and E_k^2
For j = 1 to 2
Calculate $g_{\min}\left(E_k^j\right)$
If $g_{\min}\left(E_k^j\right)<0$ then
Calculate $g_{\max}\left(E_k^j\right)$
If $g_{\max}\left(E_k^j\right)>0$ then $C = C \cup E_k^j$ and $n = n+1$
If $g_{\max}\left(E_k^j\right)\leq 0$ then $F = F \cup E_k^j$
End if (for the loop of $g_{\min}\left(E_k^j\right)<0$)
End if (for the loop of j = 1 to 2)
End if (for the loop k = 1 to m)
Set counter $m = n$
If C can be partitioned, go to step 2.
If C can not be partitioned, stop and calculate plausibility of failure from Eq. (22)

$$Pl(g<0) = \sum_{B\in F} m(B) + \sum_{B\in C} m(B) \tag{22}$$

as the sum of BPA values of all focal elements B which belong to sets F and C.

A set C which is initially equal to the entire frame of discernment FD (see step 1) is partitioned into sets E^1 and E^2. The partitioning sequence is explained at the end of this section. The minimum and maximum values of g in the E^1 and E^2 domains are calculated; i.e.

$g_{\min}(E_i) = \min g(\mathbf{X}),\ \mathbf{X} \in E_i, i = 1,2$ and $g_{\max}(E_i) = \max g(\mathbf{X}),\ \mathbf{X} \in E_i, i = 1,2$ (see step 2). If $g_{\min}(E_i) < 0$ and $g_{\max}(E_i) > 0$, E^i is placed in set *C*. If $g_{\min}(E_i) < 0$ and $g_{\max}(E_i) < 0$, E^i is placed in set *F*. Otherwise, E^i is not considered further. For a subsequent iteration *k* in step 2, each set which has been placed in *C* (denoted by E_k) is further partitioned into sets E_k^1 and E_k^2, and the process continues. If all sets put in *C* represent focal elements and therefore, can not be partitioned further, the algorithm stops and Eq. (22) is used to calculate the plausibility of failure.

The above algorithm is demonstrated with a hypothetical example. Figure 4 shows the location of the FD relative to the limit state g=0 for a particular iteration. A hypothetical BPA structure is also shown. Each "rectangle" represents a focal element. In this case, we have 20 focal elements denoted by m_i, *i*=1,2,…,20. A set which is initially equal to FD, is partitioned into sets E^1 and

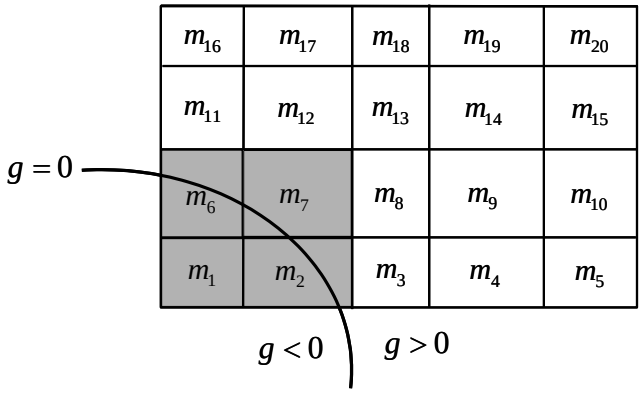


Figure 4. A hypothetical two-dimensional BPA structure.

E^2 such that $E^1 = \bigcup_i m_i\ , i = 1,2,6,7,11,12,16,17$ and $E^2 = \bigcup_i m_i\ , i = 3,4,5,8,9,10,13,14,15,18,19,20$. Subsequently, the minimum and maximum values of g in the E^1 and E^2 domains ($g_{\min}(E^i)$ and $g_{\max}(E^i)$, *i*=1,2) are calculated. Because $g_{\min}(E^1) < 0$ and $g_{\max}(E^1) > 0$, E^1 is placed in *C*. However, $g_{\min}(E^2) > 0$ and therefore, E^2 is not considered further. This is the end of the first iteration.

The second iteration starts by partitioning *C* which is equal to E^1 of the first iteration, into $E^{11} = \bigcup_i m_i\ , i = 11,12,16,17$ and $E^{12} = \bigcup_i m_i\ , i = 1,2,6,7$. Similarly to the first iteration, E^{11} is discarded and E^{12} is placed in *C* which is now composed of E^{12} only. Note that at the end of the second iteration, set *F* is empty. At the third iteration, *C* or equivalently E^{12}, is partitioned into

$E^{121} = \bigcup_i m_i \ , i = 1,6$ and $E^{122} = \bigcup_i m_i \ , i = 2,7$ which are both placed in *C*. At the fourth iteration, E^{121} is partitioned into $E^{1211} = m_1$ and $E^{1212} = m_6$ and E^{122} is partitioned into $E^{1221} = m_2$ and $E^{1222} = m_7$. Now E^{1211} is placed in *F* and E^{1212}, E^{1221} and E^{1222} are placed in *C*. Because all previous sets consist of one focal element each, they can not be partitioned further. Therefore, the algorithm stops. Finally, $F = m_1$ and $C = \bigcup_i m_i \ , i = 2,6,7$. Eq. (22) is used to calculate the plausibility of $g<0$ as the sum of BPA values of all focal elements in *F* and *C*; i.e. $Pl(g < 0) = \sum_{B \in F} m(B) + \sum_{B \in C} m(B)$.

The described algorithm uses the following partitioning scheme for an *n*-dimensional hyper-rectangle representing the FD which corresponds to *n* uncertain variables and parameters. For the k^{th} iteration (k = 1,…,n), the hyper-rectangle is partitioned into two parts with an (n-1)-dimensional hyper-plane perpendicular to the k^{th} dimension. Each part has roughly the same number of focal elements. For iteration $k > n$, the (n-1)-dimensional hyper-plane is perpendicular to the $(k\text{-}n)^{th}$ dimension.

4. BAYESIAN RELIABILITY-BASED DESIGN OPTIMIZATION

It has been mentioned that if we only know the bounds within which an uncertain variable varies, interval analysis or possibility theory can be used to quantify and propagate uncertainty. If additional information is available in terms of expert opinions for example, the evidence theory can be used. It is common however, in engineering design, to know the bounds of the uncertain variables and also have additional information in the form of a discrete but limited number of sample points based on historic data or experiment data. In this case, we can not infer a probabilistic distribution because of the limited number of sample points. However, a Bayesian approach [32, 33] can be used to estimate the probability distribution. If more information is obtained later in the form of additional sample points, a more accurate estimation of the probability distribution can be obtained. The next subsections provide the basics of Bayesian approach as well as the introduction of the extreme value distribution in the Bayesian approach in order to account for the fact that we only have a small set of sample points which are different at each experiment.

4.1. BAYESIAN RELIABILITY ESTIMATION

Let us denote available data by *D* and the probability of success by θ. We wish to improve our knowledge about the unknown quantity θ by utilizing the known information in the available data *D*. To make inferences about θ, we build a conditional probability distribution $P(\theta|D)$ that describes how we believe θ is distributed considering the existence of data *D*. Using the Bayesian rule, it can be shown that

$$P(\theta \mid D) = \frac{\Gamma(N+\alpha+\beta)}{\Gamma(D+\alpha)\Gamma(N-D+\beta)}\theta^{D+\alpha-1}(1-\theta)^{N-D+\beta-1} = Beta\left(D+\alpha, N-D+\beta\right) \quad (23)$$

where,

$$Beta(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad (24)$$

with α, and β being the *Beta* distribution parameters. As expected, the posterior distribution is also a *Beta* distribution because *Beta* is a conjugate family of distributions.

It should be noted that initially there is no prior information about θ and any values between 0 and 1 may be assumed equally probable. In this case, a uniform prior $P(\theta)=U(0,1)$ is used which is equivalent to *Beta(1,1)*.

If we have r_0 successes out of N_0 sample points, then the probability distribution $P(\theta \mid r_0)$ is proportional to Beta (r_0 +1, N_0 - r_0 +1). If additional N_1 data is obtained later, where r_1 is the number of successes, then the total number of success is $r_0 + r_1$ and the total number of failures is $N_0 + N_1 - r_0 - r_1$. In this case, the probability distribution $p(\theta \mid r_0 + r_1)$ is proportional to Beta ($r_0 + r_1$ +1, $N_0 + N_1 - r_0 - r_1$ +1). Note that if we use the constraint function *g* to divide the design space into feasible and infeasible domains, then a feasible realization of *g* is considered a success and an infeasible realization is considered a failure.

Let us define two vectors $\boldsymbol{R}$ = [$\boldsymbol{Y}$, $\boldsymbol{Z}$] and $\boldsymbol{U}$ = [$\boldsymbol{X}$, $\boldsymbol{P}$] where $\boldsymbol{Y}$ and $\boldsymbol{Z}$ denote the random variables and parameters whose PDFs are known and $\boldsymbol{X}$ and $\boldsymbol{P}$ denote the uncertain variables and parameters whose PDFs are not known. If $\boldsymbol{R}$ is not empty, each realization of $\boldsymbol{U}$ = [$\boldsymbol{X}$, $\boldsymbol{P}$] results in a distribution of *g* values. In this case, we can calculate the probability *Pr* [*g* ($\boldsymbol{Y}$, $\boldsymbol{Z}$) > 0 | ($\boldsymbol{X}$, $\boldsymbol{P}$)] that a sample point [$\boldsymbol{Y}$, $\boldsymbol{Z}$] will result in a feasible realization given the sample point [$\boldsymbol{X}$, $\boldsymbol{P}$]. This conditional probability is the expected feasible realization of one sample.

Using a limited number of sample points we obtain therefore, a probability distribution instead of a single probability value. Because there are few sample points and the samples are random, the *Beta* distribution may not be accurately representing the actual distribution. In order to increase our confidence of the predicted probability, we propose to use the extreme distribution of the smallest value using the *Beta* distribution as the basic distribution. If *X* is a Beta distributed uncertain variable and there are *n* available sample points of *X*, the CDF of the extreme minimum value Y_1 (i.e. $Y_1 = \min(X_1, X_2, ..., X_n)$) is given by

$$F_{Y_1}(y) = 1 - [1 - F_X(y)]^n . \quad (25)$$

4.2. CONSTRUCTION OF THE EXTREME SMALLEST VALUE DISTRIBUTION

Because we have a limited number of sample points [***X***, ***P***] the probability *Pr* [*g* ***(Y, Z)*** *> 0 |* ***(X, P)***] is approximated by the *Beta* distribution. To increase our confidence of constraint satisfaction (reliability) exceeding a specified target reliability R, we express each probabilistic constraint in terms of a confidence percentile [42]. For the i^{th} constraint, this is expressed as

$$\nabla\left(P\left(g_i(\mathbf{d},\mathbf{Y},\mathbf{Z})\geq 0\,|(\mathbf{X},\mathbf{P})\right)\geq R_i\right)=\int_{p_f^t}^{1}P(\theta\,|\,D)d\theta\geq\sigma, \qquad (26)$$

where σ is a specified confidence percentile, $p_f^t=-\Phi(1-R_i)$ is the target probability of failure for the i^{th} constraint, and ∇ denotes the confidence percentile. The latter is calculated based on the extreme value distribution. It provides a conservative distribution of the probability of constraint satisfaction which is not a scalar. It should be noted that the extreme value distribution provides a much smaller confidence percentile compared with the *Beta* distribution for the same reliability. This means that it is much safer (or more conservative) to use the extreme value distribution in design optimization.

For a confidence percentile σ, let us denote by P_B and P_B' the probability corresponding to σ based on the extreme value and *Beta* distributions, respectively. Also, let us assume that the number of available sample points is *N*. For the extreme value distribution $1-\sigma=1-[1-F_X(P_B)]^N$, resulting in $P_B=F_X^{-1}[1-\sqrt[N]{\sigma}]$ where *X* ~ Beta (*a, b*). Similarly for the *Beta* distribution $1-\sigma=1-[1-F_X(P_B')]$, or $P_B'=F_X^{-1}[1-\sigma]$ where *X* ~ Beta (*a, b*). It is easy to see that if *N*=1, $P_B=P_B'$. However because $\sqrt[N]{\sigma}\geq\sigma$ or $1-\sqrt[N]{\sigma}\leq 1-\sigma$, P_B is less than P_B', if *N* is larger than 1. For this reason, the extreme value distribution based confidence percentile provides a more conservative (smaller) probability compared with the *Beta* distribution.

4.3. EVALUATION OF BAYESIAN TARGET RELIABILITY

In design optimization, the target reliability must be predefined. Because we do not have however, enough data, it is not practical to set the target reliability very high (e.g. $\beta=3$). If the predefined target reliability is high, the confidence percentile will be low. In this section, we will calculate the maximum target reliability based on an existing sample size *N*.

If we have *N* sample points, the safest *Beta* distribution is Beta(N+1,1). The maximum Bayesian target reliability is therefore, equal to $P_B=F_X^{-1}[1-\sqrt[N]{\sigma}]$ where *X* ~ Beta(N+1, 1), and σ is the confidence percentile. The larger the *N*, the higher the maximum target reliability is. However, the latter must be always lower than the allowable maximum reliability. For example, if we have 50 sample points, the maximum target reliability with confidence percentile 0.8 must be lower than 90%.

A Bayesian-based design optimization process entails the following steps:

1. Construct *Beta* distribution based on existing sample data.
2. Construct an extreme smallest value distribution using the above *Beta* distribution as the basic distribution.

3. Calculate the maximum target reliability for a specified confidence percentile.
4. Solve the design optimization problem using reliabilities which are based on the extreme smallest value distribution with a specified confidence percentile.

4.4. A BAYESIAN APPROACH TO DESIGN OPTIMIZATION

Reliability-based design optimization (RBDO) provides optimum designs in the presence of only random (or aleatory) uncertainty. A typical RBDO problem is formulated

$$\begin{aligned} &\min_{\mathbf{d},\boldsymbol{\mu}_\mathbf{Y}} \; f(\mathbf{d},\boldsymbol{\mu}_\mathbf{Y},\boldsymbol{\mu}_\mathbf{Z}) \\ \text{s.t.} \quad &P(g_i(\mathbf{d},\mathbf{Y},\mathbf{Z})\geq 0)\geq R_i = 1-p_{f_i}, \quad i=1,...,n \\ &\mathbf{d}^{\mathrm{L}} \leq \mathbf{d} \leq \mathbf{d}^{\mathrm{U}} \;,\; \boldsymbol{\mu}_\mathbf{Y}^{\mathrm{L}} \leq \boldsymbol{\mu}_\mathbf{Y} \leq \boldsymbol{\mu}_\mathbf{Y}^{\mathrm{U}} \end{aligned} \tag{27}$$

where $\mathbf{Y}\in R^{\ell}$ is the vector of random design variables and $\mathbf{Z}\in R^{r}$ is the vector of random design parameters.

For a variety of practical applications, the uncertain information may be provided as a mixture of sample points and probability distributions. In this case, a Bayesian approach can be used based on the confidence percentile concept. A Bayesian Approach Design Optimization (BADO) problem with a combination of random and Bayesian uncertain variables can be formulated as

$$\begin{aligned} &\min_{\mathbf{d},\mathbf{x}^N,\boldsymbol{\mu}_Y} \; \mathrm{f}(\mathbf{d},\boldsymbol{\mu}_\mathbf{Y},\boldsymbol{\mu}_\mathbf{Z},\mathbf{x}^N,\mathbf{p}^N) \\ \text{s.t.} \quad &\nabla(P(\,g_i(\mathbf{d},\mathbf{X},\mathbf{P})\geq 0\,)\geq R_i)\geq \sigma, \quad i=1,...,n \\ &\mathbf{d}_L \leq \mathbf{d} \leq \mathbf{d}_U \;,\; \boldsymbol{\mu}_\mathbf{Y}^L \leq \boldsymbol{\mu}_\mathbf{Y} \leq \boldsymbol{\mu}_\mathbf{Y}^U \\ &\mathbf{x}_L \leq \mathbf{x}^N \leq \mathbf{x}_U \end{aligned} \tag{28}$$

where $\mathbf{d}\in R^{k}$ is the vector of deterministic design variables, $\mathbf{X}\in R^{m}$ is the vector of Bayesian uncertain design variables, $\mathbf{P}\in R^{q}$ is the vector of Bayesian uncertain design parameters, $\mathbf{Y}\in R^{\ell}$ is the vector of random design variables, $\mathbf{Z}\in R^{r}$ is the vector of random design parameters, R_i is the target reliability, σ is the confidence percentile factor, and ∇ is the confidence function.

All constraints in Problem (28) are expressed using a confidence percentile because the predicted probability is distributed based on the extreme value distribution instead of having a single value. We need the confidence percentile in order to calculate a single probability value. It should be noted that the described formulation represents a double-loop optimization sequence. The design optimization of the outer loop calls a series of Bayesian uncertain constraints. Each Bayesian uncertain constraint is in general, a global optimization problem.

It should be noted that the double-loop optimization structure of Problem (28) is different from the double-loop RBDO structure. In the outer loop, the deterministic variables $\mathbf{d}$, the mean values $\boldsymbol{\mu}_\mathbf{Y}$ of

random variables and the normal points $\mathbf{x}^N$ of Bayesian uncertain variables are used as design variables. In the inner loop, based on the distributions of some of the input design variables and the available sample points for the remaining design variables, an extreme value distribution is constructed using the Bayesian approach. Subsequently, we calculate the reliability of the constraint using the confidence percentile principle. Because the Bayesian uncertain variables are represented using discrete sample points, we can not use a gradient-based local optimizer to calculate the optima. Instead, we must use a global optimizer.

4.4.1. A PRESSURE VESSEL EXAMPLE

This example considers the design of a thin-walled pressure vessel [43] which has hemispherical ends as shown in Figure 5. The design objective is to calculate the radius *R*, mid-section length *L* and wall thickness *t* in order to maximize the volume while avoiding yielding of the material in both the circumferential and radial directions under an internal pressure *P*. Geometric constraints are also considered. The material yield strength is *Y*. A safety factor *SF* = 2 is used.

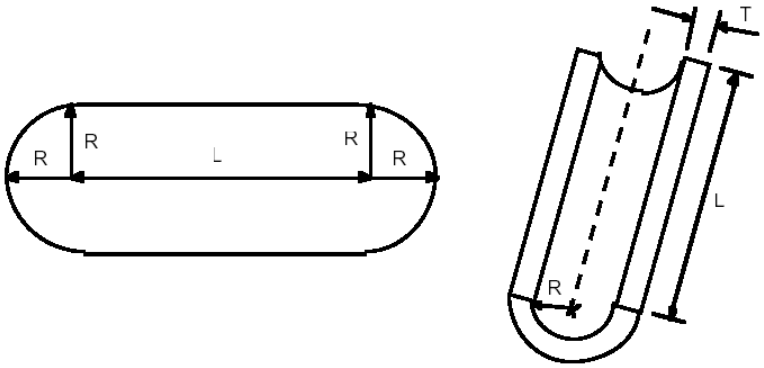


Figure 5. Thin-walled pressure vessel.

The BADO problem is stated as

$$\max_{R_N, L_N, t_N} \; f = \frac{4}{3}\pi R_N^3 + \pi R_N^2 L_N \tag{29}$$

$$\text{s.t.} \quad \nabla\left(P\left(g_i(\mathbf{X}) \geq 0\right) \geq R_i\right) \geq \sigma, \; j = 1,\ldots,5$$

$$g_1(\mathbf{X}) = 1.0 - \frac{P(R+0.5t)SF}{2tY}$$
$$g_2(\mathbf{X}) = 1.0 - \frac{P(2R^2+2Rt+t^2)SF}{(2Rt+t^2)Y}$$
$$g_3(\mathbf{X}) = 1.0 - \frac{L+2R+2t}{60}$$
$$g_4(\mathbf{X}) = 1.0 - \frac{R+t}{12}$$
$$g_5(\mathbf{X}) = 1.0 - \frac{5t}{R}$$
$$0.25 \le t_N \le 2.0$$
$$6.0 \le R_N \le 24 \quad .$$
$$10 \le L_N \le 48$$

There are three design variables (R, L, t) and two design parameters (P, Y) where P is the internal pressure and Y is the material yielding strength. The design variable R is considered a Bayesian uncertain variable. To compare results with RBDO, we sample 50 points based on the PDF of a normal distribution N (R^N ,1.5) The design variables L and t and the design parameters P and Y are normally distributed random

Table 1. Comparison of BADO and PBDO optima for the pressure vessel example.

	Design Variables			**Objective**
	R^N	L^N	t^N	$f(X)$
Det. Opt.	11.75	36	0.25	22400
Reliability Optimum (p=0.85/β=1.036)	10.1926	34.7147	0.25	15757
Bayesian Uncertainty (N=50)				
σ=0.6, p=0.85	9.50	33.2099	0.4306	13100
σ=0.6, p=0.75	10.2778	34.4444	0.2639	15970
σ=0.8, p=0.85	9.50	31.4815	0.4853	12511
σ=0.8, p=0.75	10.50	31.4815	0.3750	15745
Possibilistic Uncertainty				
σ=0.6, p=0.85	8.9464	33.0912	0.25	11314
σ=0.6, p=0.75	8.9825	34.1069	0.25	11676
σ=0.8, p=0.85	8.0464	33.0912	0.25	8908
σ=0.8, p=0.75	8.0825	34.1069	0.25	9207

variables with standard deviations equal to 3, 0.1, 50 and 13000, respectively. The mean values of P and Y are equal to 1,000 and 260,000.

For the vessel example with a combination of Bayesian and random variables, Table 1 gives the BADO results based on different target reliabilities and confidence percentiles. When the confidence percentile is σ=0.8, and the target probability is p=0.85, the BADO and RBDO results are 12511 and 15757, respectively. Because RBDO uses probabilistic distribution information, it utilizes more information compared with BADO which uses only a limited number of sample points. Thus, the BADO result should be more conservative. Because the objective is maximized in this example, the BADO result is less than the RBDO result. For the same confidence percentile of σ=0.8, if the target probability is 0.75, the BADO objective is 15745. If the target probability is 0.85, then the objective is equal to 12511. The higher the confidence percentile is, the lower the objective becomes. It should be noted that the uncertain variables in BADO are characterized only by a limited number of sample points, while only the bounds are known for the uncertain variables in PBDO. Therefore, the latter represent the least amount of information. For this reason, the PBDO design has the smallest objective value of 8908 which is obtained for a confidence percentile of σ=0.8 and a target probability of p= 0.85.

4.5. A COMBINED BAYESIAN AND EBDO APPROACH

For the above Bayesian approach, we know the range of the uncertain variables and parameters and also have a limited number of sample points. In actual engineering design however, assuming that this range is partitioned into a number of segments, we only know how many sample points are within a certain segment. In this case, we do not have an exact distribution of those sample points within the segment and we can not use therefore, the BADO methodology of section 4.4 to construct the probability distribution function of the constraint. Also, since the total number of sample points is limited, we can not assume that the probability of being within a segment is equal to the number of samples in the segment divided by the total number of samples in the whole range.

In this case, in order to utilize the existing information, we can use evidence theory to calculate the Basic Probability Assignment (BPA) for a segment of each Bayesian variable. In summary, the following tasks are performed in order to calculate the belief and plausibility of the failure region:

1) For each Bayesian input variable and parameter, construct a *Beta* distribution using the available data, and then form the extreme value distribution. Calculate the BPA structure for each variable and parameter using a predefined confidence percentile and the extreme value distribution.
2) Construct the BPA structure for the m-dimensional frame of discernment, where m is the number of input variables and parameters. Assume independent input variables and parameters.
3) Identify the failure region space based on the limit state functions (constraints).
4) Calculate the belief and plausibility measures of the failure region. The failure region must be identified only within the frame of discernment. The true probability of failure is bracketed by the belief and plausibility measures.

5) If more information becomes available, we can obtain a more accurate estimate of the BPA structure using an assumed confidence percentile.

This process is illustrated with an example in the following subsection.

4.5.1. THE PRESSURE VESSEL EXAMPLE

The same pressure vessel example of section 4.4.1 is considered here. We initially assume that we have only 100 sample points. Based on this limited available information, we only know the number of sample points within specified segments (bins) as is for example, indicated in Table 2 and shown in Figure 6 for R_N . However, we do not know the exact distribution of the sample points within each segment.

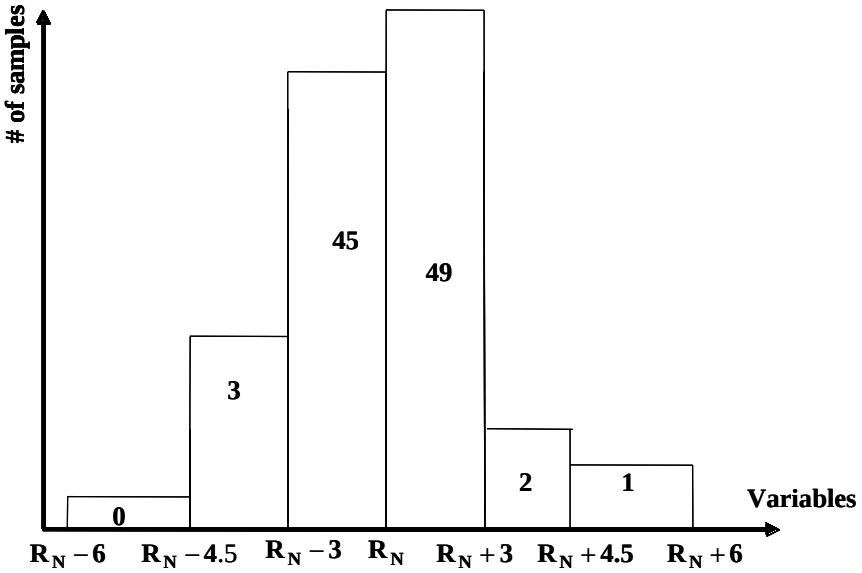


Figure 6. Histogram of sample points.

According to Figure 6, there are three sample points of R_N within the [R_N - 4.5, R_N - 3] segment. We cannot assume that the probability of having samples in that segment is equal to 3/100=0.03, because there are not enough sample points. However, we know that the extreme probability distribution for the smallest value will be $F_{Y_1}(p) = 1 - [1 - F_X(p)]^{100}$, where p denotes probability and X ~ Beta (3+1, 100-3+1) =Beta (4, 98). If we use a predefined confidence percentile of σ=0.8, then $Pr(R_N - 4.5 < R < R_N - 3 \mid$

σ=0.8) = $p = F_X^{-1}[1-\sqrt[100]{0.8}]$ =0.0054, which is smaller than the real CDF (approximately equal to 3/50=0.06). Similarly, $Pr(R_N - 3.0 < R < R_N \mid \sigma=0.8)$ = 0.3155, which is also smaller than 45/100=0.45. Because of the existing uncertainty, if the confidence percentile is large enough, the values of the BPA structure calculated using the Bayesian approach of section 4.2 will be smaller than the actual values.

Table 2. BPA structure for Bayesian Variable *R* (100 sample points).

R	# of Sample Points	BPA(Extreme Value)
$[R_N - 6.0, R_N - 4.5]$	0	2.2e-5
$[R_N - 4.5, R_N - 3.0]$	3	0.0054
$[R_N - 3.0, R_N]$	45	0.3155
$[R_N, R_N + 3.0]$	49	0.3523
$[R_N + 3.0, R_N + 4.5]$	2	0.0025
$[R_N + 4.5, R_N + 6.0]$	1	0.00068
$[R_N - 6.0, R_N + 6.0]$	---	**0.3236**

If we have more sample points, the BPA structure can be estimated more accurately. In Tables 3 and 4, we utilize 300 and 1000 sample points, respectively. For the same confidence percentile of σ=0.8, for 300 samples, the estimated probability is $Pr(R_N - 3.0 < R < R_N \mid \sigma=0.8) = p = F_X^{-1}[1-\sqrt[300]{0.8}]$=0.393. For 1000 samples, the estimated probability is equal to 0.4457, which is very close to the CDF of normal distribution 0.475. It should be noted that for 100 sample points, the same probability is equal to 0.3155. Using the calculated BPA structure, we can use steps 2 to 5 of section 4.5 to determine the optimal design using the EBDO algorithm. The more accurate the BPA structure is, the less conservative (smaller objective in this example) the optimum design will be. At the limit, the design approaches the RBDO design.

Table 3. BPA structure for Bayesian Variable *R* (300 sample points).

R	# of Sample Points	BPA(Extreme Value)
$[R_N - 6.0, R_N - 4.5]$	2	0.00057
$[R_N - 4.5, R_N - 3.0]$	6	0.0048
$[R_N - 3.0, R_N]$	145	0.393
$[R_N, R_N + 3.0]$	138	0.371
$[R_N + 3.0, R_N + 4.5]$	8	0.0078
$[R_N + 4.5, R_N + 6.0]$	1	0.00013
$[R_N - 6.0, R_N + 6.0]$	---	**0.2228**

Table 4. BPA structure for Bayesian Variable *R* (1000 Sample points).

R	# of Sample Points	BPA(Extreme Value)
[R_N - 6.0 , R_N - 4.5]	7	0.00157
[R_N - 4.5 , R_N - 3.0]	22	0.00985
[R_N - 3.0 , R_N]	501	0.4457
[R_N , R_N + 3.0]	445	0.3906
[R_N + 3.0 , R_N + 4.5]	16	0.0062
[R_N + 4.5 , R_N + 6.0]	9	0.00244
[R_N - 6.0 , R_N + 6.0]	---	**0.1436**

Based on the evidence theory the sum of all BPA values should be equal to one. However in Tables 2, 3 and 4, the sums of BPA are 0.6764, 0.7772 and 0.8564, respectively. The difference is due to unavailable information because of the limited number of sample points. It represents the uncertain belief of being somewhere between R_N - 6.0 and R_N + 6.0 without knowing the exact segment. Figure 7 shows the BPA values based on 100 samples. The uncertain belief is equal to 1-0.6764=0.3236 for the 100 sample point case, equal to 1-0.7772=0.2228 for the 300 sample point case, and equal to 1-0.8564=0.1436 for the 1000 sample point case. The uncertain belief will contribute to the belief measure (see section 3.1) if the range [R_N - 6.0, R_N + 6.0] is within the feasible area.

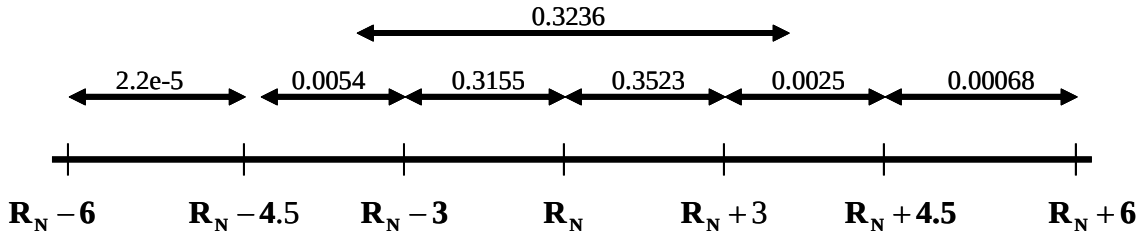


Figure 7. BPA of *R* for 100 sample points.

Considering the information from the above example, Table 5 compares the results between the Bayesian approach, EBDO and RBDO for a reliability index of $\beta = 0.385$ (p_f =0.35) and a confidence

percentile of σ=0.8. The EBDO results are based on the assumption that a very large number of sample points is available from which the BPA structure is calculated. According to Table 5, the Bayesian approach (BADO of section 4.4) provides the most conservative result (smallest objective of 15098 for p_f =0.35) compared with the EBDO and RBDO optima of 16802 and 19610, respectively, because it utilizes the least amount of information among the three approaches. For comparison purposes the PBDO optimum of 7269 is also shown in Table 5 for the zero α -cut (worst-case design) as well as the Bayesian Evidence optimum of 9805. As expected, the Bayesian Evidence optimum is better than the PBDO optimum because it uses more information. However, the Bayesian Evidence optimum of 9805 is smaller than the Bayesian optimum of 15098 because the BPA structure of the former is more conservative than the extreme value distribution of the latter. It should also be noted that although the Bayesian Evidence approach is the most conservative compared with the RBDO, EBDO and BADO approaches, it is less conservative than the worst-case scenario of PBDO, as expected. Table 5 also compares results for p_f =0.45 with similar trends observed.

Table 5. Comparison of design optimization approaches.

	Reliability Optimum (RBDO)	**Bayesian Optimum (BADO)**		**Bayesian Evidence Optimum**		**Possibility Optimum (PBDO)**		**Evidence Optimum (EBDO)**	
Design Variables	p_f **=0.35,** **(** β **=0.385)**	p_f **=0.45**	p_f **=0.35**	p_f **=0.45**	p_f **=0.35**	***a*=0,** p_f **=0.45**	***a* =0,** p_f **=0.35**	p_f **=0.45**	p_f **=0.35**
R_N	11.153	10.574	10.166	8.654	8.481	7.346	7.211	10.778	10.555
L_N	35.330	33.539	32.963	34.032	32.098	34.905	34.905	33.703	33.950
t_N	0.264	0.300	0.291	0.254	0.254	0.25	0.25	0.263	0.263
Objective									
$f(R_N, L_N)$	19610	16725	15098	10718	9805	7574	7269	17535	16802

5. SUMMARY AND CONCLUSIONS

If only the bounds are available within which an uncertain variable varies, interval analysis or possibility theory can be used to quantify and propagate uncertainty. If additional information is known in terms of expert opinions for example, the evidence theory can be used. If in addition to the bounds of the uncertain variables, there is information in the form of a discrete but limited number of sample points, we can not infer a probabilistic distribution because of the limited number of sample points. However, a Bayesian

approach can be used to estimate the probability distribution which can be subsequently, utilized in a Reliability-Based Design Optimization algorithm.

This paper has presented a method called Bayesian Approach Design Optimization (BADO) to solve design problems with uncertain variables in the form of both finite sample points and probability distributions. Also, a Bayesian approach was proposed to estimate the Basic Probability Assignment (BPA) for a specified confidence percentile, using only the number of available sample points within ranges. Subsequently, the evidence theory was used to obtain the optimal design.

A pressure vessel example was used to demonstrate the proposed Bayesian approach in design optimization and compare the results with known design methods such as reliability-based, possibility-based and evidence-based (RBDO, PBDO and EBDO) design optimization. It was clearly demonstrated that reducing the amount of available information in quantifying uncertainty, results in a more conservative design. We showed that the proposed Bayesian approach as well as the existing RBDO, PBDO and EBDO methods can quantify the tradeoff between available information and less optimal design (loss of optimality).

ACKNOWLEDGEMENTS

This study was performed with funding from the General Motors Research and Development Center and the Automotive Research Center (ARC), a U.S. Army Center of Excellence in Modeling and Simulation of Ground Vehicles at the University of Michigan. The support is gratefully acknowledged. Such support does not however, constitute an endorsement by the funding agencies of the opinions expressed in the paper.

REFERENCES

1. Tu, J., Choi, K. K. and Park, Y. H., "A New Study on Reliability-Based Design Optimization", ASME Journal of Mechanical Design, 121, 557-564, 1999.
2. Liang, J., Mourelatos, Z. P., and Tu, J., "A Single-Loop Method for Reliability-Based Design Optimization," Proceedings of ASME Design Engineering Technical Conferences, Paper# DETC2004/ DAC-57255, 2004.
3. Wu, Y.-T., Shin, Y., Sues, R. and Cesare, M., "Safety – Factor Based Approach for Probabilistic – Based Design Optimization," 42nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, Seattle, WA, 2001.
4. Lee, J. O., Yang, Y. O. and Ruy, W. S., "A Comparative Study on Reliability Index and Target Performance Based Probabilistic Structural Design Optimization," Computers and Structures, 80, 257-269, 2002.

5. Youn, B. D., Choi, K. K. and Park, Y. H. "Hybrid Analysis Method for Reliability-Based Design Optimization," ASME Journal of Mechanical Design, 125(2), 221-232, 2001.
6. Oberkampf, W., Helton, J. and Sentz, K., "Mathematical Representations of Uncertainty," AIAA Non-Deterministic Approaches Forum, AIAA 2001-1645, Seattle, WA, April 16-19, 2001.
7. Sentz, K. and Ferson, S., "Combination of Evidence in Dempster – Shafer Theory," Sandia National Laboratories Report SAND2002-0835, April 2002.
8. Klir, G. J. and Yuan, B., Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice Hall, 1995.
9. Klir, G. J. and Filger, T. A., Fuzzy Sets, Uncertainty, and Information, Prentice Hall, 1988.
10. Yager, R. R., Fedrizzi, M. and Kacprzyk, J. (Editors), Advances in the Dempster – Shafer Theory of Evidence, John Wiley & Sons, Inc., 1994.
11. Dubois, D. and Prade, H., Possibility Theory, Plenum Press, New York, 1988.
12. Moore, R. E., Interval Analysis, Prentice-Hall, 1966.
13. Zadeh, L. A., "Fuzzy Sets," Information and Control, 8, 338-353, 1965.
14. Muhanna, R. L. and Mullen, R. L., "Uncertainty in Mechanics Problems – Interval-Based Approach," Journal of Engineering Mechanics, 127(6), 557-566, 2001.
15. Mullen, R. L. and Muhanna, R. L., "Bounds of Structural Response for all Possible Loadings," ASCE Journal of Structural Engineering, 125(1), 98-106, 1999.
16. Penmetsa, R. C. and Grandhi, R. V., "Efficient Estimation of Structural Reliability for Problems with Uncertain Intervals," Computers and Structures, 80, 1103-1112, 2002.
17. Du, X. and Sudjianto, A., "Reliability-Based Design with a Mixture of Random and Interval Variables," Proceedings of ASME Design Engineering Technical Conferences, Paper# DETC2003/ DAC-48709, 2003.
18. Rao, S. S. and Cao, L., "Optimum Design of Mechanical Systems Involving Interval Parameters," ASME Journal of Mechanical Design, 124, 465-472, 2002.
19. Gu, X., Renaud, J. E. and Batill, S. M., "An Investigation of Multidisciplinary Design Subject to Uncertainties," 7th AIAA/USAF/NASA/ISSMO Multidisciplinary Analysis & Optimization Symposium, St. Louis, Missouri, 1998.
20. Elishakoff, I. E., Haftka, R. T. and Fang, J. "Structural Design under Bounded Uncertainty – Optimization with Anti-Optimization," Computers and Structures, 53, 1401-1405, 1994.
21. Lombardi, M. and Haftka, R. T., "Anti-Optimization Technique for Structural Design under Load Uncertainties," Computer Methods in Applied Mechanics and Engineering, 157, 19-31, 1998.
22. Du, X. and Chen, W., "An Integrated Methodology for Uncertainty Propagation and Management in Simulation-Based Systems Design," AIAA Journal, 38(8), 1471-1478, 2000.
23. Mourelatos, Z. P. and Zhou, J., "Reliability Estimation with Insufficient Data Based on Possibility theory", AIAA Journal, 43(8), 1696-1705, 2005.
24. Zhou, J. and Mourelatos, Z. P., "A Sequential Algorithm for Possibility-Based Design Optimization," ASME Journal of Mechanical Design, 130(1), 2008.
25. Choi, K. K., Du, L. and Youn, B. D., "A New Fuzzy Analysis Method for Possibility-Based Design Optimization," 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, AIAA 2004-4585, Albany, NY, 2004.

26. Nikolaidis, E., Chen, S., Cudney, H., Haftka, R. T. and Rosca, R., "Comparison of Probability and Possibility for Design Against Catastrophic Failure Under Uncertainty," ASME Journal of Mechanical Design, 126, 2004.
27. Oberkampf, W. L. and Helton, J. C., "Investigation of Evidence Theory for Engineering Applications," AIAA Non-Deterministic Approaches Forum, AIAA 2002-1569, Denver, CO, April, 2002.
28. Bae, H-R, Grandhi, R. V. and Canfield, R. A., "An Approximation Approach for Uncertainty Quantification Using Evidence Theory," Reliability Engineering and System Safety, 86, 215-225, 2004.
29. Bae, H-R, Grandhi, R. V. and Canfield, R. A., "Epistemic Uncertainty Quantification Techniques Including Evidence Theory for Large-Scale Structures," Computers and Structures, 82, 1101-1112, 2004.
30. Agarwal, H., Renaud, J. E., Preston, E. L. and Padmanabhan, D., "Uncertainty Quantification Using Evidence Theory in Multidisciplinary Design Optimization," Reliability Engineering and System Safety, 85, 281-294, 2004.
31. Mourelatos, Z. P. and Zhou, J., "A Design Optimization Method using Evidence Theory," ASME Journal of Mechanical Design, 128(4), 901-908, 2006.
32. Box, G. and Tiao, G., Bayesian Inference in Statistical Analysis, John Wiley & Sons, New York, 1992.
33. Dempster, A. P., "A Generalization of Bayesian Inferences based on a Sample from a Finite Univariate Population," Biometrics, 54(2-3), 515-528, 1967.
34. Siu, N. O. and Kelly, D. L., "Bayesian Parameter Estimation in Probabilistic Risk Assessment," Reliability Engineering System Safety, 62(1-2), 89-116, 1998.
35. Ross, T. J., Fuzzy Logic with Engineering Applications, McGraw Hill, 1995.
36. Sentz, K. and Ferson, S., "Combination of Evidence in Dempster – Shafer Theory," Sandia National Laboratories Report SAND2002-0835, April 2002.
37. Dong, W. M. and Shah, H. C., "Vertex Method for Computing Functions of Fuzzy Variables," Fuzzy Sets and Systems, 24, 65-78, 1987.
38. Choi, K. K., Du, L. and Youn, B. D., "A New Fuzzy Analysis Method for Possibility-Based Design Optimization," 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, AIAA 2004-4585, Albany, NY, 2004.
39. Tu, J. and Jones, D. R., "Variable Screening in Metamodel Design by Cross-Validated Moving Least Squares Method", Proceedings 44th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, AIAA-2003-1669, Norfolk, VA, April 7-10, 2003.
40. Ye, K. Q., Li, W. and A. Sudjianto, "Algorithmic Construction of Optimal Symmetric Latin Hypercube Designs", Journal of Statistical Planning and Inference, 90, 145-159, 2000.
41. Jones, D. R., Perttunen, C. D. and Stuckman, B. E., "Lipschitzian Optimization without the Lipschitz Constant," Journal of Optimization Theory and Applications, 73(1), 157-181, 1993.
42. Subroto, G. and Papalambros, P.Y., "A Bayesian Approach to Reliability-Based Optimization with Incomplete Information," ASME Journal of Mechanical Design, 128(4), 909-918, 2006.

43. Lewis, K. and Mistree, F., "Collaborative, Sequential and Isolated Decisions in Design," Proceedings of ASME Design Engineering Technical Conferences, Paper# DETC1997/ DTM-3883, 1997.

Propagation and Provenance of Probabilistic and Interval Uncertainty in Cyberinfrastructure-Related Data Processing and Data Fusion

Paulo Pinheiro da Silva[1], Aaron Velasco[2], Martine Ceberio[1], Christian Servin[1], Matthew G. Averill[2], Nicholas Del Rio[1], Luc Longpré[1], and Vladik Kreinovich[1]
Departments of [1] Computer Science and [2] Geological Sciences
University of Texas, El Paso, TX 79968, USA, contact vladik@utep.edu

Abstract. In the past, communications were much slower than computations. As a result, researchers and practitioners collected different data into huge databases located at a single location such as NASA and US Geological Survey. At present, communications are so much faster that it is possible to keep different databases at different locations, and automatically select, transform, and collect relevant data when necessary. The corresponding cyberinfrastructure is actively used in many applications. It drastically enhances scientists' ability to discover, reuse and combine a large number of resources, e.g., data and services.

Because of this importance, it is desirable to be able to gauge the the uncertainty of the results obtained by using cyberinfrastructure. This problem is made more urgent by the fact that the level of uncertainty associated with cyberinfrastructure resources can vary greatly – and that scientists have much less control over the quality of different resources than in the centralized database. Thus, with the cyberinfrastructure promise comes the need to analyze how data uncertainty *propagates* via this cyberinfrastructure.

When the resulting accuracy is too low, it is desirable to produce the *provenance* of this inaccuracy: to find out which data points contributed most to it, and how an improved accuracy of these data points will improve the accuracy of the result. In this paper, we describe algorithms for propagating uncertainty and for finding the provenance for this uncertainty.

Keywords: cyberinfrastructure, uncertainty, interval uncertainty, probabilistic uncertainty, provenance

1. Cyberinfrastructure: A Brief Overview

Practical problem: need to combine geographically separate computational resources. In different knowledge domains in science and engineering, there is a large amount of data stored in different locations, and there are many software tools for processing this data, also implemented at different locations. Users may be interested in different information about this domain.

Sometimes, the information required by the user is already stored in *one of* the *databases*. For example, if we want to know the geological structure of a certain region in Texas, we can get this

information from the geological map stored in Austin. In this case, all we need to do to get an appropriate response from the query is to get this data from the corresponding database.

In other cases, different pieces of the information requested by the user are *stored at different locations*. For example, if we are interested in the geological structure of the Rio Grande Region, then we need to combine data from the geological maps of Texas, New Mexico, and the Mexican state of Chihuahua. In such situations, a correct response to the user's query requires that we access these pieces of information from different databases located at different geographic locations.

In many other situations, the appropriate answer to the user's request requires that we not only collect the relevant data $x_1, \ldots, x_n$, but that we also use some *data processing* algorithms $f(x_1, \ldots, x_n)$ to process this data. For example, if we are interested in the large-scale geological structure of a geographical region, we may also use the gravity measurements from the gravity databases. For that, we need special algorithms to transform the values of gravity at different locations into a map that describes how the density changes with location. The corresponding data processing programs often require a lot of computational resources; as a result, many such programs reside on computers located at supercomputer centers, i.e., on computers which are physically separated from the places where the data is stored.

The need to combine computational resources (data and programs) located at different geographic locations seriously complicates research.

Centralization of computational resources – traditional approach to combining computational resources; its advantages and limitations. Traditionally, a widely used way to make these computational resources more accessible was to move all these resources to a *central location*. For example, in the geosciences, the US Geological Survey (USGS) was trying to become a central repository of all relevant geophysical data. However, this centralization requires a large amount of efforts: data is presented in different formats, the existing programs use specific formats, etc. To make the central data repository efficient, it is necessary:

- to reformat all the data,
- to rewrite all the data processing programs – so that they become fully compatible with the selected formats and with each other, etc.

The amount of work that is needed for this reformatting and rewriting is so large that none of these central repositories really succeeded in becoming an easy-to-use centralized database.

Cyberinfrastructure – a more efficient approach to combining computational resources. Cyberinfrastructure technique is a new approach that provides the users with the efficient way to submit requests without worrying about the geographic locations of different computational resources – and at the same time avoid centralization with its excessive workloads. The main idea behind this approach is that *we keep all (or at least most) the computational resources*

- *at their current locations*,
- *in their current formats*.

To expedite the use of these resources:

- we supplement the local computational resources with the "metadata", i.e., with the information about the formats, algorithms, etc.,
- we "wrap up" the programs and databases with auxiliary programs that provide data compatibility into *web services*,

and, in general, we provide a cyberinfrastructure that uses the metadata to automatically combine different computational resources.

For example, if a user is interested in using the gravity data to uncover the geological structure of the Rio Grande region, then the system should automatically:

- get the gravity data from the UTEP and USGS gravity databases,
- convert them to a single format (if necessary),
- forward this data to the program located at San Diego Supercomputer Center, and
- move the results back to the user.

This example is exactly what we have been designing under the NSF-sponsored Cyberinfrastructure for the Geosciences (GEON) project; see, e.g., (Aguiar et al., 2004; Aldouri et al., 2004; Averill et al., 2005; Ceberio et al., 2006; Ceberio et al., 2005; Keller et al., 2006; Platon et al., 2005; Schiek et al., 2007; Sinha, 2006; Torres et al., 2004; Wen et al., 2001; Xie et al., 2003), and what we are currently doing under the NSF-sponsored Cyber-Share project. This is similar to what other cyberinfrastructure projects are trying to achieve.

Technical advantages of cyberinfrastructure: a brief summary. In different knowledge domains, there is a large amount of data stored in different locations; algorithms for processing this data are also implemented at different locations. Web services – and, more generally, cyberinfrastructure – provide the users with an efficient way to submit requests without worrying about the geographic locations of different computational resources (databases and programs) – and avoid centralization with its excessive workloads (Gates at el., 2006). Web services enable the user to receive the desired data $x_1, \ldots, x_n$ and the results $y = f(x_1, \ldots, x_n)$ of processing this data.

Main advantage of cyberinfrastructure: the official NSF viewpoint. Up to now, we concentrated on the technical advantages of cyberinfrastructure. However, its advantages (real and potential) go beyond technical. According to the final report of the National Science Foundation (NSF) Blue Ribbon Advisory Panel on Cyberinfrastructure, "a new age has dawned in scientific and engineering research, pushed by continuing progress in computing, information, and communication technology, and pulled by the expanding complexity, scope, and scale of today's challenges. The capacity of this technology has crossed thresholds that now make possible a comprehensive 'cyberinfrastructure' on which to build new types of scientific and engineering knowledge environments and organizations and to pursue research in new ways and with increased efficacy.

Such environments and organizations, enabled by cyberinfrastructure, are increasingly required to address national and global priorities, such as understanding global climate change, protecting

our natural environment, applying genomics-proteomics to human health, maintaining national security, mastering the world of nanotechnology, and predicting and protecting against natural and human disasters, as well as to address some of our most fundamental intellectual questions such as the formation of the universe and the fundamental character of matter."

Main advantage of cyberinfrastructure: in short. Cyberinfrastructure greatly enhances the ability of scientists to discover, reuse and combine a large number of resources, including data and services.

2. Data Processing vs. Data Fusion

Practically important situation: it is difficult to directly measure the desired quantity with a given accuracy. In practice, we are often interested in a quantity y which is difficult (or even impossible) to directly measure with the desired accuracy.

In this situation, there are two ways to estimate the value of the desired quantity y with the desired accuracy:

- measuring *other* (related) easier-to-measure quantities and then extracting the value y from these measurements; this is called *data processing*; and
- measuring the same quantity y many times and combining the results of these measurements; this is called *data fusion*.

Important terminological comment. To avoid confusion, we would like to emphasize that sometimes, the term "data processing" refers to *all* possible processing of data by computers. In this more general sense, data fusion can be viewed as a particular case of data processing. In this paper, we limit ourselves to the narrow sense of the term "data processing".

First idea: data processing. One possible way of estimating the desired quantity y with a given accuracy is to look for easier-to-measure quantities $x_1, \ldots, x_n$ which are related to the desired y by a known dependence $y = f(x_1, \ldots, x_n)$. Based on the results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of measuring these auxiliary quantities, we can then compute an estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ for the quantity y.

The entire process of measurement followed by estimation is called an *indirect measurement* of y; see, e.g., (Rabinovich, 2005). The actual computation of $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is known as *data processing*.

Comment. Data processing is one of the main reasons why computers were invented in the first place, and it is still one of the major uses of computers.

Data processing: example from the geosciences. In geosciences, we want to know the structure at different depth. To determine this structure, we need to know the density y of the material at different depths. It is very difficult (and very expensive) to *directly* measure this density. Therefore, geoscientists measure this density *indirectly*.

For example, during an earthquake, geoscientists record the seismic waves at sensors located at different points on the Earth surface. As a result, we obtain the travel times $x_1, \ldots, x_n$ of the

seismic signal from the earthquake location to the sensor location. Based on these travel times, we determine the structure of the Earth along the paths of the corresponding seismic waves.

The main limitations of this analysis is that earthquakes are unpredictable, they occur only at some locations and as a result, several important areas of the earth are not well covered by the corresponding paths. Thus, in addition to such *passive* (earthquake-related) seismic analysis, geoscientists also perform *active* seismic experiments, in which they start small-scale explosions in specially allocated areas and measure the travel times of the generated seismic waves. Based on these travel times, we can also determine the desired Earth structure, i.e., to be more precise, the values of the density at different depths and different locations; see, e.g., (Averill, 2007; Hole, 1992; Parker, 1994).

Specifics of data processing in cyberinfrastructure. In *traditional* data processing, when we want to know the value of a difficult-to-measure quantity y, and we know the relation between this quantity and easier-to-measure quantities x_i, we then *measure* the values x_i and use the results $\widetilde{x}_i$ of these measurements to compute the estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ for the desired quantity y.

As we have mentioned earlier, the main idea of a *cyberinfrastructure* is to keep all the existing measurement results readily available. Thus, with cyberinfrastructure in place, first we *look for the results* $\widetilde{x}_i$ of *measuring* x_i in the existing databases. Only when we do not find these results $\widetilde{x}_i$ – or when these results are not accurate enough – only then we actually start measuring.

Specifics of data processing in cyberinfrastructure: example from the geosciences. For example, if we want to know the geophysical structure in a certain area, instead of performing active seismic experiments we first try to combine all the known results of active seismic experiments which are related to this area. If this information is not sufficient, then we will, of course, have to perform new experiments.

Second idea: data fusion. The second idea is also very straightforward: since we cannot achieve the desired accuracy in the desired quantity y by a *single* measurement, we perform *several* independent measurements of this same quantity, and then combine ("fuse") the resulting (less accurate) values $\widetilde{y}_1, \ldots, \widetilde{y}_n$ into a single (more accurate) estimate $\widetilde{y}$ for y.

This combination can be as simple as taking an arithmetic average $\widetilde{y} = \dfrac{1}{n} \cdot (\widetilde{y}_1 + \ldots + \widetilde{y}_n)$, or it can be more complicated: e.g., taking a weighted average or applying some non-linear combination technique. Several such techniques will be described and analyzed later in this paper.

Data fusion: examples. Data fusion is, in effect, a standard procedure that is routinely done in engineering and scientific practice (see, e.g., (Rabinovich, 2005)):

- the super-precise time is obtained by using three (or more) independent precise clocks and combining the results of these measurements;
- in medical practice, important quantities such as high blood pressure are often performed at least twice, etc.

Specifics of data processing in cyberinfrastructure. In the *traditional* engineering and scientific practice, we actually *measure* the desired quantity y several times. With *cyberinfrastructure* in

place, we first *look for the existing results of measuring* the desired quantity, and try to fuse them into a single estimate.

Only if the accuracy of the resulting estimate is not good enough, then we perform additional measurements.

Combination of data processing and data fusion. In real life, to achieve the desired accuracy, it is often necessary both to use multiple measurement *and* to perform indirect measurements. In other words, in many situations, we need to combine data processing and data fusion.

For example, for many geological regions, we already have several density distributions obtained by processing different seismic data. To get a more accurate picture, it is reasonable to combine (fuse) the resulting approximate values of density, i.e., to fuse the existing data processing results.

3. Need for Uncertainty Propagation, and for Provenance of Uncertainty

Need for uncertainty propagation. As we have mentioned, one of the main reasons why we need data processing (i.e., indirect measurements) and data fusion (i.e., multiple measurements) in the first place is that the accuracy of the original direct measurement is not high enough. It is therefore important to make sure that after the proposed data processing and/or data fusion, we get the desired accuracy. In other words, we must find out how the uncertainty (inaccuracy) of the direct measurement results *propagates* via the infrastructure.

The need for uncertainty propagation is enhanced by the fact that the level of uncertainty associated with cyberinfrastructure resources can vary greatly – as well as the level of uncertainty of any response derived from such resources. Also, in contrast to the centralized platform, in cyberinfrastructure, scientists have less control about the quality of different resources. Thus, the cyberinfrastructure promise comes along with the need to support the associated uncertainty analysis uncertainty propagation.

Need for the provenance of uncertainty. When the resulting accuracy is sufficient, we get the desired estimate $\widetilde{y}$. However, sometimes, the resulting accuracy is still too low. In this situation, it is desirable to produce the *provenance* of this inaccuracy: to find out which data points contributed most to it, and how an improved accuracy of these data points will improve the accuracy of the result.

Comment. In this paper, we mainly deal with the provenance of *uncertainty*. It is worth mentioning that in general, other aspects of provenance are also very important: e.g., to be able to adequately gauge the *reliability* of different measurement results (and thus, to form a decision on how much we trust these results), we must take into account the provenance of these results – i.e., which team performed these measurements, what auxiliary data was used in pre-processing these measurement results, etc.

4. Uncertainty of the Results of Direct Measurements: Probabilistic and Interval Approaches

Measurement uncertainty: general description. To find out how the inaccuracies $\Delta x_i = \widetilde{x}_i - x_i$ of direct measurements (= measurement errors) propagate through the cyberinfrastructure, we need to recall how these inaccuracies Δx_i are usually described.

The manufacturer of the measuring instrument must supply us with an upper bound Δ_i on the measurement error. If no such upper bound is supplied, this means that no accuracy is guaranteed, and the corresponding "measuring instrument" is practically useless. In this case, once we performed a measurement and got a measurement result $\widetilde{x}_i$, we know that the actual (unknown) value x_i of the measured quantity belongs to the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$, where $\underline{x}_i = \widetilde{x}_i - \Delta_i$ and $\overline{x}_i = \widetilde{x}_i + \Delta_i$.

Probabilistic uncertainty. In many practical situations, we not only know the interval $[-\Delta_i, \Delta_i]$ of possible values of the measurement error; we also know the probability of different values Δx_i within this interval. This knowledge underlies the traditional engineering approach to estimating the error of indirect measurement, in which we assume that we know the probability distributions for measurement errors Δx_i.

These probabilities are often described by a normal distribution, so in standard engineering textbook on measurement, it is usually assumed that the distribution of Δx_i is normal, with 0 average and known standard deviation σ_i; see, e.g. (Fuller, 1987; Rabinovich, 2005).

In general, we can determine the desired probabilities of different values of Δx_i by comparing the results of measuring with this instrument with the results of measuring the same quantity by a standard (much more accurate) measuring instrument. Since the standard measuring instrument is much more accurate than the one use, the difference between these two measurement results is practically equal to the measurement error; thus, the empirical distribution of this difference is close to the desired probability distribution for measurement error.

Interval uncertainty. There are two cases, however, when in practice, we do not determine the probabilities:

- First is the case of cutting-edge measurements, e.g., measurements in fundamental science. When a Hubble telescope detects the light from a distant galaxy, there is no "standard" (much more accurate) telescope floating nearby that we can use to calibrate the Hubble: the Hubble telescope is the best we have.
- The second case is the case of measurements on the shop floor. In this case, in principle, every sensor can be thoroughly calibrated, but sensor calibration is so costly – usually costing ten times more than the sensor itself – that manufacturers rarely do it.

In both cases, we have no information about the probabilities of Δx_i; the only information we have is the upper bound on the measurement error.

In this case, after we performed a measurement and got a measurement result $\widetilde{x}_i$, the only information that we have about the actual value x_i of the measured quantity is that it belongs to the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

What we consider in this paper. For each of the 2 techniques for improving accuracy (data processing and data fusion), we must therefore consider 2 possible situations:

– when we know the probabilities of inaccuracies of direct measurements, and

– when we only know upper bounds (intervals) for these inaccuracies.

So, we have $2 \times 2 = 4$ possible situations. In this paper, we will consider all four situations. We start with data processing under probabilistic and interval uncertainty, and then we cover data fusion under both types of uncertainty.

For three of these four situations, the answer is reasonably straightforward; for the fourth one, we will come up with new formulas.

5. Typical Situation: Measurement Errors are Reasonably Small

Before we start analyzing different situations, let us mention that in this paper, we will only consider a typical situation in which the direct measurements are accurate enough, so that the resulting approximation errors Δx_i are small, and terms which are quadratic (or of higher order) in Δx_i can be safely neglected. In such situations, for data processing, the dependence of the desired value $y = f(x_1, \ldots, x_n) = f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n)$ on Δx_i can be safely assumed to be linear.

When approximation errors are small, we can simplify the expression for $\Delta y = \widetilde{y} - y = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(x_1, \ldots, x_n)$, if we expand the function f in Taylor series around the point $(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ and restrict ourselves only to linear terms in this expansion. As a result, we get the expression

$$\Delta y = c_1 \cdot \Delta x_1 + \ldots + c_n \cdot \Delta x_n,$$

where by c_i we denoted the value of the partial derivative $\dfrac{\partial f}{\partial x_i}$ at the point $(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

In the linear approximation, for small $h > 0$, we have $f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i + h, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n) \approx f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n) + c_i \cdot h$, hence we can determine c_i as

$$c_i = \frac{1}{h} \cdot (f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i + h, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n) - \widetilde{y}).$$

Comment. There are practical situations when the accuracy of the direct measurements is not high enough, and hence, quadratic terms cannot be safely neglected (see, e.g., (Jaulin, 2001) and references therein). In this case, the problem of error estimation for indirect measurements becomes computationally difficult (NP-hard) even when the function $f(x_1, \ldots, x_n)$ is quadratic (Kreinovich et al., 1998; Vavasis, 1991). However, in most real-life situations, the possibility to ignore quadratic terms is a reasonable assumption, because, e.g., for an error of 1% its square is a negligible 0.01%.

6. Case of Data Processing

Propagation of uncertainty through data processing: case of probabilistic uncertainty. In the statistical setting, the desired measurement error Δy is a linear combination of independent

Gaussian variables Δx_i. Therefore, Δy is also normally distributed, with 0 average and the standard deviation

$$\sigma = \sqrt{c_1^2 \cdot \sigma_1^2 + \ldots + c_n^2 \cdot \sigma_n^2}.$$

Comment. A similar formula holds if we *do not* assume that Δx_i are normally distributed: it is sufficient to assume that they are independent variables with 0 average and known standard deviations σ_i.

Uncertainty provenance in data processing: case of probabilistic uncertainty. The above formula not only describes the *propagation* of uncertainty, it also describes the *provenance* of uncertainty. Indeed, for every i, since $\sigma^2 = \sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2$, we know which component of the resulting variance σ^2 comes from the inaccuracy σ_i of the i-th measurement. We can therefore easily predict how replacing the i-th measurement with a more accurate one (with $\sigma_i^{\text{new}} \ll \sigma_i$) will affect the resulting variance σ^2.

Propagation of uncertainty through data processing: case of interval uncertainty. In the interval setting, we do not know the probability of different errors Δx_i; instead, we only know that $|\Delta x_i| \le \Delta_i$. In this case, the sum $\sum_{i=1}^{n} c_i \cdot \Delta x_i$ attains its largest possible value if each term $c_i \cdot \Delta x_i$ in this sum attains the largest possible value:

- If $c_i \ge 0$, then this term is a monotonically non-decreasing function of Δx_i, so it attains its largest value at the largest possible value $\Delta x_i = \Delta_i$; the corresponding largest value of this term is $c_i \cdot \Delta_i$.
- If $c_i < 0$, then this term is a decreasing function of Δx_i, so it attains its largest value at the smallest possible value $\Delta x_i = -\Delta_i$; the corresponding largest value of this term is $-c_i \cdot \Delta_i = |c_i| \cdot \Delta_i$.

In both cases, the largest possible value of this term is $|c_i| \cdot \Delta_i$, so, the largest possible value of the sum Δy is

$$\Delta = |c_1| \cdot \Delta_1 + \ldots + |c_n| \cdot \Delta_n.$$

Similarly, the smallest possible value of Δy is $-\Delta$.

Hence, the interval of possible values of Δy is $[-\Delta, \Delta]$, and the interval of possible values of the actual value y is $[\widetilde{y} - \Delta, \widetilde{y} + \Delta]$.

Uncertainty provenance in data processing: case of interval uncertainty. The above formula not only describes the *propagation* of uncertainty, it also describes the *provenance* of uncertainty. Indeed, for every i, since $\Delta = \sum_{i=1}^{n} |c_i| \cdot \Delta_i$, we know which component of the resulting approximation error Δ comes from the inaccuracy Δ_i of the i-th measurement. We can therefore easily predict how replacing the i-th measurement with a more accurate one (with $\Delta_i^{\text{new}} \ll \Delta_i$) will affect the resulting approximation error Δ.

7. Case of Data Fusion

Propagation of uncertainty through data fusion: case of probabilistic uncertainty. In data fusion, we know several results $\widetilde{y}_1, \ldots, \widetilde{y}_n$ of measuring the same quantity y. Under probabilistic uncertainty, we assume that the corresponding n measurements errors are independent normally distributed random variables with 0 mean and known standard deviations σ_i. In this case, for each possible value y, the probability density ρ_i of getting $\widetilde{y}_i$ is equal to

$$\rho_i(y) = \frac{1}{\sqrt{2\pi} \cdot \sigma_i} \cdot \exp\left(-\frac{(y - \widetilde{y}_i)^2}{2\sigma_i^2}\right),$$

and thus, the probability density $\rho(y)$ of having the given n measurements is equal to

$$\rho(y) = \rho_1(y) \cdot \ldots \cdot \rho_n(y) = \text{const} \cdot \exp\left(-\sum_{i=1}^{n} \frac{(y - \widetilde{y}_i)^2}{2\sigma_i^2}\right).$$

As a resulting estimate $\widetilde{y}$ for the desired (unknown) quantity y, it is then reasonable to select the most probable value, i.e., the value for which the probability density $\rho(y)$ is the largest.

Maximizing $\rho(y)$ is equivalent to minimizing the quadratic function $-\ln(\rho(y))$; differentiating this quadratic expression with respect to y and equating the derivative to 0, we conclude that

$$\widetilde{y} = \frac{1}{\sum\limits_{i=1}^{n} \frac{1}{\sigma_i^2}} \cdot \sum_{i=1}^{n} \frac{\widetilde{y}_i}{\sigma_i^2}.$$

This estimate is a linear combination of normally distributed estimates $\widetilde{y}_i$ with mean y and standard deviation σ_i, with coefficients $c_i = \sigma_i^{-1} / \left(\sum\limits_{j=1}^{n} \sigma_j^{-2}\right)$. Thus, $\widetilde{y}$ is also normally distributed, with the same mean y and the standard deviation $\sigma^2 = \sum\limits_{i=1}^{n} c_i^2 \cdot \sigma_i^2$, i.e., with standard deviation

$$\sigma^2 = \frac{1}{\sum\limits_{i=1}^{n} \frac{1}{\sigma_i^2}}.$$

This formula can also be rewritten as

$$\frac{1}{\sigma^2} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}.$$

Uncertainty provenance in data fusion: case of probabilistic uncertainty. The above formula not only describes the *propagation* of uncertainty, it also describes the *provenance* of uncertainty. Indeed, for every i, since $\sigma^{-2} = \sum\limits_{i=1}^{n} \sigma_i^{-2}$, we know which component of the resulting variance σ^2 comes from the inaccuracy σ_i of the i-th measurement.

We can therefore easily predict how replacing the i-th measurement with a more accurate one (with $\sigma_i^{\rm new} \ll \sigma_i$) will affect the resulting variance σ^2. Good news is we can predict this accuracy beforehand, without actually performing the measurements – since the resulting accuracy σ depends only on the accuracies σ_i of individual measurements and not on the results of these measurements.

Case of unknown probabilistic uncertainty. Sometimes, we do not know the accuracy of the fused measurements. In this case, we can use the differences between the measurement results $\widetilde{y}_1, \ldots, \widetilde{y}_n$ to estimate the standard deviation $\sigma_1 = \ldots = \sigma_n$ of the corresponding measurements by using the usual statistical formula $\sigma_1^2 = \dfrac{1}{n-1} \cdot \sum\limits_{i=1}^{n} (\Delta y_i - E)^2$, where $E \stackrel{\rm def}{=} \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} \Delta y_i$.

Propagation of uncertainty through data fusion: case of interval uncertainty. Under interval uncertainty, we know n results $\widetilde{y}_1, \ldots, \widetilde{y}_n$ of measuring the same quantity y, and we know the accuracy Δ_i of each measurement. Thus, for each i, we know that the actual (unknown) value y of the desired quantity must belong to the interval $\mathbf{y}_i \stackrel{\rm def}{=} [\widetilde{y}_i - \Delta_i, \widetilde{y}_i + \Delta_i]$.

Fusion here is straightforward: the set of all the values y which belong to all n intervals is equal to the *intersection* $\mathbf{y} = [\underline{y}, \overline{y}] = \mathbf{y}_1 \cap \ldots \cap \mathbf{y}_n$ of these intervals. Here, $\underline{y} = \max(\widetilde{y}_1 - \Delta y_1, \ldots, \widetilde{y}_n - \Delta y_n)$, $\overline{y} = \min(\widetilde{y}_1 + \Delta y_1, \ldots, \widetilde{y}_n + \Delta y_n)$, and the accuracy $\Delta = \dfrac{\overline{y} - \underline{y}}{2}$ of the fused estimate can be computed as

$$\Delta = \frac{1}{2} \cdot (\min(\widetilde{y}_1 + \Delta y_1, \ldots, \widetilde{y}_n + \Delta y_n) - \max(\widetilde{y}_1 - \Delta y_1, \ldots, \widetilde{y}_n - \Delta y_n)).$$

Case of unknown interval uncertainty: a reasonable approach. Sometimes, we do not know the accuracy Δ_i of the fused measurements. In this case, it is reasonable to get a single estimate for $\Delta_1 = \ldots = \Delta_n$ for all these measurements. We know that the intervals $[\widetilde{y}_i - \Delta_1, \widetilde{y}_1 + \Delta_1]$ must intersect – since they all contain the actual (unknown) value of the desired quantity y.

For the intersection to be non-empty, every lower bound $\widetilde{y}_i - \Delta_1$ must be smaller than or equal to every upper bound $\Delta y_j + \Delta_1$. Thus, we must have $\widetilde{y}_i - \widetilde{y}_j \le 2\Delta_1$. So, we can conclude that $\Delta_1 \ge \dfrac{1}{2} \cdot (\max\limits_i \Delta y_i - \min\limits_i \Delta y_i)$.

Case of unknown interval uncertainty: seemingly reasonable proposal and its limitations. The actual value Δ_1 can be larger than this half-difference, but as a first approximation, it may be reasonable to take $\Delta_1 \approx \dfrac{1}{2} \cdot (\max\limits_i \Delta y_i - \min\limits_i \Delta y_i)$. This particular choice may not be the most adequate, since in this case, the intersection of the corresponding intervals $[\widetilde{y}_i - \Delta_1, \widetilde{y}_1 + \Delta_1]$ consists of a single point – the midpoint $y_{\rm mid} \stackrel{\rm def}{=} \dfrac{1}{2} \cdot (\max\limits_i \Delta y_i + \min\limits_i \Delta y_i)$. This conclusion is somewhat misleading because it erroneously suggests that we know the exact value of the estimated quantity.

In the following text, we will return to this problem and show how to get a somewhat more adequate estimate.

Uncertainty provenance in data fusion: case of interval uncertainty. The above formula describes the *propagation* of uncertainty. From the viewpoint of uncertainty propagation, this for-

mula is even simpler than in the probabilistic case. So, from the computational viewpoint, we can say that as far as propagation of uncertainty is concerned, the situation with interval uncertainty is easier-to-handle than the situation with probabilistic uncertainty.

With provenance, however, the situation is exactly opposite. For probabilistic uncertainty, we can predict the resulting accuracy beforehand, without actually performing the measurements – since the resulting accuracy σ depends only on the accuracies σ_i of individual measurements and not on the results of these measurements. In contrast, for the interval uncertainty, for the same accuracies $\Delta_1, \ldots, \Delta_n$ of the individual measurements, we can get different accuracy Δ of the fusion result – depending on the actual measurement results.

Let us illustrate this problem on the simplest example, when the actual (unknown) value is $y = 0$, and we fuse two measurements with the exact same accuracy $\Delta_1 = \Delta_2 = 1$. All we know about the results of these two measurements is that the resulting intervals contain the actual value y. Since this is the only restriction, we can two radically different extreme situations:

- It is possible that in both measurements, we get the same interval $[-1, 1]$ containing 0. In this case, the intersection is exactly the same interval, so the resulting accuracy is $\Delta = 1$, the same accuracy with which we started.

- It is also possible that in the first measurement, we get the interval $[-1, 0]$ and in the second measurement, we get the interval $[0, 1]$. In this case, as a result of data fusion, we get the exact value of the measured quantity, with $\Delta = 0$.

We can also have all possible values in between. In general, if we have n measurements with accuracies $\Delta_1, \ldots, \Delta_n$, then the half-width Δ of the intersection of the corresponding intervals can take any values from 0 to $\min(\Delta_1, \ldots, \Delta_n)$.

Planning data fusion under interval uncertainty: formulation of the problem. If the accuracy of the result of data fusion is not sufficient, we should then supplement the existing measurements with one or several more accurate ones. How accurate should these new measurement be? how many of these more accurate measurements should we make? It is desirable to have some answers to these questions before we go into the time- and resources-consuming process of actually buying the corresponding sensors and performing the measurements – because if we do not get the desired accuracy again, this time-consuming process will be mostly wasting time.

In other words, it is desirable to produce an estimate for the accuracy Δ of the result of fusing measurements with accuracies $\Delta_1, \ldots, \Delta_n$, an estimate that we can obtain before we start the actual measurements. How can we solve this problem?

Planning data fusion under interval uncertainty: main idea. Our main idea of solving the above problem is as follows. We know that the i-th measurement has accuracy Δ_i. This means that the only information that we have about the possible values of the i-th measurement error Δy_i is that this error belongs to the interval $[-\Delta_i, \Delta_i]$.

We have no information about the probabilities of different values of Δy_i within this interval. According to Laplace's principle of indifference, in this situation, it is reasonable to assume that all possible values have the same probability, i.e., that the distribution of Δy_i on the interval $[-\Delta_i, \Delta_i]$ is uniform.

For each combinations of choices of Δy_i, we get different measurement results $\widetilde{y}_i = y + \Delta y_i$ and thus, different intersections $\mathbf{y} = [\underline{y}, \overline{y}] = \mathbf{y}_1 \cap \ldots \cap \mathbf{y}_n$ of the corresponding intervals $\mathbf{y}_i = [\widetilde{y}_i - \Delta_i, \widetilde{y}_i + \Delta_i] = [y + \Delta y_i - \Delta_i, y + \Delta y_i + \Delta_i]$.

We are interested in the accuracy of the fused results. This accuracy can be gauged by the largest possible absolute value Δ of the different between the actual value y and values from the fused interval $\mathbf{y}$.

As we have mentioned, in principle, this accuracy Δ can be as small as large as $\min(\Delta_1, \ldots, \Delta_n)$. However, the probability of such a large inaccuracy Δ is reasonably small; in our estimates of Δ, we would like to ignore small-probability events. In other words, we would like to select an allowable small probability p_0 of mis-estimation, and estimate Δ as the smallest value for which the probability to have $\overline{y} \le y + \Delta$ is at least $1 - p_0$ and the probability to have $\underline{y} \ge y - \Delta$ is also $\ge 1 - p_0$.

Thus, we arrive at the following precise problem.

Planning data fusion under interval uncertainty: precise formulation of the problem. Let $p_0 > 0$ be a fixed real number. We start with an arbitrary value y. Let $\Delta y_1, \ldots, \Delta y_n$ be n independent random variables such that each variable Δy_i is uniformly distributed on the interval $[-\Delta_i, \Delta_i]$. By Δ, we mean that smallest value for which the probability that for the intersection $\mathbf{y} = [\underline{y}, \overline{y}] = \mathbf{y}_1 \cap \ldots \cap \mathbf{y}_n$ of the intervals $\mathbf{y}_i = [\widetilde{y}_i - \Delta_i, \widetilde{y}_i + \Delta_i]$, where $\widetilde{y}_i = y + \Delta y_i$, the following two properties hold:

- the probability to have $\overline{y} \le y + \Delta$ is at least $1 - p_0$, and
- the probability to have $\underline{y} \ge y - \Delta$ is also $\ge 1 - p_0$.

Towards an estimate for Δ. The condition that $\overline{y} \le y + \Delta$ means that

$$\min(y + \Delta y_1 + \Delta_1, \ldots, y + \Delta y_n + \Delta_n) \le y + \Delta.$$

Which number is smaller and which is larger does not change when we shift all these numbers by the same shift y. Thus, $\min(y + \Delta y_1 + \Delta_1, \ldots, y + \Delta y_n + \Delta_n) = y + \min(\Delta y_1 + \Delta_1, \ldots, \Delta y_n + \Delta_n)$ and hence, the above inequality takes the form

$$\min(\Delta y_1 + \Delta_1, \ldots, \Delta y_n + \Delta_n) \le \Delta.$$

The probability p_{opp} for the opposite inequality

$$\min(\Delta y_1 + \Delta_1, \ldots, \Delta y_n + \Delta_n) > \Delta$$

should be $\le p_0$.

The minimum of several sums is $> \Delta$ if and only if each of these sums is $> \Delta$. Thus, the above opposite inequality holds if all n inequalities $\Delta y_i + \Delta_i > \Delta$ hold. Since the variables Δy_i are independent, we thus conclude that $p_{\text{opp}} = p_1 \cdot \ldots \cdot p_n$, where $p_i \stackrel{\text{def}}{=} \text{Prob}(\Delta y_i + \Delta_i > \Delta) = \text{Prob}(\Delta y_i > \Delta - \Delta_i)$. Since Δy_i is uniformly distributed on the interval $[-\Delta_i, \Delta_i]$, the probability p_i is equal to the ratio of

- the size of the set $(\Delta - \Delta_i, \Delta_i]$ where the corresponding inequality holds to

– the size of the overall set $[-\Delta_i, \Delta_i]$ on which the distribution is defined,

i.e., to $p_i = \dfrac{\Delta_i - (\Delta - \Delta_i)}{2\Delta_i} = \dfrac{2\Delta_i - \Delta}{2\Delta_i} = 1 - \dfrac{\Delta}{2\Delta_i}$. Thus,

$$p_{\text{opp}} = \prod_{i=1}^{n} \left(1 - \frac{\Delta}{2\Delta_i}\right).$$

This product decreases with Δ; thus, the smallest possible value Δ for which $p_{\text{opp}} \le p_0$ can be determined from the condition $p_{\text{opp}} = p_0$, i.e., $\prod\limits_{i=1}^{n} \left(1 - \dfrac{\Delta}{2\Delta_i}\right) = p_0$.

Taking logarithms of both sides, we get

$$\sum_{i=1}^{n} \ln\left(1 - \frac{\Delta}{2\Delta_i}\right) = \ln(p_0).$$

We are interested in the case when data fusion is efficient, i.e., when $\Delta \ll \Delta_i$. In this case, $\dfrac{\Delta}{2\Delta_i} \ll 1$, and we can use an approximate linearized formula $\ln(1-x) \approx -x$ which is true for small x. This formula leads to $\sum\limits_{i=1}^{n} \dfrac{\Delta}{\Delta_i} = 2|\ln(p_0)|$, i.e., to $\Delta \cdot \left(\sum\limits_{i=1}^{n} \dfrac{1}{\Delta_i}\right) = 2|\ln(p_0)|$ and

$$\Delta = \frac{\text{const}}{\sum\limits_{i=1}^{n} \dfrac{1}{\Delta_i}},$$

or, equivalently, $\dfrac{1}{\Delta} = \text{const} \cdot \sum\limits_{i=1}^{n} \dfrac{1}{\Delta_i}$.

The second inequality leads to the exact same formula for Δ.

Data fusion under interval uncertainty: result. When we fuse n measurement results with accuracies Δ_i, the accuracy Δ of the fused estimate can be estimated based on the formula

$$\frac{1}{\Delta} = \text{const} \cdot \sum_{i=1}^{n} \frac{1}{\Delta_i},$$

in which the constant $\text{const} = 2|\ln(p_0)|$ depends on the allowed probability p_0 that the actual inaccuracy of the fused value is higher than this estimate.

Case of unknown interval uncertainty: revisited. Let us recall that in the case of data fusion under unknown interval uncertainty, a reasonable choice for the accuracy $\Delta_1 = \ldots = \Delta_n$ of the fused measurements is the smallest value Δ_1 for which the corresponding intervals $[\widetilde{y}_i - \Delta_1, \widetilde{y}_i + \Delta_1]$ intersect. The problem with this approach is that for this smallest value, the intersection consists of a single point y_{mid} – making it sound as if we knew the exact value of the estimated quantity y.

To avoid this erroneous impression, a reasonable idea is to estimate the accuracy Δ of the fused result – for $\Delta_1 = \ldots = \Delta_n$ we get $\Delta = \Delta_1/n$ – and "add" this accuracy Δ to this point, i.e., return the interval $[y_{\text{mid}} - \Delta, y_{\text{mid}} + \Delta]$ as the interval estimate for the desired quantity y.

Comparison between data fusion under probabilistic and interval uncertainty. The above formula for Δ is similar to the formula $\frac{1}{\sigma^2} = \text{const} \cdot \sum_{i=1}^{n} \frac{1}{\sigma_i^2}$ which describes the accuracy σ of data fusion under probabilistic uncertainty. The main difference is that instead of the variances σ_i^2 and σ^2 we now have upper bounds Δ_i and Δ.

In practical terms, this formal difference can be described as follows.

- If we apply data fusion to n results known with the same probabilistic uncertainty $\sigma_1 = \ldots = \sigma_n$, then we result of data fusion is known with the uncertainty $\sigma = \frac{\sigma_i}{\sqrt{n}}$.
- On the other hand, if we we apply data fusion to n results known with the same interval uncertainty $\Delta_1 = \ldots = \Delta_n$, then we result of data fusion is known with the uncertainty $\Delta = \frac{\sigma_i}{n}$.

Thus, with interval uncertainty, we get a much faster ($\sim 1/n$) decrease in approximation error than for the probabilistic uncertainty ($\sim 1/\sqrt{n}$). This fact is in line with similar estimates from (Walster, 1988; Walster and Kreinovich, 1996).

Comment. It is worth mentioning that there is a similar difference for data processing: in the interval case, we have $\Delta = \sum_{i=1}^{n} |c_i| \cdot \Delta_i$, whereas in the probabilistic case, we have $\sigma^2 = \sum_{i=1}^{n} |c_i|^2 \cdot \sigma_i^2$.

The difference between the formulas for data fusion and data processing is similar to the formulas for the resistance R of of an electric circuit consisting of resistances $R_1, \ldots, R_n$: when the resistances are placed sequentially, we get $R = R_1 + \ldots + R_n$ (as for data processing); when the resistance are placed in parallel to each other, we get $\frac{1}{R} = \frac{1}{R_1} + \ldots + \frac{1}{R_n}$ (as for data fusion).

8. Propagation of Uncertainty When We Have Both Data Processing and Data Fusion

Motivations. As we have mentioned earlier, in many real-life situations, to get the desired accuracy, we must apply both data fusion and data processing.

Example. For example, we can fuse several values $y_1, \ldots, y_n$ each of which is obtained by data processing.

Main idea. In this case, to find the accuracy of the final result, we propagate the uncertainty through all these data fusion/data processing steps.

Example. In the above example,

- we first use the formulas for propagating uncertainty under data processing to come up with accuracy values (Δ_i or σ_i) for y_i, and then
- we use the formulas for uncertainty propagation under data fusion to combine these values Δ_i (correspondingly, σ_i) into a single estimate Δ (corr., σ).

9. Towards Optimal Data Processing and Data Fusion

Motivations. Up to now, we concentrated on the *analysis* of given data processing and data fusion scenarios. However, since we have explicit (and simple) formulas for the propagation of uncertainty under data processing and data fusion, we can actually solve the problem of finding the least expensive way to guarantee the given accuracy.

To perform this *optimization*, we must know how the cost of measuring related quantities with different accuracies.

Towards optimal data fusion: preliminary description. For data fusion, let $c^{\text{prob}}(\sigma)$ denote the cost of measuring the desired quantity with standard deviation σ, and let $c^{\text{int}}(\Delta)$ denote the cost of measuring the desired quantity with the guaranteed upper bound Δ on the measurement error. Typically, $c^{\text{prob}}(\sigma) = \dfrac{C}{\sigma^\alpha}$ and $c^{\text{int}}(\Delta) = \dfrac{C}{\Delta^\alpha}$ for some constants C and $\alpha > 0$; see, e.g., (Nguyen et al., 2008; Nguyen and Kreinovich, 2008) and references therein.

Towards optimal data fusion: probabilistic case. In the probabilistic case, we must find the values σ_i for which $\sum\limits_{i=1}^{n} c^{\text{prob}}(\sigma_i) \to \min$ under the constraint that the sum $\sum\limits_{i=1}^{n} \sigma_i^{-2}$ is equal to the given value σ^{-2}. By applying Lagrange multiplier method to this constraint optimization problem, we get an unconstraint optimization problem $\sum\limits_{i=1}^{n} c^{\text{prob}}(\sigma_i) + \lambda \cdot \sum\limits_{i=1}^{n} \sigma_i^{-2} \to \min$. Differentiating w.r.t. σ_i and equating the derivative to 0, we conclude that $c'(\sigma_i) \cdot \sigma_i^3 = \text{const} = 2\lambda$.

For a function $c^{\text{prob}}(\sigma) = \dfrac{C}{\sigma^\alpha}$, the expression $c'(\sigma_i) \cdot \sigma_i^3$ is monotonic in σ_i and thus, the equality occurs only for one value σ_i – hence in the optimal plan, $\sigma_1 = \ldots = \sigma_n$. To get the desired value σ, we must have $\sigma_i = \sqrt{n} \cdot \sigma$.

Towards optimal data fusion: interval case. Similarly, in the interval case, the problem of minimizing $\sum\limits_{i=1}^{n} c^{\text{int}}(\Delta_i)$ under the constraint $\sum\limits_{i=1}^{n} \Delta_i^{-1} = \Delta^{-1}$ leads to the equation $c'(\Delta_i) \cdot \Delta_i^2 = \text{const} = \lambda$.

For a function $c^{\text{int}}(\Delta) = \dfrac{C}{\Delta^\alpha}$, the expression $c'(\Delta_i) \cdot \Delta_i^2$ is monotonic in Δ_i and thus, the equality occurs only for one value Δ_i – hence in the optimal plan, $\Delta_1 = \ldots = \Delta_n$. To get the desired value Δ, we must have $\Delta_i = n \cdot \Delta$.

Towards optimal data processing: preliminary description. For data processing, let $c_i^{\text{prob}}(\sigma_i)$ denote the cost of measuring the i-th quantity with standard deviation σ_i, and let $c_i^{\text{int}}(\Delta_i)$ denote the cost of measuring the i-th quantity with the guaranteed upper bound Δ_i on the measurement error. Just like in the case data fusion, typically, we have $c_i^{\text{prob}}(\sigma_i) = \dfrac{C_i}{\sigma_i^{\alpha_i}}$ and $c_i^{\text{int}}(\Delta_i) = \dfrac{C_i}{\Delta^{\alpha_i}}$ for some constants C_i and α_i.

Towards optimal data processing: probabilistic case. In the probabilistic case, the problem of minimizing $\sum\limits_{i=1}^{n} c_i^{\text{prob}}(\sigma_i)$ under the constraint $\sum\limits_{i=1}^{n} c_i^2 \cdot \sigma_i^2 = \sigma^2$ leads to the equation $\dfrac{c'(\sigma_i)}{c_i^2 \cdot \sigma_i} = \text{const} = -2\lambda$.

For $c_i^{\rm prob}(\sigma_i) = \dfrac{C_i}{\sigma_i^{\alpha_i}}$, we get $\sigma_i = \left(\dfrac{\alpha_i \cdot C_i}{2\lambda \cdot c_i^2}\right)^{1/(2+\alpha_i)}$, where λ can be determined from the equation

$$\sum_{i=1}^{n} c_i^2 \cdot \left(\frac{\alpha_i \cdot C_i}{2\lambda \cdot c_i^2}\right)^{2/(2+\alpha_i)} = \sigma^2.$$

Towards optimal data processing: interval case. In the interval case, the problem of minimizing $\sum\limits_{i=1}^{n} c_i^{\rm int}(\Delta_i)$ under the constraint $\sum\limits_{i=1}^{n} |c_i| \cdot \Delta_i = \Delta$ leads to the equation $\dfrac{c_i'(\Delta_i)}{|c_i|} = {\rm const} = -\lambda$.

For $c_i^{\rm int}(\Delta_i) = \dfrac{C_i}{\Delta_i^{\alpha_i}}$, we get $\Delta_i = \left(\dfrac{\alpha_i \cdot C_i}{\lambda \cdot |c_i|}\right)^{1/(1+\alpha_i)}$, where λ can be determined from the equation

$$\sum_{i=1}^{n} |c_i| \cdot \left(\frac{\alpha_i \cdot C_i}{\lambda \cdot |c_i|}\right)^{2/(2+\alpha_i)} = \Delta.$$

10. Combining Probabilistic and Interval Uncertainty

Motivations. In the previous sections, we assumed that in data processing an in data fusion, either all measurement results are known with probabilistic uncertainty, or all measurement results are known with interval uncertainty.

In practice, some measurement results are known with probabilistic uncertainty (i.e., we know the probabilities of the corresponding measurement errors), and some are only known with interval uncertainty (i.e., we only know the upper bounds on the corresponding measurement errors). In this case, how can we estimate the accuracy of the result of data processing or data fusion?

Case of data processing. For data processing, it is possible to provide an answer to the above question. Indeed, suppose that we produce an estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ for the desired quantity y which is based on the results $\widetilde{x}_i$ of directly measuring n related quantities $x_1, \ldots, x_n$.

We are interested in situations in which some of the measurement errors are known with probabilistic uncertainty, and some with interval uncertainty. Without losing generality, we can assume that the values $x_1, \ldots, x_k$ are known with probabilistic uncertainty and the values $x_{k+1}, \ldots, x_n$ are known with interval uncertainty. In other words, we know the standard deviations $\sigma_1, \ldots, \sigma_k$ of the first k measurements, and we know the upper bounds $\Delta_{k+1}, \ldots, \Delta_n$ of the others. In this case, the above linearized formula $\Delta y = \sum\limits_{i=1}^{n} c_i \cdot \Delta x_i$ can be rewritten as $\Delta y = \Delta y^{\rm prob} + \Delta y^{\rm int}$, where $\Delta y^{\rm prob} = \sum\limits_{i=1}^{k} c_i \cdot \Delta x_i$ and $\Delta y^{\rm int} = \sum\limits_{i=k+1}^{n} c_i \cdot \Delta x_i$.

As we already know, Δy^{prob} is a normally distributed random variable with 0 mean and standard deviation $\sigma = \sqrt{\sum_{i=1}^{k} c_i^2 \cdot \sigma_i^2}$ and Δy^{int} is a variable about which we only know that it belongs to the interval $[-\Delta, \Delta]$, where $\Delta = \sum_{i=k+1}^{n} |c_i| \cdot \Delta_i$.

So, we conclude that the approximation error Δy is the sum of two error components: a random one with a known σ and an interval one with a known Δ.

The resulting two-component description of measurement and approximation error is in line with the measurement practice. The above two-component description of an approximation error is in line with the standard practice in measurement theory (see, e.g., (Rabinovich, 2005)), where a measurement error Δx is often described by its two component:

- a *random* error component $\Delta_r x \stackrel{\text{def}}{=} \Delta x - E[\Delta x]$ with 0 mean ($E[\Delta_r x] = 0$), for which we usually know the standard deviation σ, and
- a *systematic* error component $\Delta_s x \stackrel{\text{def}}{=} E[\Delta x]$ for which we only know the upper bound Δ on its absolute value.

The situation in which we only know the upper bound Δ on the (absolute value of) the total measurement error can be viewed as a degenerate case of this two-component description, with $\sigma = 0$.

Data processing: case when we have a two-component description of all the measurement errors. In view of the prevalence of the two-component error description in measurement practice, it is reasonable to consider the following situation.

We want to estimate the value of the desired difficult-to-measure quantity y. We know the relation $y = f(x_1, \ldots, x_n)$ between this quantity y and easier-to-measure quantities $x_1, \ldots, x_n$. For each of these auxiliary quantities x_i, we know the measurement result $\widetilde{x}_i$ and we know that the corresponding measurement error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ can be represented as a sum of two components $\Delta x_i = \Delta_s x_i + \Delta_r x_i$, where:

- the component $\Delta_s x_i$ is a random variable with 0 mean and known standard deviation σ_i;
- about the component $\Delta_r x_i$, we only know the upper bound Δ_i on the (absolute value of the) measurement error, i.e., we know that $|\Delta_r x_i| \le \Delta_i$.

Based on the measurement results $\widetilde{x}_1, \ldots, \widetilde{x}_n$, we compute an estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ for y. What can we say about the approximation error $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$ of this estimate?

In the linearization case, we can conclude that $\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i$. Since $\Delta x_i = \Delta_s x_i + \Delta_r x_i$, we can conclude that $\Delta y = \Delta_s y + \Delta_r y$, where $\Delta_r y = \sum_{i=1}^{n} c_i \cdot \Delta_r x_i$ and $\Delta_s y = \sum_{i=1}^{n} c_i \cdot \Delta_s x_i$. We already know how to handle each of these two sums, so we conclude that the approximation error Δy also consists of two components:

– the component $\Delta_s y$ is a random variable with 0 mean and known standard deviation

$$\sigma = \sqrt{\sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2};$$

– about the component $\Delta_r y$, we only know the upper bound $\Delta = \sum\limits_{i=1}^{n} |c_i| \cdot \Delta_i$ on the (absolute value of the) measurement error, i.e., we know that $|\Delta_r y| \le \Delta$.

Case of data fusion. In data fusion, we have n results $\widetilde{y}_1, \ldots, \widetilde{y}_n$ of measuring the same quantity y. In the previous sections, we assume that either all of these measurement errors are known with probabilistic uncertainty, or that all of them are know with interval uncertainty.

Let us now consider the case when some of the measurement errors are known with probabilistic uncertainty, and some with interval uncertainty. Without losing generality, we can assume that the values $y_1, \ldots, y_k$ are known with probabilistic uncertainty and the values $y_{k+1}, \ldots, y_n$ are known with interval uncertainty. In other words, we know the standard deviations $\sigma_1, \ldots, \sigma_k$ of the first k measurements, and we know the upper bounds $\Delta_{k+1}, \ldots, \Delta_n$ of the others. In this case,

– we can use the data fusion formula for the probabilistic uncertainty to fuse the measurement $\widetilde{y}_1, \ldots, \widetilde{y}_k$ into a single result $\widetilde{y}$ with a standard deviation σ, and

– we can fuse the interval-valued measurements by taking the intersection

$$[\underline{y}, \overline{y}] \stackrel{\text{def}}{=} [\widetilde{y}_{k+1} - \Delta_{k+1}, \widetilde{y}_{k+1} + \Delta_{k+1}] \cap \ldots \cap [\widetilde{y}_n - \Delta_n, \widetilde{y}_n + \Delta_n]$$

of the corresponding intervals.

It is therefore important to fuse the interval estimate with accuracy Δ and the probabilistic estimate with the accuracy σ The result of the fusion depends on the relation between Δ and σ:

– If $\Delta \gg \sigma$, this means that the interval estimate is much much wider than what we can simply conclude based on the probabilistic information. Thus, in this case, the fused information consists simply of the probabilistic estimate.

– If $\sigma \gg \Delta$, this means that the probabilistic estimate is much worse than the interval one. In this case, the fused information consists simply of the interval estimate.

– If the estimates σ and Δ are approximately of the same order, this means that we can keep either one of them.

Comment. Our recommendation for the case when the estimates σ and Δ are approximately of the same order is somewhat vague. For this case, it would be nice to come up with a better answer to the fusion question.

11. Adding Reliability and Trust: Results and Open Problems

Formulation of the problem. In the previous sections, we concentrated on the measurement uncertainty, i.e., on the difference between the measurement result and the actual value of the corresponding quantity. We also assumed that these differences are relatively small.

This smallness assumption holds in many practical situations. However, sometimes, we have values which are completely off: a measuring instrument can malfunction, a computer may have misread this information, etc. Such values are often called *outliers*. In short, in addition to being not 100% accurate, the measurement results are also not 100% *reliable*. How can we take this possible unreliability into account in data processing and data fusion?

How we can describe the reliability of different measurement results. A natural way to describe the reliability of different measurement results is to provide the probability p_i that the i-th measurement result is an outlier. It is usually assumed that in terms of reliability, different measurement results are independent – so that, e.g., the probability that both the i-th and the j-th results are outliers is equal to the product $p_i \cdot p_j$.

These probabilities can be gauged, e.g., based on our knowledge of what team performed these measurements, what is the track records of this particular team, what auxiliary values have been used in pre-processing these results, etc. In other words, these probabilities can be gauged based on the provenance of the corresponding measurement results.

Based on these probabilities, we need to estimate the reliability p of the results of data processing and data fusion.

Case of data processing. Let us first consider the case of data processing, when we transform n measurement results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of n auxiliary quantities into an estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of the desired quantity y.

For this estimate to be valid, all n measurement results must be valid (i.e., none of them is an outlier). The probability that the i-th measurement result is not an outlier is equal to $1 - p_i$. Since we assumed independence, the probability $1 - p$ that all n measurement results are not outliers is equal to the product

$$1 - p = \prod_{i=1}^{n} (1 - p_i).$$

Hence,

$$p = 1 - \prod_{i=1}^{n} (1 - p_i).$$

If all the probability p_i are small, we can ignore quadratic and higher order terms in this formula and conclude that $1 - p \approx 1 - \sum\limits_{i=1}^{n} p_i$, i.e., that

$$p \approx \sum_{i=1}^{n} p_i.$$

Comment. Since $p_i > 0$, we have $1 - p < 1 - p_i$ for all i and hence, $p > p_i$. It is worth mentioning that for data processing, the un-reliability p of the result of data processing is larger than each individual probability p_i. Thus, if one of the input measurement results is highly unreliable, the result of data processing is highly unreliable as well.

In particular, if we process n values with the same un-reliability $p_1 = \ldots = p_n$, then the un-reliability p of the result of data processing is n times larger: $p \approx n \cdot p_1 \gg p_1$.

Case of data fusion. The above-described data fusion techniques assume that we use all n results $\widetilde{y}_1, \ldots \widetilde{y}_n$ of measuring the desired quantity y. Thus, for these techniques, the reliability p of the resulting estimate $\widetilde{y}$ can be estimated by using a similar formula

$$p = 1 - \prod_{i=1}^{n} (1 - p_i).$$

For small p_i, we can use a linearized version of this formula $p \approx \sum\limits_{i=1}^{n} p_i$.

So here, just like for data processing, the un-reliability p of the result of data fusion is (much) higher than the un-reliability of individual measurement results.

Case of interval uncertainty. Let us show that in the case of linear uncertainty, we can get much better reliability values than in the general data fusion situation.

Indeed, in the case of interval uncertainty, we start with n intervals $[\underline{y}_i, \overline{y}_i]$ which contain the desired value y. In the "reliable" data fusion, we simply take the intersection of these n intervals, i.e., we take the interval $[\underline{y}, \overline{y}]$, where $\underline{y} = \max(\underline{y}_1, \ldots, \underline{y}_n)$ and $\overline{y} = \min(\overline{y}_1, \ldots, \overline{y}_n)$. The minimum and the maximum are attained for some specific values i and j; thus, we always have $\underline{y} = \underline{y}_i$ for some i and $\overline{y} = \overline{y}_j$ for some appropriate value j. Hence,

- the reliability for the lower endpoint $\underline{y}$ is simply equal to the reliability p_i of the i-th measurement, and
- the reliability of the upper endpoint $\overline{y}$ is equal to the reliability p_j of the j-th measurement.

The corresponding values $p = p_i$ and $p = p_j$ are much better than in the general case when $p \gg p_i$ for all i.

Data fusion can also improve reliability: towards an algorithm. Interval data fusion can lead to even more reliable results: namely, an appropriate data fusion can drastically improve the reliability of the result.

Indeed, let us assume that we have n interval measurements $[\underline{y}_1, \overline{y}_1], \ldots, [\underline{y}_n, \overline{y}_n]$ with reliabilities $p_1, \ldots, p_n$. If these reliability are too high, how can we combine these values to get an estimate for y for which the corresponding probability p does not exceed a given threshold p_0?

Let us illustrate this possibility on the example of the upper endpoint $\overline{y}$ of the desired reliable bound for y. For that, let us sort the values $\overline{y}_i$ into an increasing sequence

$$\overline{y}_{(1)} \le \overline{y}_{(2)} \le \ldots \le \overline{y}_{(n)}.$$

The corresponding probabilities will now be $p_{(1)}, p_{(2)}, \ldots, p_{(n)}$. In the "reliable" data fusion, we simply take the smallest value $\overline{y}_{(1)}$ as the desired estimate $\overline{y}$. For fusing possibly un-reliable data, we can no longer do that.

Instead, let us choose, as $\overline{y}$, the k-th value $\overline{y}_{(k)}$ for some k. This estimate is not valid only in one case: when all k estimates $\overline{y}_{(1)}, \ldots, \overline{y}_{(k)}$ are un-reliable. Since we assumed independence, the probability for this is equal to the product $p_{(1)} \cdot \ldots \cdot p_{(k)}$. Thus, ro guarantee reliability $p \le p_0$, we can select the first k for which $p_{(1)} \cdot \ldots \cdot p_{(k)} \le p_0$. Thus, we arrive at the following algorithm.

Data fusion which improves reliability of interval estimates: an algorithm. We start with n intervals $[\underline{y}_i, \overline{y}_i]$ with are reliable with probabilities p_i. Our objective is to fuse them into a single interval $[\underline{y}, \overline{y}]$ which is the most accurate under the constraint that each of its endpoints $\underline{y}$ and $\overline{y}$ has an un-reliability $\le p_0$ for some given value p_0.

To get the desired value $\overline{y}$, we sort the upper endpoints $\overline{y}_i$ into an increasing sequence $\overline{y}_{(1)} \le \overline{y}_{(2)} \le \ldots \le \overline{y}_{(n)}$, select the smallest k for which $p_{(1)} \cdot \ldots \cdot p_{(k)} \le p_0$, and take $\overline{y} = \overline{y}_{(k)}$.

Similarly, to get the desired value $\underline{y}$, we sort the lower endpoints $\underline{y}_i$ into a decreasing sequence $\underline{y}_{(1)} \ge \underline{y}_{(2)} \ge \ldots \ge \underline{y}_{(n)}$, select the smallest k for which $p_{(1)} \cdot \ldots \cdot p_{(k)} \le p_0$, and take $\underline{y} = \underline{y}_{(k)}$.

Comment. When all input intervals have the same reliability $p_1 = \ldots = p_n$, the condition

$$p_{(1)} \cdot \ldots \cdot p_{(k)} \le p_0$$

takes the form $p_1^k \le p_0$. The smallest k for which this inequality holds can be easily computed as $k = \left\lceil \dfrac{|\ln(p_0)|}{|\ln(p_1)|} \right\rceil$.

It is worth mentioning that in this case, we do not need to spend $O(n \cdot \log(n))$ times on sorting the bounds, since the k-th value in the ordered sequence can be computed in linear time; see, e.g., (Cormen et al., 2001).

12. Case Study: Seismic Inverse Problem in the Geosciences

12.1. Description of the Case Study

Seismic inverse problem in the geosciences: brief reminder. As a case study, we consider the seismic inverse problem in the geosciences; see, e.g., (Averill, 2007; Hole, 1992; Parker, 1994). In this problem, we measure the travel times $x_1, \ldots, x_n$ of the seismic signals and based on these travel times, we reconstruct the velocity of sound $y = f(x_1, \ldots, x_n)$ at different points inside the Earth. There exist several algorithms for reconstructing this velocity. In our research, we use one of most widely used algorithms proposed by J. Hole (Hole, 1992).

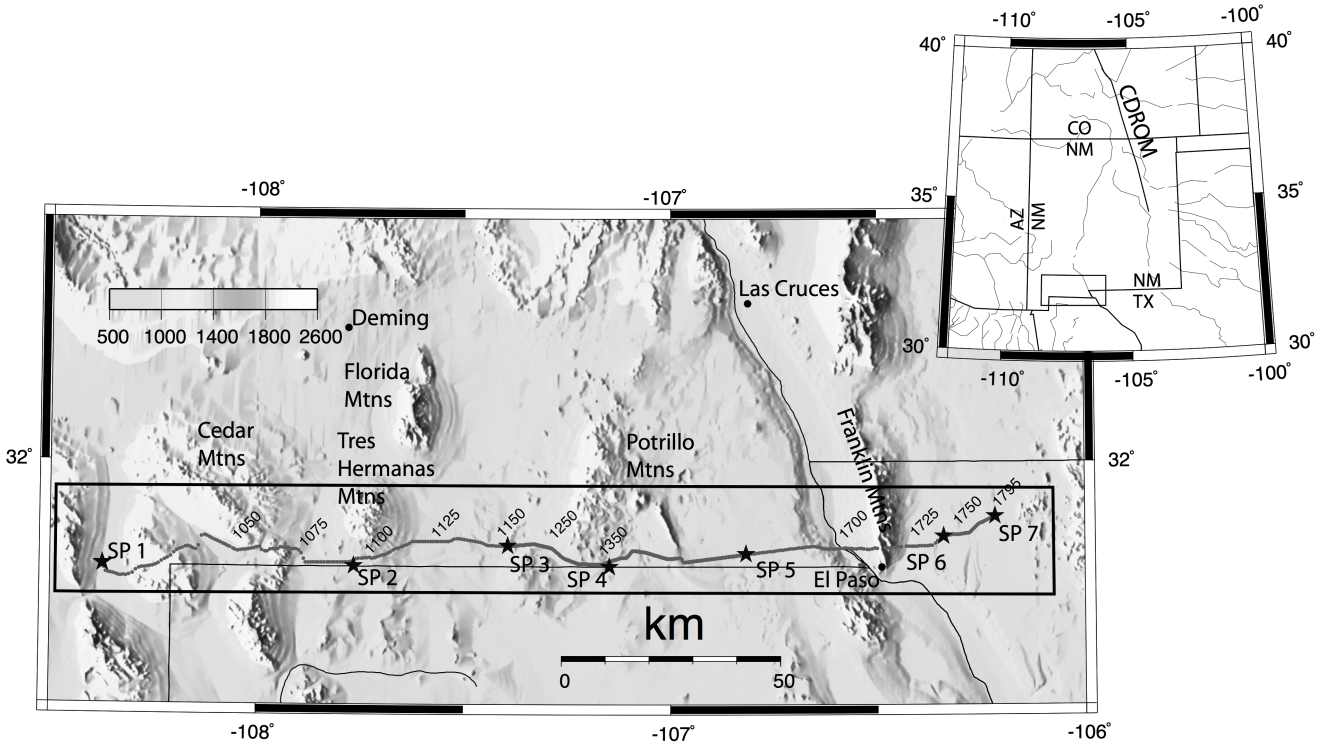
Our objective is to estimate the uncertainty of the resulting velocity estimates.

What we plan to do. The problem of estimating uncertainty has been actively researched in geosciences; see, e.g., (Averill et al., 2005; Averill et al., 2007; Doser et al., 1998; Maceira et al., 2005). In this section, we will apply the above-described techniques to this problem, explain the results and their limitations, and provide a heuristic method of overcoming these limitations, a method which can be applied to other problems as well.

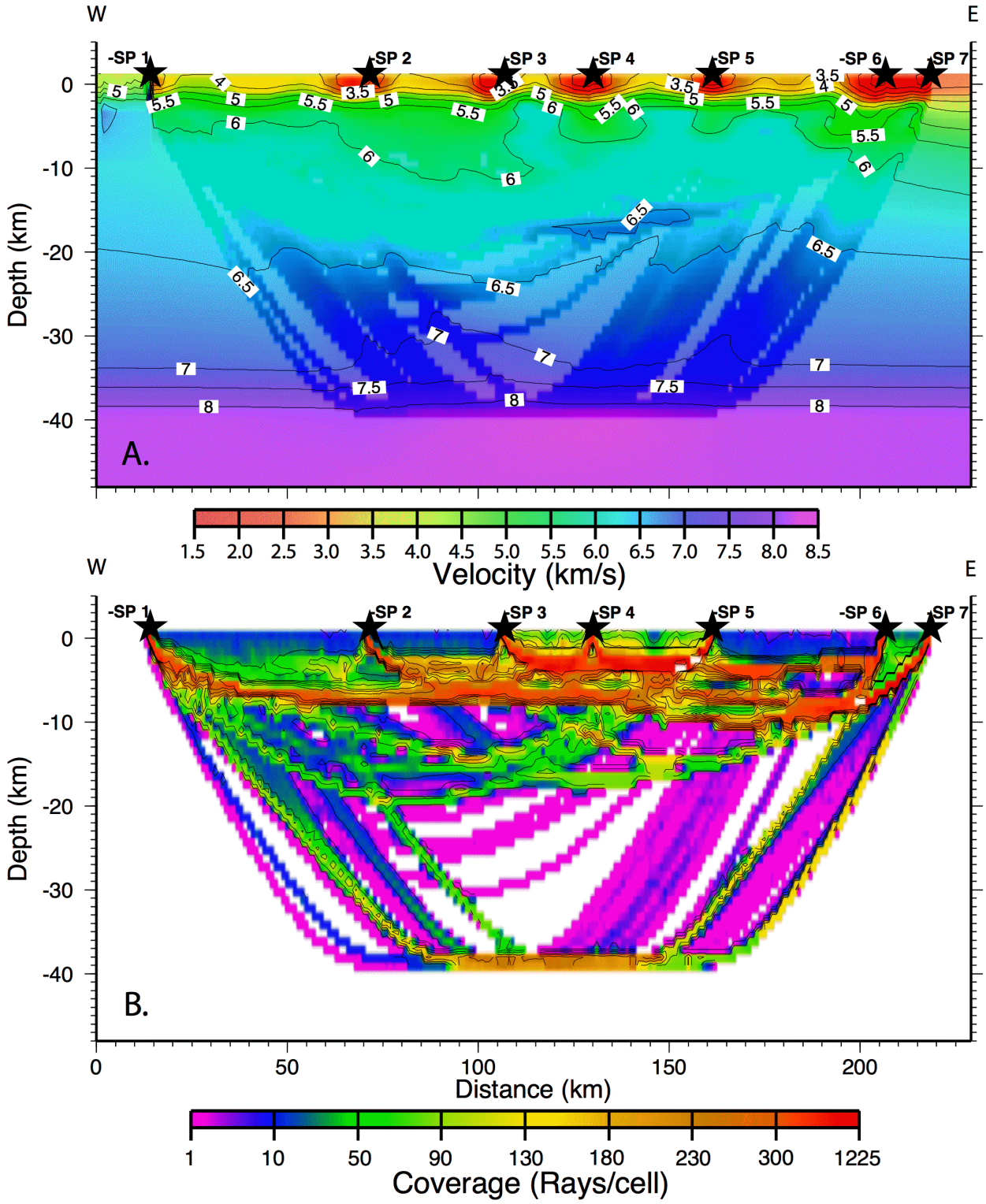
Details of this application are presented in (Averill, 2007).

Case study: brief description. The data used for our analysis was obtained from the Potrillo Volcanic Experiment (PVF), a large-scale active source seismology profile designed to investigate the crustal structure across southern New Mexico and Far West Texas. This field experiment was conducted in 2003.

The PVF experiment was composed of 8 shots of 1000–2000 lbs.; 793 seismic recorders (TEXANS) were deployed at variable spacing of 100 m, 200 m and 600 m over 205 km. The location map for the (PVF) experiment is give below. Stars show shot point locations. Small gray dots represent receiver locations. Black box outlines model space for tomography.



The resulting velocity distribution is given presented in the following picture. Velocity model is gridded at 1×1 km spacing. Illuminated coloring shows location of ray coverage within the model. Coverage model showing location and coverage density for rays traced within the model is presented in the next picture.


W
E
-SP 1
SP 2
SP 3
SP 4
SP 5
-SP 6
SP 7
Depth (km)
A.
1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5
Velocity (km/s)
B.
0 50 100 150 200
Distance (km)
1 10 50 90 130 180 230 300 1225
Coverage (Rays/cell)

What is the accuracy with which we know these values of velocity?

Main source of direct measurement errors in the seismic inverse problem. The input to the seismic inverse problem consists of travel times x_i. Each travel time is the time that a seismic wave takes to travel from the location of the explosion to the corresponding sensor. It is determined as the first moment of time at which we detect the incoming seismic wave (on top of the noise). For the low-energy artificial explosions which are used in seismic experiments, the signal-to-noise ratio is rather small, especially for sensors located several dozens kilometers away from the experiment location. As a result, we may miss the first peak of the seismic wave and erroneously identify the second peak as the arrival time of the seismic wave.

This "picking error" is the main source of error in measuring travel time. A typical size of a picking error is the time distance between the two peaks of the seismic wave, i.e., about 150 ms.

12.2. First Try: Probabilistic Approach

First try: probabilistic approach. As we have mentioned, traditionally in science and engineering, the probabilistic approach is used to estimate the uncertainty of the result of data processing. In this approach, we assume that the errors of different direct measurements are independent random variables, with 0 mean and known standard deviations σ_i.

This method have been successfully used in geosciences. In particular, it was used in the analysis of the passive seismic inverse problem, when we use only the travel times of the seismic waves generated by the earthquakes; see, e.g., (Maceira et al., 2005). For this problem, the independence assumption makes sense since different earthquakes at different locations are indeed independent.

In view of these past successes, we decided to apply this technique to our active seismic inverse problems.

In principle, we can use the above formula. In principle, to find the desired value σ, we can use the above formula $\sigma = \sqrt{\sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2}$, where each partial derivative can be determined by numerical differentiation, as $c_i = \frac{1}{h} \cdot (f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i + h, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n) - \widetilde{y})$.

Limitations of directly using the above formula. The direct use of the above formula requires that we call the program f (in our case, the program for solving the seismic inverse problem) $n+1$ times, where n is the total number of inputs: once to compute $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$, and n more times to compute the values $c_1, \ldots, c_n$.

The problem with directly using the above formula is that the program f requires several hours time to compute (e.g., in the geological applications, computing f may involve solving an inverse problem), and the number n of inputs x_i is in the thousands. Thus, calling the program f $n+1$ times requires 1,000 times longer than several hours – i.e., several months.

Monte-Carlo simulations: main idea. In the probabilistic setting, we can use straightforward (Monte-Carlo) simulation, and drastically save the computation time. In this approach, we use a computer-based random number generator to simulate the normally distributed error. A standard normal random number generator usually produces a normal distribution with 0 average and standard deviation 1. So, to simulate a distribution Δx_i with a standard deviation σ_i, we multiply

the result α_i of the standard Gaussian random number generator by σ_i. In other words, we take $\Delta_i = \sigma_i \cdot \alpha_i$, and we simulate x_i as $\widetilde{x}_i - \Delta x_i$.

As a result of N Monte-Carlo simulations, we get N values

$$\Delta y^{(1)} = c_1 \cdot \Delta x_1^{(1)} + \ldots + c_n \cdot \Delta x_n^{(1)}, \ldots, \Delta y^{(N)} = c_1 \cdot \Delta x_1^{(N)} + \ldots + c_n \cdot \Delta x_n^{(N)}$$

which are normally distributed with the desired standard deviation σ. So, we can determine σ by using the standard statistical estimate

$$\sigma = \sqrt{\frac{1}{N-1} \cdot \sum_{k=1}^{N} \left(\Delta y^{(k)}\right)^2}. \tag{1}$$

Computation time required for Monte-Carlo simulation. The relative error of the above statistical estimate depends only on N (as $\approx 1/\sqrt{N}$), and not on the number of variables n. Therefore, the number N_f of calls to f that is needed to achieve a given accuracy does not depend on the number of variables at all.

The error of the above algorithm is asymptotically normally distributed, with a standard deviation $\sigma_e \sim \sigma/\sqrt{2N}$. Thus, if we use a "two sigma" bound, we conclude that with probability 95%, this algorithm leads to an estimate for σ which differs from the actual value of σ by $\leq 2\sigma_e = 2\sigma/\sqrt{2N}$.

This is an error with which we estimate the error of indirect measurement; we do not need too much accuracy in this estimation, because, e.g., in real life, we say that an error is $\pm 10\%$ or $\pm 20\%$, but *not* that the error is, say, $\pm 11.8\%$. Therefore, in estimating the error of indirect measurements, it is sufficient to estimate the characteristics of this error with a relative accuracy of, say, 20%.

For the above "two sigma" estimate, this means that we need to select the smallest N for which $2\sigma_e = 2\sigma/\sqrt{2N} \leq 0.2 \cdot \sigma$, i.e., to select $N_f = N = 50$.

In many practical situations, it is sufficient to have a standard deviation of 20% (i.e., to have a "two sigma" guarantee of 40%). In this case, we need only $N = 13$ calls to f.

On the other hand, if we want to guarantee 20% accuracy in 99.9% cases, which correspond to "three sigma", we must use N for which $3\sigma_e = 3 \cdot \sigma/\sqrt{2N} \leq 0.2 \cdot \sigma$, i.e., we must select $N_f = N = 113$, etc.

For $n \approx 10^3$, all these values of N_f are much smaller than $N_f = n$ required for numerical differentiation.

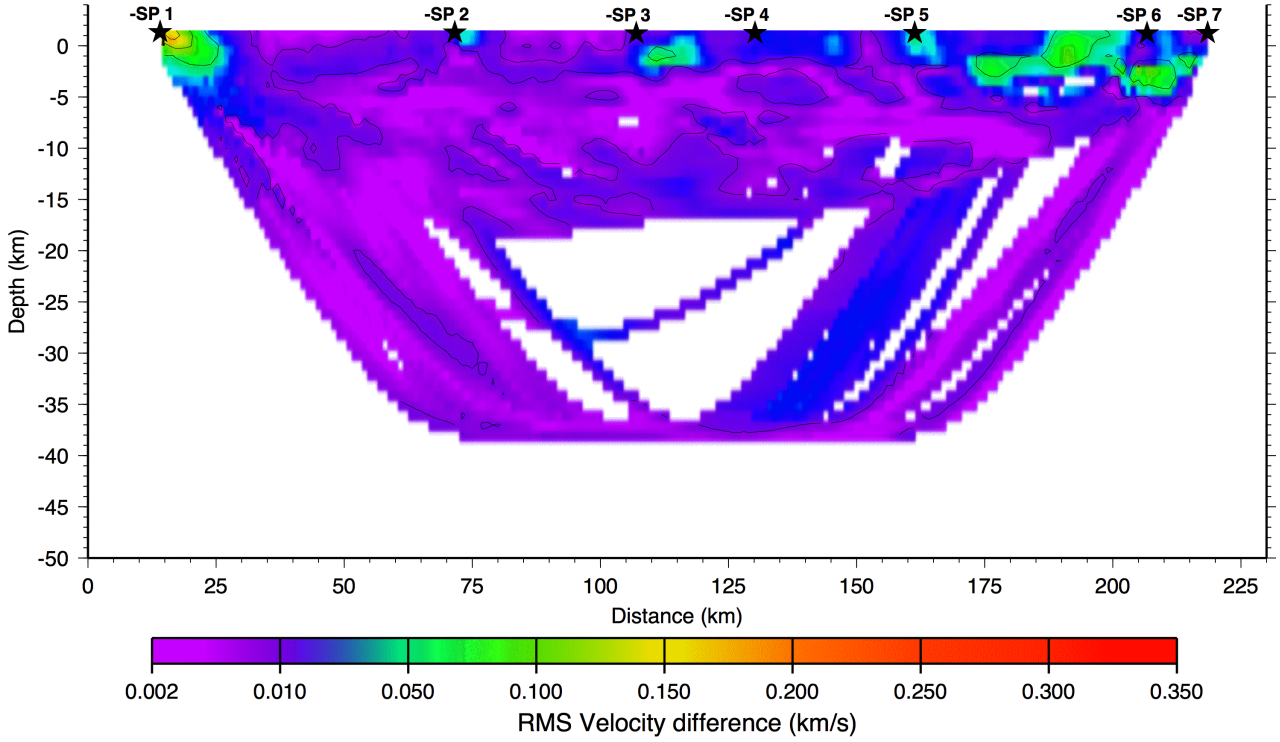
Additional advantage: parallelization. In Monte-Carlo algorithm, we need 50 calls to f. If each call requires a minute, the resulting time takes about an hour, which may be too long for on-line results. Fortunately, different calls to the function f are independent on each other, so we can run all the simulations in parallel.

The more processors we have, the less time the resulting computation will take. If we have as many processors as the required number of calls, then the time needed to estimate the error of indirect measurement becomes equal to the time of a single call, i.e., to the time necessary to compute the result $\widetilde{y}$ of this indirect measurement. Thus, if we have enough processors working in parallel, we can compute the result of the indirect measurement *and* estimate its error during the same time that it normally takes just to compute the result.

In particular, if the result $\widetilde{y}$ of indirect measurement can be computed in real time, we can estimate the error of this result in real time as well.

Probabilistic case: results. Following the above algorithm, we randomly perturbed the travel-time data by a Gaussian distribution with a standard deviation equal to the "picking error" of 150 ms. The perturbed data was used to generate a new velocity model. This process was repeated multiple times, and the resulting velocity models are used to calculate the RMS difference σ in velocity.

The majority of the values are less than 0.01 km/s (see below).



From our experience of comparing different results, we know that the actual difference between different estimates $\widetilde{y}$ for the velocity is much higher. Thus, these results are misleadingly low.

Also, the results seem to be qualitatively misleading: the values of σ are the highest near the shots (where the reconstruction is more accurate) and smaller elsewhere.

Comment. When instead of single value $\sigma_i = 150$ ms, we used more realistic different values at different sensor locations (obtained by using a technique from (Zelt and Forsyth, 1994)), we got similar results.

Comment. In addition to the Monte-Carlo approach, we also tried the jack-knife approach (see, e.g., (Lees and Crosson, 1989; Tihelaar and Ruff, 1989)) in which the data is divided into two sets (we divided into even and odd sensors). Each of these two data sets was inverted to generate

the velocity distribution. The resulting distributions were compared to the result of processing the combined data set; the differences are taken as an estimate for σ.

The resulting values σ are similar to the probabilistic case: the values σ are too low (generally below 0.06 km/s), and qualitatively wrong: the highest values are located near the ends of the profile, adjacent to the shot points and in regions of lower ray coverage.

12.3. Second Try: Interval Approach

Toward interval estimates. In our probabilistic estimates, we made a simplifying assumption that the measurement errors of different measurements are independent random variable. Since this assumption is false, this means that there is a correlation between these errors.

We do not know the value of this correlation. It is therefore reasonable to now try the more general interval case, which makes no assumption about the correlations.

Interval approach: brief reminder. In this approach, we assume that we know the upper bounds Δ_i on the measurement errors Δx_i, and we compute the upper bounds Δ on the resulting error Δy.

In our example, we take $\Delta_i = 150$ ms.

In principle, we can use the above explicit formula. In principle, to find the desired value Δ, we can use the above formula $\Delta = \sum_{i=1}^{n} |c_i| \cdot \Delta_i$, where each partial derivative can be determined by numerical differentiation. However, similarly to the probabilistic case, this method requires that we call f $4n + 1$ times – which can lead to months of computations.

To avoid these computations, we use the Cauchy-based method described in (Kreinovich et al., 2007; Kreinovich et al., 2004).

Mathematics behind the Cauchy method. In our simulations, we use *Cauchy distribution* – i.e., probability distributions with the probability density $\rho(z) = \dfrac{\Delta}{\pi \cdot (z^2 + \Delta^2)}$; the value Δ is called the *(scale) parameter* of this distribution.

Cauchy distribution has the following property that we will use: if $z_1, \ldots, z_n$ are independent random variables, and each of z_i is distributed according to the Cauchy law with parameter Δ_i, then their linear combination $z = c_1 \cdot z_1 + \ldots + c_n \cdot z_n$ is also distributed according to a Cauchy law, with a scale parameter $\Delta = |c_1| \cdot \Delta_1 + \ldots + |c_n| \cdot \Delta_n$.

Therefore, if we take random variables δ_i which are Cauchy distributed with parameters Δ_i, then the value

$$\delta \stackrel{\text{def}}{=} f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(\widetilde{x}_1 - \delta_1, \ldots, \widetilde{x}_n - \delta_n) = c_1 \cdot \delta_1 + \ldots + c_n \cdot \delta_n$$

is Cauchy distributed with the desired parameter $\Delta = \sum_{i=1}^{n} |c_i| \cdot \Delta_i$. So, repeating this experiment N times, we get N values $\delta^{(1)}, \ldots, \delta^{(N)}$ which are Cauchy distributed with the unknown parameter, and from them we can estimate Δ.

The bigger N, the better estimates we get.

There are two questions to be solved:

– how to simulate the Cauchy distribution;

– how to estimate the parameter Δ of this distribution from a finite sample.

Simulation can be based on the functional transformation of uniformly distributed sample values: $\delta_i = \Delta_i \cdot \tan(\pi \cdot (r_i - 0.5))$, where r_i is uniformly distributed on the interval $[0, 1]$.

In order to estimate Δ, we can apply the Maximum Likelihood Method

$$\rho(\delta^{(1)}) \cdot \rho(\delta^{(2)}) \cdot \ldots \cdot \rho(\delta^{(N)}) \to \max,$$

where $\rho(z)$ is a Cauchy distribution density with the unknown Δ. When we substitute the above-given formula for $\rho(z)$ and equate the derivative of the product with respect to Δ to 0 (since it is a maximum), we get an equation

$$\frac{1}{1 + \left(\dfrac{\delta^{(1)}}{\Delta}\right)^2} + \ldots + \frac{1}{1 + \left(\dfrac{\delta^{(N)}}{\Delta}\right)^2} = \frac{N}{2}. \tag{2}$$

The left-hand side of (2) is an increasing function that is equal to $0 (< N/2)$ for $\Delta = 0$ and $> N/2$ for $\Delta = \max \left|\delta^{(k)}\right|$; therefore the solution to the equation (2) can be found by applying a bisection method to the interval $\left[0, \max \left|\delta^{(k)}\right|\right]$.

It is important to mention that we assumed that the function f is reasonably linear within the box $[\widetilde{x}_1 - \Delta_1, \widetilde{x}_1 + \Delta_1] \times \ldots \times [\widetilde{x}_n - \Delta_n, \widetilde{x}_n + \Delta_n]$. However, the simulated values δ_i may be outside the box. When we get such values, we do not use the function f for them, we use a normalized function that is equal to f within the box, and that is extended linearly for all other values (we will see, in the description of an algorithm, how this is done).

As a result, we arrive at the following algorithm.

Algorithm.

– Apply f to the results of direct measurements: $\widetilde{y} := f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$;

– For $k = 1, 2, \ldots, N$, repeat the following:

- use the standard random number generator to compute n numbers $r_i^{(k)}$, $i = 1, 2, \ldots, n$, that are uniformly distributed on the interval $[0, 1]$;
- compute Cauchy distributed values $c_i^{(k)} := \tan(\pi \cdot (r_i^{(k)} - 0.5))$;
- compute the largest value of $|c_i^{(k)}|$ so that we will be able to normalize the simulated measurement errors and apply f to the values that are within the box of possible values: $K := \max_i |c_i^{(k)}|$;
- compute the simulated measurement errors $\delta_i^{(k)} := \Delta_i \cdot c_i^{(k)} / K$;
- compute the simulated "actual values" $x_i^{(k)} := \widetilde{x}_i - \delta_i^{(k)}$;

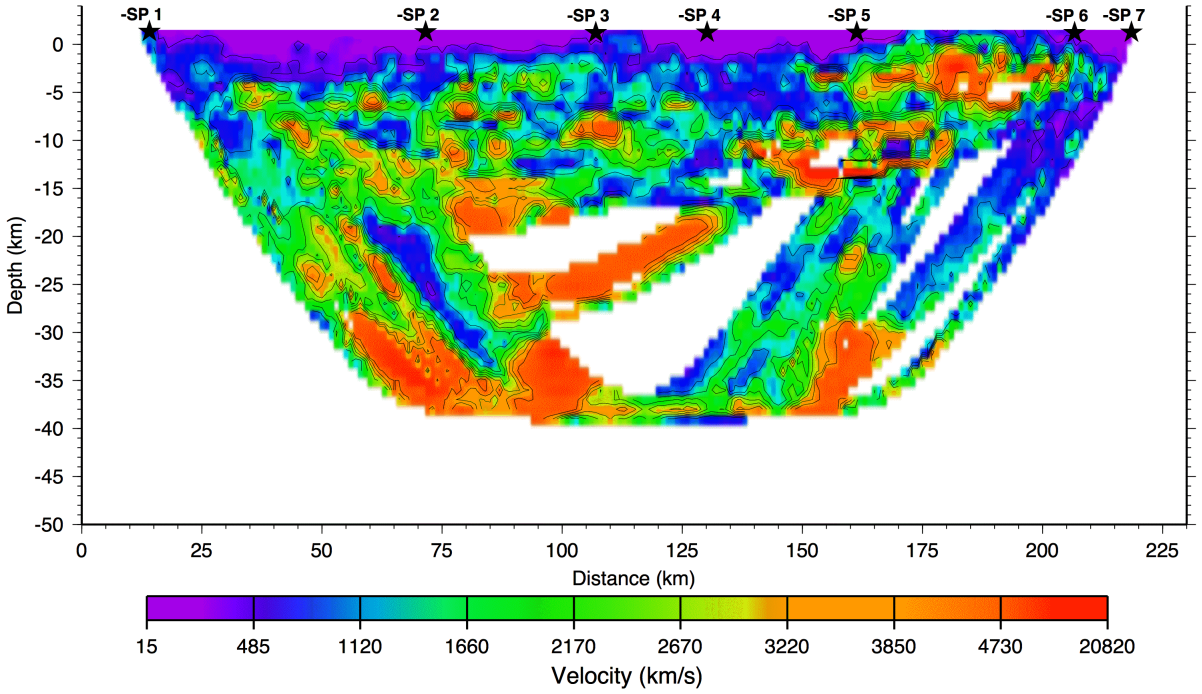
- apply the program f to the simulated "actual values" and compute the simulated error of the indirect measurement:

$$\delta^{(k)} := K \cdot \left(\widetilde{y} - f\left(x_1^{(k)}, \ldots, x_n^{(k)}\right)\right);$$

– Compute Δ by applying the bisection method to solve the equation (2).

Comment. To avoid confusion, we should emphasize that, in contrast to the Monte-Carlo solution for the probabilistic case, the use of Cauchy distribution in the interval case is a computational trick and *not* a truthful simulation of the actual measurement error Δx_i: indeed, we know that the actual value of Δx_i is always inside the interval $[-\Delta_i, \Delta_i]$, but a Cauchy distributed random attains values outside this interval as well.

Interval case: results. The results (given below) show some interesting features which we can use to qualitatively interpret the accuracy of the velocity values. In general, the values correspond well to the density and geometry of ray coverage in the model (see an earlier picture). The lowest values are in the upper part of the model and along paths of greatest ray coverage. The highest values or regions of lowest resolution are deeper in the model, near the center of the model with low ray coverage, and beneath the El Paso area (between shotpoints 5 and 6), where urban noise has decreased the number of travel-time picks and their quality. Whereas theses values do provide a good assessment of reliability for different regions of the model, they are clearly not useful in absolute terms.



Comment. When instead of single value $\Delta_i = 150$ ms, we used more realistic different values at different sensor locations (Zelt and Forsyth, 1994), we got similarly over-large results.

Clarifying comment. The above negative result is easy to explain by the following back-of-the-envelope calculations. Let us take two neighboring sensors at a distance $d = 600$ m from each other. Let x_1 be the time by which the seismic wave arrived at the first sensor, and let x_2 be the time by which this wave arrived at the second sensor. This means that this wave took time $t_2 - t_1$ to travel a distance d between the two sensors and thus, its velocity in this area can be estimated as $v = d/(x_2 - x_1)$.

The actual velocity near the surface is about $v \approx 2$ km/s, so the actual time difference is $x_2 - x_1 = d/v \approx 0.3$ sec. The observed value $\widetilde{x}_2 - \widetilde{x}_1$ is thus 0.3 sec, and the upper bound on each measurement error Δx_1 and Δx_2 is 0.15 sec. Thus, the upper bound on the error $\Delta x_2 - \Delta x_1$ is 0.3 sec. So, by using interval uncertainty, we conclude that the actual (unknown) value of $x_2 - x_1$ can take any value from 0 to 0.6 sec. When $x_2 - x_1$ is close to 0, for the corresponding velocity $v = d/(x_2 - x_1)$ we get meaningless thousands of km/s.

12.4. A New Heuristic Approach

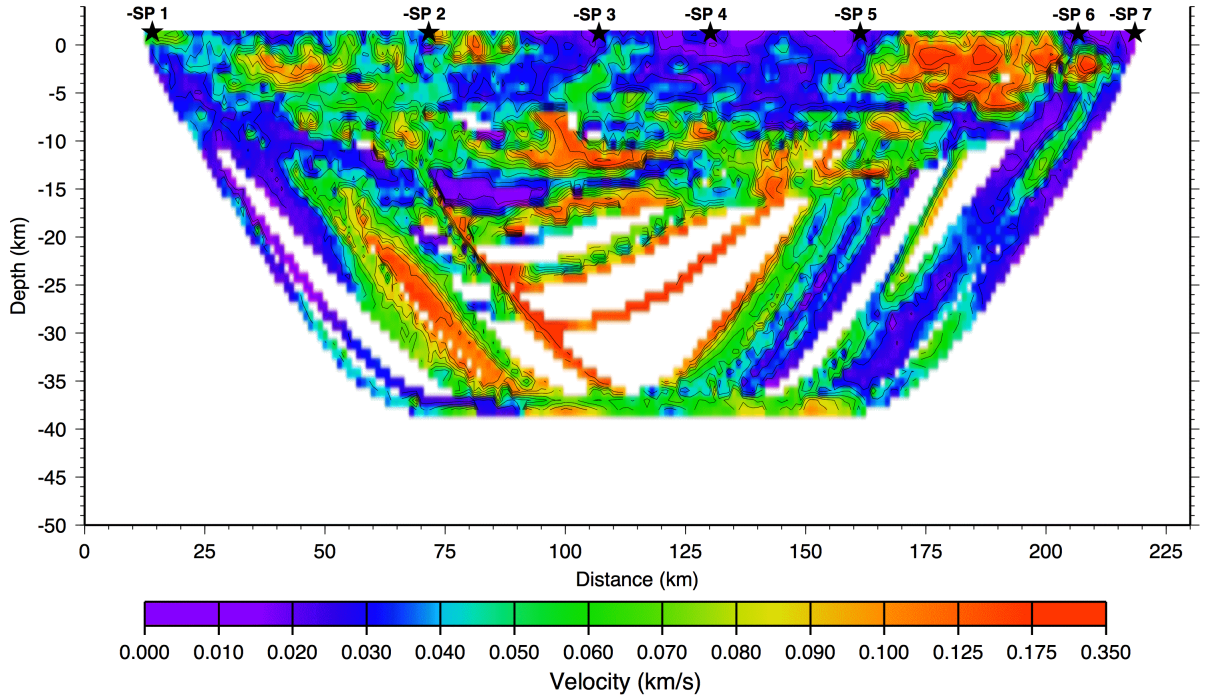
Towards the main idea. The *guaranteed* bounds provided by the interval approach are too high. How can we improve these bounds?

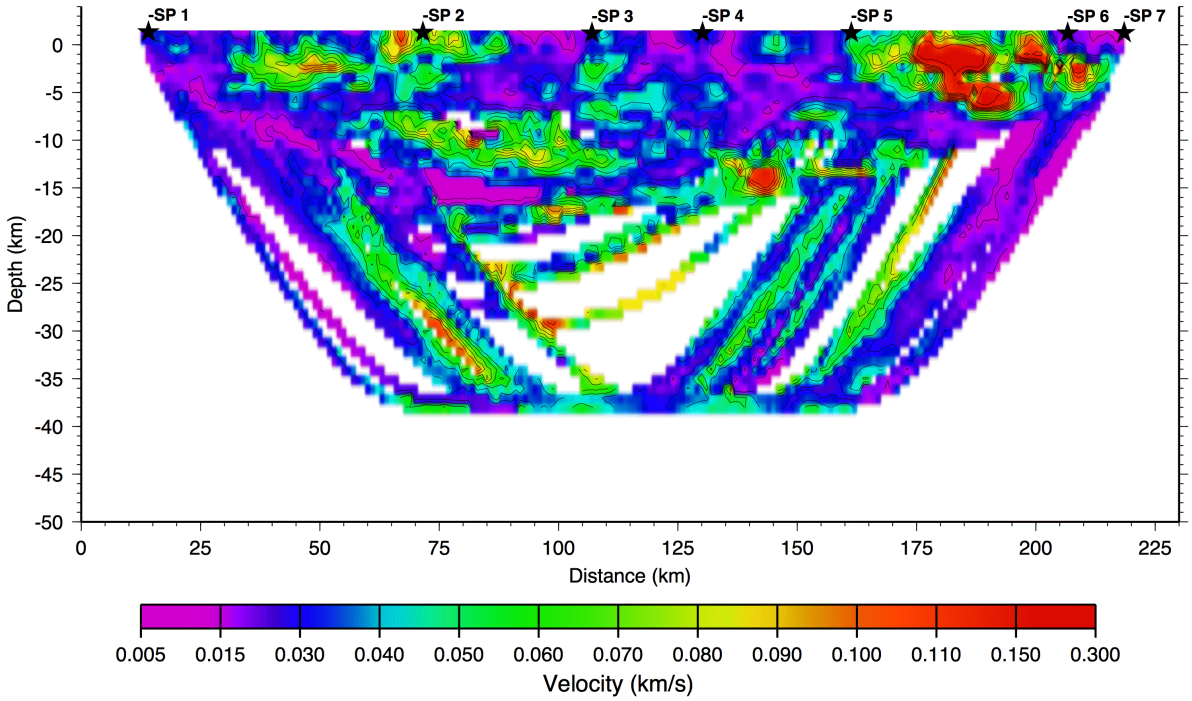
One possible solution comes from the following simple observation. For the normally distributed random variable with 0 mean and standard deviation σ, the only guaranteed upper bound is ∞. In practice, however, we can say that with confidence 90%, the actual value of this variable does not exceed 2σ, with confidence 99.9%, it does not exceed 3σ, etc. To get a bound with 90% confidence, we "cut-off" the top 10% of the normal distribution. To get the bound with 99.9% confidence, we "cut-off" the top 0.1% of the normal distribution, etc.

Main idea. Since guaranteed bounds are too high, it is reasonable to restrict ourselves to bounds guaranteed with a given confidence, e.g., bounds which are guaranteed with a confidence of 95% and dismiss the top 5% of uncertainty values. To find such bounds in the Cauchy method, we "cut-off" the top 5% of the corresponding Cauchy distribution. To be more precise, we find the threshold value x_0 for which the probability of exceeding this value is 5% (or any other desired cut-off probability p_0), and then replace values x for which $x > x_0$ with x_0 and for $x < -x_0$ with $-x_0$. For the Cauchy distribution, we have found that a 95% confidence level is obtained for the bounds of $-12.706 \le x_0 \le 12.706$ (see Appendix).

So, to get more realistic estimates for Δ, in the Cauchy approach, we use the "cut-off" Cauchy distribution instead of the original one.

Heuristic approach: results. The results of applying the Cauchy approach with 95% and 90% confidence are presented on the next page. Good news is that, in contrast to the practically useless interval-case values of uncertainty, here, velocity uncertainties Δ are exactly as expected. At the 95% confidence, the values Δ range from 0.01 to 0.3 km/s, and at a 90% confidence level, they range from 0.005 to 0.23 km/s.


-SP 1
-SP 2
-SP 3
-SP 4
-SP 5
-SP 6
-SP 7
Depth (km)
0
-5
-10
-15
-20
-25
-30
-35
-40
-45
-50
0
25
50
75
100
125
150
175
200
225
Distance (km)
0.000 0.010 0.020 0.030 0.040 0.050 0.060 0.070 0.080 0.090 0.100 0.125 0.175 0.350
Velocity (km/s)


-SP 1
-SP 2
-SP 3
-SP 4
-SP 5
-SP 6
-SP 7
Depth (km)
0
-5
-10
-15
-20
-25
-30
-35
-40
-45
-50
0
25
50
75
100
125
150
175
200
225
Distance (km)
0.005 0.015 0.030 0.040 0.050 0.060 0.070 0.080 0.090 0.100 0.110 0.150 0.300
Velocity (km/s)

On the qualitative level, the values Δ are still as geophysically reasonable as the values computed by the original interval-case method:

- the lowest values of Δ are found near the shotpoints, and along paths of highest ray coverage;
- the highest uncertainties are near the center of the model with lowest ray coverage and beneath the El Paso region between shotpoints 5 and 6.

Conclusions

In the past, communications were much slower than computations. As a result, researchers and practitioners collected different data into huge databases located at a single location such as NASA and US Geological Survey. At present, communications are so much faster that it is possible to keep different databases at different locations, and automatically select, transform, and collect relevant data when necessary. The corresponding cyberinfrastructure is actively used in many applications. It drastically enhances scientists' ability to discover, reuse and combine a large number of resources, e.g., data and services.

Because of this importance, it is desirable to be able to gauge the the uncertainty of the results obtained by using cyberinfrastructure. This problem is made more urgent by the fact that the level of uncertainty associated with cyberinfrastructure resources can vary greatly – and that scientists have much less control over the quality of different resources than in the centralized database. Thus, with the cyberinfrastructure promise comes the need to analyze how data uncertainty *propagates* via this cyberinfrastructure.

When the resulting accuracy is too low, it is desirable to produce the *provenance* of this inaccuracy: to find out which data points contributed most to it, and how an improved accuracy of these data points will improve the accuracy of the result. In this paper, we describe algorithms for propagating uncertainty and for finding the provenance for this uncertainty.

The above results mainly deal either with the *probabilistic* situations, when we either know the probability distributions of different measurement errors (and different errors are independent), or with *interval* situations, when we only know the upper bounds on the measurement errors. Probabilistic estimates tend to *underestimate* the resulting error – since in reality, different measurement errors are correlated (e.g., they have the same systematic error components). Interval estimates tend to *overestimate* because they are based on – often unrealistic – worst-case scenarios. It is thus desirable to combine these estimates to get more realistic error bounds. We describe several such combination methods, their mathematical justifications, and their successful use in processing geospatial data.

Acknowledgements

This work was supported in part by NSF grants HRD-0734825, EAR-0225670, and EIA-0080940, by Texas Department of Transportation grant No. 0-5453, by the Japan Advanced Institute of

Science and Technology (JAIST) International Joint Research Grant 2006-08, and by the Max Planck Institut für Mathematik.

References

Aguiar, M. S., G. P. Dimuro, A. C. R. Costa, R. K. S. Silva, F. A. Costa, and V. Kreinovich. The multi-layered interval categorizer tesselation-based model. In: C. Iochpe and G. Câmara, Editors. *IFIP WG2.6 Proceedings of the 6th Brazilian Symposium on Geoinformatics Geoinfo'2004*, Campos do Jordão, Brazil, November 22–24, 2004, pages 437–454.

Aldouri R., G. R. Keller, A. Q. Gates, J. Rasillo, L. Salayandia, V. Kreinovich, J. Seeley, P. Taylor, and S. Holloway. GEON: Geophysical data add the 3rd dimension in geospatial studies. In: *Proceedings of the ESRI International User Conference 2004*, San Diego, California, August 9–13, 2004, Paper 1898

Averill, M. G. *A Lithospheric Investigation of the Southern Rio Grande Rift*, University of Texas at El Paso, Department of Geological Sciences, PhD Dissertation, 2007.

Averill, M. G., K. C. Miller, G. R. Keller, V. Kreinovich, R. Araiza, and S. A. Starks. Using expert knowledge in solving the seismic inverse problem. In: *Proceedings of the 24nd International Conference of the North American Fuzzy Information Processing Society NAFIPS'2005*, Ann Arbor, Michigan, June 22–25, 2005, pages 310–314

Averill, M. G., K. C. Miller, G. R. Keller, V. Kreinovich, R. Araiza, and S. A. Starks. Using Expert Knowledge in Solving the Seismic Inverse Problem. *International Journal of Approximate Reasoning*, 453):564–578, 2007.

Ceberio, M., S. Ferson, V. Kreinovich, S. Chopra, G. Xiang, A. Murguia, and J. Santillan. How to take into account dependence between the inputs: from interval computations to constraint-related set computations, with potential applications to nuclear safety, bio- and geosciences. In: *Proceedings of the Second International Workshop on Reliable Engineering Computing*, Savannah, Georgia, February 22–24, 2006, pages 127–154.

Ceberio, M., V. Kreinovich, S. Chopra, and B. Ludäscher. Taylor model-type techniques for handling uncertainty in expert systems, with potential applications to geoinformatics. In: *Proceedings of the 17th World Congress of the International Association for Mathematics and Computers in Simulation IMACS'2005*, Paris, France, July 11–15, 2005.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 2001.

Doser, D. I., K. D. Crain, M. R. Baker, V. Kreinovich, and M. C. Gerstenberger. Estimating uncertainties for geophysical tomography. *Reliable Computing*, 4(3):241–268, 1998.

Fuller, W. A. *Measurement error models*. J. Wiley & Sons, New York, 1987.

Gates A. Q., V. Kreinovich, L. Longpré, P. Pinheiro da Silva, and G. R. Keller. Towards secure cyberinfrastructure for sharing border information. In: *Proceedings of the Lineae Terrarum: International Border Conference*, El Paso, Las Cruces, and Cd. Juárez, March 27–30, 2006.

Hole, J. A. Nonlinear High-Resolution Three-Dimensional Seismic Travel Time Tomography. *J. Geophysical Research*, 97(B5):6553–6562, 1992.

Jaulin, L., M. Kieffer, O. Didrit, and E. Walter. *Applied Interval Analysis*, Springer Verlag, London, 2001.

Keller, G. R., T. G. Hildenbrand, R. Kucks, M. Webring, A. Briesacher, K. Rujawitz, A. M. Hittleman, D. J. Roman, D. Winester, R. Aldouri, J. Seeley, J. Rasillo, T. Torres, W. J. Hinze, A. Gates, V. Kreinovich, and L. Salayandia. A community effort to construct a gravity database for the United States and an associated Web portal. In: A. K. Sinha, Editor. *Geoinformatics: Data to Knowledge*, pages 21–34, Geological Society of America Publ., Boulder, Colorado, 2006.

Kreinovich, V., J. Beck, C. Ferregut, A. Sanchez, G. R. Keller, M. G. Averill, and S. A. Starks. Monte-Carlo-type techniques for processing interval uncertainty, and their potential engineering applications. *Reliable Computing*, 13(1):25–69, 2007.

Kreinovich, V., and S. Ferson. A new Cauchy-Based black-box technique for uncertainty in risk analysis. *Reliability Engineering and Systems Safety*, 85(1–3):267–279, 2004.

Kreinovich, V., A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1998.

Longpré, L., and V. Kreinovich. How to efficiently process uncertainty within a cyberinfrastructure without sacrificing privacy and confidentiality", In N. Nedjah, A. Abraham, and L. de Macedo Mourelle, Editors. *Computational Intelligence in Information Assurance and Security*, pages 155–173, Springer-Verlag, 2007.

Lees, J. M., and R. S. Crosson. Tomographic inversion for three-dimensional velocity structure at Mount St. Helens using earthquake data. *Journal of Geophysical Research*, 94:5716–5728, 1989.

Maceira, M., S. R. Taylor, C. J. Ammon, X. Yang, and A. A. Velasco, High-resolution Rayleigh wave slowness tomography of Central Asia. *Journal of Geophysical Research*, Vol. 110, paper B06304, 2005.

Nguyen, H. T., O. Kosheleva, V. Kreinovich, and S. Ferson. Trade-Off Between Sample Size and Accuracy: Case of Dynamic Measurements under Interval Uncertainty. *Proceedings of International Workshop on Interval/Probabilistic Uncertainty and Non-Classical Logics UncLog'08*, JAIST, Japan, March 25–28, 2008 (to appear).

Nguyen, H. T., and V. Kreinovich. Trade-Off Between Sample Size and Accuracy: Case of Static Measurements under Interval Uncertainty. *Proceedings of International Workshop on Interval/Probabilistic Uncertainty and Non-Classical Logics UncLog'08*, JAIST, Japan, March 25–28, 2008 (to appear).

Parker, R. L. *Geophysical Inverse Theory*, Princeton University Press, Princeton, New Jersey, 1994.

Platon, E., K. Tupelly, V. Kreinovich, S. A. Starks, and K. Villaverde. Exact bounds for interval and fuzzy functions under monotonicity constraints, with potential applications to biostratigraphy. In: *Proceedings of the 2005 IEEE International Conference on Fuzzy Systems FUZZ-IEEE'2005*, Reno, Nevada, May 22–25, 2005, pages 891–896.

Rabinovich, S. G. *Measurement Errors and Uncertainty. Theory and Practice*, Springer Verlag, Berlin, 2005.

Schiek, C. G., R. Araiza, J. M. Hurtado, A. A. Velasco, V. Kreinovich, and V. Sinyansky. Images with Uncertainty: Efficient Algorithms for Shift, Rotation, Scaling, and Registration, and Their Applications to Geosciences. In: M. Nachtegael, D. Van der Weken, E. E. Kerre, and Wilfried Philips (eds.), Soft Computing in Image Processing: Recent Advances, Springer Verlag, 2007, pp. 35-64.

Sinha, A. K., Editor. *Geoinformatics: Data to Knowledge*, Geological Society of America Publ., Boulder, Colorado, 2006.

Tichelaar, B. W., and L. R. Ruff. How good are our best models? *EOS*, 70:593–606, 1989.

Torres R., G. R. Keller, V. Kreinovich, L. Longpré, and S. A. Starks. Eliminating duplicates under interval and fuzzy uncertainty: an asymptotically optimal algorithm and its geospatial applications. *Reliable Computing*, 10(5):401–422, 2004.

Vavasis, S. A. *Nonlinear Optimization: Complexity Issues*. Oxford University Press, New York, 1991.

Walster, G. W. Philosophy and practicalities of interval arithmetic. In: *Reliability in Computing*, pages 309–323, Academic Press, N.Y., 1988.

Walster, G. W., and V. Kreinovich. For unknown-but-bounded errors, interval estimates are often better than averaging. *ACM SIGNUM Newsletter*, 31(2):6–19, 1996.

Wen Q., A. Q. Gates, J. Beck, V. Kreinovich, J. R. Keller. Towards automatic detection of erroneous measurement results in a gravity database. In: *Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference*, Tucson, Arizona, October 7–10, 2001, pages 2170–2175.

Xie H., N. Hicks, G. R. Keller, H. Huang, and V. Kreinovich. An IDL/ENVI implementation of the FFT based algorithm for automatic image registration. *Computers and Geosciences*, 29(8):1045–1055, 2003.

Zelt, C. A., and P. J. Barton. Three-dimensional seismic refraction tomography: A comparison of two methods applied to data from the Faeroe Basin. *J. Geophysical Research*, 103(B4):7187–7210, 1998.

Zelt, C. A., and D. A. Forsyth. Modeling wide-angle seismic data for crustal structure Grenville province. *J. of Geophys. Res.* 99:11687-11704, 1994.

Appendix

The standard Cauchy distribution is characterized by the probability density function $\rho(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$. We are given a small probability p_0 (e.g., $p_0 = 5\%$), and we want to find the value x_0 such that the probability that $|x| \geq x_0$ is exactly p_0. In other words, we want the probability that $x \geq x_0$ or $x \leq -x_0$ to be equal to p_0. Since the Cauchy distribution is symmetric, the probability that $x \leq -x_0$ is equal to the probability that $x \geq x_0$. Therefore, the probability that $|x| \geq x_0$ is equal to twice the probability that $x \geq x_0$: $p_0 = \text{Prob}(|x| \geq x_0) = 2 \cdot \text{Prob}(x > x_0)$.

For the Cauchy distribution,

$$\frac{p_0}{2} = \text{Prob}(x > x_0) = \frac{1}{\pi} \cdot \int_{x_0}^{\infty} \frac{1}{1+x^2} = \frac{1}{\pi} \cdot (\arctan(\infty) - \arctan(x_0)) =$$

$$\frac{1}{\pi} \cdot \left(\frac{\pi}{2} - \arctan(x_0)\right) = \frac{1}{2} - \frac{1}{\pi} \cdot \arctan(x_0).$$

Thus, we must take $\arctan(x_0) = \frac{\pi}{2} \cdot (1 - p_0)$ and

$$x_0 = \tan\left(\frac{\pi}{2} \cdot (1 - p_0)\right). \tag{3}$$

For small p_0, we can get an even simpler formula. Indeed, in general, $x_0 = \tan\left(\frac{\pi}{2} - \frac{\pi}{2} \cdot p_0\right) = \frac{\sin\left(\frac{\pi}{2} - \frac{\pi}{2} \cdot p_0\right)}{\cos\left(\frac{\pi}{2} - \frac{\pi}{2} \cdot p_0\right)}$. We know that $\sin\left(\frac{\pi}{2} - \alpha\right) = \cos(\alpha)$ and $\cos\left(\frac{\pi}{2} - \alpha\right) = \sin(\alpha)$, so $x_0 = \frac{\cos\left(\frac{\pi}{2} \cdot p_0\right)}{\sin\left(\frac{\pi}{2} \cdot p_0\right)}$. For small α, we have $\sin(\alpha) \approx \alpha$ and $\cos(\alpha) \approx 1$, hence for small p_0, we get $x_0 \approx \frac{1}{\frac{\pi}{2} \cdot p_0}$ and

$$x_0 \approx \frac{2}{\pi \cdot p_0}. \tag{4}$$

Stochastic wave groups in weakly nonlinear random waves

Francesco Fedele
School of Civil & Environmental Engineering, Georgia Institute of Technology, Georgia USA

Abstract. A stochastic model of wave groups is presented to explain the occurrence of large waves in nonlinear random seas. The model leads to the description of the non-Gaussian statistics of oceanic waves and to a new asymptotic distribution of crest heights over large waves in a form that generalizes the Tayfun model. Comparisons based on a first wave data set collected at the Tern platform in the northern North Sea during an extreme storm, and a second set collected in the southern North Sea (WACSIS) show good agreement with the new theoretical wave distributions. In particular, for broad band seas, the Tayfun model seems to fit the data, and thus it can be regarded suitable for describing crest statistics for engineering applications.

Keywords: crest height; stochastic wave group; second order effects; probability of exceedance; Gaussian sea; quasi-determinism, Slepian model.

1. INTRODUCTION

To the leading order of approximation, the free surface displacement $\eta(t)$ is a Gaussian process of time. Lindgren (1970,1972) showed that locally near a very high crest, the surface displacement tends to assume the same shape as the covariance function $\psi(T) = \langle \eta(t)\eta(t+T)\rangle$. This is the Slepian model (Kac & Slepian 1959) whose time-domain formulation was used by Tromans et al. (1991) to analyze wave measurements.

An alternative view of the Slepian model was offered in the eighties by Boccotti (1989,2000). His theory of quasi determinism revealed the mechanics of three dimensional wave groups and their relation to the occurrence of extreme waves in a Gaussian sea and confirmed with field experiments (Boccotti et al., 1993a,1993b, Phillips et al. 1993a, 1993b).

In Gaussian sea waves, both crest and trough distributions follow the same Rayleigh law for narrow-band spectra (Longuet-Higgins, 1952). In the more general case of Gaussian waves with finite-band spectra, the Rayleigh distribution serves as an upper bound for the exceedance probability of crest heights.

In reality, water waves are nonlinear, and the probability density function of the surface displacement tends to deviate from the Gaussian form. In particular, due to second order nonlinearities the water surface presents sharper crests and shallower rounded troughs. Thus, the skewness λ_3 of surface elevations is not zero (Longuet-Higgins 1963). The exact theoretical form of the corresponding distribution of nonlinear wave crests is not known under general conditions. A series expansion based on the Edgeworth's form the Gram-Charlier distribution was proposed by Longuet-Higgins (1963), but can lead to expressions that violate the non-negativity condition on probability densities. Crest heights of large waves can be over predicted unrealistically in steep storm seas in deep or transitional

water depths. Convenient and simple narrow-band approximation for deep-water waves was given by Tayfun (1980, 1986a, 2006) in the early eighties based upon weakly second order wave theory. As a corollary, Tayfun (1980) also derived an analytical distribution for the crest statistics and a least-upper-bound (*lub*) distribution of crest heights (Tayfun and Al-Humoud, 2002). Comparisons of such models with various deep and shallow water second-order simulations have been carried out by Forristall (2000) and Prevosto & Forristall (2002).

The recent experimental results of Onorato et al. (2006) and the numerical simulations of Socquet-Juglard et al. (2005) both show that for the case of multidirectional random waves, the nonlinear effects are due dominantly to bound waves and the Tayfun distribution explains very well the crest statistics. Deviations from the Tayfun distribution may occur only in long-crested narrow-band waves due to third order nonlinear effects, such as the Benjamin- Feir type modulation instability (Zakharov 1999, Janssen 2003) as shown by Onorato et al. (2006) and Socquet-Juglard et al. (2005). Thus, for practical engineering applications where realistic oceanic conditions are characterized by multidirectional spectra, the second order Stokes theory, and thus the Tayfun model, still offers a valid theoretical framework for the wave statistics.

In this paper, we propose an alternative view of second order wave theory and a generalization of the Tayfun model. We first present an extension of the theory of quasi-determinism of Boccotti (1989,2000), defining a stochastic wave group that describes the dynamics of the wave surface around a randomly chosen very large crest (Lindgren 1970,1972). The stochastic wave group can be thought as a first order regression approximation according to Rychlik (1987) and Lindgren & Rychlik (1991).

In the second part of the paper, we shall study the nonlinear evolution of the stochastic wave group in the context of second order Stokes waves. This analysis will reveal the expected shape of large nonlinear crests and their statistics. In particular, we prove that the distribution of second order extreme crests is uniquely defined by the skewness λ_3 of the nonlinear surface displacement. This result is in perfect agreement with the narrow-band model of Tayfun (1980,1986a,2006), and it is valid for waves at deep and transitional water depths in a manner free of any constraints on their directionality or spectral bandwidth in agreement with the analytical results of Fedele & Arena (2005). In addition, a generalization of the Tayfun model (1980, 1986a) is proposed. Both the models are free of any bandwidth constraints and depends only on the global properties of the spectrum available from wave hindcasts. We also consider the Weibull model of Forristall (2000), and an exact closed form solution of the crest distribution based on the asymptotics for the h-upcrossings in Gaussian multivariate processes derived by Breitung and Richter (1996) which yields to the First Order Reliability Method (FORM).

Comparisons based on a first wave data set collected at the Tern platform in the northern North Sea during an extreme storm, and a second set collected in the southern North Sea (WACSIS) are presented. In particular, for broad band seas, the new theoretical models do not improve upon the Tayfun distribution (Tayfun 1980,1986, 2006), which thus can be regarded suitable for describing crest statistics for engineering applications.

2. Second order random waves

Consider weakly nonlinear random waves propagating in water of uniform depth d. The second order sea surface displacement ζ from the mean sea level at a fixed point $\mathbf{x}$ is given by

$$\zeta(\mathbf{x},t) = \zeta_1(\mathbf{x},t) + \zeta_2(\mathbf{x},t) \tag{1}$$

where the first order linear Guassian component ζ_1 is of the form

$$\zeta_1(\mathbf{x},t) = \sum_{i=1}^{N} z_i \cos(\boldsymbol{\theta}_i) \tag{2}$$

and the second order correction ζ_2 is given by

$$\zeta_2(\mathbf{x},t) = \frac{1}{4}\sum_{i,j=1}^{N} z_i z_j \left[A_{ij}^{+}\cos(\boldsymbol{\theta}_i+\boldsymbol{\theta}_j) + A_{ij}^{-}\cos(\boldsymbol{\theta}_i-\boldsymbol{\theta}_j)\right], \tag{3}$$

with

$$\boldsymbol{\theta}_i = \mathbf{k}_i\cdot\mathbf{x} - \omega_i t + \varepsilon_i = k_i x\cos\phi_i + k_i y\sin\phi_i - \omega_i t + \varepsilon_i.$$

Here, A_{ij}^{+} and A_{ij}^{+} are second order interaction coefficients (see e.g. Sharma & Dean 1979, Forristall 2000), $\mathbf{k}_i$ are horizontal wave-number vectors, with $k_i = |\mathbf{k}_i|$, the directional angles ϕ_i refer to the x axis, $\mathbf{x} = (x,y)$ is the horizontal spatial vector coincident with the mean water surface, ω_i is the wave frequency related to $\mathbf{k}_i$ through the dispersion relation $k_i \tanh k_i d = \omega_i^2/g$. We assume that frequencies ω_i are different from each other, the number N is infinitely large and that the phase angles ε_i are independent and uniformly distributed in $[0, 2\pi]$. The linear wave amplitudes z_i are related to the wave spectral density $S(\mathbf{k})$ as

$$S(\mathbf{k})d\mathbf{k}{=}S(k,\phi)k\ \delta k\ \delta\phi = \sum_i \frac{z_i^2}{2},$$

where the sum is over i's for which $(k_i,\phi_i) \in ([k,k+\delta k],[\phi,\phi+\delta\phi])$.

2.1. Basic definitions and assumptions

The jth order moment of the linear spectrum is

$$m_j = \int_0^\infty \omega^j S(\mathbf{k})\, d\mathbf{k}.$$

The validity of the form assumed for ζ is measured by the smallness of the *rms* surface gradient (Tayfun 1993)

$$\mu_1 = \sqrt{\left\langle |\nabla\zeta_1|^2 \right\rangle} = m_4/g^2 << 1 \tag{4}$$

where $\langle\cdot\rangle$ means time average. The spectral mean frequency ω_m, the mean zero-upcrossing frequency ω_0 of the underlying linear process ζ_1 and the bandwidth ν of the spectral density $S(\mathbf{k})$ are defined respectively as

$$\omega_m = \frac{m_1}{m_0}, \quad \omega_0 = \sqrt{\frac{m_2}{m_0}}, \quad \nu = \sqrt{\frac{m_0 m_2}{m_1^2} - 1}. \tag{5}$$

Moreover, $EX_+ = \omega_0/2\pi$ is the expected number per unit time of zero up-crossings of ζ, correct to $O(\mu_1)$. To the same order, the space-time covariance $\Psi(\mathbf{X}, T)$ of ζ is given by

$$\Psi(\mathbf{X}, T) = \langle \zeta_1(\mathbf{X}, t)\, \zeta_1(\mathbf{X}, t+T) \rangle = \int S(\mathbf{k}) \cos(\mathbf{k}\cdot\mathbf{X} - \omega T) d\mathbf{k}$$

where $\mathbf{X} = (X, Y)$ and $\psi(T) = \Psi(\mathbf{0}, T)$ for brevity. Hereafter, the first absolute minimum of $\psi(T)$ occurs at time $T = T^*$ and that $\psi(T)$ decreases monotonically between $T = 0$ (when the absolute maximum is attained) and $T = T^*$.

The first moment $\langle\zeta\rangle = 0$, and the higher order moments $\langle\zeta^p\rangle$ with $p = 2, ..4$ are given, correct to $O(\mu_1)$, by

$$\begin{aligned}
\left\langle \zeta^2 \right\rangle &= m_0 + O\left(\mu_1^2\right), \qquad (6)\\
\left\langle \zeta^3 \right\rangle &= \frac{3}{2} \int S(\mathbf{k}_1) S(\mathbf{k}_2) \left[A^+(\mathbf{k}_1, \mathbf{k}_2) + A^-(\mathbf{k}_1, \mathbf{k}_2) \right] d\mathbf{k}_1 d\mathbf{k}_2 + O\left(\mu_1^2\right),\\
\left\langle \zeta^4 \right\rangle &= 3m_0^2 + O\left(\mu_1^2\right),
\end{aligned}$$

where $A^{\pm}(\mathbf{k}_i, \mathbf{k}_j) = A^{\pm}_{ij}$. The spectral mean frequency and the mean zero-upcrossing frequency of the nonlinear process ζ are given by ω_m and ω_0 in Eq. (5) and they are correct to $O(\mu_1)$.

3. Large crests in Gaussian seas

Assume for the moment that a large wave crest of amplitude h is observed at $\mathbf{x} = \mathbf{x}_0 = (x_0, y_0)$ and $t = t_0$. Boccotti (2000) and Fedele (2006b) showed that as $h/\sigma \to \infty$, with probability approaching 1, a well defined wave group passes through the point $\mathbf{x} = \mathbf{x}_0$, with the apex of its development stage occurring at time $t = t_0$. As $h/\sigma \to \infty$, the surface displacement ζ_c around $\mathbf{x} = \mathbf{x}_0$ and $t = t_0$ is asymptotically described by the sum of a deterministic part $\zeta_{\rm det}$ of $O(h)$ and a residual random process R_ζ of $O(1)$, viz.

$$\zeta_c(\mathbf{X}, T) = \zeta_{\rm det}(\mathbf{X}, T) + R_\zeta(\mathbf{X}, T), \tag{7}$$

where

$$\zeta_{\rm det}(\mathbf{X}, T) = \langle \zeta_1(\mathbf{X}, T) \,|\, \zeta_1(\mathbf{0}, 0) = h \rangle = \frac{h}{\sigma^2} \Psi(\mathbf{X}, T). \tag{8}$$

Thus, ζ_c represents the conditional process $\zeta_1(\mathbf{X}, T) \,|\, \zeta_1(\mathbf{0}, 0) = h$ and $\zeta_{\rm det}$ is its conditional expectation. As $h/\sigma \to \infty$ in (7), the residual R_ζ becomes negligible relative to the first term, leading to

$$\zeta_c(\mathbf{X}, T) = \zeta_{\rm det}(\mathbf{X}, T) + O(h^0). \tag{9}$$

Thus, a high local maximum also corresponds to a local wave crest since $\zeta_{\rm det}$ attains its absolute maximum at $(T=0, \mathbf{X}=\mathbf{0})$. Moreover, ζ_c can be also interpreted as the wave surface around a randomly chosen large crest (Lindgren 1970,1972; Boccotti 2000) if h is assumed to be a random variable described by the Rayleigh probability density

$$p_R(h) = \frac{EX(h)}{EX_+} = \exp\left(-\frac{h^2}{2\sigma^2}\right)\frac{h}{\sigma^2}, \tag{10}$$

where $EX(h)dh$ represents the expected number per unit time of local maxima of the surface displacement recorded at $\mathbf{X}=0$ and $T=0$, whose amplitudes lie between h and $h+dh$. This model is the first order regression approximation of the wave process locally near a randomly chosen large crest (Rychlik 1987, Lindgren & Rychlik 1991). The random process (9) represents a family of wave groups which evolves in space and time attaining the largest crest at $\mathbf{X}=0$ and $T=0$. Thus, $\zeta_c \approx \zeta_{\rm det}$ is asymptotically correct to $O(h)$, and it either represents the wave field locally to a given crest height h, or it defines the conditional process for the dynamics in space-time around a randomly chosen crest if h is interpreted as a Rayleigh distributed random variable.

Our principal interest is in two-dimensional crests of the surface displacement, viz. the largest maxima of a surface time series recorded at a fixed point. Therefore, h is Rayleigh distributed. In general though, (9) can be also interpreted as a snapshot of the wave surface locally around a three-dimensional crest at a particular instant of time. In this case, the variable h is not distributed according to the Rayleigh law. In fact, in Gaussian processes the crest height follows the Rayleigh distribution by virtue of the one-to-one correspondence between each h-upcrossing point and a maximum of amplitudes greater than a large threshold h. In multi-dimensional Gaussian fields, this one-to-one correspondence is lost since h-upcrossings are level curves. In this case, an appropriate definition of a h-upcrossing is necessary, yielding an asymptotic form of the crest distribution different from the Rayleigh law (Adler 1981, Adler & Hasofer 1976, Wilson & Adler 1982, Piterbarg 2003).

4. Stochastic wave groups

We now extend and generalize some results of Boccotti (1989) to wave groups with large crests. Boccotti considers, as $H/\sigma \to \infty$, the conditional process

$$\zeta_b(\mathbf{X},T) = (\zeta_1(\mathbf{X},T)\,|\zeta_1(\mathbf{X},0) = H/2,\ \zeta_1(\mathbf{X},T_w) = -H/2\,)$$

where H represents the largest wave height in the group, and $T_w = T^* + O(H^{-1})$ is the time-lag between the crest of the wave and the following trough. In particular, Boccotti derives the asymptotic form of the statistical distribution of H (Boccotti 1989, 2000, see also Tayfun & Fedele 2007b). Boccotti (2000) and later Fedele (2007b) both show that largest wave heights occur not as waves reach the apex of a group, but just after they pass it. In the present case, we draw upon Boccotti's concepts but consider the largest crest which occurs at the apex of a wave group. Specifically, we examine the conditional process $\zeta_c(\mathbf{X},T)$ around a large crest, and analyse its $O(h^0)$-random residual R_ζ and thus devise a new formulation of wave groups in Gaussian seas. First, the

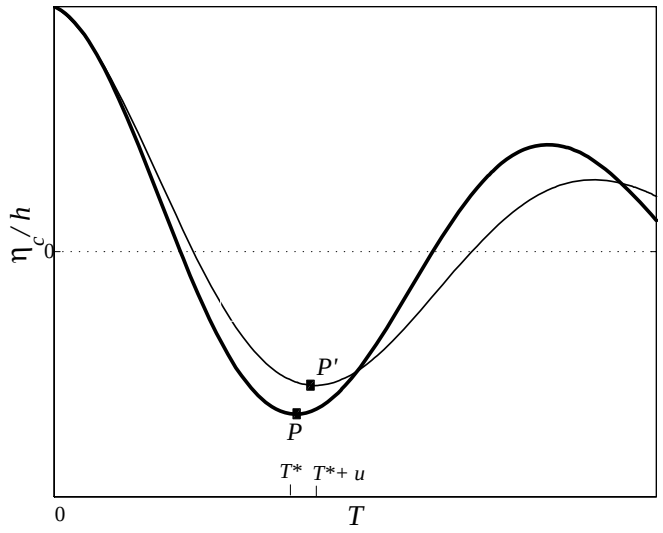


Figure 1.

wave profile $\eta_c(T)$ at $\mathbf{X} = \mathbf{0}$ is expressed in terms of an $O(h)$ contribution $\eta_{\rm det}(T) = \zeta_{\rm det}\,(\mathbf{0}, T)$ and the random residual $r(T) = R_\zeta(\mathbf{0}, T)$ of $O(h^0)$ as

$$\eta_c(T) = \eta_{\rm det}(T) + r(T) \tag{11}$$

where

$$\eta_{\rm det}(T) = \zeta_{\rm det}(\mathbf{0}, T) = h\frac{\psi(T)}{\sigma^2}.$$

We can now determine the effects of the residual $r(T)$ on η_c. Specifically, as $h/\sigma \to \infty$, with probability approaching 1, the surface profile locally near a large crest tends to assume the shape given by $\eta_{\rm det}(T)$ (see Lindgren 1972, Boccotti 2000). The latter represents a wave profile with a crest of amplitude h at time $T = 0$ followed by a local minimum of amplitude $\eta_{\rm det}(T^*)$ at $T = T^*$, with T^* being the abscissa of the first local minimum of $\psi(T)$ (point P in figure 1). Further, when the absolute minimum of $\psi(T)$ occurs at $T = T^*$, then $\eta_{\rm det}(T)$ represents a large wave with period $T_h \approx 2T^*$ and a crest-to-trough amplitude H given by

$$H = h\left(1 - \frac{\psi(T^*)}{\sigma^2}\right).$$

For large h, the wave trough of the profile $\eta_c(T)$ following the crest of amplitude h shall now occur at time $T = T^* + u$, shown as point $P\prime$ in figure 1, with u being random. To obtain an explicit expression for u, we set the time derivative of the profile η_c equal to zero at $T = T^* + u$ and use the expansion

$$\dot\eta_c(T^* + u) = \ddot\eta_{\rm det}(T^*)u + \dot r(T^*) + O(u^2) = 0.$$

Thus,

$$u = -\frac{\dot{r}(T^*)}{\ddot{\eta}_{\rm det}(T^*)} + O(u^2 h^{-1}). \tag{12}$$

Note that u is of $O(h^{-1})$ because the residual process $\dot{r}(T^*)$ is of $O(h^0)$ and $\ddot{\eta}_{\rm det}(T^*)$ is of $O(h)$. Thus, the residual terms in (12) are of $O(h^{-3})$ and negligible. By expansion, the value of the surface displacement $\eta_c(T)$ at $T^* + u$ is then given by

$$\eta_c(T^* + u) = \eta_{\rm det}(T^*) + \frac{1}{2}\ddot{\eta}_{\rm det}(T^*)u^2 + r(T^*) + O(h^{-2}). \tag{13}$$

Because u is of $O(h^{-1})$, it follows that

$$\eta_c(T^* + u) = \eta_{\rm det}(T^*) + \Delta + O\left(h^{-1}\right),$$

where $\Delta = r(T^*)$ is the residual at T^* of $O(h^0)$. Correct to the same order, $\eta_c(T^*) = \eta_c(T^* + u)$. Thus, as $h/\sigma \to \infty$, a crest of amplitude h that occurs at $T = 0$, is followed after a time lag $T^* + u$ by a trough, and $\eta_c(T)$ and its first time derivative $\dot{\eta}_c(T)$ at $T = T^*$ attain values given, correct to $O\left(h^0\right)$, by

$$\begin{aligned} \eta_c(T^*) &= \eta_{\rm det}(T^*) + \Delta + O\left(h^{-1}\right), \\ \dot{\eta}_c(T^*) &= -\ddot{\eta}_{\rm det}(T^*)u + O\left(h^{-1}\right). \end{aligned} \tag{14}$$

Conversely, if the conditions in (14) hold, then a crest of amplitude h at time $T = 0$ is followed by a trough at time $T = T^* + u$.

Next, we describe $\eta_c(T)$ locally near a randomly chosen crest, using a regression approximation (Rychlik 1987, Lindgren & Rychlik 1991). In particular, such an approximation must satisfy the conditions in (14), viz. it must have a local maximum of amplitude h at time $T = 0$ followed by a trough of amplitude $\eta^*_{\rm det} + \Delta$ at $T = T^* + u$. For linear Gaussian functions, an approximation to $\eta_c(T)$ satisfying both conditions exactly is given by

$$\eta_c(T) = A\psi(T) + B\psi(T - T^* - u), \tag{15}$$

where

$$A = \frac{\psi(0)h - \psi(T^* + u)\cdot(\psi(T^*)h + \Delta)}{\psi^2(0) - \psi^2(T^* + u)}, \qquad B = \frac{\psi(0)\cdot[\psi(T^*)h + \Delta] - \psi(T^* + u)h}{\psi^2(0) - \psi^2(T^* + u)}.$$

To $O(h^0)$, u drops out, and $\eta_c(T)$ becomes

$$\eta_c(T) = \eta_{\rm det}(T) + \frac{\Delta}{\sigma^2}\frac{-\psi^*\psi(T) + \psi(T - T^*)}{1 - \psi^{*2}} + O(h^{-1}), \tag{16}$$

ignoring terms of $O(h^{-1})$, and $\psi^* \equiv \psi(T^*)/\psi(0)$. With the random residual r of $O(1)$ explicitly determined now, it can be differentiated from $\eta_{\rm det}(T)$ of $O(h)$ in (11).

The necessary conditions for the existence of a local maximum at $T=0$, i.e. $\ddot{\eta}_c(0)<0$, and a local minimum at $T=T^*$, i.e. $\ddot{\eta}_c(T^*+u)>0$, yield the following inequality constraint:

$$h>\Delta\ \min\left(\frac{\psi^*+\ddot{\psi}^*}{1-\psi^{*2}},\frac{\psi^*+1/\ddot{\psi}^*}{1-\psi^{*2}}\right), \tag{17}$$

where $\ddot{\psi}^*\equiv\ddot{\psi}(T^*)/\left|\ddot{\psi}(0)\right|$. If the surface spectral density is defined over a compact support in the frequency domain, then the moments m_j for $j>3$ are finite, and $\eta(t)$ is differentiable at least twice. Thus, the terms appearing in (17) are bounded, and since Δ is of $O(1)$, h can be chosen sufficiently large to satisfy the above inequality, viz. $\Delta/h\sim O(h^{-1})$.

It is straightforward to extend the above time formulation to the space-time domain obtaining a new approximation of the stochastic wave group ζ_c in (7) in the form

$$\zeta_c(\mathbf{X},T)=\zeta_{\rm det}(\mathbf{X},T)+\frac{\Delta}{\sigma^2}\frac{-\psi^*\ \Psi(\mathbf{X},T)+\Psi(\mathbf{X},T-T^*)}{1-\psi^{*2}}+O(h^{-1}). \tag{18}$$

Evidently, this is an improved expression of the wave surface locally around a large crest correct to $O(h^0)$, where the random residual R_ζ in (7) is explicitly determined as $\Delta/h\to 0$, and terms of $O(h^{-1})$ have been neglected.

For a given h, ζ_c is the conditional processes locally around a given crest, i.e. $\zeta_1(\mathbf{X},T)\,|\zeta_1(\mathbf{0},0)=h$. If we instead interpret h and Δ as random, then ζ_c identifies a *stochastic wave group*, describing the dynamics locally around a randomly chosen crest.

The joint pdf of the random variables h, Δ and u, as $h/\sigma\to\infty$, is given by (Boccotti 1989)

$$p(h,\Delta,u)=h\frac{\exp\left(-\frac{h^2}{2\sigma^2}-\frac{\Delta^2}{2\sigma^2(1-\psi^{*2})}-\frac{h^2u^2\left|\ddot{\psi}(0)\right|}{2\sigma^4\gamma^2}\right)}{\sigma^2 2\pi\sqrt{\sigma^2(1-\psi^{*2})\frac{\sigma^4\gamma^2}{h^2\left|\ddot{\psi}(0)\right|}}}. \tag{19}$$

The probability $p(h,\Delta,u)dhd\Delta du$ can be interpreted as the fraction of realizations of linear ζ_1 with a large crest of amplitude h occurring at some t_0, proceeded by a trough of amplitude $\eta^*_{\rm det}+\Delta$ at T^*+u . As $h/\sigma\to\infty$, each realization of ζ_1 resembles a wave group evolving in accordance with (18).

The joint probability density of h and Δ follows from (19), with $\xi\to\infty$, as

$$p_{\xi,\tilde{\Delta}}(\xi,\tilde{\Delta})=\int_{-\infty}^{\infty}p(\xi,\tilde{\Delta},u)du=p_\xi(\xi)p_{\tilde{\Delta}}(\tilde{\Delta}), \tag{20}$$

where

$$p_\xi(\xi)=\xi\exp\left(-\frac{\xi^2}{2}\right),\qquad p_{\tilde{\Delta}}(\tilde{\Delta})=\frac{\exp\left(-\frac{\tilde{\Delta}^2}{2(1-\psi^{*2})}\right)}{\sqrt{2\pi(1-\psi^{*2})}}, \tag{21}$$

and $\xi=h/\sigma$ and $\tilde{\Delta}=\Delta/\sigma$ are dimensionless variables. Thus, ξ and $\tilde{\Delta}$ are independent. Note that with h given in (18), averaging over $\tilde{\Delta}$ yields the conditional mean

$$\langle\zeta_1(\mathbf{X},T)\,|\zeta_1(\mathbf{0},0)=h\rangle=\langle\zeta_c(\mathbf{X},T)\rangle_{\tilde{\Delta}}=\zeta_{\rm det}(\mathbf{X},T),$$

as expected.

5. Nonlinear stochastic groups and large crests

Herein, we examine the nonlinear evolution of the stochastic wave group $\zeta_c(\mathbf{X}, T)$ in second order random seas, explaining how its linear structure is distorted by nonlinearities. We argue that, prior to focussing, the nonlinear wave group tends to reflect the characteristics of a well defined Gaussian group that can be defined by (18). Due to nonlinearities, the Gaussian group will nonlinearly evolve forming an extreme crest with a different amplitude $h_{nl} > h$, h being the linear crest height. The relationship between h and h_{nl} is given by the nonlinear conditional process $\zeta_{nc} = (\zeta \,|\zeta_1 = \zeta_c)$. For large waves, ζ_{nc} is equivalent to $\zeta(\mathbf{X}, T)\,|\zeta_1(\mathbf{0}, 0) = h$, drawing upon Fedele & Arena (2005). The nonlinear mapping $f(\zeta_1)$ between ζ_1 and ζ is known from (1),(2) and (3), and it yields

$$\zeta_{nc} = (\zeta(\mathbf{X}, T)\,|\zeta_1(\mathbf{0}, 0) = h) = f(\zeta_c). \tag{22}$$

We recall that h and $\tilde{\Delta}$ are random variables with the joint pdf (20), and $\zeta_{nc} = f(\zeta_c)$ is the nonlinear stochastic group which describes the wave dynamics locally around a randomly chosen crest. To compute $f(\zeta_c)$, we note that (1) along with (2) and (3) *not only* defines weakly nonlinear random waves *but also* the general analytical solution for the second order surface displacement, if the amplitudes c_i and the phases $\boldsymbol{\theta}_i$ are regarded as deterministic variables. Thus, if we set in (1) the linear component ζ_1 of the surface ζ equal to ζ_c in (18), it follows that

$$\zeta_{nc} = f(\zeta_c) = \zeta_c + \frac{h^2}{4\sigma^4}\mathcal{F} + \frac{h\Delta}{2\sigma^4}\frac{-\psi^*\mathcal{F} + \mathcal{G}}{1 - \psi^{*2}} + O\left(\Delta^2\right), \tag{23}$$

where

$$\mathcal{F}(\mathbf{X}, T) = \int S_1 S_2 \left(A_{12}^+ \cos\left(\beta_{12}^+\right) + A_{12}^- \cos\left(\beta_{12}^-\right)\right) d\mathbf{k}_1 d\mathbf{k}_2, \tag{24}$$

$$\mathcal{G}(\mathbf{X}, T) = \int S_1 S_2 \left[A_{12}^+ \cos\left(\beta_{12}^+ + \omega_1 T^*\right) - A_{12}^- \cos\left(\beta_{12}^- + \omega_1 T^*\right)\right] d\mathbf{k}_1 d\mathbf{k}_2,$$

with the abbreviated notation $S_j = S(\mathbf{k}_j)$, $j = 1, 2$, and

$$A_{12}^{\pm} = A^{\pm}(\mathbf{k}_1, \mathbf{k}_2), \qquad \beta_{12}^{\pm} = (\mathbf{k}_1 \pm \mathbf{k}_2) \cdot \mathbf{X} - (\omega_1 \pm \omega_2)\, T.$$

6. Crest Statistics from nonlinear groups

The highest crest of the nonlinear stochastic wave group ζ_{nc} also occurs at $\mathbf{X} = 0$ and $T = 0$ correct to $O(\mu_1)$, with a dimensionless amplitude $\xi_{\max} = h_{nl}/\sigma$ given by

$$\xi_{\max} = \xi + \frac{\mu}{2}\xi^2 + \frac{\mu K}{2}\tilde{\Delta}\xi, \tag{25}$$

where

$$\mu = \frac{\lambda_3}{3} = \frac{\langle \zeta(t)^3 \rangle}{3\sigma^3}, \qquad K = 2\frac{-\psi^* + \kappa_1}{1 - \psi^{*2}} \tag{26}$$

with

$$\kappa_1 = \frac{\mathcal{G}(\mathbf{0},0)}{\mathcal{F}(\mathbf{0},0)} = \frac{\langle \zeta_1(\mathbf{0},t)\zeta_2(\mathbf{0},t)\ \zeta_1(\mathbf{0},t+T^*)\rangle}{2\mu}, \tag{27}$$

and λ_3 stands for the skewness coefficient of surface elevations correct to $O(\mu_1)$.

6.1. Recovering the Tayfun model

As $\xi \to \infty$, and ignoring terms of $O(\tilde{\Delta})$ in (25) we obtain

$$\xi_{\max} = \xi + \frac{\mu}{2}\xi^2. \tag{28}$$

Thus, the probability of exceedance for the nonlinear wave crest height $\xi_{\max}$ readily follows from the Rayleigh distribution of ξ as

$$\Pr\{\xi_{\max} > \lambda\} = \exp\left(-\frac{\xi(\lambda)}{2}^2\right), \tag{29}$$

where ξ follows from (28) with $\xi_{\max} = \lambda$. The result stated in (28) is valid for directional waves in waters of finite depth irrespective of the spectral bandwidth. It also agrees with the original narrow-band model of Tayfun (1980) appropriate to long-crested deep-water waves. In fact, Tayfun proposed the same expression for the crest height ξ_c, replacing μ with

$$\mu_m = m_0^{1/2}\frac{\omega_m^2}{g}. \tag{30}$$

This parameter is also a measure of steepness for unidirectional short-crested waves in deep water if one neglects the frequency-difference contributions. If the latter are included, then the parameter μ in (26) can be expressed explicitly for various theoretical spectra in the form

$$\mu = \mu_m(1 - \gamma\nu + \nu^2), \tag{31}$$

where, for example, $\gamma = 2/\sqrt{3} = 1.1547$ for rectangular spectra, and $\gamma = 2/\sqrt{\pi} = 1.1284$ for Gaussian spectra. For oceanic applications we shall assume that $\gamma = 1$ and define for the deep-water case

$$\mu_a = \mu_m(1 - \nu + \nu^2) \tag{32}$$

both for unidirectional waves and as an approximate upper bound for directional waves. As an alternative, Tayfun (2006) estimates μ from Forristall's Weibull model (Forristall 2000) as

$$\mu_{Fj} = 16\frac{\alpha_j^3}{\beta_j}\Gamma\left(\frac{3}{\beta_j}\right) - \frac{1}{4}\sqrt{\frac{\pi}{2}}, \tag{33}$$

where α_j and β_j represent the parameters of the Weibull distribution

$$\Pr\{\xi_{\max} > x\} = \exp\left[-\left(\frac{x}{4\alpha_j}\right)^{\beta_j}\right] \tag{34}$$

used by Forristall to fit (34) to simulations of second order random seas, and $j = 2$ or 3 corresponding to unidirectional (2D) or directional (3D) waves, respectively. Thus, not only for narrow-band waves, but also for high crest amplitudes, i.e. as $h/\sigma \to \infty$, crest heights are described by (28), with μ defined as $\lambda_3/3$ under the most general conditions. Moreover, all crest-height statistics depend clearly on a few integral properties such as m_0 (or σ), ω_m, ν and/or λ_3. These are easily estimated from a surface time series.

6.2. Generalizing the Tayfun model

As $\xi \to \infty$, and when we retain all the terms in (25), then

$$\Pr(\xi_{\max} > \lambda) = \int_{-\infty}^{\infty} \Pr\left\{\xi > \xi(\lambda, w) \left| \tilde{\Delta} = w \right.\right\} \; p_{\tilde{\Delta}}(\tilde{\Delta} = w) dw,$$

where $\xi(\lambda, w)$ follows from (25) with $\xi_{\max} = \lambda$ and

$$\Pr\left\{\xi > \xi(\lambda, w) \left| \tilde{\Delta} = w \right.\right\} = \exp\left[-\frac{\xi(\lambda, w)^2}{2}\right].$$

As $\lambda \to \infty$, an asymptotic solution to the preceding integral can be obtained, if we set

$$\xi\left(\lambda, \tilde{\Delta}\right) = \xi_0(\lambda) + a(\lambda)\,\tilde{\Delta} + O\left(\tilde{\Delta}^2\right) \tag{35}$$

where $\lambda = \xi_0 + \frac{\mu}{2}\xi_0^2$ and

$$a(\lambda) = -\frac{K}{2}\frac{\mu\ \xi_0}{1 + \mu\ \xi_0}. \tag{36}$$

Because ξ and $\tilde{\Delta}$ are statistically independent and by neglecting $O(\tilde{\Delta}^3)$, it follows after some algebra that

$$\Pr\{\xi_{\max} > \lambda\} = \frac{\exp\left[-\frac{1-\beta(\lambda)}{2}\xi_0^2\right]}{\sqrt{1 + \left(1 - \psi^{*2}\right) a(\lambda)^2}}, \tag{37}$$

where $\lambda >> 1$ and

$$\beta(\lambda) = \frac{\left(1 - \psi^{*2}\right) a(\lambda)^2}{1 + \left(1 - \psi^{*2}\right) a(\lambda)^2}.$$

We shall refer to this asymptotic result as the generalized Tayfun distribution. Evidently, it is not normalized to unity at the origin since its intended range of validity is over large waves. In the narrow-band limit as $\nu \to 0$, $K \to 0$, and the Tayfun distribution is recovered. An exact expression

for K in terms of spectral parameters can be obtained because the frequency-difference terms have been ignored. Under this condition, we recall that α vanishes and K takes the form

$$K = K^+ = -\frac{\ddot{\psi}^* + \psi^*}{1 - \psi^{*2}} \tag{38}$$

since $\kappa_1 = \left(\psi^* - \ddot{\psi}^*\right)/2$. Note further that in general, $\psi^* \to -1 + O(\nu)$ and $\ddot{\psi}^* \to 1 - O(\nu)$. Thus, if we include the frequency-difference terms, then $|K| \leq |K^+|$, and as $\nu \to 0$, $K^+ \to K \to 0$.

7. Crest statistics from Breitung's asymptotics

Recently, Baxevani et al. (2005) improved the asymptotic formula of h-upcrossings in Gaussian multivariate processes derived by Breitung and Richter (1996). They presented a rigorous view of the FORM (first order feliability method) and SORM (second order reliability method) used in applications to compute crest exceedances. We restrict our attention to FORM, and consider the hypersurface in the Euclidean space R^{2N} defined by the second order surface displacement of Eq.(1) written in terms of the column vectors $\mathbf{p} = (p_1, p_2, ..., p_N)$ and $\mathbf{q} = (q_1, q_2, ..., q_N)$, where $\{p_n\}$ and $\{q_n\}$ represent the sets of the spectral components of the linear surface displacement ζ_1 and its Hilbert transform respectively (see Baxevani et al. 2005 for details), that is

$$\lambda = \zeta_1(\mathbf{p}, \mathbf{q}) + \zeta_2(\mathbf{p}, \mathbf{q}), \tag{39}$$

with λ being a fixed threshold. Moreover the components of the vectors $\mathbf{p}, \mathbf{q}$ are independent Gaussian variables with zero mean and unit variance. Then the crest exceedance in FORM is given by

$$\Pr\{\xi_{\max} > \lambda\} = \exp\left[-\frac{g(\lambda)^2}{2}\right] \tag{40}$$

where $g(\lambda) = \|\mathbf{z}_{\min}\|$ is the minimal distance between the origin and the point $P_{\min} \in \mathbb{R}^{2N}$ identified by the column vector $\mathbf{z}_{\min} = [\tilde{\mathbf{p}}, \tilde{\mathbf{q}}]$ on the hypersurface Γ defined by (39). Here, $\|\mathbf{z}_{\min}\| = \sqrt{\tilde{\mathbf{p}}^T\tilde{\mathbf{p}} + \tilde{\mathbf{q}}^T\tilde{\mathbf{q}}}$ is the classical Euclidean norm of the vector $\mathbf{d} \in \mathbb{R}^{2N}$, and T signifies the transpose. In this case, the solution for $\mathbf{z}_{\min}$ can be obtained numerically by using standard optimization techniques (Tromans and Vanderschuren, 2004). If we compare the crest exceedance distribution of (29) with the FORM distribution of (40), it is seen that the vector entries $(\tilde{\mathbf{p}}, \tilde{\mathbf{q}})$ of the optimal vector $\mathbf{z}_{\min}$ for very large N, can be written, with a little abuse of notation, as

$$\tilde{\mathbf{p}} = \left[\xi_0(\lambda)\frac{\sqrt{2S(\mathbf{k}_1)d\mathbf{k}}}{\sigma}, ..., \xi_0(\lambda)\frac{\sqrt{2S(\mathbf{k}_N)d\mathbf{k}}}{\sigma}\right], \quad \tilde{\mathbf{q}} = \mathbf{0} \tag{41}$$

where $\lambda = \xi_0 + \frac{\mu}{2}\xi_0^2$. The Lagrange multiplier method and some algebra will show that $P_{\min} \in \mathbb{R}^{2N}$ pointed by the vector $\tilde{\mathbf{d}}$ is indeed the point on the hypersurface (39) at minimal distance from the origin, correct to $O(\mu\xi_0)$.

For simplicity, we shall prove the above statement for narrow-band waves only. In this case, the wave surface is given by (Tayfun 1980)

$$\zeta = \zeta_1 + \frac{\mu}{2}\left(\zeta_1^2 - \hat{\zeta}_1^2\right)$$

where $\hat{\zeta}_1$ is the Hilbert transform respect to time of ζ_1. Thus, from (39)

$$\zeta_1(\mathbf{p},\mathbf{q}) = \mathbf{z}^T\mathbf{p}, \qquad \zeta_2(\mathbf{p},\mathbf{q}) = \frac{\mu}{2}\left(\mathbf{p}^T\mathbf{z}\mathbf{z}^T\mathbf{p} - \mathbf{q}^T\mathbf{z}\mathbf{z}^T\mathbf{q}\right)$$

where μ is the steepness of the waves, and the column vector $\mathbf{z}$ has entries given by the spectral components $(\mathbf{z})_j = \sqrt{2S(\mathbf{k}_j)d\mathbf{k}}/\sigma$ such that $\mathbf{z}^T\mathbf{z} = 1$. Consider now the Lagrangian function

$$\mathcal{L} = \frac{1}{2}\left(\mathbf{p}^T\mathbf{p} + \mathbf{q}^T\mathbf{q}\right) + \chi\left(x - \mathbf{z}^T\mathbf{p} - \frac{\mu}{2}\left(\mathbf{p}^T\mathbf{z}\mathbf{z}^T\mathbf{p} - \mathbf{q}^T\mathbf{z}\mathbf{z}^T\mathbf{q}\right)\right)$$

where the Lagrange multiplier χ is introduced in order to minimize over the hypersurface Γ in (39). Some nontrivial algebra shows that the gradients $\frac{\partial\mathcal{L}}{\partial\mathbf{p}}$ and $\frac{\partial\mathcal{L}}{\partial\mathbf{q}}$ vanish for the critical vectors $(\tilde{\mathbf{p}},\tilde{\mathbf{q}})$ given by

$$\tilde{\mathbf{p}} = \rho\xi_c\mathbf{z}, \qquad \tilde{\mathbf{q}} = \mathbf{0}. \tag{42}$$

where

$$\rho = \frac{1}{1+\frac{\mu\xi_0}{2}} + \left(1+\frac{\mu\xi_0}{2}\right)\frac{\mu\xi_0}{2} = 1 + \frac{1}{2}\mu^2\xi_0^2 + O\left(\mu^3\xi_0^3\right).$$

Thus, the crest exceedance distribution is then given by

$$\Pr\{\xi_{\max} > \lambda\} = \exp\left[-\frac{\tilde{\mathbf{p}}^T\tilde{\mathbf{p}} + \tilde{\mathbf{q}}^T\tilde{\mathbf{q}}}{2}\right] = \exp\left[-\frac{\xi_0^2}{2}\rho^2\right]. \tag{43}$$

Also, one can show that the critical point $(\tilde{\mathbf{p}},\tilde{\mathbf{q}})$ on the hypersurface Γ of Eq. (42) is at minimal distance $g_{\min}(\lambda) = \rho\xi_0$ from the origin. Note finally that the Breitung distribution (43) coincides with the Tayfun distribution (29) correct to $O(\mu\xi_0)$.

8. Data Comparisons

In the following we shall present results of the analysis of two data sets. The first set comprises 9 hours of measurements gathered during a severe storm in January, 1993 with a Marex radar from the Tern platform located in the northern North Sea in 167 m water depth. The second set represents nearly 9 hours of measurements gathered in January, 1998 with a Baylor wave staff from Meetpost Noordwijk in 18 m average water depth in the southern North Sea. Forristall elaborates the nature of the first data, hereafter simply referred to as Tern. The second set is from Wave Crest Sensor Intercomparison Study and we shall call it as WACSIS for brevity (Forristall et al. 2002). The spectral properties of Tern are characterized by $\sigma = 3.02\ m$, $\nu = 0.629$ and $\lambda_3 = 0.174$ observed, and for WACSIS by $\sigma = 0.981\ m$, $\nu = 0.490$ and observed $\lambda_3 = 0.231$.

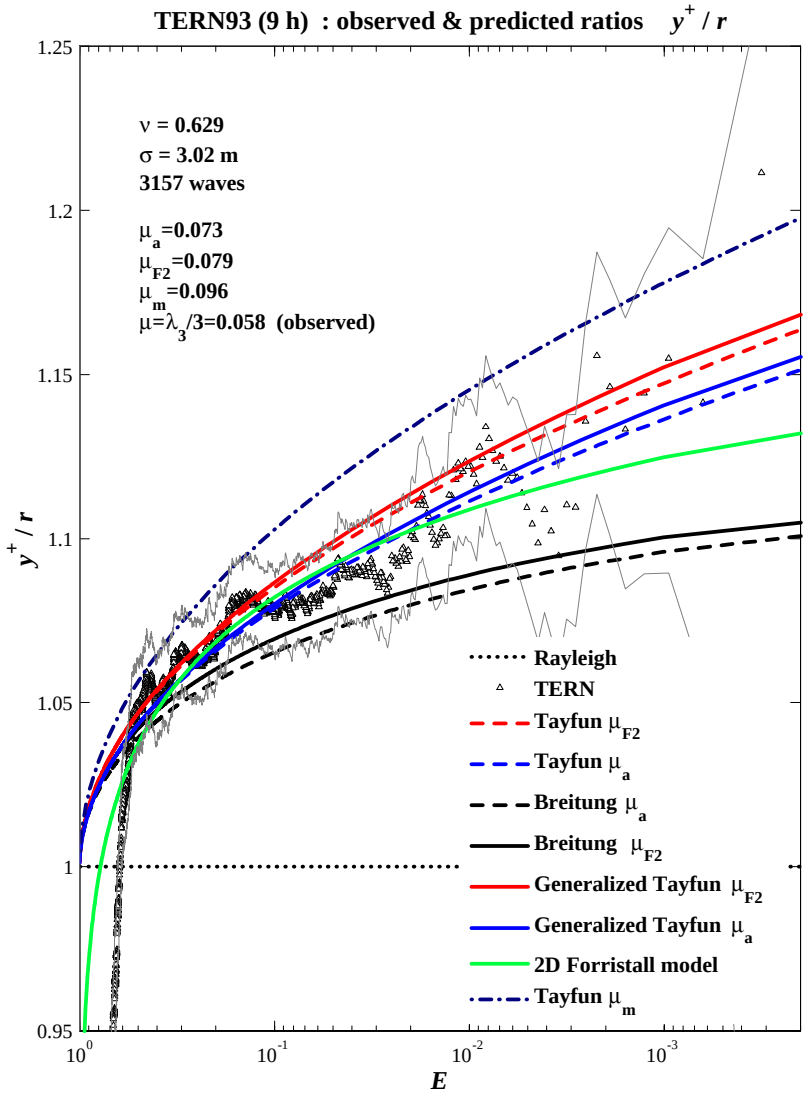

TERN93 (9 h) : observed & predicted ratios y^+ / r
ν = 0.629
σ = 3.02 m
3157 waves
μ_a=0.073
μ_{F2}=0.079
μ_m=0.096
$\mu=\lambda_3/3$=0.058 (observed)
Rayleigh
TERN
Tayfun μ_{F2}
Tayfun μ_a
Breitung μ_a
Breitung μ_{F2}
Generalized Tayfun μ_{F2}
Generalized Tayfun μ_a
2D Forristall model
Tayfun μ_m
y^+ / r
E
1.25
1.2
1.15
1.1
1.05
1
0.95
10^0
10^{-1}
10^{-2}
10^{-3}

Figure 2.

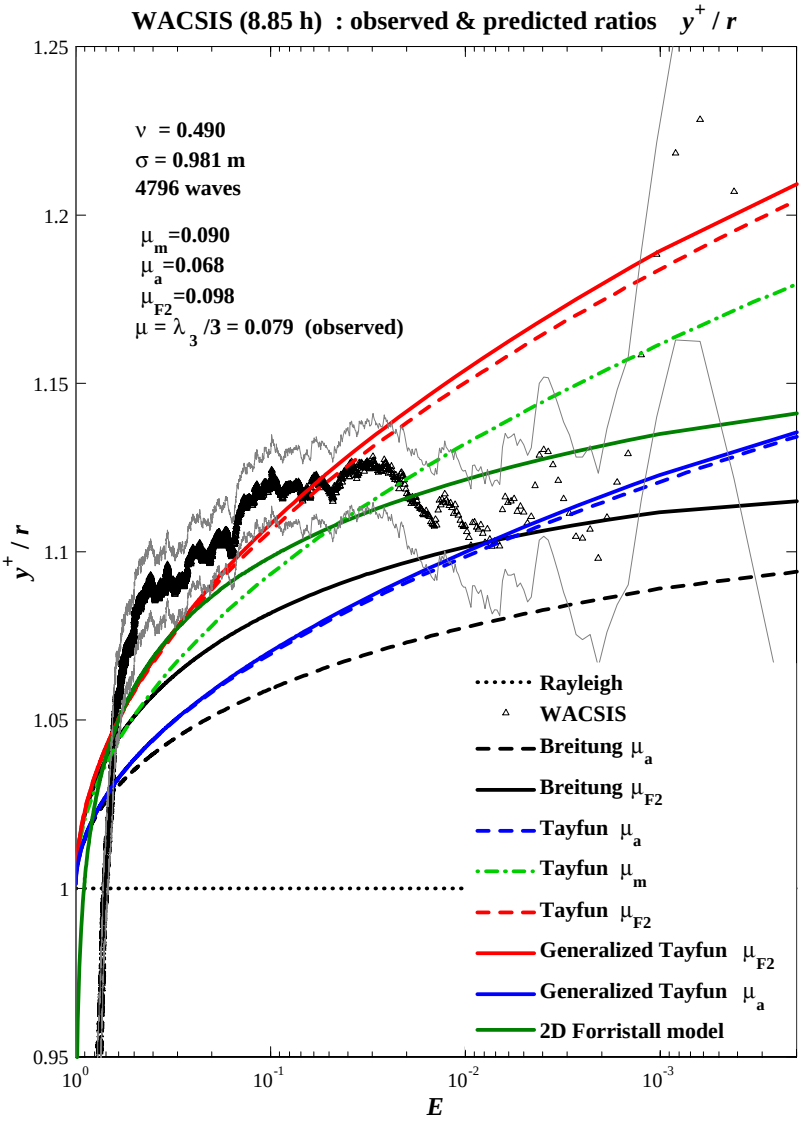

WACSIS (8.85 h) : observed & predicted ratios y^+ / r
ν = 0.490
σ = 0.981 m
4796 waves
μ_m=0.090
μ_a=0.068
μ_{F2}=0.098
$\mu = \lambda_3 /3 = 0.079$ (observed)
Rayleigh
WACSIS
Breitung μ_a
Breitung μ_{F2}
Tayfun μ_a
Tayfun μ_m
Tayfun μ_{F2}
Generalized Tayfun μ_{F2}
Generalized Tayfun μ_a
2D Forristall model
y^+ / r
E
1.25
1.2
1.15
1.1
1.05
1
0.95
10^0
10^{-1}
10^{-2}
10^{-3}

Figure 3.

In figure 2, the ratio y^+/r of nonlinear crests y^+ to the corresponding linear Rayleigh-distributed crests defined as $r = \sigma\sqrt{-2\ln P}$, is plotted for Tern and compared against the original Tayfun model ($\mu \simeq \mu_m = 0.096$ from (30)), the approximate model ($\beta = 1$ in (31) and $\mu \simeq \mu_a = 0.073$), the 2D Tayfun-Forristall model ($\mu \simeq \mu_{F_2} = 0.079$, see (33)), the 2D Weibull model of Forristall (see 34, $\alpha_2 = 0.3715$, $\beta_2 = 1.8683$), the generalized Tayfun models ($K = 0.394$) from (37) based on the estimates μ_{F_2} and μ_a, respectively, and finally the Breitung's approximation of (43). It is evident that the 2D Tayfun-Forristall model describes the observed data extremely well, whereas the original Tayfun model overestimates the observed crest heights, but it also serves as a somewhat conservative upper bound to the distribution of crest heights over high waves. Evidently, the improvement of the new distributions (Breitung and generalized Tayfun models) is essentially negligible. Similar results also hold for WACSIS, as shown in figure 3.

9. CONCLUSIONS

We have presented a complete theory for second order random waves and their statistics based on the concept of stochastic wave group. This theory provides a framework for predicting the expected shape of large waves and the statistics of large wave crests quite accurately within the context of second-order random wave theory and it can be extended to analyze the properties of third order nonlinear random waves (Fedele 2006a,2006c). We have proposed a generalization of the Tayfun model valid under general conditions in transitional or deep water depths, and that depends upon spectral parameters easily estimated from wave hindcasts. Furthermore, we derive an exact closed form solution for the crest distribution of FORM based on the Breitung's asymptotics (Breitung and Richter, 1996).

The generalized Tayfun model and the FORM model although compare well with oceanic measurements gathered from the Tern platform in the northern North Sea (Tern) and with a Baylor wave staff in the southern North Sea (WACSIS), do not really improve upon the original model of Tayfun, which thus can be regarded suitable for describing crest statistics for engineering applications.

10. Acknowledgements

The authors thank George Forristall for providing both the Tern and WACSIS data.

References

Adler R, Hasofer AM 1976. Level crossing for random fields. *The annals of Probability*, 4(1), 1-12.

Adler R, 1981. The geometry of random fields. Wiley, London.

Baxevani A., Hgaberg O., Rychlik I. 2005. Note on the distribution of extreme wave crests.*ASME Proc. 24th International Conference on Offshore Mechanics and Arctic Engineering* (OMAE 2005) paper OMAE2005-67571.

Boccotti P. On mechanics of irregular gravity waves. *Atti Acc. Naz. Lincei*, Memorie, 1989;19:11-170.

Boccotti P, Barbaro G and Mannino L. 1993a. A field experiment on the mechanics of irregular gravity waves. *J. Fluid Mech.*;252:173-186.

Boccotti P, Barbaro G, Fiamma V et al. 1993b. An experiment at sea on the reflection of the wind waves. *Ocean Engng.*;20:493-507.

Boccotti P. *Wave mechanics for ocean engineering*. Elsevier Science 2000, Oxford.

Breitung K. & Richter W.D. 1996. A geometric approach to an asymptotic Expansion for Large Deviation Probabilities of Gaussian Random vectors. J. Multivariate Analysis 58, 1-20 article no. 0036.

Fedele, F, Arena F. 2005. Weakly Nonlinear Statistics of High Non-linear Random Waves. *Physics of fluids*;17:1, 026601.

Fedele, F. 2005. Successive wave crests in Gaussian seas. *Prob. Eng. Mechanics* 20(4), 355-363.

Fedele, F. 2006a. Extreme Events in Nonlinear Random Seas. *ASME Journal Offshore Mechanics and Arctic Engineering* 128(1):11-16

Fedele F. 2006b. Wave Groups in a Gaussian Sea. Ocean Engineering, 33:17-18;2225-2239

Fedele F. 2006c. Explaining extreme waves by a theory of stochastic wave groups. *Computer & structures* special issue on computational stochastic mechanics (in press)

Forristall GZ. 2000. Wave Crest Distributions: Observations and Second-Order Theory. *Journal of Physical Oceanography*;30(8):1931-1943.

Forristall, GZ, Krogstad, HE, Taylor, PH, Barstow SS, Prevosto M, Tromans P. 2002. Wave crest sensor intercomparison study: an overview of WACSIS. *Proceedings, 21st International Conference on Offshore Mechanics and Arctic Engineering*, ASME, paper no. OMAE2002-28438, pp. 1-11.

Janssen, P A. E. M. 2003. Nonlinear four-wave interactions and freak waves. *J. Phys. Oceanogr.* 33, no. 4, 863–884.

Kac M & Slepian D. 1959. Large excursions of Gaussian processes. Ann. Math. Statist. 30,1215-1228.

Lindgren G. 1970. Some properties of a normal process near a local maximum. *Ann. Math. Statist.* 4(6):1870-1883.

Lindgren G. 1972. Local maxima of Gaussian fields. *Ark. Mat.* 10:195-218.

Lindgren G., Rychlik I, 1991. Slepian models and regression approximations in crossing and extreme value theory. *International Statistical Review/Revue International de Statistique*, 59(2), 195-225.

Longuet-Higgins MS. On the statistical distribution of the heights of sea waves, *J. Mar. Res.* 1952;11:245-266.

Longuet-Higgins MS 1957. The statistical analysis of a random, moving surface. Phil. Trans. Roy. Soc. A, 249, 321-387

Longuet-Higgins MS. 1963. The effects of non-linearities on statistical distributions in the theory of sea waves. J. *Fluid Mech.*;17:459-480.

Phillips OM, Gu D and Donelan M. 1993a. On the expected structure of extreme waves in a Gaussian sea, I. Theory and SWADE buoy measurements. *J. Phys. Oceanogr.*;23:992-1000.

Phillips OM, Gu D and Walsh EJ. 1993b. On the expected structure of extreme waves in a Gaussian sea, II. SWADE scanning radar altimeter measurements. *J. Phys. Oceanogr.*;23:2297-2309.

Onorato M., Osborne AR,Serio L., Cavaleri L.,Brandini C., Stansberg CT 2006. Extreme waves, modulational instability and second order theory: wave flume experiments on irregular waves. European Journal of Mechanics - B/Fluids 25:5;586-601

Prevosto M,Forristall GZ. 2002. Statistics of wave crests from models vs. Measurements. *ASME Proc. 21st International Conference on Offshore Mechanics and Arctic Engineering Oslo* OMAE 2002;OMAE 28443 paper.

Rychlik I. 1987, Joint distribution of successive zero crossing distances for stationary Gaussian processes. *J. Appl. Prob.* 24, 378-385.

Sharma JN & Dean RG. 1979. Developement and Evaluation of a Procedure for Simulating a Random Directional Second Order Sea Surface and Associated Wave Forces. *Ocean Engineering Report* n.20, University of Delaware.

Socquet-Juglard, H., Dysthe, K., Trulsen, K., Krogstad, H.E. & Liu, J. 2005. Probability distributions of surface gravity waves during spectral changes, J. Fluid Mechanics 542, 195 - 216

Tayfun, M.A. 1980. Narrow-Band Nonlinear Sea Waves. *J. Geophys. Res.*;85(C3):1548-1552.

Tayfun, M.A. 1986a. On Narrow-Band Representation of Ocean Waves. Part I: Theory. *J. Geophys. Res.*;91(C6):7743-7752.

Tayfun, MA. 2006. Statistics of nonlinear wave crests and groups. *Ocean Engineering* 33:11-12;1589-1622

Tayfun, M. A. and Al-Humoud, J. 2002. Least Upper Bound Distribution for Nonlinear Wave Crests. *Journal of Waterway, Port, Coastal, and Ocean Engineering*;128(4):144-151.

Tayfun A. & Fedele F. 2007 Wave-height distributions and nonlinear effects. Ocean Engineering (in press). A shorter version is in Proceedings of the 25th International Conference on Offshore Mechanics and Arctic Engineering, Hamburg, Germany 2006, paper no. OMAE2006-92019

Tromans PS, Anaturk AR and Hagemeijer P. 1991. A new model for the kinematics of large ocean waves - application as a design wave -. *Shell International Research* publ. 1042.

Tromans P.S. & Vanderschuren L. 2004. A spectral Response Surface Method for Calculating Crest Elevation Statistics. *ASME Journal Offshore Mechanics and Arctic Engineering* 126(1):51-53

Wilson RJ, Adler R 1982. The structure of Gaussian fields near a level crossing. Advance Applied Prob. 14, 543-565.

Zakharov VE. 1999. Statistical theory of gravity and capillary waves on the surface of a finite-depth fluid. Journal of European Mechanics B-fluids;18(3):327-344.

Structural Integrity Prediction via Stochastic Local Regression

Seung-Kyum Choi
Systems Realization Laboratory
G. W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology
Savannah, GA, 31407
email: schoi@me.gatech.edu

Abstract: A primary challenge of stochastic analysis is to discover rigorous ways to forecast the low probability of failure which is critical to reliability constraints. In this paper, a new framework is proposed for the accurate estimation of the low failure probability. Combining the excellent advantages of the polynomial chaos expansion, and local regression method will result in a new simulation-based modeling technique that enables the accuracy of the structural integrity prediction. The proposed procedure can allow for realistic modeling of sophisticated statistical variations and facilitate in order to achieve improved reliability by eliminating unnecessary conservative approximations. An example problem is depicted to illustrate how the method is used to provide a quantitative basis for developing robust designs associated with the low probability of failure.

Keywords: Polynomial Chaos Expansion, Moving Least-Squares, Local Regression, Low Failure Probability

1. Introduction

In recent years, the rapid development and improvement of novel design concepts, especially utilizing novel material systems, is a major request of the aerospace and automobile industry. In addition, new digital and information science technologies are creating the potential for new high-level design fields, such as micro-electro-mechanical systems (MEMS) and multi-scale engineering systems. However, introducing this state-of-the-art technology and new material systems is rapidly increasing the complexity of most engineered systems. There exist significant difficulties in anticipating, understanding, designing, and controlling both normal and abnormal behaviors of the complex systems. In addition, uncertainties in material properties, geometry, manufacturing processes, and operational environments of the complex engineered systems are clearly critical at all scales (nano-, micro-, meso-, and macro-scale). For example, the typical tolerances of geometric accuracy and surface finish are on the order of tenths of microns during the fabrication processes (Maluf, 2004), and the common microfabrication material (i.e. polycrystalline silicon) has 9~15% variation in its Young's modulus and tensile strength (Sharpe, Turner, and Edwards, 1999).

To compensate the ignorance of uncertainties in input parameters, safety factors have traditionally been incorporated approximately in engineering designs. Generally, the factor of safety is understood to be the ratio of the expected strength to response to the expected load (Choi, Grandhi, and Canfield, 2006). In practice, both the strength and load are variables, the values of which are scattered about their respective mean values. When the scatter in the variables is considered, the factor of safety could potentially be less than unity, and the traditional factor of safety-based design would fail. More likely, the factor of safety is too conservative, leading to an overly expensive design for a given level of safety. Probabilistic methods are convenient tools to describe or model physical phenomena that are too complex to treat with the present level of scientific knowledge. The probabilistic method explicitly incorporates given statistical data into the design algorithms and provides safer designs at given cost, whereas conventional deterministic design with the safety factor discards such data. However, the probabilistic-based approach often requires repeated evaluations of the probability of failure and it induces the computational challenge associated with the large number of computer simulations when the system requires extremely low failure probability, such as 10^{-5}~10^{-7}.

A common approach to the computationally-expensive procedure of the probabilistic methods is to approximate the system response using relatively inexpensive surrogate modeling techniques. In the approximation of the response function, the accuracy depends on the choice of the basis function and the sampling method including the choice of the sampling region and the position of the sampling points. An effective choice of the basis function for the uncertainty analysis is the direct use of stochastic expansions, i.e. Polynomial Chaos Expansion (PCE) (Ghanem and Spanos, 1991), since the stochastic expansions provide analytically appealing convergence properties based on the concept of a random process. The PCE can reduce computational effort of uncertainty quantification in engineering design applications where the system response is computed implicitly. Choi et al. (2006) recently developed an uncertainty analysis framework which can account for nonlinear fluctuations of large-scale system responses by integrating the PCE, the Karhunen-Loeve (KL) transform, and Latin Hypercube Sampling (LHS). This research utilized the stochastic expansion and the dimension reduction procedure to generate the random field and showed the applicability of the method to the complex engineered systems.

The objective of the current study is to provide the accurate estimation of the low failure probability of complex engineered systems by utilizing efficient probabilistic methods which can realistically model complicated statistical variations. To achieve a high quality surrogate model, a local regression method, namely Moving Least-Squares (MLS) method (Lancaster and Salkauskas, 1981), is integrated to a previously developed probabilistic decision support framework (Choi, Canfield, and Grandhi, 2006). The main advantage of the MLS method is that the regression coefficients are not constant, but rather parameter dependent. This quality allows the data analysis to not be constrained to a specific global function in order to fit a model to the data. Instead, the fitting segments spawn a local-global approximation allowing the data to acclimate to the function over a wide range of parameters. The stochastic modeling process repeats and recalibrates the PCE model with the local regression scheme until sufficient model adequacies are achieved. This will allow for an accurate estimation of the low probability

of failure with limited sampling points. The following sections provide a brief description and main ideas behind the local regression method and then focus on the technical details integrating the stochastic approximation procedure to provide the accuracy of the structural integrity prediction of complex engineered systems.

2. Mathematical Basis for Solution Concept

2.1. LOCAL REGRESSION

The efficacy of local regression schemes such as MLS method, lazy learning method, and locally weighted regression method have been successfully shown in recent engineering applications (Lancaster and Salkauskas, 1981; Stone, 1977; Cleveland, 1979; Katkovnik, 1979; Toropov, Scharamm, Sahai, Jones, and Zeguer, 2005). The basic idea of the local regression is to fit curves and surfaces to localized subsets of the data by a multivariate smoothing procedure with moving processes. The detailed steps of the MLS approximation are described in Figure 1. First, we define a local domain based on the domain influence factor or bandwidth, r. In the second step, we construct an approximation at a calculation point, x_i. These procedures can be repeated to each different calculation point by moving the local domain. Therefore, the regression coefficients of the MLS are not constant but a function of the calculation position or location. The "moving" process is analogous to a weighted moving average method, which is a common method in a time series analysis. In fact, applying zero degree polynomials in the local regression yields a weighted moving average. The advantage of the local approximation compared to the classical global fitting methods is that the method does not require a global function of any form to fit a given model and can generate accurate and smooth fitting of nonlinear responses without significant distortions.

Consider the linear regression model

$$y(x) = \beta_0 + \beta_1 p_1(x) + ... + \beta_m p_m(x) + \varepsilon \quad (1)$$

where $p_j(x)$, $j = 0,1,2,\ldots,m$, are the basis polynomial of order m, β_j are the regression coefficients, and ε, the error of the model equation, is assumed to be normally distributed with mean zero and variance σ_e^2.

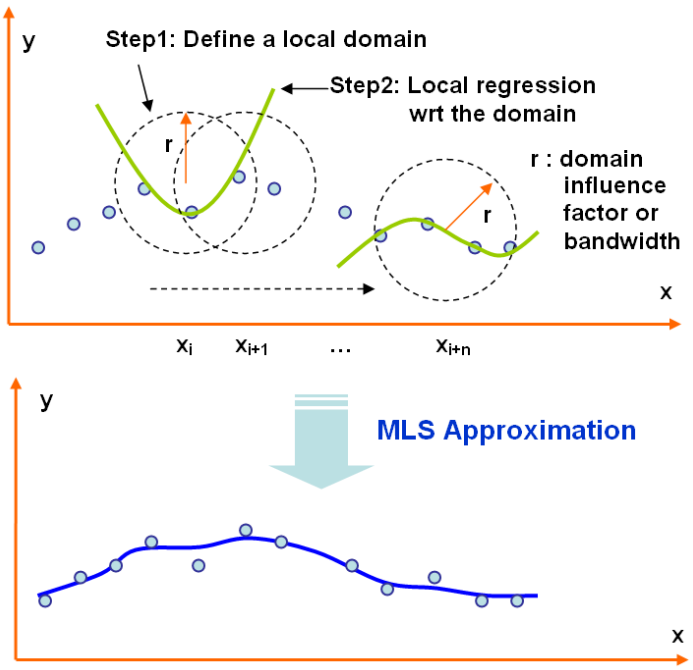


Figure 1. Moving Least-Squares Approximation

Equation (1) can be written in matrix notation for *n* sample values of *x* and *y* as

$$Y = X\hat{\beta} + e \tag{2}$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & p_1(x_1) & p_2(x_1) \ldots & p_k(x_1) \\ 1 & p_1(x_2) & p_2(x_2) \ldots & p_k(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & p_1(x_n) & p_2(x_n) \ldots & p_k(x_n) \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \textit{and} \quad e = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Here, the simplest polynomial model is the monomials of x^m, i.e., $p^T(x) = [1, x, x^2, ..., x^m]$ and in 2D space, $p^T(x, z) = [1, x, z, x^2, xz, z^2, ..., x^m, z^m]$.

The least-squares procedure results in obtaining the regression coefficients:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{3}$$

The fitted model and the residuals are

$$\hat{Y} = X\hat{\beta} \text{ and } e = Y - \hat{Y} \tag{4}$$

In the method of the Moving Least-Squares (MLS) approximation, the regression coefficient vector, $b(x)$, can be calculated as,

$$b(x) = [X^T W(x) X]^{-1} X^T W(x) Y \tag{5}$$

where X is a n x p matrix of the levels of the regressor variables, Y is a n x 1 vector of the responses, and $W(x)$ is a weight matrix and it is a none zero diagonal matrix:

$$W(x) = \begin{bmatrix} w_1(x-x_1) & 0 & \dots & 0 \\ 0 & w_2(x-x_2) & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & w_n(x-x_n) \end{bmatrix} \tag{6}$$

Consequently, the model Y in Eq. (2) can be approximated by MLS approximants $u^h(x)$ as follows

$$u^h(x) = \sum_{j=0}^{m} p_j(x) b_j(x) = p^T(x) b(x) \tag{7}$$

The weight matrix, Eq.(6), is a function of the location or position of x and there are several types of weighting functions:

(a) Exponential weight function

$$w_i(x-x_i) = w(d_i) = \begin{cases} \exp(-(d_i/r_i)^2), & if \quad d_i/r_i \le 1 \\ 0, & if \quad d_i/r_i > 1 \end{cases} \tag{8a}$$

(b) Conical weight function

$$w(d_i) = \begin{cases} 1-(d_i / r_i)^2, & if \quad d_i / r_i \le 1 \\ 0, & if \quad d_i / r_i > 1 \end{cases} \tag{8b}$$

(c) Spline weight function

$$w(d_i) = \begin{cases} 1-6(d_i / r_i)^2 + 8(d_i / r_i)^3 - 3(d_i / r_i)^4, & if \quad d_i / r_i \le 1 \\ 0, & if \quad d_i / r_i > 1 \end{cases} \tag{8c}$$

where $d_i = \|x - x_i\|$ is the distance from the sample point x_i to x, and the domain influence factor, r_i, is directly related to the smoothing length; namely, the size of the support for the weight function. It is also called the bandwidth.

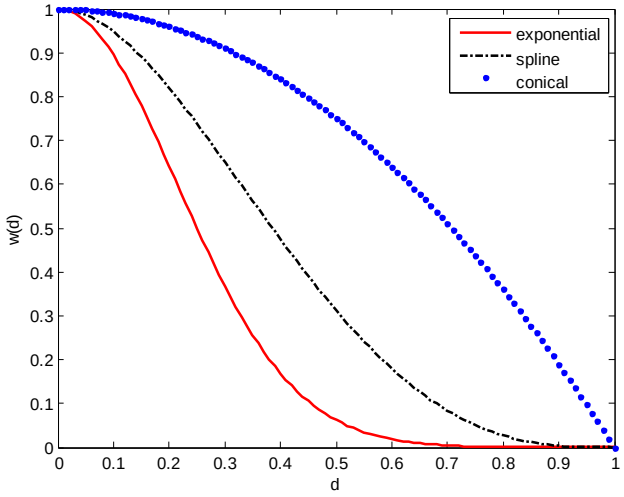


Figure 2. Weight Functions

Figure 2 depicts the three types of the weight functions discussed in this section. It is important to note that the shape of the fitted curve is not critically sensitive to the precise selection of the weight function. However, the careful adjustment of the domain influence factor of the weight function is critical so that the interval should contain enough data points to obtain the regression coefficients. Otherwise, the regression procedure will envisage a singular matrix. The additional discussion on the effects of several weighting functions and the resulting local approximation can be found in Ref. (Dolbow and Belytschko, 1998).

2.2. STOCHASTIC APPROXIMATION

The Polynomial Chaos Expansion (PCE) stemmed from both Wiener and Ito's work on mathematical descriptions of irregularities (Wiener, 1938). A simple definition of the PCE for a Gaussian random response, $u(\theta)$, as a convergent series is as follows:

$$u(\theta) = a_0\Gamma_0 + \sum_{i=1}^{\infty} a_{i1}\Gamma_1(\xi_{i1}(\theta)) + \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1} a_{i_1 i_2}\Gamma_2(\xi_{i1}(\theta),\xi_{i2}(\theta)) + \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1}\sum_{i_3=1}^{i_2} a_{i_1 i_2 i_3}\Gamma_3(\xi_{i1}(\theta),\xi_{i2}(\theta),\xi_{i3}(\theta)) + \ldots \quad (9)$$

where $\{\xi_i(\theta)\}_{i=1}^{\infty}$ is a set of Gaussian random variables, $\Gamma_p(\xi_{i1},...,\xi_{ip})$ is the generic element of a set of multidimensional Hermite polynomials, usually called homogeneous chaos of order *p*, $a_{i1},...,a_{ip}$ are deterministic constants, and θ represents an outcome in the space of possible outcomes of a random event.

Equation (9) can be written more simply as

$$u(\theta) = \sum_{i=0}^{P} b_i \Psi_i(\vec{\xi}(\theta)) \quad (10)$$

where b_i and $\Psi_i(\vec{\xi}(\theta))$ are one-to-one correspondences between the coefficients $a_{i1},.....,a_{ip}$ and the functions $\Gamma_p(\xi_{i1},.....,\xi_{ip})$, respectively. If *u* is a function of a normally distributed random variable *x*, which has the known mean μ_x and variance σ_x^2, ξ is a normalized variable: $\xi = (x-\mu_x)/\sigma_x$. For example, the two-dimensional case of Eq. (9) can be expanded as:

$$\begin{aligned} u(\theta) = {} & a_0\Gamma_0 + a_1\Gamma_1(\xi_1) + a_2\Gamma_1(\xi_2) \\ & + a_{11}\Gamma_2(\xi_1,\xi_1) + a_{12}\Gamma_2(\xi_2,\xi_1) + a_{22}\Gamma_2(\xi_2,\xi_2) \\ & + a_{111}\Gamma_3(\xi_1,\xi_1,\xi_1) + a_{211}\Gamma_3(\xi_2,\xi_1,\xi_1) + a_{221}\Gamma_3(\xi_2,\xi_2,\xi_1) + a_{222}\Gamma_3(\xi_2,\xi_2,\xi_2)\ldots \end{aligned} \quad (11)$$

Equation (11) can be recast in terms of $\Psi_i[.]$ and b_i as follows:

$$u(\theta) = b_0\Psi_0 + b_1\Psi_1 + b_2\Psi_2 + b_3\Psi_3 + b_4\Psi_4 + b_5\Psi_5 + \ldots \tag{12}$$

Thus, the term $a_{11}\Gamma_2(\xi_1, \xi_1)$ becomes $b_3\Psi_3$ for this two-dimensional case.

The general expression to obtain the multidimensional Hermite polynomials is given by (Ghanem, and Spanos, 1991)

$$\Gamma_p(\xi_{i1}, \ldots\ldots, \xi_{ip}) = (-1)^n \frac{\partial^n e^{-\frac{1}{2}\vec{\xi}^T\vec{\xi}}}{\partial\xi_{i1}, \ldots\ldots, \partial\xi_{ip}} e^{\frac{1}{2}\vec{\xi}^T\vec{\xi}} \tag{13}$$

where the vector $\vec{\xi}$ consists of *n* Gaussian random variables $(\xi_{i1}, \ldots\ldots, \xi_{in})$. Generally, the one-dimensional Hermite polynomials are defined by

$$\Psi_n(\xi) = (-1)^n \frac{\varphi^{(n)}(\xi)}{\varphi(\xi)} \tag{14}$$

where $\varphi^{(n)}(\xi)$ is the n^{th} derivative of the normal density function, $\varphi(\xi) = 1/\sqrt{2\pi}\ e^{-\xi^2/2}$. This is simply the single-variable version of Eq. (13). From Eq. (14), we can readily find

$$\{\Psi_i\} = \{1,\ \xi,\ \xi^2 - 1,\ \xi^3 - 3\xi,\ \xi^4 - 6\xi^2 + 3,\ \xi^5 - 10\xi^3 + 15\xi, \ldots\} \tag{15}$$

Thus, a second order, 2-D PCE is given by

$$u(\theta) = b_0 + b_1\xi_1(\theta) + b_2\xi_2(\theta) + b_3(\xi_1^2(\theta) - 1) + b_4\xi_1(\theta)\xi_2(\theta) + b_5(\xi_2^2(\theta) - 1) \tag{16}$$

where $\xi_1(\theta)$ and $\xi_2(\theta)$ are two independent random variables.

PCE can be used to represent the response of an uncertain system in the non-intrusive formulation (Pettit, Canfield, and Ghanem, 2002; Choi, Grandhi, Canfield, and Pettit, 2004). The basic idea of this approach is to project the response and stochastic system operator onto the stochastic space spanned by PCE with the projection coefficients, b_i, being evaluated through an efficient sampling scheme. We first define the vector x at a particular point $(\xi_1, \xi_2, \ldots, \xi_m)$ of random variables

$$x^T = [1 \;\; \Psi_1(\xi_1) \;\; \Psi_2(\xi_1) ... \Psi_p(\xi_1) \;\; \Psi_1(\xi_2) \;\; \Psi_2(\xi_2) ... \Psi_p(\xi_2) \;\; \Psi_1(\xi_m) \;\; \Psi_2(\xi_m) ... \Psi_p(\xi_m)] \tag{17}$$

where p is the order of polynomial and $\Psi_j(\xi_i)$ are PCE. The estimated response at this point is

$$y(x) = x^T \hat{\beta} \tag{18}$$

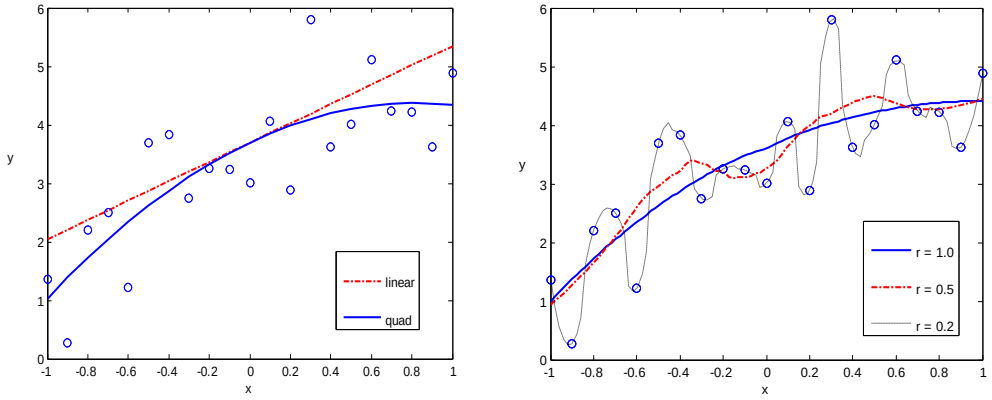
where $\hat{\beta}$ is a set of undetermined coefficients of PCE and it can be obtained from Equation (5).

2.3. SOLUTION STRATEGIES

For the utilization of the local regression method in practice, the selection of its basic components, such as the basis function, the weighting function, and the domain influence factor, *r*, is critical to provide the reliable model adequacy of the approximation. For instance, the domain influence factor has a significant effect on the fitted shape. Depending upon the size of the domain influence factor or the bandwidth, the user can adjust the closeness of fit, and this flexibility can also enable the user to achieve the same result of the interpolation and the global regression as shown in Figure 3b. Further discussions on the fixed bandwidth and nearest-neighbor bandwidth selection in the local regression are available in Refs. (Cleveland, 1979; Katkovnik, 1979). Figure 3a shows the fitted model for the same data by using the global regression method. In the case of the global regression, the data analysis is constrained to a specific global function to fit a model data. It is clear that the local regression method provides sufficient flexibility to achieve good model adequacy. However, when the size of the domain influence factor is small, the obtained response approximation can be unstable against the effects of random fluctuations, or noise phenomena. Therefore, it is important to develop criterions for the selection of the basic components of the local regression method.

In this study, the PCE is employed as a basis function. More satisfactory solutions can be expected because of the orthogonal property of the PCE. For the common polynomial regression model of the monomials of x^m, the columns of X in Eq. (2) can sometimes be nearly collinear, which causes an ill-conditioned problem, because negative values of x produce negative values for all odd powers, and positive values of x produce large positive values for all of the function. Hence, small changes in the basis function lead to relatively large changes in the regression coefficients. Another important issue with the polynomial regression is in determining an appropriate order of polynomials. By using a linear basis function (first-order polynomial) in the local regression often induces rapid changes in the slope. In the local regression method, increasing the degree of polynomials can typically enlarge the bandwidth without introducing intolerable bias; it eventually produces smoother fitting shapes compared to the linear basis (Lancaster and Salkauskas, 1981; Stone, 1977; Cleveland, 1979). In order to determine the appropriate degree of the polynomials and the size of the domain influence factor, several possible criterions, which involve R^2, C_p statistics (Montgomery, 1997), and the graphical diagnostics, can be considered. The graphical diagnostics, such as the plot of residuals ε versus $\hat{y}$, or y versus $\hat{y}$, can

provide a visual assessment of model effectiveness. The visual inspections of residuals are preferable to understand certain characteristics of the regression results because analysts can easily construct the plots and reveal useful information from the unorganized data. However, the visual inspection is a labor intensive process and it is difficult to automate. An advantage of the R^2 and C_p statistics is that the procedure can be automated. It does not require labor intensive processes. Since the automated procedure can underestimate a peak in a surface and sometimes produces a poor solution, an ideal criterion can be a cross-validation by using both the graphical diagnostics and R^2 or C_p statistics.



(a) Global Regression (b) Local Regression with r = 0.2, r = 5.0, and r = 1.0

Figure 3. Effect of Local Regression and Domain Influence Factor

Figure 4 shows the flowchart of the solution strategies for determining the appropriate parameters of the basic components in the MLS approximation. In this procedure, the utilization of the stratified sampling technique known as LHS is expected to decrease the number of simulations needed. To determine the parameters for the MLS approximation, the R^2 value has been checked along with the graphical diagnostics of the regression result as shown in Figure 4. The formula for R is defined by

$$R_{X_1,X_2} = \frac{Cov(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}} \tag{19}$$

where Cov(·) is the measure of correlation of the fluctuations of the two different quantities; namely, covariance and σ_{X_1} represents the standard deviations for X_1.

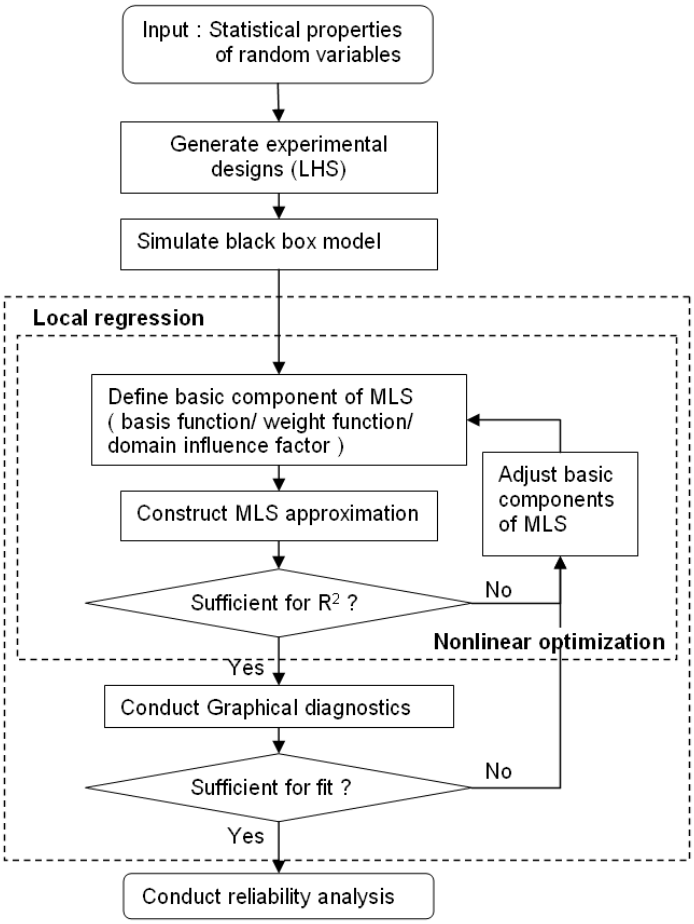


Figure 4. Solution Strategies for Local Regression

R^2 can vary from 0.0 to 1.0, where a R^2 value of 1.0 indicates the regression perfectly fits the data. R^2 is a good measure to automate the determining procedure of the basic components for the MLS approximation using nonlinear optimization. However, when R^2 is misused, the user can produce an undesirable interpolation with very high order polynomial models. The introduction of the graphical diagnostics step, such as the residual analysis, can detect this undesirable and uncontrollable result with little additional effort. For instance, the abnormality of the residual plots indicates that the selected model is inadequate or that an error exists in the analysis. There are no significant computational costs to obtain statistical properties of the responses after constructing the PCE representation of stochastic responses.

3. Structural Integrity Prediction

3.1. THREE-BAR TRUSS EXAMPLE

Reliability analysis evaluates various statistical properties and the probability of system failure by determining whether the limit-state functions are exceeded. Generally, the limit state indicates the margin of safety between the resistance and the load of structures. The limit state function, $g(\cdot)$, and probability of failure, P_f, can be defined as

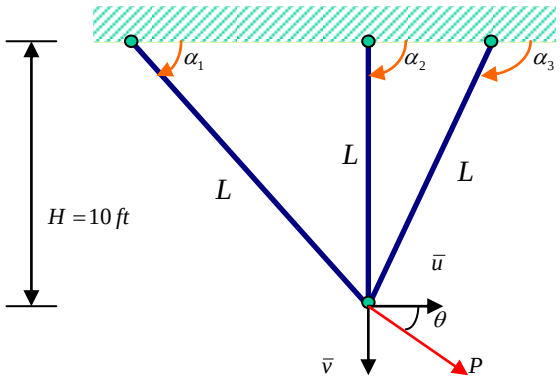
$$g(X) = R(X) - S(X)$$
$$P_f = P\ [g(\cdot) \le 0] \quad (20)$$

where R is the resistance and S is the loading of the system. Both $R(.)$ and $S(.)$ are functions of the random variable X. The notation $g(.) < 0$ denotes the failure region. Likewise, $g(.) = 0$ and $g(.) > 0$ indicate the failure surface and safe region, respectively.

In this section, the estimation of the low failure probability will be discussed by comparisons of a sampling method and the proposed method. An indeterminate, asymmetric system of a three pin-connected truss structure is illustrated in Figure 5. The unloaded length, L_m, and orientation, α_m, of each member are deterministic. Young's modulus, E_m, of each member is also assumed to be deterministic. The load has a random magnitude, P, and direction, θ. The cross-sectional area A for all members is also random. The random quantities are initially considered normally distributed and uncorrelated:

A~N(1 in^2, 0.1 in^2)
P~N(1000 *lb*, 250 *lb*)
θ~N(45^o, 7.5^o)

where the symbol $x \sim N(\mu_x, \sigma_x)$ denotes that the random variable x is treated as a normal distribution and has the mean of μ_x and standard deviation of σ_x.



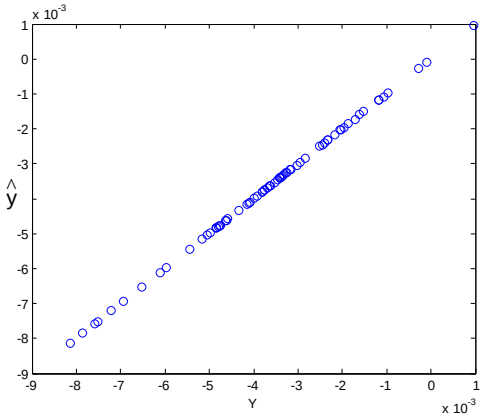
$$\alpha_1 = 45^\circ, \alpha_2 = 90^\circ, \alpha_3 = 110^\circ, E = 2.9 \times 10^6 \; psi$$

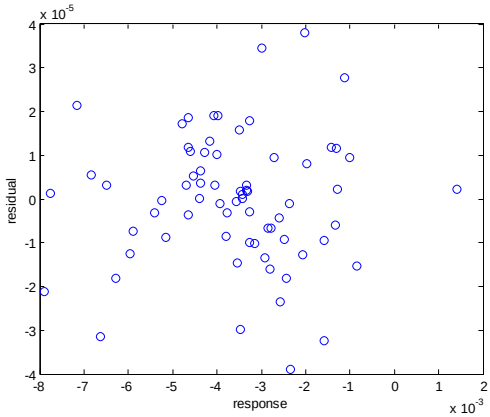
Figure 5. Pin-connected Three-bar Truss

The principle of virtual work is used to calculate the displacement vector $[\bar{u}, \bar{v}]^T$ of the joint at which the load is applied and is given by the solution of the following system of equations:

$$P\cos\theta = \sum_{m=1}^{3} (\bar{u}\cos^2\alpha_m + \bar{v}\cos\alpha_m \sin\alpha_m)\frac{E_m A_m}{L_m}$$
$$P\sin\theta = \sum_{m=1}^{3} (\bar{v}\sin^2\alpha_m + \bar{u}\cos\alpha_m \sin\alpha_m)\frac{E_m A_m}{L_m} \qquad (21)$$

The horizontal deflection of the structure should be $\bar{u}$ < 0.0015 in. This restriction is considered as a limit state. To obtain the probability of failure, P_f, one million simulations were conducted to obtain a converged result in MCS. 200 samples of LHS were used to obtain the surrogate model of the limit state by using the third-order PCE model with the exponential weight function of Eq. (8a). The plots of y versus $\hat{y}$ (Figure 6a) or residuals versus $\hat{y}$ (Figure 6b) provide a visual assessment of model effectiveness in regression analysis. Since the residual plot of Figure 6b exhibits white noise behavior which means there is no abnormality and the residual plot in Figure 6a yields points around the 45° line, the estimated regression function shows accurate predictions of the values that are actually observed. Therefore, the selected PCE and the weight function model of MLS are sufficient for fitting the given data. After conducting the local regression, P_f is calculated using one million MCS simulations with the obtained PCE model.





(a) Residual Plot: y

versus $\hat{y}$ (b) Residual Plot: $\hat{y}$ versus Residual

Figure 6. Residual Analysis

Table 1. Comparison of Methods for Reliability Analysis

	P_f	95% Confidence Interval
MCS	4.70×10^{-6}	$[3.64\times10^{-6}, 5.76\times10^{-6}]$
PCE+MLS	4.34×10^{-6}	$[4.01\times10^{-6}, 4.67\times10^{-6}]$

The corresponding results of the current example are summarized in Table 1. The PCE result converged to $P_f = 4.34\times10^{-6}$ and 95% confidence interval is also obtained. The confidence interval indicates a range of values that likely contains the analysis results. For this case, the user can be 95% confident that the true mean of P_f will be between 4.01×10^{-6} and 4.67×10^{-6}. The confidence interval of MCS is larger than the result of PCE, but it has an overlapping region with the PCE's. The interval can be reduces as the sampling size increases in the case of MCS. The obtained result exhibits that the use of PCE along with MLS is applicable to the estimation of the low failure probability.

4. Summary

A new framework is proposed for the accurate estimation of the low failure probability of common engineering problems by utilizing efficient probabilistic methods which can realistically model complicated statistical variations. A local regression method, MLS, is integrated to a previously developed probabilistic decision support framework which combines the PCE and LHS. The stochastic modeling process repeats and recalibrates the PCE model with the local regression scheme until sufficient model adequacies are achieved. This allows for an accurate estimation of the low probability of failure with limited sampling sets. This increased capability has the potential to provide significant robust designs with a minimal amount of computational cost.

References

Choi, S., Canfield, R.A, and Grandhi, R.V., "Estimation of Structural Reliability for Gaussian Random Fields," *Structure & Infrastructure Engineering*, Vol. 2, No. 3-4., pp. 161-173, 2006.

Choi, S., Grandhi, R.V., and Canfield, R.A., "Robust Design of Mechanical Systems via Stochastic Expansion," International Journal of Materials and Product Technology, Vol. 25, pp. 127-143, 2006.

Choi, S., Grandhi, R.V., and Canfield, R.A., "Structural Reliability under non-Gaussian Stochastic Behavior," Computers and Structures, Vol. 82, pp.1113-1121, 2004.

Choi, S., Grandhi, R.V., and Canfield, R.A., *Reliability-Based Structural Design*, Springer, London, 2006.

Choi, S., Grandhi, R.V., Canfield, R.A. and Pettit, C.L., "Polynomial Chaos Expansion with Latin Hypercube Sampling for Estimating Response Variability," *AIAA Journal*, Vol. 42, No. 6, pp. 1191-1198, 2004.

Cleveland, W., "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, Vol. 74, pp. 829-836, 1979.

Cureton, E., and D'Agostinao, R., *Factor Analysis: An Applied Approach*, Lawrence Erlbaum Associates, NJ, 1983.

Dolbow J., and Belytschko T., "An Introduction to Programming the Meshless Element Free Galerkin Method," *Archives of Computational Methods in Engineering*, Vol. 5, No. 3, pp. 207-241, 1998.

Ghanem, R., and Spanos, P.D., *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag, NY, 1991.

Katkovnik,V., "Linear and Nonlinear Methods of Nonparametric Regression Analysis," *Soviet Automatic Control*, Vol. 5, pp. 25-34, 1979.

Lancaster, P. and Salkauskas, K., "Surfaces Generated by Moving Least-Squares Methods," *Math. Comput.*, Vol. 37, pp. 141–158, 1981.

Maluf, N., "An Introduction to Microelectromechanical Systems Engineering," *Artech House*, Boston, 2004.

Montgomery, D.C., *Design and Analysis of Experiments*, Wiley, New York, 1997.

Pettit, C.L., Canfield, R.A., and Ghanem, R., "Stochastic Analysis of an Aeroelastic System," *15th ASCE Engineering Mechanics Conference*, Columbia University, New York, 2002.

Sharpe, W. N., Turner, K.T., and Edwards. R.L., "Tensile Testing of Polysilicon," *Experimental Mechanics*, Vol. 39, No. 3, pp. 162-170, 1999.

Stone, C.J., "Consistent Nonparametric Regression," *The Annals of Statistics*, Vol. 5, pp. 595-645, 1977.

Toropov, V., Scharamm, U., Sahai, A., Jones, R., and Zeguer, T., "Design Optimization and Stochastic Analysis based on the Moving Least Squares Method," *6th World Congresses of Structural and Multidisciplinary Optimization*, Brazil, 2005.

Watanabe, S., "Karhunen-Loève Expansion and Factor Analysis," *In Transactions of the 4th Prague Conference on Information Theory*, Prague, pp. 635-660, 1965.

Wiener, N., "The Homogeneous Chaos," *American Journal of Mathematics*, Vol. 60, pp.897-936, 1938.

Comparison of Interval and Convex Analyses

Isaac Elishakoff[1], Xiaojun Wang[2], Zhiping Qiu[2]
[1]*Florida Atlantic University, U. S. A*
email: elishako@fau.edu
[2]*Beijing University of Aeronautics and Astronautics, China*
email: xjwang@buaa.edu.cn and zpqiu@buaa.edu.cn

Abstract: This study shows that the type of the analytical treatment that should be adopted for non-probabilistic analysis of uncertainty depends upon the available experimental data. The main idea is based on the consideration that the maximum structural response predicted by the preferred theory ought be minimal, and the minimum structural response predicted by the preferred theory ought be maximal, to constitute a lower overestimation. Prior to the analysis the existing data ought be enclosed by the minimum volume hyper-rectangle V_1 that contains all experimental data. The experimental data also have to be enclosed by the minimum volume ellipsoid V_2. If V_1 is smaller than V_2 and the response calculated based on it $R(V_1)$ is smaller than $R(V_2)$, then one has to prefer interval analysis. However, if V_1 is in excess of V_2 and $R(V_1)$ is greater than $R(V_2)$, then the analyst ought to utilize convex modeling. If V_1 equals V_2 or these two quantities are in close vicinity, then two approaches can be utilized with nearly equal validity. Some numerical examples are given to illustrate the efficacy of the proposed methodology.

Keywords: uncertainty description, convex modeling, interval analysis, ellipsoid, hyper-rectangle

1. Introduction

Probabilistic approaches are used by numerous analysts for the safety assessment of structures whose parameters or loadings on them are modeled as uncertain variables or functions. In recent decades, some alternatives of it have been suggested. Fuzzy-sets based approaches gain much popularity. There are many discussions on philosophical implications of each of these approaches. Whereas the probabilistic methodology requires the knowledge of probability densities, the fuzzy-sets based approaches demand the knowledge of membership functions. More recently, yet another alternative is embraced by the investigators, that is not based upon any specified measure, either probabilistic or fuzzy, of uncertain variables. It presupposes the knowledge only of bounds of uncertain quantities. These are then called as unknown-but-bounded or uncertain-but-bounded variables. This analysis is both old and new. It is old chronologically but new by its revived use. Apparently the first work on response of a single-degree-of-freedom system under uncertain-but-bounded excitation was written by Bulgakov in 1946. He specially mentioned that the task is to calculate the upper bounds of structural response "under unfavorable circumstances", when the "disturbing action $y_p(t)(p=1,2,\ldots,r)$ satisfy the condition $|y_p(t)|\le l_p(l_p$constant)

but are otherwise arbitrary one-valued continuous functions of the time *t* possessing as many derivatives as necessary ". This problem was dubbed by Bulgakov (1946) as the "problem of accumulation of disturbances" (see also his other paper, in 1940, which considers a special case).

There is a considerable literature in the Russian language on the Bulgakov's problem. Independently, in late sixties, Schweppe (1968) developed an analogous thinking based on ellipsoidal modeling, representing the uncertain variables as belonging to an ellipsoid.

Recently, some researchers in uncertain mechanics are developing interval analysis whereas others follow convex modeling (Ben-Haim and Elishakoff, 1990; Rao and Berke, 1997; Lombardi, 1998; Pantelides and Ganzerli, 1998, 1999; Mullen and Muhanna, 1999; Manson, 2005; McWilliam, 2001; Moens and Vandepitte, 2007). The question arises if these analyses are interrelated specifically, should one perform both analyses, or one of them in preferable? This work tries to elucidate the possible reply to this question. Some researchers performed a comparison of results derived by both methods. Elishakoff, Li and Starnes (2001) derived a minimum volume ellipsoid that encloses the minimum volume parallelepiped for buckling analysis. Elishakoff, Cai and Starnes (1994) studied the buckling of elastic column on non-linear elastic foundation by interval analysis whereas Qiu, Ma and Wang (2006) dealt with the same problem via convex modeling. Qiu and Wang (2003) specially distinguished between these two non-probabilistic set theoretical models.

Although convex modeling and interval analysis have been used extensively, in practice, which of the non-probabilistic uncertain descriptions, convex modeling or interval analysis should be preferred? In this study, this problem will be answered. The experimental data are shown to be of the cardinal influence on which of these methods ought be given a preference.

Consider the case that due to high cost of the measurements the experimental points are too scant to determine their statistical information on uncertain parameters: if we choose non-probabilistic set-theoretical convex methods, convex modeling or interval analysis, for uncertain modeling, then the precondition is to seek or determine the suitable set containing the limited experimental points. In fact, there is more than one set to be able to enclose the limited experimental points. However, too big set will produce over-conservative bounds on the structural responses. Of course, it is impossible for us to know the real bounds on uncertain parameters based on the limited experimental points. The enclosing set with minimal volume property may be a better selection, which will produce lower overestimation on the bounds of the structural responses. We can only act on what we know.

2. Description of the Method by Zhu, Elishakoff and Starnes

In this section, the description of the method by Zhu, Elishakoff and Starnes (1996), in which the smallest hyper-rectangle and the smallest ellipsoid containing the given experimental data are determined, is stated in brief.

Suppose that there are *m* uncertain parameters $a_i (i = 1, 2, \cdots, m)$ describing either the structural properties or the excitation. These parameters constitute an *m*-dimensional parameter space, namely,

$a=(a_1,a_2,\cdots,a_m)$. Suppose that we have limited information on these parameters, represented by M experimental points, $a^{(r)}(r=1,2,\cdots,M)$ in this m-dimensional space. Convex modeling assumes that all these experimental points belong to an ellipsoid

$$(a-a_0)^T W(a-a_0)\le 1 \tag{1}$$

where a_0 is the state vector of the central point of the ellipsoid, and W is the weight matrix. Interval analysis assumes that all experimental points belong to a hyper-rectangle.

By using transformation matrix $T_m(\theta_1,\theta_2,\cdots,\theta_{m-1})$ given in Ref. Zhu et al.(1996), the above M points in the rotated coordinate system will have their new coordinates denoted by $b^{(r)}(r=1,2,\cdots,M)$. To obtain the smallest ellipsoid, let us first examine an m-dimensional box of the form

$$\left|b-b_0\right|\le d \tag{2}$$

which contains all M points. The vector of semi-axes $d=(d_1,d_2,\cdots,d_m)^T$ and the vector of central points $b_0=(b_{10},b_{20},\cdots,b_{m0})^T$ of the "box" in the rotated coordinate system are given by

$$\begin{aligned} d_k &= \frac{1}{2}\left(\max_r(b_k^{(r)})-\min_r(b_k^{(r)})\right), \\ b_{k0} &= \frac{1}{2}\left(\max_r(b_k^{(r)})+\min_r(b_k^{(r)})\right), \end{aligned} \qquad (r=1,2,\cdots,M;\ k=1,2,\cdots,m) \tag{3}$$

We now enclose this box by an ellipsoid

$$\sum_{k=1}^{m}\frac{(b_k-b_{k0})^2}{g_k^2}\le 1 \tag{4}$$

where g_k are the semi-axes of the ellipsoid. There are infinite number of ellipsoids which contain the box given in Eq.(2). Clearly, the best choice is the one with minimum volume. The volume of an m-dimensional ellipsoid is given by

$$V_e=C_m\prod_{k=1}^{m}g_k \tag{5}$$

where C_m is a constant.

From the monograph by Elishakoff, Li and Starnes (2001) and paper by Qiu (2003), corresponding to the smallest ellipsoid, the semi-axes of the smallest ellipsoid should be

$$g_i=\sqrt{m}d_i,\quad (i=1,2,\cdots,m) \tag{6}$$

Thus, once the size of the box Eq.(2) is known, the semi-axes of the minimum-volume ellipsoid enclosing the box of the experimental data are readily determined by utilizing Eq.(6). If there are no experimental points at the corner of the box, the size of such an ellipsoid may further be reduced until one

of the experimental points reaches the surface of the ellipsoid. The semi-axes of the ellipsoid in this case may be replaced by ηg_k, where the factor is determined from the condition

$$\eta = \sqrt{\max_r \sum_{k=1}^{m} \frac{\left(b_k^{(r)} - b_{k0}\right)^2}{g_k^2}} \le 1, \quad (r = 1, 2, \cdots, M) \tag{7}$$

If there are some experimental points in the corner of the multidimensional box, the factor η equals unity. The ellipsoid (4) can be written in the form

$$(b - b_0)^T D (b - b_0) \le 1 \tag{8}$$

in which b_0 is the vector of central points whose components are given by Eq.(3), and D is a diagonal matrix

$$D = diag\left((\eta g_1)^{-2}, (\eta g_2)^{-2}, \cdots, (\eta g_m)^{-2}\right) \tag{9}$$

The volume of the ellipsoid now reads

$$V_e = C_m \eta^m \prod_{k=1}^{m} g_k \tag{10}$$

which is a function of a set of parameters $\theta_k (k = 1, 2, \cdots, m-1)$. Therefore, the best ellipsoid among these ellipsoids is the one which contains all given points and possesses the minimum volume, i.e.,

$$V_e = \min_{\theta_1, \theta_2, \cdots, \theta_{m-1}} \left\{V_e(\theta_1, \theta_2, \cdots, \theta_{m-1})\right\} \tag{11}$$

A possible approach to determine this ellipsoid is to search among all possible cases by increasing $\theta_k (k = 1, 2, \cdots, m-1)$ from 0 to π/2 in sufficiently small increments $\Delta\theta_k$, and to compare the volumes of so obtained ellipsoids. Once one finds the ellipsoid with minimum volume in one direction, say $\theta_{k0} (k = 1, 2, \cdots, m-1)$, the ellipsoid can be transformed back into the original coordinate system by applying the transformation matrix T_m. Hence, the vector a_0 of central point and the weight matrix W in Eq.(1) become

$$a_0 = T_m^T b_0, \quad W = T_m^T D T_m \tag{12}$$

where $T_m = T_m(\theta_{10}, \theta_{20}, \cdots, \theta_{m0})$. So Eq.(12) constitutes the smallest ellipsoid containing all experimental points. The "box" corresponding to the smallest ellipsoid is the smallest hyper-rectangle.

3. Convex Modeling and Interval Analysis for the Structural Response

For convenience, in this section, convex modeling method and interval analysis method for the static response analysis of structures with uncertain parameters are reformulated (see Ref. Qiu (2003)). In fact, the presented concept in this study also can be applied to other linear elastic structural mechanics problem with uncertainty, such as the natural frequency analysis, the dynamic response analysis etc.

The matrix equation of static equilibrium in the finite element method can be written as

$$K(a)u(a) = f(a) \tag{13}$$

where $K = (k_{ij})$ is the $n \times n$ -dimensional stiffness matrix, $u = (u_i)$ is the *n*-dimensional nodal displacement vector and $f = (f_i)$ is the *n*-dimensional external load vector; $a = (a_1, a_2, \cdots, a_m)^T$ is the structural parameters, such as the physical, material and geometric properties in structures.

Consider a realistic situation in which not enough information is available on the structural parameters to justify an assumption on their probabilistic characteristics. It is assumed that by use of Zhu, Elishakoff and Starnes's method (1996), the derived smallest ellipsoid and the derived smallest hyper-rectangle on the structural parameters can be obtained as, respectively,

$$Z(W, \theta) = \{a : a \in R^m, (a - a_0)^T W (a - a_0) \le \theta^2\} \tag{14}$$

and

$$\underline{a} \le a \le \overline{a} \ \text{ or } \ a_0 - \Delta a \le a \le a_0 + \Delta a \tag{15}$$

where $a_0 = (a_{i0}) \in R^m$ is the nominal value vector of the structural parameter vector a, W is a positive definite matrix and is called the weight matrix, θ is a positive constant and is called the radius of the ellipsoid; $\underline{a}$ and $\overline{a}$ are the lower bound and upper bound of the hyper-rectangle, Δa is the radius of the hyper-rectangle.

The structural parameter of a value slightly different from this nominal value can be denoted as

$$a = a_0 + \delta a \ \text{ or } \ a_i = a_{i0} + \delta a_i \quad , \quad i = 1, 2, \cdots, m \tag{16}$$

where $\delta a = (\delta a_i) \in R^m$ is a small quantity.

By Taylor's series expansion, the static displacement of the structure with uncertain parameter vector $a = a_0 + \delta a$, to first order in δa, is

$$u_i(a) = u_i(a_0 + \delta a) = u_i(a_0) + \sum_{j=1}^{m} \frac{\partial u_i(a_0)}{\partial a_j} \delta a_j, \quad i = 1, 2, \cdots, n \tag{17}$$

For convenience of notation, let us define

$$\varphi^T = \left(\frac{\partial u_i(a_0)}{\partial a_1}, \frac{\partial u_i(a_0)}{\partial a_2}, \cdots, \frac{\partial u_i(a_0)}{\partial a_m} \right) = \left(\frac{\partial u_{i0}}{\partial a_1}, \frac{\partial u_{i0}}{\partial a_2}, \cdots, \frac{\partial u_{i0}}{\partial a_m} \right) \tag{18}$$

By combination of Eq.(17) and Eq.(14), the most and least favourable response for convex modeling method can be obtained as (see Ref. Ben-Haim and Elishakoff (1990))

$$\underline{u}_C = u_0 - \theta \sqrt{\varphi^T W^{-1} \varphi} \ \text{ and } \ \overline{u}_C = u_0 + \theta \sqrt{\varphi^T W^{-1} \varphi} \tag{19}$$

By combination of Eq.(17) and Eq.(15), the most and least favourable responses for interval analysis method can be obtained as (see Ref. Qiu (2003))

$$\underline{u}_{iI} = u_{i0} - \sum_{j=1}^{m} \left| \frac{\partial u_{i0}}{\partial a_j} \right| \Delta a_j \quad \text{and} \quad \overline{u}_{iI} = u_{i0} + \sum_{j=1}^{m} \left| \frac{\partial u_{i0}}{\partial a_j} \right| \Delta a_j \tag{20}$$

Thus, in the case that the smallest intervals or hyper-rectangle containing uncertain parameters are known, interval analysis method can be adopted to obtain the most and least favorable responses. In the case that the smallest ellipsoid containing uncertain parameters are known, convex modeling method can be adopted to obtain the most and least favorable responses.

So, a question will arise. Which method is better? In other words, which method will give the tighter bounds on the structural responses? In the following, a 7-bar planar truss structure and a 60-bar space truss structure are used to reply to this quest.

4. Seven-Bar Planar Truss Structure

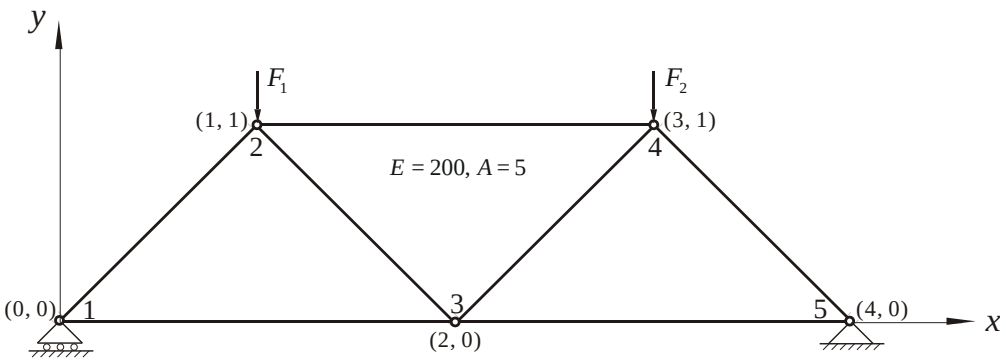


Figure 1. A 7-bar planar truss structure

Let us consider a 7-bar planar truss structure with linear elastic properties depicted in Figure 1. Here, $A = 5$ is the cross-sectional area, $E = 200$ is Young's modulus, F_1 is an external load at node No.2, F_2 is an external load applied at node No.4. The parameters of the truss are given as dimensionless numbers, since the physical values are not relevant to our analysis.

This truss is the same as adopted by Skalna (2003) but here the loads F_1 and F_2 are considered to be uncertain, and the other properties of the truss, such as A and E, are deterministic. Namely, the truss members have deterministic stiffness.

In the following, several sets of hypothesized data for uncertain parameters will be given. By use of the Zhu, Elishakoff and Starnes's method (1996), the smallest ellipse and rectangle can be derived. Based on the derived ellipse and rectangle, the most and least favorable responses of the structure can be calculated by convex modeling method and interval analysis method, respectively.

We will discuss this problem in the following two cases: one is that the principal axes of the derived ellipse and rectangle are parallel to the global coordinate system; the other is that the principal axes of the derived ellipse and rectangle are *not* parallel to the global coordinate system.

4.1. THE PRINCIPAL AXES OF THE DERIVED ELLIPSE AND RECTANGLE ARE PARALLEL TO THE GLOBAL COORDINATE SYSTEM

Case I: Consider a set of hypothesized data for uncertain parameters as shown in Figure 2, and they are listed in Table 1. Here these hypothesized data are randomly generated in order to proceed to the numerical simulations, but in practice the samples for uncertain parameters can be generally obtained by the experiments.

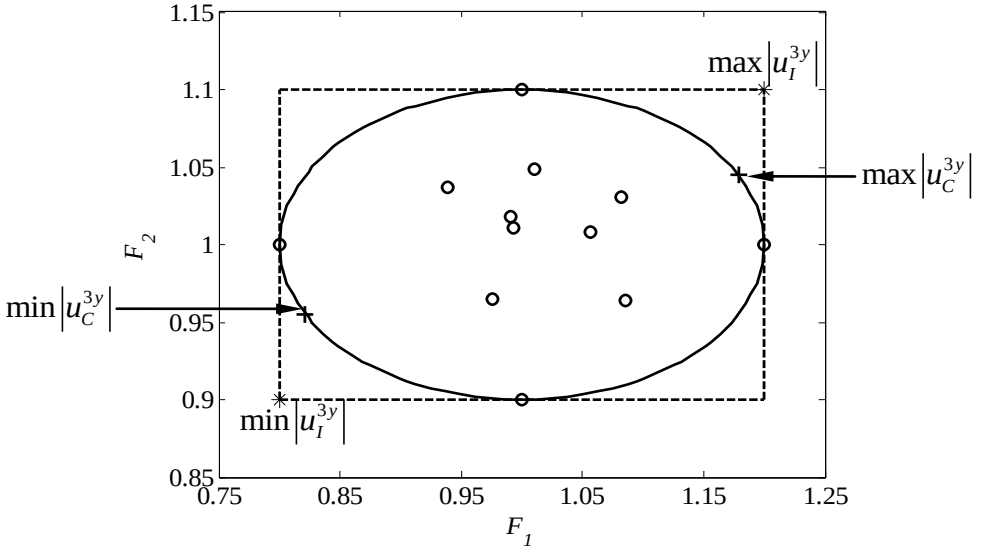


Figure 2. Rectangle and ellipse containing the data on uncertain parameters F_1 and F_2

The smallest rectangle obtained from the set of data by using of Zhu, Elishakoff and Starnes's method (1996) is

$$F_1^I = [0.80, 1.20], \quad F_2^I = [0.90, 1.10] \tag{21}$$

Based on Eq.(21), we conclude that the central values of F_1 and F_2 are, respectively,

$$F_{1c} = (0.80 + 1.20)/2 = 1.0, \quad F_{2c} = (0.90 + 1.10)/2 = 1.0 \tag{22}$$

and the values of radii F_1 and F_2 are, respectively,

$$\Delta F_1 = (1.20 - 0.80)/2 = 0.2,\ \ \Delta F_2 = (1.10 - 0.90)/2 = 0.1 \tag{23}$$

Thus, one can analyze the system as subjected to an interval load vector with nominal values $(1.0, 1.0)$ and scatter of $(20\%, 10\%)$.

Table 1. The values of uncertain parameter F_1 and F_2

k	1	2	3	4	5	6	7	8	9	10	11	12
F1	0.991	1.082	1.085	0.938	0.976	0.993	1.011	1.056	0.800	1.200	1.000	1.000
F2	1.018	1.031	0.964	1.037	0.965	1.011	1.048	1.008	1.000	1.000	0.900	1.100

On the other hand, the smallest ellipse can be obtained from the set of data by using of Zhu, Elishakoff and Starnes's method (1996). The optimal rotation angle θ_{10} obtained is 0°, so the transformation matrix T_2 is

$$T_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{24}$$

In the case of $\theta_{10} = 0^\circ$, the vector of semi-axes and the vector of central point of the "box" in the optimal rotated coordinate system are, respectively, $d = (d_1, d_2)^T = (0.2, 0.1)^T$ and $b_0 = (b_{10}, b_{20})^T = (1.0, 1.0)^T$. The semi-axes of the smallest ellipsoid are $g_1 = \sqrt{2} d_1 = 0.2828$ and $g_2 = \sqrt{2} d_2 = 0.1414$. The diagonal matrix D is

$$D = diag\left((\eta g_1)^{-2}, (\eta g_2)^{-2}\right) = diag\left(25, 100\right) \tag{25}$$

where $\eta = \sqrt{2}/2$. Thus, we can get

$$a_0 = T_2^T b_0 = (1.0, 1.0)^T, \quad W = T_2^T D T_2 = \begin{bmatrix} 25 & 0 \\ 0 & 100 \end{bmatrix} \tag{26}$$

It can be seen from Figure 2 that the derived rectangle contains the derived ellipse based on the hypothesized data listed in Table 1.

We can find that the higher-order derivatives of static responses of the 7-bar planar truss structure with respect to uncertain parameters are all zeros. Thus, Eq.(17) based on the first-order Taylor series for this example will be linear and exact, i.e.

$$\begin{aligned} u_i(F_1, F_2) &= u_i(F_{1c} + \delta F_1, F_{2c} + \delta F_2) \\ &= u_i(F_{1c}, F_{2c}) + \frac{\partial u_i(F_c)}{\partial F_1}\delta F_1 + \frac{\partial u_i(F_c)}{\partial F_2}\delta F_2, \quad i = 1, 2, \cdots, n \end{aligned} \tag{27}$$

This is the reason why only the external loads are taken as the uncertain parameters in this study.

Taking the derivative of both sides of Eq.(13) yields

$$\frac{\partial K}{\partial F_j} u + K \frac{\partial u}{\partial F_j} = \frac{\partial f}{\partial F_j}, \quad j = 1, 2 \tag{28}$$

Due to the vanishing of $\frac{\partial K}{\partial F_j}$ for this problem, the sensitivity derivative of the structural response with respect to uncertain parameters becomes

$$\frac{\partial u}{\partial F_j} = K^{-1} \frac{\partial f}{\partial F_j}, \; j = 1, 2 \tag{29}$$

Substitution of Eqs.(22), (23) and (29) into Eq.(20) yields the most and least favorable responses in y-direction of node 3 of the 7-bar planar truss structure obtained from interval analysis method as follows

$$\min \left| u_I^{3y} \right| = 0.005803, \quad \max \left| u_I^{3y} \right| = 0.007852 \tag{30}$$

Substitution of Eqs.(26) and (29) into Eq.(19) provides us with the most and least favorable responses in *y*-direction of node 3 of the 7-bar planar truss structure obtained from convex modeling method as follows

$$\min \left| u_C^{3y} \right| = 0.006064, \quad \max \left| u_C^{3y} \right| = 0.007591 \tag{31}$$

The "∗" points on the derived rectangle in Figure 2 are the most and least favorable points for interval analysis method. The "+" points on the derived ellipse in Figure 2 are the most and least favorable points for convex modeling method. The two markers "∗" and "+" have the same meaning in sequel figures.

Thus, it can be seen from Eqs.(30) and (31) that interval analysis method gives tighter bounds of responses than convex modeling method in the case of data points listed in Table 1.

Case II: Consider another set of hypothesized data for uncertain parameters as shown in Figure 3, and they are listed in Table 2.

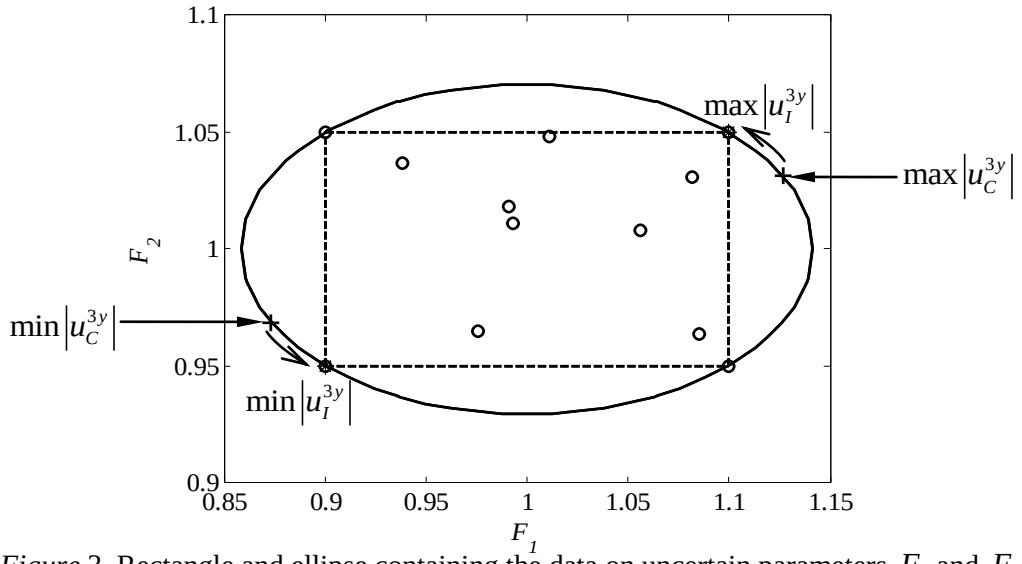


Figure 3. Rectangle and ellipse containing the data on uncertain parameters F_1 and F_2

Table 2. The values of uncertain parameter F_1 and F_2

k	1	2	3	4	5	6	7	8	9	10	11	12
F_1	0.991	1.082	1.085	0.938	0.976	0.993	1.011	1.056	0.900	1.100	1.100	0.900
F_2	1.018	1.031	0.964	1.037	0.965	1.011	1.048	1.008	0.950	0.950	1.050	1.050

The smallest rectangle obtained from the set of data by using of Zhu, Elishakoff and Starnes's method (1996) is

$$F_1^I = [0.90, 1.10], \quad F_2^I = [0.95, 1.05] \tag{32}$$

Based on Eq.(32), we conclude that the central values and the values of radii of F_1 and F_2 are, respectively,

$$F_{1c} = 1.0, \quad F_{2c} = 1.0 \text{ and } \Delta F_1 = 0.1, \quad \Delta F_2 = 0.05 \tag{33}$$

Thus, one can analyze the system as subjected to an interval load vector with nominal values $(1, 1)$ and scatter of $(10\%, 5\%)$.

On the other hand, the smallest ellipse can be obtained from the set of data by using of Zhu, Elishakoff and Starnes's method (1996). The optimal rotation angle θ_{10} obtained is 0°. Similar to Eqs.(24)~(26), the vector a_0 of central point and the weight matrix W can be obtained as

$$a_0 = T_2^T b_0 = (1.0, 1.0)^T, \quad W = T_2^T D T_2 = \begin{bmatrix} 50 & 0 \\ 0 & 200 \end{bmatrix} \tag{34}$$

It can be seen from Figure 3 that the derived ellipse contains the derived rectangle based on the hypothesized data listed in Table 2.

By substituting Eqs.(33) and (29) into Eq.(20) and substituting Eqs.(34) and (29) into Eq.(19), the most and least favorable responses in y-direction of node 3 of the 7-bar planar truss structure can be, respectively, obtained from interval analysis method and convex modeling method as follows

$$\min \left| u_I^{3y} \right| = 0.006316, \quad \max \left| u_I^{3y} \right| = 0.007340 \tag{35}$$

and

$$\min \left| u_C^{3y} \right| = 0.006288, \quad \max \left| u_C^{3y} \right| = 0.007367 \tag{36}$$

Thus, it can be seen from Eqs.(35) and (36) that convex modeling method gives tighter bounds of responses than interval analysis method in the case of data points listed in Table 2.

Under this circumstance, an interesting phenomenon can be seen. For convex modeling method, the extreme value points on the ellipse in Figure 3 may be different based on different structural parameters. Namely, the locations of the extreme value points of convex modeling method will change by changing the structural parameters. In certain particular case, the extreme value points of convex modeling method and interval analysis method will coincide.

4.2. THE PRINCIPAL AXES OF THE DERIVED ELLIPSE AND RECTANGLE ARE NOT PARALLEL TO THE GLOBAL COORDINATE SYSTEM.

Case I: Consider a set of hypothesized data for uncertain parameters as shown in Figure 4, and they are listed in Table 3.

The smallest rectangle obtained from the set of data by using of Zhu, Elishakoff and Starnes's method (1996) is shown as Figure 4. The smallest ellipse can be obtained from the set of data by using of Zhu, Elishakoff and Starnes's method (1996). The optimal rotation angle θ_{10} obtained is 30°. Similarly, the vector a_0 of central point and the weight matrix W can be obtained as

$$a_0 = T_2^T b_0 = (0.366, 1.366)^T, \quad W = T_2^T D T_2 = \begin{bmatrix} 43.75 & -32.48 \\ -32.48 & 81.25 \end{bmatrix} \tag{37}$$

As above mentioned, Eq.(17) based on the first-order Taylor series will be exact and linear for this example. Due to the convexity of the derived smallest rectangle, the most and least favorable responses in *y*-direction of node 3 of the 7-bar planar truss structure for interval analysis method will reach on the four vertexes of the smallest rectangle. By calculating and comparing the four responses, the most and least favorable responses or the minimum and maximum values of them are, respectively,

$$\min \left| u_I^{3y} \right| = 0.004855, \quad \max \left| u_I^{3y} \right| = 0.006970 \tag{38}$$

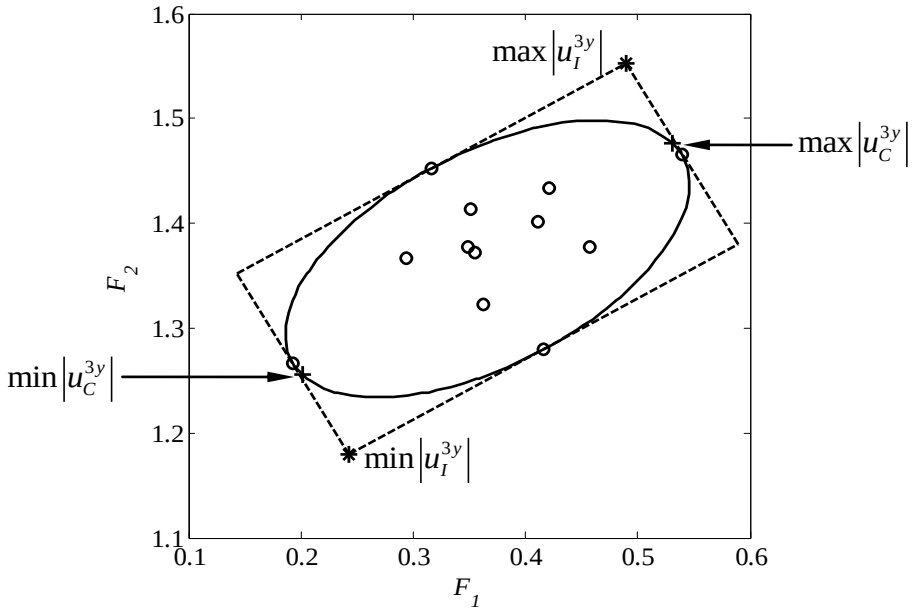


Figure 4. Rectangle and ellipse containing the data on uncertain parameters F_1 and F_2

Table 3. The values of uncertain parameter F_1 and F_2

k	1	2	3	4	5	6	7	8	9	10	11	12
F_1	0.349	0.422	0.458	0.294	0.362	0.355	0.351	0.411	0.193	0.539	0.416	0.316
F_2	1.377	1.434	1.377	1.367	1.323	1.372	1.413	1.401	1.266	1.466	1.279	1.453

By substituting of Eqs.(37) and (29) into Eq.(19), we obtain the most and least favorable responses in *y*-direction of node 3 of the 7-bar planar truss structure obtained from convex modeling method as follows

$$\min\left|u_C^{3y}\right| = 0.004972, \quad \max\left|u_C^{3y}\right| = 0.006854 \tag{39}$$

Thus, it can be seen from Eqs.(38) and (39) that interval analysis method gives tighter bounds of responses than convex modeling method in the case of data points listed in Table 3.

Case II: Consider another set of hypothesized data for uncertain parameters as shown in Figure 5, and they are listed in Table 4.

The smallest rectangle obtained from the set of data by using of Zhu, Elishakoff and Starnes's method (1996) is shown as Figure 5. The smallest ellipse can be obtained from the set of data by using of Zhu,

Elishakoff and Starnes's method (1996). The optimal rotation angle θ_{10} obtained is 30°. Similarly, the vector a_0 of central point and the weight matrix W can be obtained as

$$a_0 = T_2^T b_0 = (0.366, 1.366)^T, \quad W = T_2^T D T_2 = \begin{bmatrix} 87.50 & -64.95 \\ -64.95 & 162.50 \end{bmatrix} \tag{40}$$

In perfect analogy with Eq.(38), the most and least favorable responses in *y*-direction of node 3 of the 7-bar planar truss structure for interval analysis method can be obtained as follows

$$\min\left|u_I^{3y}\right| = 0.005384, \ \max\left|u_I^{3y}\right| = 0.006441 \tag{41}$$

We substitute of Eqs.(40) and (29) into Eq.(19) to get the most and least favorable responses in *y*-direction of node 3 of the 7-bar planar truss structure obtained from convex modeling method as follows

$$\min\left|u_C^{3y}\right| = 0.005247, \ \max\left|u_C^{3y}\right| = 0.006578 \tag{42}$$

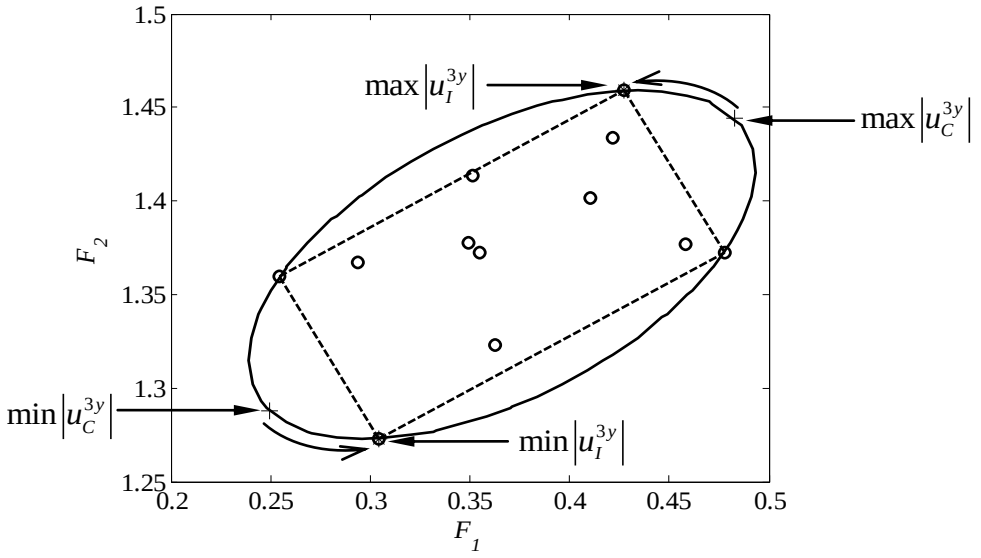


Figure 5. Rectangle and ellipse containing the data on uncertain parameters F_1 and F_2

Table 4. The values of uncertain parameter F_1 and F_2

k	1	2	3	4	5	6	7	8	9	10	11	12
F_1	0.349	0.422	0.458	0.294	0.362	0.355	0.351	0.411	0.304	0.478	0.428	0.254
F_2	1.377	1.434	1.377	1.367	1.323	1.372	1.413	1.401	1.273	1.373	1.459	1.359

Thus, it can be seen from Eqs.(41) and (42) that convex modeling method gives tighter bounds of responses than interval analysis method in the case of data points listed in Table 4. Although only the displacement responses in *y*-direction of node 3 of the 7-bar planar truss structure are compared, the analysis will not change qualitatively if a different aspect of response of the truss structure were used to carry out the comparisons of convex modeling with interval analysis due to the linear elastic properties.

We can find from the above analysis that the choose for two methods, convex modeling or interval analysis, is decided by the distribution of sample data points on uncertain parameters.

5. Sixty-Bar Space Truss Structure

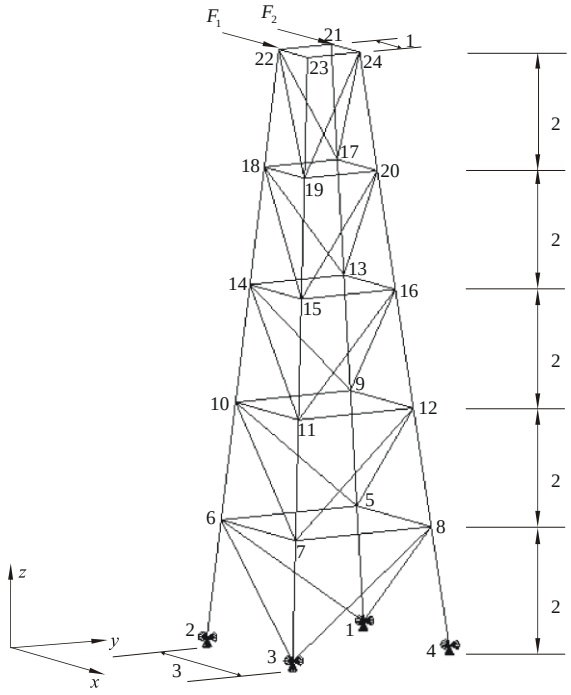


Figure 6. A 60-bar space truss structure

Consider a 60-bar space truss structure with linear elastic properties subject to two *x*-directional loads as shown in Figure 6. The external loads F_1 and F_2, respectively, act on nodes No.21 and No.22. Young's moduli of the bars are $E_i = 2.1\times10^{11}$ $(i = 1, 2\cdots, 60)$. The cross-sectional areas of the bars are $A_i = 1.0\times10^{-3}$ $(i = 1, 2\cdots, 60)$.

Suppose that the external loads F_1 and F_2 are still considered to be uncertain, and the other properties of the truss, such as A and E, are deterministic. Namely, the truss members have deterministic stiffness.

In previous section, the case that there exists the inclusion relation between the derived ellipse and rectangle is studied. In this section, we will consider the non-inclusion relation between them.

Case I: Consider a set of hypothesized data for uncertain parameters as shown in Figure 7, and they are listed in Table 5.

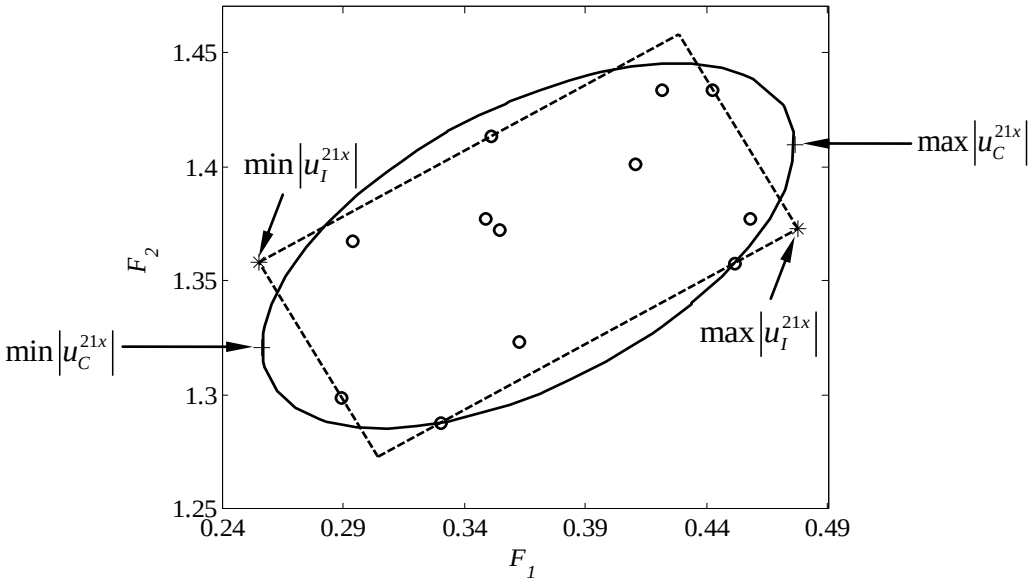


Figure 7. Rectangle and ellipse containing the data on uncertain parameters F_1 and F_2

Table 5. The values of uncertain parameter F_1 and F_2

k	1	2	3	4	5	6	7	8	9	10	11	12
F_1	0.349	0.422	0.458	0.294	0.362	0.355	0.351	0.411	0.330	0.452	0.443	0.289
F_2	1.377	1.434	1.377	1.367	1.323	1.372	1.413	1.401	1.288	1.358	1.433	1.299

The smallest rectangle obtained from the set of data by using of Zhu, Elishakoff and Starnes's method (1996) is shown as Figure 7. The smallest ellipse can be obtained from the set of data by using of Zhu, Elishakoff and Starnes's method (1996). The optimal rotation angle θ_{10} obtained is 30°. Similarly, the vector a_0 of central point and the weight matrix W can be obtained as

$$a_0 = T_2^T b_0 = (0.3664, 1.3653)^T, \quad W = T_2^T D T_2 = \begin{bmatrix} 119.71 & -91.10 \\ -91.10 & 224.91 \end{bmatrix} \tag{43}$$

Similar to Eq.(38) and Eq.(41), the most and least favorable responses in *x*-direction of node 21 of the 60-bar space truss structure for interval analysis method can be obtained as follows

$$\min \left| u_I^{21x} \right| = 1.6491\text{E-}7, \quad \max \left| u_I^{21x} \right| = 3.0862\text{E-}7 \tag{44}$$

Substitution of Eqs.(43) and (29) into Eq.(19) yields the most and least favorable responses in *x*-direction of node 21 of the 60-bar space truss structure obtained from convex modeling method as follows

$$\min \left| u_C^{21x} \right| = 1.6575\text{E-}7, \quad \max \left| u_C^{21x} \right| = 3.0777\text{E-}7 \tag{45}$$

Thus, it can be seen from Eqs.(44) and (45) that convex modeling method gives tighter bounds of responses than interval analysis method in the case of data points listed in Table 5.

Case II: Consider another set of hypothesized data for uncertain parameters as shown in Figure 8, and they are listed in Table 6.

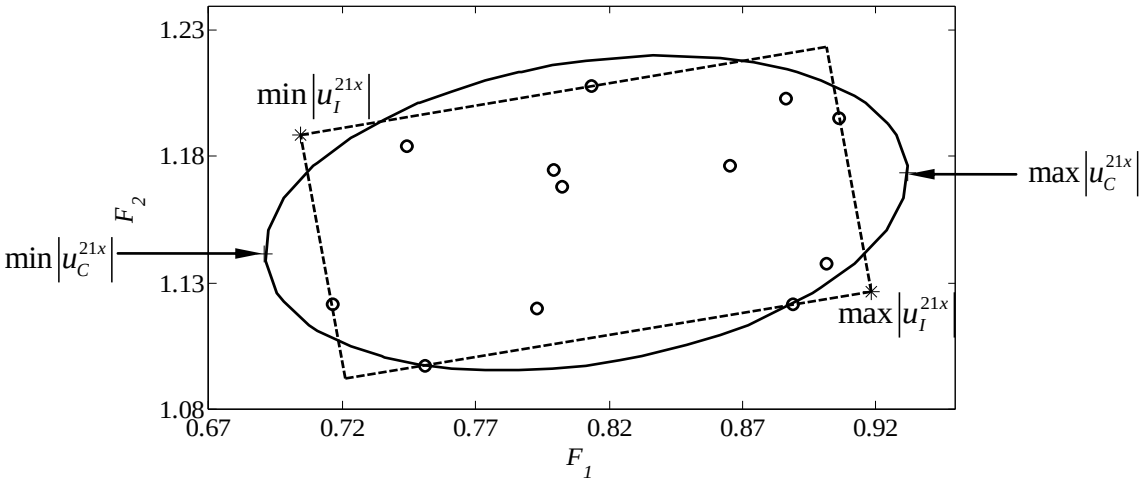


Figure 8. Rectangle and ellipse containing the data on uncertain parameters F_1 and F_2

Table 6. The values of uncertain parameter F_1 and F_2

k	1	2	3	4	5	6	7	8	9	10	11	12
F_1	0.7991	0.887	0.901	0.744	0.793	0.803	0.813	0.865	0.751	0.889	0.906	0.716
F_2	1.175	1.203	1.138	1.184	1.119	1.168	1.208	1.176	1.097	1.121	1.196	1.121

The smallest rectangle obtained from the set of data by using of Zhu, Elishakoff and Starnes's method (1996) is shown as Figure 8. The smallest ellipse is obtained from the set of data by using of Zhu, Elishakoff and Starnes's method (1996). The optimal rotation angle θ_{10} obtained is 10°. Similarly, the vector a_0 of central point and the weight matrix W can be obtained as

$$a_0 = T_2^T b_0 = (0.8113, 1.1576)^T, \quad W = T_2^T D T_2 = \begin{bmatrix} 73.46 & -35.98 \\ -35.98 & 271.17 \end{bmatrix} \tag{46}$$

Similar to Eq.(38), the most and least favorable responses in *x*-direction of node 21 of the 60-bar space truss structure for interval analysis method can be obtained as follows

$$\min \left| u_I^{21x} \right| = 4.5511\text{E-}7, \quad \max \left| u_I^{21x} \right| = 5.9339\text{E-}7 \tag{47}$$

Substitution of Eqs.(46) and (29) into Eq.(19) results in the most and least favorable responses in *x*-direction of node 21 of the 60-bar space truss structure obtained from convex modeling method as follows

$$\min \left| u_C^{21x} \right| = 4.4628\text{E-}7, \quad \max \left| u_C^{21x} \right| = 6.0222\text{E-}7 \tag{48}$$

Thus, it can be seen from Eqs.(47) and (48) that interval analysis method gives tighter bounds of responses than convex modeling method in the case of data points listed in Table 6.

From the analysis of this section, we still can find that the sample data points decide which of the non-probabilistic uncertainty descriptions, convex modeling or interval analysis, to be prefered.

6. Conclusion

In this study, through numerical examples convex modeling and interval analysis are extensively compared based on the same experimental points. Some explanations are given for the problem that which of the non-probabilistic uncertainty descriptions, convex modeling or interval analysis, ought be utilize. Given the experimental points, the smallest hyper-rectangle and the smallest ellipsoid containing them can be obtained. From these numerical examples it can be concluded that (1) If V_1 is smaller than V_2, then one has to prefer interval analysis; (2) If V_1 is in excess of V_2, then the analyst ought to utilize convex modeling; (3) If V_1 equals V_2 or these two quantities are in close vicinity, then two approaches can be utilized with nearly equal validity. Therefore, the type of the analytical treatment that should be adopted for non-probabilistic analysis of uncertainty depends upon the available experimental data.

Of course, the purpose of the paper is not to replace the probabilistic approach by the non-probabilistic set-theoretic convex methods. The latter is a possible alternative or a supplementary way of the uncertainty analysis when scarce data is available to justify the probabilistic analysis. We conclude that the type of the analysis of uncertainty depends on the type and amount of available information.

References

Ben-Haim, Y. and I. Elishakoff. Convex Models of Uncertainty in Applied Mechanics. *Amsterdam: Elsevier Science Publishers*, 1990.

Bulgakov, B. V. On the accumulation of disturbances in linear oscillatory systems with constant coefficients. *Compotes Rendus de l'Academic des Sciences de UURSS*, 1946, LI(5): 343-345.

Bulgakov, B V. Fehleranhaeufung Bei Kreiselapparaten. *Ingenieur-Archiv*, 1940, 11: 461-469. (in German).

Elishakoff, I, Li, Y W, and Starnes, J H, Jr. Non-Classical Problems in the Theory of Elastic Statbility. *Cambridge: Cambridge University Press*, 2001.

Elishakoff. I., Cai, G. Q. and Starnes J H Jr. Non-linear buckling of a column with initial imperfection via stochastic and non-stochastic convex models. *International Journal of Non-linear Mechanics*, 1994, 29(1): 71-82.

Ganzerli, S. and Pantelides, C. P. Load and resistance convex models for optimum design. *Journal of Structural Optimization*, 1999, 17: 259-268.

Lombardi, M. Optimization of uncertain structures using non-probabilistic models. *Computers and Structures*, 1998, 67(1-3): 99-103.

Manson, G. Calculating frequency response functions for uncertain systems using complex affine analysis. *Journal of Sound and Vibration*, 2005, 288: 487-521.

McWilliam, S. Anti-optimization of uncertain structures using interval analysis. *Computers and Structures*, 2001, 79: 421-430.

Moens, D. and Vandepitte, D. Interval sensitivity theory and its application to frequency response envelope analysis of uncertain structures. *Computer Methods in Applied Mechanics and Engineering*, 2007, 196: 2486-2496.

Mullen, R. L. and Muhanna, R. L. Bounds of structural response for all possible loading combinations. *Journal of Structural Engineering*, 1999, 125(1): 98-106.

Pantelides, C. P. and Ganzerli, S. Design of trusses under uncertain loads using convex models. *Journal of Structural Engineering*, 1998, 124(3): 318-329.

Qiu, Z. P. Comparison of static response of structures using convex models and interval analysis method. *International Journal for Numerical Methods in Engineering*, 2003, 56(12): 1735-1753.

Qiu, Z. P., Ma, L. H. and Wang, X. J. Ellipsoidal-bound convex model for the non-linear buckling of a column with uncertain initial imperfection. *International Journal of Non-linear Mechanics*, 2006, 41: 919-925.

Qiu, Z. P. and Wang, X. J. Comparison of dynamic response of structures with uncertain-but-bounded parameters using non-probabilistic interval analysis method and probabilistic approach. *International Journal of Solids and Structures*, 2003, 40(20): 5423-5439.

Rao, S. S. and Berke, L. Analysis of uncertain structural system using interval analysis. *AIAA Journal*, 1997, 35: 727-735.

Schweppe, F. C. Recusive state estimation: Unkown but bounded errors and system inputs. *IEEE Transactions On Automatic Control*, 1968, AC-13: 22-28.

Skalna, I. Methods for solving systems of linear equations of structure mechanics with interval parameters. *Computer Assisted Mechanics and Engineering Sciences*, 2003, 10: 281-293.

Zhu, L. P., Elishakoff, I. and Strarnes, J. H., Jr. Derivation of multi-dimensional ellipsoidal convex model for experimental data. *Mathematical and Computer Modelling*, 1996, 24(2): 103-114.

How to Estimate, Take Into Account, and Improve Travel Time Reliability in Transportation Networks

Ruey L. Cheu, Vladik Kreinovich, François Modave, Gang Xiang, Tao Li, and Tanja Magoc
Center for Transportation Infrastructure Systems, University of Texas, El Paso, TX 79968, USA, contact vladik@utep.edu

Abstract. Many urban areas suffer from traffic congestion. Intuitively, it may seem that a road expansion (e.g., the opening of a new road) should always improve the traffic conditions. However, in reality, a new road can actually worsen traffic congestion. It is therefore extremely important that before we start a road expansion project, we first predict the effect of this project on traffic congestion.

Traditional approach to this prediction is based on the assumption that for any time of the day, we know the exact amount of traffic that needs to go from each *o*rigin city zone A to every other *d*estination city zone B (these values form an *OD-matrix*), and that we know the exact capacity of each road segment. Under this assumption, known efficient algorithms produce the equilibrium traffic flows.

In reality, the road capacity may unpredictably change due to weather conditions, accidents, etc. Drivers take this uncertainty into account when planning their trips: e.g., if a driver does not want to be late, he or she may follow a slower route but with a guaranteed arrival time instead of a (on average) faster but unpredictable one. We must therefore take this uncertainty into account in traffic simulations. In this paper, we describe algorithms that take this uncertainty into account.

Keywords: transportation networks, traffic assignment, reliability, risk-taking behavior

1. Decreasing Traffic Congestion: Formulation of the Problem

Decreasing traffic congestion: a practical problem. Many urban areas suffer from traffic congestion. It is therefore desirable to decrease this congestion: e.g., by building new roads, or by adding new lanes to the existing roads.

Important difficulty: a new road can worsen traffic congestion. Intuitively, it may seem that a road expansion (e.g., the opening of a new road) should always improve the traffic conditions. However, in reality, a new road can actually worsen traffic congestion. Specifically, if too many cars move to a new road, this road may become even more congested than the old roads initially were, and so the traffic situation will actually decrease – prompting people to abandon this new road. This possible negative effect of a new road on congestion is a very well known "paradox" of transportation science, a paradox which explains the need for a detailed analysis in the planning of the new road; see, e.g, (Ahuja et al., 1993; Sheffi, 1985). This paradox was first discovered by A.

Doig (see (Appa, 1973)) and first published in (Braess, 1968; Charnes and Klingman, 1971; Szwarc, 1971).

Importance of the preliminary analysis of the results of road expansion. Our objective is to decrease traffic congestion. We have just mentioned that an addition of a new road can actually worsen the traffic congestion. It is therefore extremely important that before we start a road expansion project, we first predict the effect of this project on traffic congestion.

Traditional approach to predicting the results of road expansion. Traditional approach to predicting the results of road expansion is based on the assumption that for any time of the day, we know the exact amount of traffic that needs to go from each *o*rigin city zone A to every other *d*estination city zone B (these values form an *OD-matrix*), and that we know the exact capacity of each road segment. Under this assumption, known efficient algorithms produce the equilibrium traffic flows; see, e.g., (Sheffi, 1985).

Limitations of the traditional approach to predicting the results of road expansion. In reality, the road capacity may unpredictably change due to weather conditions, accidents, etc. Drivers take this uncertainty into account when planning their trips: e.g., if a driver does not want to be late, he or she may follow a slower route but with a guaranteed arrival time instead of a (on average) faster but unpredictable one.

We must therefore take this uncertainty into account in traffic simulations.

What we do in this paper. In this paper, we describe algorithms that take the above uncertainty into account.

Comment. Some of the results presented in this paper first appeared in our research report (Cheu et al., 2007). This report also describes a software package that implements our algorithms.

2. Traffic Assignment: Brief Reminder

Road assignment problem: informal description. In order to select the best road expansion project, we must be able to predict how different projects will affect road congestion. For that, we need to be able, based on the traffic demand and on the road capacities, to predict the traffic on different places of different roads at different times of the day. This prediction problem is called the *traffic assignment* problem.

To describe this problem in precise terms, we need to describe how exactly the traffic demand is described, how the road capacities are described, and what exactly assumptions do we make about the drivers' behavior.

Granulation. To describe traffic demand, we divide the urban area into *zones* and describe how many drivers need to get from one zone to another.

Similarly, to describe road capacity, we divide all the roads into road *segments* (*links*), and describe the capacity of each link.

The time of the day is similarly divided into *time intervals*.

Comment. How to select an appropriate size of a zone, of a road link, and of a time interval?

– On the one hand, the finer the division, we more accurate is the resulting traffic picture.

– On the other hand, the finer the division, the more zones and links we need to consider and hence, the more computations we need to perform.

Thus, the granularity of the traffic problem should be determined by the trade-off between accuracy and computational complexity.

For example, for the city of El Paso with a population of 700,000, a standard road network model consists of 681 zones and 4836 road links.

How to describe traffic demand? Once we divided the urban area into n zones, we must describe, for every two zones i and j, the number of drivers d_{ij} who need to go from zone i to zone j. The corresponding $n \times n$ matrix is called an *origin-to-destination* matrix, or an O-D matrix, for short.

So, to the traffic demand is described by the O-D matrices corresponding to different times of the day.

How to describe road capacity? For each road link, the road capacity is usually described by the number c of cars per hour which can pass through this road link.

How to describe travel time along a road link? Every road link has a posted speed limit. When there are few cars of this road, then these few cars can safely travel at the speed limit s. The resulting travel time t^f along this road link can be estimated as L/s, where L is the length of this road link. This travel time t^f is called a *free flow* travel time.

When the traffic volume v increases, congestions starts, the cars start slowing each other down. As a result, the travel time t along the road link increases. The dependence of the travel time on the volume is usually described by the Bureau of Public Roads (BPR) formula

$$t = t^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^{\beta}\right].$$

The parameters a and β are determined experimentally; usually, $a \approx 0.15$ and $\beta \approx 4$.

Equilibrium. When a new road is built, some traffic moves to this road to avoid congestion on the other roads; this causes congestion on the new road, which, in its turn, leads drivers to go back to their previous routes, etc. These changes continue until there are alternative routes in which the overall travel time is larger.

Eventually, this process converges to an *equilibrium*, i.e., to a situation in which the travel time along all used alternative routes is exactly the same – and the travel times along other un-used routes is higher; see, e.g., (Sheffi, 1985).

There exist efficient algorithms which, given the traffic demand (i.e., the O-D matrices) and the road capacity, computes the corresponding equilibrium (Sheffi, 1985). This algorithm computes the traffic volume along each road link, the travel time between every two zones, etc.

3. How We Can Use the Existing Traffic Assignment Algorithms to Solve Our Problem: Analysis

Our main objective: reminder. Our main objective is to predict how different road projects will affect future traffic congestion – so that we will be able to select a project which provides the best congestion relief.

To be able to do that, we must predict the traffic congestion resulting from the implementation of each of the road projects.

How we can predict the traffic congestion resulting from different road projects. As we have mentioned, to apply the existing traffic assignment algorithms, we need to know the traffic capacities and traffic demands.

The traffic capacities of the improved road network come directly from the road project – we know which new road links we build, what is their capacity, and which existing links are expanded. So, to solve our problem, we need to find the traffic demands.

Future traffic demands: what is known. There exist tools and techniques for predicting population growth in different zones, and for describing how this population growth will affect the overall traffic demand. Texas Department of Transportation (TxDOT) have been using the resulting predictions of daily O-D matrices corresponding to different future times (such as the year 2030).

Future traffic demands: what is lacking. To get a better understanding of the future traffic patterns, we must be able to describe how this daily traffic is distributed over different time intervals, in particular, how much of this traffic occurs during the critical time intervals corresponding to the morning rush hour. In other words, we need to "decompose" the daily O-D matrix into O-D matrices corresponding to different time intervals, e.g., 1 hour or 15 minute intervals.

How to find traffic demands corresponding to different times of the day: first approximation. In the first approximation, we can determine these O-D matrices by simply assuming that the proportion of drivers starts their trip at different times (such as 7 to 7:15 am, 7:15 to 7:20 am, etc.) as now. This first approximation is described in the next chapter.

Limitation of the first approximation predictions– and the need for better predictions. the problem with this first approximation is that the existing traffic pattern is based on the current traffic congestion. For example, if traveling from zone A to zone B takes a long time (say, 1 hour), drivers who need to drive from A to B and reach B by 9 am leave early, at 8 am, so as to be at their destination on time. As a result, in the existing traffic pattern, we have a lot of drivers leaving from A to B at 8 am.

If we simply use the existing travel pattern, we will therefore predict that in the future, a similarly big portion of drivers going from A to B also leaves at 8 am.

If we build a new road segment that eases this congestion, then there is no longer a need for these drivers to leave earlier. As a result, the actual O-D value corresponding to leaving at 8 am will be much smaller than according to our first approximation prediction.

To provide a more accurate prediction of the future traffic demand, we must therefore take into account the road improvements. In the following sections,w e describe how this can be taken into account.

Taking uncertainty into account. Finally, as we mentioned earlier, we need to take into account the uncertainty which which we can predict travel times. This is taken into account in the final sections of this paper.

4. How to Predict Future Traffic Demand: First Approximation

Main idea behind the first approximation: reminder. To predict the effect of different road projects on the future traffic congestion, we need to know future traffic demand, i.e., we need to know how many drivers will go from every zone to every other zones at different moments of time.

We usually have *daily* predictions, i.e., predictions describing the overall daily traffic for every origin-destination (O-D) pair. Based on these daily O-D matrices, we must predict O-D matrices corresponding to different time intervals.

It is reasonable to assume that in the planned future, the distribution of departure times will be approximately the same as at present. Under this assumption, we can estimate the O-D matrix corresponding to a certain time interval by simply multiplying the (future) daily O-D matrix by the corresponding *K-factor* – portion of traffic which occurs during this time interval. These K-factors can be determined by an empirical analysis of the current traffic: a K-factor corresponding to a certain time interval can be estimated as a ratio between

- the number of trips which start at this time interval, and
- the overall number of trips.

Use of empirical K-factors and linear interpolation. At present, the empirical values of the K-factor are only available for hourly intervals. If we want to find the K-factors corresponding to half-hours or 15 minute intervals, it is reasonable to use linear interpolation. Let us illustrate linear interpolation on a simple example. Let us assume that we know K-factors corresponding to the hourly traffic, in particular, we know that:

- at 7:00 am, the K-factor is 6.0%, meaning that at this moment of time, the traffic volume (in terms of vehicles per hour) is equal to 6.0% of the daily traffic volume (in terms of vehicles per day); and
- at 8:00 am, the K-factor is 8.0%, meaning that at this moment of time, the traffic volume (in terms of vehicles per hour) is equal to 8.0% of the daily traffic volume (in terms of vehicles per day).

For example, if for some O-D pair, the daily traffic volume is 1,000 vehicles per day, then:

- at 7:00 am, the traffic volume will be $6.0\% \cdot 1000 = 60$ vehicles per hour, and

– at 8:00 am, the traffic volume will be $8.0\% \cdot 1000 = 80$ vehicles per hour.

If we are interested in half-hour intervals, then we need to also estimate the traffic volume at the intermediate moment of time 7:30 am. Linear interpolation means that as such an estimate, we use the value $(6.0 + 8.0)/2 = 7\%$. So, we get the following K-factors for the half-hour time intervals:

– at 7:00 am, the K-factor is 6.0%;

– at 7:30 am, the K-factor is 7.0%;

– at 8:00 am, the K-factor is 8.0%.

Similarly, to extrapolate into 15 minute intervals, we use $(6.0 + 7.0)/2 = 6.5\%$ for 7:15 am and $(7.0 + 8.0)/2 = 7.5\%$ for 7:45 am. So, we get the following K-factors for the 15 minute time intervals:

– at 7:00 am, the K-factor is 6.0%;

– at 7:15 am, the K-factor is 6.5%;

– at 7:30 am, the K-factor is 7.0%;

– at 7:45 am, the K-factor is 7.5%;

– at 8:00 am, the K-factor is 8.0%.

In the above example, in which for some O-D pair the daily traffic volume is 1,000 vehicles per day:

– at 7:00 am the traffic volume is $6.0\% \cdot 1000 = 60$ vehicles per hour,

– at 7:15 am the traffic volume is $6.5\% \cdot 1000 = 65$ vehicles per hour,

– etc.

5. How to Take Departure Time Choice into Account

Need to take departure time choice into consideration. To understand how different road projects will affect the future traffic, we need to estimate the O-D matrices for different time intervals. At present, we usually only have estimates for the daily O-D matrices. In the previous section, we described how to use the current K-factors to divide the daily O-D matrices into O-D matrices for different time intervals.

The resulting O-D matrices are, however, only a first approximation to the actual O-D matrices. Indeed, the existing O-D matrices and the existing values of the K-factor are based on the experience of the drivers under current driving conditions. A driver selects his or her departure time based on the time that the driver needs to reach the destination (e.g., the work-start time), and the expected

travel time. For example, if the driver needs to be at work at 8:00am, and the travel time to his or her destination is 30 minutes, then the driver leaves at 7:30 am.

Population changes and new roads will change expected travel time. For example, if due to the increased population and the resulting increase road congestion the expected travel time increases to 45 minutes, then the same driver leaves at 7:15 am instead of the previous 7:30 am. So, the corresponding entry in O-D matrix corresponding to 7:30 am will decrease while a similar entry in the O-D matrix corresponding to 7:15 am will increase.

Similarly, if a new freeway decreases the expected travel time to 15 minutes, then the driver will leave at 7:45 am instead of the original 7:30 am. In this case, the corresponding entry in O-D matrix corresponding to 7:30 am will decrease while a similar entry in the O-D matrix corresponding to 7:45 am will increase.

In general, the change in a transport network and/or the change in travel time will change the departure time choice and thus, change the resulting O-D matrix. Let us describe how we can take this departure time choice into consideration.

The use of logit model: general idea. In transportation engineering, the most widely used model for describing the general choice (especially the choice in transportation-related situations) is the logit model. In the logit model, the probability of departure in different time intervals is determined by the utility of different departure times to the driver. According to this model, the probability P_i that a driver will choose the i-th time interval is proportional to $\exp(u_i)$, where u_i is the expected utility of selecting this time interval. The coefficient at $\exp(u_i)$ must be chosen from the requirement that the sum of these probabilities be equal to 1. So, the desired probability has the form $P_i = \exp(u_i)/s$, where $s \stackrel{\text{def}}{=} \exp(u_1) + \ldots + \exp(u_n)$. (Motivation for this model is presented in Appendix A.)

To apply the logit model, we must be able to estimate the utilities of different departure time choices. According to (Noland and Small, 1995; Noland et al., 1998), the utility u_i of choosing the i-th time interval is determined by the following formula:

$$u_i = -0.1051 \cdot E(T) - 0.0931 \cdot E(SDE) - 0.1299 \cdot E(SDL) - 1.3466 \cdot P_L - 0.3463 \cdot \frac{S}{E(T)},$$

where $E(T)$ is the expected value of travel time T, $E(SDE)$ is the expected value of the wait time SDE when arriving early, $E(SDL)$ is the expected value of the delay SDL when arriving late, P_L is the probability of arriving late, and S is the variance of the travel time. If we denote departure time by t_d, and the desired arrival time by t_a, then we can express SDE as $SDE = \max(t_a - (t_d + T), 0)$, and SDL as $SDL = \max((t_d + T) - t_a, 0)$. So, to estimate the values of the utilities, we must be able to estimate the values of all these auxiliary characteristics.

How to estimate the expected travel time, expected wait and delay times, and the probability of arriving late. The first of these auxiliary values – the expected value $E(T)$ of the traffic time T – is the most straightforward to compute: we can find it by simply applying a standard traffic assignment procedure (e.g., the one implemented in the standard package TransCAD) to the original O-D matrices.

To estimate the expected value $E(SDE)$ of the wait time SDE and the expected value $E(SDL)$ of the time delay SDL, in addition to the travel time, we must also know the departure time t_d and the desired arrival time t_a.

Let us start our analysis with the departure time t_d. For simplicity, for all the traffic originating during a certain time interval, as a departure time, we take the midpoint of the corresponding time interval. For example, for all the traffic originating between 7:00 am and 7:15 am, we take 7:07.5 am as the departure time.

The analysis of the desired arrival time t_a is slightly more complicated. The desired arrival time depends on the time of the day. In the morning, the desired arrival time is the time when the drivers need to be at work or in school. During the evening rush hour, the desired arrival time is the time by which the drivers want to get back home, etc.

In terms of traffic congestion, the most crucial time interval is the morning rush hour, when for most drivers, the desired arrival time is the work-start time. In view of this, in the following text, we will refer to all desired arrival times as work-start times.

The work-start time usually depends on the destination zone. For example, in El Paso, most zones have the same work-start time with the exception of a few zones such as:

- the Fort Bliss zones where the military workday starts earlier, and
- the University zone(s) where the school day usually starts somewhat later.

For every zone, we therefore usually know the (average) work-start time, i.e., the (average) desired arrival time for all the trips with the destination in this zone.

Of course, the actual work-start time for different drivers arriving in the zone may somewhat differ from the average work-start time for this zone. To take this difference into consideration, we assume that the distribution of the actual works-start time follows a bell-shaped distribution around the average. We only consider discrete time moments, e.g., time moments separated by 15 minute time intervals. It makes sense to assume that:

- for the 40% of the drivers, the actual work-start time is the average for this zone,
- for 20%, the work-start time is 15 minute later,
- for another 20%, the work-start time is 15 minutes earlier,
- for 10%, it is 30 minutes later, and
- for the remaining 10%, it is 30 minutes earlier.

For example, if the average work-start time for a zone is 8:00 am, then the assumed work-start times are as follows

- for 10% of the drivers, the work-start time is 7:30 am;
- for 20% of the drivers, the work-start time is 7:45 am;
- for 40% of the drivers, the work-start time is 8:00 am;

– for 20% of the drivers, the work-start time is 8:15 am; and, finally,

– for 10% of the drivers, the work-start time is 8:30 am.

For each of these 5 groups, we can estimate the corresponding value of SDE as $SDE(t_a) = \max(t_a - (t_d + T), 0)$. To get the desired value of the expected wait time $E(SDE)$, we need to combine these values $SDE(t_a)$ with the corresponding probabilities. For example, when the average work-start time is 8:00 am, the expected value of SDE is equal to

$$E(SDE) = 0.1 \cdot SDE(7{:}30) + 0.2 \cdot SDE(7{:}45) + 0.4 \cdot SDE(8{:}00) + 0.2 \cdot SDE(8{:}15) + 0.1 \cdot SDE(8{:}15).$$

Similarly, we can estimate the expected value $E(SDL)$ of the delay SDL. By adding the probabilities corresponding to different work-start times, we can also estimate the probability P_L of being late.

How to estimate the variance of the travel time. In the previous paragraphs, we described how to estimate the expected values $E(T)$, $E(SDE)$, $E(SDL)$, and the probability P_L. To compute the desired utility value, we only need one more characteristic: the variance S of the travel time. Let us analyze how we can estimate the variance S.

In the deterministic traffic assignment model, once we know the capacities of all the road links and the traffic flows (i.e., the values of the O-D matrix), we can uniquely determine the traffic times for all O-D pairs. In practice, the travel time can change from day to day. Some changes in travel time are caused by a change in weather, by special events, etc.; the resulting deviations from travel time are usually minor. The only case when travel times change drastically is when there is a serious road incident somewhere in the network. Since incidents are the major source of travel time delays, it is reasonable to analyze incidents to estimate the variance S of the travel time.

For this analysis, we need to have a record of incidents which occurred during a certain period of time (e.g., 90 days). The record of each incident typically includes the location and time of this incident, and the number of lanes of the corresponding road which were closed because of this incident. To estimate the variance S corresponding to a certain time interval (e.g., from 8:00 to 8:15 am), we should only consider the incidents which occurred during that time interval. Based on the incident location, we can find the link on which this incident occurred. The incident decreases the capacity of this link. This decrease can be estimated based on the original number of lanes and on the number of lanes closed by this incident.

Comment. If all the lanes were closed by the incident, then the capacity of the link goes down to 0. A reader should be cautioned that the TransCAD software tool does not allow us to enter 0 value of a link capacity. To overcome this problem, we set the capacity to the smallest possible value (such 1 vehicle per hour). For all practical purposes, this is equivalent to setting this capacity to 0.

Let us now provide heuristic arguments for estimating the decrease in capacity in situations in which some lanes remain open. Let us start with the simplest case of a 1-lane road. In reality, depending on the severity of an incident, the factor from 0 to 1 describing the decreased capacity can take all possible values from the interval $[0, 1]$. In the incident record, we only mark whether the incident actually led to the lane closure or not. In other words, instead of the actual value of the capacity-decrease factor, we only keep, in effect, 0 or 1, with

– 0 corresponding to the closed lane, and

– 1 corresponding to the open lane.

In yet another terms, we approximate the actual value of the factor by 0 or 1. It is reasonable to assume that:

– factors 0.5 or higher get approximated by 1 (lane open), while

– factors below 0.5 are approximated by 0 (lane closed).

So, the incident records in which the lane remained open correspond to all possible values of the capacity-decrease factor from the interval $[0.5, 1]$. As a reasonable average value of this factor for the case when the lane remained open, we can therefore take the midpoint of this interval, i.e., the value 0.75.

In multi-lane roads, an incident usually disrupts the traffic on all the lanes. It is therefore reasonable to assume that if no lanes were closed, then the capacity of each lane was decreased to 75% of its original value. Thus, for minor incidents in which no lanes were closed, we set the resulting capacity to 3/4 of the original capacity of the link.

For a 2-lane road, if one lane is closed and another lane remain open, then we have one lane with 0 capacity and one lane with 3/4 of the original capacity; the resulting capacity is 3/4 of the capacity of a single lane, i.e., 3/8 of the original capacity of the 2-lane road.

For a 3-lane road, if one lane is closed this means that we retain only 2/3 of the incident-reduced 75% capacity, i.e., 1/2 of the original capacity. If two lanes are closed, this means that we retain only 1/3 of the reduced capacity, i.e., 1/4 of the original capacity.

Similar values can be estimated for 4-lane roads and, if necessary, for roads with a larger number of lanes.

For each recorded incident occurring at a given time interval, we replace the original capacity in the incident-affected link by the correspondingly reduced value, and solve the traffic assignment problem for thus reduced capacity. As a result, for each O-D pair, we get a new value of the travel time.

– when the incident is far away from the route, this travel time may be the same as in the original (no-incidents) traffic assignment;

– however, if the incident is close to the route (or on this route), this travel time is larger than in the no-incidents case.

Thus, for each O-D pair and for each time interval, for each day d during the selected time period P (e.g., 90 days), we have a value of the travel time $t(d)$:

– if there was no incident on this day, the value of the travel time comes from the original traffic assignment;

– for the days on which there was an incident during the given time interval, the travel time comes from the analysis of the network with the correspondingly reduced capacity.

Based on these values $t(d)$, we compute the mean value E of the travel time as $E = \frac{1}{P} \cdot \sum_{d=1}^{P} t(d)$, and then the desired variance S as $S = \frac{1}{P} \cdot \sum_{d=1}^{P} (t(d) - E)^2$.

How to take into account departure time choice when making traffic assignments: a seemingly natural idea and its limitations. In the two previous text, we described how we can compute the characteristics which are needed to estimate the utility related to each departure time. Let us now assume that we know the original O-D matrices for each time interval i. For each time interval i, we can use the corresponding O-D matrix and solve the traffic assignment problem corresponding to this time interval. From the resulting traffic assignment, we can compute the values of the desired auxiliary characteristics, and thus, estimate the expected utility u_i of departing at this time interval i. The logit formula $P_i = \exp(u_i)/s$, where $s = \exp(u_1) + \ldots + \exp(u_n)$, enables us to compute the probability P_i that the driver will actually select departure time interval i.

The probability P_i means that out of N drivers who travel from the given origin zone to the given destination zone, $N \cdot P_i$ leave during the i-th time interval. The overall number of drivers who leave from the given origin zone to the given destination zone can be computed by adding the corresponding values in the original O-D matrices for all time intervals. Multiplying this sum by P_i, we get the new value. These new values form the new O-D matrices for different time intervals i.

These new O-D matrices take into account the departure time choice. However, they are not the ultimate O-D matrices. Indeed, since we have changed the O-D matrices, we thus changed the traffic assignments at different moments of time; this will lead to different values of utilities u_i and probabilities P_i.

As an example, let us assume that there is an O-D pair for which the free-flow travel time is 30 minutes. Let us also assume that for the corresponding destination, everyone needs to be at work at 8 am. Let us also assume that at present, there is not much traffic congestion between the origin and destination zones, so everyone leaves around 7:30 am and gets to work on time. Since we are estimating the distribution of traffic flow over time intervals based on the existing traffic, we will thus conclude that

- in the O-D matrix corresponding to 7:30 am, we will have all the drivers, while
- in the O-D matrices corresponding to earlier time intervals, we will have no drivers at all.

Let us now apply these O-D matrices to the future traffic, when due to the population increase, the traffic volume becomes much higher. Due to this higher traffic volume, the traffic time will drastically exceed 30 minutes, so all the drivers leaving at 7:30 am will be, e.g., 15 minutes late.

On the other hand, drivers who happen to leave at 7:15 am encounter practically no traffic – because there was no one needing to drive at this time in the original O-D matrix, so their travel time is exactly 30 minutes, and they get to work by 7:45 am, 15 minutes earlier. As we have seen in the above empirical formula (and in full accordance with common sense), the penalty for being 15 minutes late is much higher than the penalty of being 15 minutes early. As a result, the utility corresponding to leaving at 7:15 am is higher than the probability of leaving at 7:30 am. Hence, in

accordance with the logit formula, the probability that a driver will select to leave at 7:15 am is much higher than the probability that this driver will leave at 7:30 am.

So, in the new O-D matrices, most drivers will leave at 7:15 am, and the values corresponding to leaving at 7:30 am will be much lower. If the drivers really follow the pattern corresponding to the new O-D matrix, then the traffic congestion corresponding to 7:30 am will be much lighter than before, so the utility of leaving at 7:30 am will become higher and thus, the probability of leaving at 7:30 am will increase again. It is reasonable to expect that if we repeat this procedure several times, we will eventually reach the desired stable values of the O-D matrix.

Let us describe these ideas in precise term. In essence, we have described a procedure which transforms the original set M of O-D matrices into a new set $F(M)$ of O-D matrices, a set which takes into account departure time choice based on the traffic assignments generated by the original O-D matrices. To completely take into account the departure time choice means to find the O-D matrices which already incorporate the departure time choice, i.e., the matrices M which do not change after this transformation: $F(M) = M$.

At first glance, it seems reasonable to find these "stable" O-D matrices M by using a reasonable iterative procedure:

- we start with the set of first-approximation O-D matrices M_1 which are obtained by multiplying the new O-D daily matrix by the original K-factors;
- then, we apply the transformation F again and again: $M_2 = F(M_1)$, $M_3 = F(M_2)$, ..., until the procedure converges, i.e., until the new set of matrices M_{i+1} becomes close to the previous set M_i.

This procedure seems even more reasonable if we recall that a similar iterative procedure is successfully used in TransCAD to find the traffic assignment. However, we found out that this seemingly reasonable procedure often does not converge.

This lack of convergence can be illustrated on a "toy" example in which we have a single origin, single destination, and two possible departure times. Similarly to the above example, let us assume that the work starts at 8 am, that the free-flow traffic time is 30 minutes, and that we consider two possible departure times 7:30 am and 7:15 am. Again, just like in the above example, we assume that the original O-D matrices are based on the existing low-congestion networks in which everyone leaves at 7:30 am and nobody leaves at 7:15 am. In other words, we assume that the K-factor for 7:30 am is 1, and the K-factor for 7:15 am is 0. We also assume that there are high penalties for being late and for spending too much time in traffic.

In accordance with the above iterative procedure, we start with the O-D matrices M1 in which everyone leaves for work at 7:30 am, and nobody leaves for work at 7:15 am. The only difference with the current situation is that we are applying the same K-factors to the future, more heavy traffic.

- For those departing at 7:15 am, there is no traffic, so the travel time is equal to the free-flow time of 30 minutes.
- The drivers departing at 7:30 am face a much heavier traffic, so we get a traffic congestion. As a result of this congestion, the travel time increases to 45 minutes.

So:

- drivers who leave at 7:15 am spend only 30 minutes in traffic and arrive 15 minutes early, while
- drivers who leave at 7:30 am spend 45 minutes on the road and are 15 minutes late.

Since we assumed that the penalties for being late are heavy, the expected utility of leaving at 7:15 am is much higher than the expected utility of leaving at 7:30 am. Thus, the probability of leaving at 7:15 am is overwhelmingly higher than the probability of leaving at 7:30 am. As a result, we arrive at the new O-D matrices $M_2 = F(M_1)$ in which almost everyone leaves at 7:15 am and practically no one leaves at 7:30 am.

For these new O-D matrices M_2:

- for those departing at 7:30 am, there is no traffic, so the travel time is equal to the free-flow time of 30 minutes;
- the drivers departing at 7:15 am face a much heavier traffic, so we get a traffic congestion; as a result of this congestion, the travel time increases to 45 minutes.

So:

- drivers who leave at 7:30 am spend only 30 minutes in traffic and arrive on time, while
- drivers who leave at 7:15 am spend 45 minutes on the road.

Since we assumed that the penalties for spending extra time on the road are heavy, the expected utility of leaving at 7:30 am is much higher than the expected utility of leaving at 7:15 am. Thus, the probability of leaving at 7:30 am is overwhelmingly higher than the probability of leaving at 7:15 am. As a result, we arrive at the new O-D matrices $M_3 = F(M_2)$ in which almost everyone leaves at 7:30 am and practically no one leaves at 7:15 am.

In other words, we are back to the original O-D matrices $M_3 \approx M_1$. These "flip-flop" changes continue without any convergence. How can we modify the above idea so as to enhance convergence?

How to take into account departure time choice when making traffic assignments: a more realistic approach. We started with the O-D matrices M_1 which describe the existing traffic behavior. We want to predict how a change in traffic volume and in road network will affect the driver's behavior. To do that, let us analyze

- how the actual drivers change their behavior if the road congestion and road conditions change, and
- how we can simulate this behavior in a computer model so as to predict these changes.

At first, the drivers simply try to follow the same traffic patterns as before, i.e., depart at the same times as before. In terms of the computer representation of the drivers' behavior, this means that the proportion of the drivers departing at different time intervals remains the same as in the original traffic. In other words, this behavior corresponds to what we described as the first approximation

M_1 – when we take the new daily O-D matrix and multiply it by the K-factors corresponding to the original traffic.

As we have mentioned, due to the change in traffic volume and in road capacity, this first-approximation behavior may lead to congestions and delays. When drivers realize this, they will change their departure time so as to avoid these new delays. The drivers will use the traffic patterns and delays caused by M_1 to decide on the new departure times. The resulting change in the O-D matrix is what we described in the previous section as a transformation F. In other words, the resulting O-D matrix is $M_2 = F(M_1)$.

The change of departure times, as reflected by the move from the original O-D matrices M_1 to the new O-D matrices M_2, will again change the traffic patterns and delay times, so again, there will be a need to change the departure times based on the new traffic delays.

In these terms, the above iterative process $M_{i+1} = F(M_i)$ corresponds to the situation when the drivers only use the experience of their most recent traffic behavior and ignore the rest of the traffic history. Let us illustrate this idea on the above "toy" example.

In this example, the drivers used to go to work at 7:30 am. For the original traffic volume, this was a reasonable departure time because it allowed them to be at work exactly at the desired time 8:00 am, and to spend as little time on the road as possible – exactly 30 minutes, the free-flow traffic time.

When the traffic volume increases, in Day 1 of this new arrangement, the drivers follow the same departure time as before, i.e., they all leave for work at 7:30 am. Since the traffic volume has increased, this departure time no longer lead to the desired results – most of the drivers are 15 minutes late for work.

Since in the first day, most drivers were 15 minutes late, on the second day they leave 15 minutes earlier, at 7:15 am, so as to be at work on time. They do reach work on time, but at the expense of driving 15 minutes longer than they used to. A few drivers, however, still leave at 7:30 am. To their pleasant surprise, they experience a smooth and fast ride and arrive at work exactly on time.

The other drivers learn about the negative experience of those who left at 7:15 am and of the positive experience of those who left at 7:30 am. In our iterative model, we assume that when the drivers decide on departure time at Day 3, they only take into account delays on the previous Day 2. Under this assumption, to select the departure time on Day 3, the drivers only use the Day 2 experience. On Day 2, departing at 7:30 am certainly led to much better results that leaving for work at 7:15 am. So, under this assumption, on Day 3, most drivers will switch to 7:30 am departure time. As a result, most of them will be again 15 minutes late for work, with the exception of those who left home earlier, at 7:15 am. Since on Day 3, leaving at 7:15 am was clearly much preferable than leaving for work at 7:30 am, on the next Day 4, most drivers will again leave at 7:15 am, etc.

In this analysis, we get the same non-converging fluctuations as we had in the previous section, but this time, we understand the reason for these fluctuations: the fluctuations are caused by the simplifying assumption that the drivers' behavior is determined only by the previous moment of time.

In reality, when the drivers choose departure times, they take into account not only the traffic congestions on the day before, but also traffic congestions on several previous days. When a driver adjusts to the new environment (e.g., to the new city), he or she takes into account not just a single previous day, but rather all the previous days of driving in this new environment.

It is reasonable to assume that all these previous days are weighted equally. Let us describe this assumption in precise terms. We start with the set M_1 of O-D matrices which describe the number of drivers leaving at different time intervals on Day 1, when the drivers follow their original departure times. Similarly to the above text, let us denote the set of O-D matrices describing the drivers on Day i by M_i.

Suppose that we already know the O-D matrices M_1, M_2, ..., M_i which describe the number of drivers leaving at different time intervals at days $1, \ldots, i$. Since the drivers weigh all these previous days equally, they estimate the expected traffic E_i as the average of the previous traffics: $E_i = \frac{1}{i} \cdot (M_1 + \ldots + M_i)$.

The drivers use this expected traffic E_i to make their departure time choices. We have already described the corresponding procedure, and we have denoted the resulting transformation of O-D matrices by F. So, we can conclude that the O-D matrices M_{i+1} corresponding to the new departure times have the form $M_{i+1} = F(E_i)$.

Thus, we arrive at a new iterative procedure that takes into account departure time choice when making traffic assignments. In this procedure,

- we start with the O-D matrices M_1 which describe the original departure times; these O-D matrices can be obtained if we multiply the daily O-D matrix by the original values of the K-factors;
- then, for $i = 2, 3, \ldots$, we repeat the following procedure: first, we compute the average $E_i = \frac{1}{i} \cdot (M_1 + \ldots + M_i)$, and then we compute $M_{i+1} = F(E_i)$;
- after the iterations stop, we use the resulting set of O-D matrices to describe the resulting traffic assignments.

Our experiments on the "toy" road network and on the actual El Paso road network confirmed that this procedure converges. An important question is when to stop iterations:

- The more iterations we perform, the closer we are to the desired "equilibrium" traffic assignment.
- However, each iteration requires a reasonably large computation time on TransCAD, so it is desirable to limit the number of iterations.

To find a reasonable stopping criterion, let us recall that the main objective of our task is to help with traffic planning decisions. To help with these decisions, we must be able to predict future consequences of different road improvement plans. Thus, the objective is to deal with the O-D matrices which describe future drivers' behavior. The only way to get such future matrices is by prediction. Prediction cannot be very accurate. At best, we can predict the accuracy of the future traffic with the accuracy of 10–15%. Thus, it makes sense to stop iterations when we have already achieved this accuracy, i.e., when the difference between the O-D matrices E_i (based on which we make the plans at moment $i + 1$) and the resulting matrices M_{i+1} is smaller than (or equal to) 10–15% of the size of the matrices themselves.

As a measure of the difference between the matrices E_i and M_{i+1}, it is reasonable to take the root mean square difference, i.e., the value $d(E_i, M_{i+1})$ determined by the formula $d^2(E_i, M_{i+1}) = \frac{1}{N} \cdot \sum_{j=1}^{N} (e_j - m_j)^2$, where N is the total number of components in the corresponding matrices (i.e., of all tuples consisting of a time interval and an O-D pair), and e_j and m_j are these components. Similarly, as a measure of the size of a set E of matrices, it is reasonable to take its root mean square value, i.e., the value $v(E)$ determined by the formula $v^2(E) = \frac{1}{N} \cdot \sum_{j=1}^{N} e_j^2$. To speed up computations, we only compute the sizes $v(M_1)$ and $v(M_2)$ for the first two iterations, and use the largest of the two resulting sizes as an estimate for the size in general. In other words, we stop when $d(E_i, M_{i+1}) \le 0.1 \cdot \max(v(M_1), v(M_2))$.

How to take into account departure time choice when making traffic assignments: final idea and the resulting algorithm. In the previous text, we described the algorithm for taking into account departure time choice when making traffic assignments. The advantage of this algorithm is that it converges. However, from the computational viewpoint, this algorithm has a serious limitation. To implement the above algorithm, we must store the sets of O-D matrices M_1, M_2, ..., M_i corresponding to different iterations. For a large city-wide road network, we need to store information about many O-D pairs at several different time intervals. For example, the standard El Paso network has 681 zones, so we need to store the information about each of the 681×681 O-D pairs at each of, say, 12 time intervals, and we most store as many different pieces of this information as there are iterations – which may be in dozens. Storing, accessing, and processing all this information requires a large amount of computation time.

It is therefore desirable to reformulate the above algorithm in such a way as to avoid this excessive storage. We will show that such a simplification is indeed possible. The idea for this simplification comes from the fact that once we know the previous average value $E_i = \frac{1}{i} \cdot (M_1 + \ldots + M_i)$, and we have computed the new matrices $M_{i+1} = F(E_i)$, we do not need to repeat all the additions to compute the new average $E_{i+1} = \frac{1}{i+1} \cdot (M_1 + \ldots + M_i + M_{i+1})$.

Indeed, the expression for E_{i+1} can be reformulated as follows:

$$E_{i+1} = \frac{1}{i+1} \cdot ((M_1 + \ldots + M_i) + M_{i+1}),$$

and, by definition of E_i, we have $M_1 + \ldots + M_i = i \cdot E_i$. Thus, to compute the new average E_{i+1}, we can use the simplified formula

$$E_{i+1} = \frac{1}{i+1} \cdot (i \cdot E_i + M_{i+1}) = E_i \cdot \left(1 - \frac{1}{i+1}\right) + M_{i+1} \cdot \frac{1}{i+1}.$$

Since $M_{i+1} = F(E_i)$, we can reformulate the iterative procedure in terms of the average matrices E_i as follows: $E_{i+1} = E_i \cdot \left(1 - \frac{1}{i+1}\right) + F(E_i) \cdot \frac{1}{i+1}$. Taking into account that $E_1 = M_1$, we arrive at the following algorithm:

- we start with the O-D matrices E_1 which describe the original departure times; these O-D matrices can be obtained if we multiply the daily O-D matrix by the original values of the K-factors;
- then, for $i = 2, 3, \ldots$, we repeat the following procedure: first, we compute $F(E_i)$, and then $E_{i+1} = E_i \cdot \left(1 - \dfrac{1}{i+1}\right) + F(E_i) \cdot \dfrac{1}{i+1}$;
- we stop when $d(E_i, F(E_i)) \le 0.1 \cdot \max(v(E_1), v(E_2))$.
- after the iteration stop, we use the resulting set of O-D matrices E_i to describe the resulting traffic assignments.

Comment. As we show in Appendix B, this iterative procedure is, in some reasonable sense, an optimal algorithm for computing the fixed point of the mapping F.

6. Taking Uncertainty into Account

Need to consider uncertainty. In the previous text, we consider deterministic traffic models, in which the link travel time is uniquely determined by the traffic volume. Real-life traffic, however, is non-deterministic. To have more accurate predictions of travel times, we must take this non-determinism into account and consider stochastic traffic models.

In a stochastic traffic model, the BPR formula only describes the *average* travel time $\bar{t}$:

$$\bar{t} = t^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^{\beta}\right].$$

The stochastic nature of traffic means the actual travel time t may differ from this average value $\bar{t}$. We must therefore describe not only how the *average* travel time $\bar{t}$ depends on t_f, v, and c, but also how the *deviations* $t - \bar{t}$ from this average depend on these parameters. For example, we may want to describe how the standard deviation of the travel time t – or some other statistical characteristic – depends on these parameters.

It turns out that several seemingly reasonable models of this dependence are faulty because the predicted travel times drastically change when we simply subdivide the road links without making any changes in the actual traffic.

In this text, we describe this phenomenon, and we describe how to set up this dependence in such a way that a simple subdivision of a road link will no longer affect the resulting travel times.

We can have different subdivision into road links. Traffic networks in a big city are usually very complicated, with lots of small roads. As a result, the fully detailed simulation of a traffic network would require a large amount of computation time.

It is well known, however, that in practice, there is no need for such a detailed simulation: it is well known that it is sufficient to divide the city into zones and consider only traffic between the zones. The size of the zone depends on the amount and direction of traffic in this zone.

Once we decided how to divide the city into zones, each major road is then naturally subdivided into *road links*, i.e., pieces of this road within each zone.

In busy downtown areas, we may have a popular restaurant in one block and a big office in a neighboring block, with completely different traffic patterns. So, in order to accurately predict downtown-related traffic, we may need to have zones of the size of a few city blocks.

On the other hand, e.g., in a large residential area, we usually get the same pattern of traffic in all its parts: traffic leaving to work in the morning and traffic coming back in the afternoon. As a result, for such areas, it is sufficient to consider larger residential communities as single zones.

For deterministic traffic models, the resulting travel times do not change much if we switch to a finer subdivision into zones: a known fact. Once we come up with zones which provide a reasonable description of the traffic patterns, we can get reasonably good predictions of the traffic volumes and travel times.

If we still have additional computational power, we can consider smaller-size zones. In this case, the original road links are further subdivided into smaller-size links. If we use such a refined model, we get an even more accurate prediction of the travel times.

However, we know that the estimates coming from the original model still provide a reasonably accurate description of the travel times.

For deterministic traffic models, the resulting travel times do not change much if we switch to a finer subdivision into zones: a mathematical explanation. For a deterministic model, one of the reasons for this accuracy is that, because of the above formula for t, the travel time t predicted by the model does not depend on how exactly we subdivide the road into road links – as long as this subdivision remains reasonable in the sense that the traffic volume and the traffic capacity does not change much within this link.

Indeed, let us assume that we start with a single link of length L in the original model and then decided to subdivide it into several sublinks of length $L_1, \ldots, L_n$ – for which $L = L_1 + \ldots + L_n$. In the original model, the travel time t along this link is predicted directly – by using the above formula

$$t = t^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^{\beta}\right].$$

In the new model, we predict individual travel times $t_1, \ldots, t_n$ along different sublinks and then predict the resulting overall travel time as $t_1 + \ldots + t_n$.

Let us show that in this case, the originally predicted travel time t is equal to the total travel time $t_1 + \ldots + t_n$ predicted by the new model.

We assume that the traffic volume v and traffic capacity c are the same for all these sublinks, the only think which is different is the free flow travel time. In other words, the predicted travel times along sublinks take the form

$$t_i = t_i^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^{\beta}\right].$$

In general, the free flow travel time t^f is determined by the length L of the road link and the speed limit s along this link: $t^f = \dfrac{L}{s}$. Similarly, for each sublink, we have $t_i^f = \dfrac{L_i}{s}$.

Since $L = L_1 + \ldots + L_n$, we conclude that $t^f = t_1^f + \ldots + t_n^f$. Thus, we conclude that

$$t_1 + \ldots + t_n = t_1^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^\beta\right] + \ldots + t_n^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^\beta\right] =$$

$$(t_1^f + \ldots + t_n^f) \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^\beta\right] = t^f \cdot \left[1 + a \cdot \left(\frac{v}{c}\right)^\beta\right].$$

So, the originally predicted travel time t is indeed equal to the total travel time $t_1 + \ldots + t_n$ predicted by the new model.

Stochastic case: brief introduction. In the deterministic case, the driver selects a route for which the expected travel time is the shortest.

According to decision theory, in the general situation with stochastic uncertainty case, preferences of a person can be described by a special *utility function* which assigns, to each possible result x, a number $U(x)$ describing the "utility" of this result for this person; a person then selects an action for which the expect value of utility is the largest.

In transportation situations, the main parameter of interest to the drive is the overall travel time, so the utility depends on the travel time t: $U = U(t)$. To make the stochastic formulation of the transportation problems similar to the deterministic ones (in which the objective is to *minimize* travel time), researchers usually replace the problem of maximizing utility with an equivalent problem of minimizing disutility $u(t)$ which is defined as $u(t) = -U(t)$. Usually, an exponential disutility function is used $u(t) = A \cdot \exp(\alpha \cdot t)$; see, e.g., (Mirchandani and Soroush, 1987; Tatineni, 1996; Tatineni et al., 1997). The justification for using such functions is given in Appendix C.

Random deviations $t_i - \overline{t}_i$ for different links are usually caused by different reasons; so traditionally, the travel times t_i on different links $t_1, \ldots, t_n$ along the path are assumed to be independent random variables. Thus, the expected disutility of a path

$$\overline{u} = E[\exp(\alpha \cdot t)] = E[\exp(\alpha \cdot (t_1 + \ldots + t_n)] = E[\exp(\alpha \cdot t_1) \cdot \ldots \cdot \exp(\alpha \cdot t_n)]$$

can be represented as a product

$$\overline{u} = E[\exp(\alpha \cdot t_1)] \cdot \ldots \cdot E[\exp(\alpha \cdot t_n)].$$

Minimizing the product is equivalent to minimizing its logarithm, i.e., the sum

$$s = \ln(E[\exp(\alpha \cdot t_1)]) + \ldots + \ln(E[\exp(\alpha \cdot t_n)]).$$

In the deterministic case, $E[\exp(\alpha \cdot t)] = \exp(\alpha \cdot t)$ hence $\ln(E[\exp(\alpha \cdot t)]) = \alpha \cdot t$. So, to make the problem more similar to the deterministic one, we can divide each logarithm by α – dividing the minimizing function by a positive function does not change where the minimum is attained.

Thus, selecting of a route can be described in a form which is very similar to selecting a deterministic route, but with $\widetilde{t}_i \stackrel{\text{def}}{=} \frac{1}{\alpha} \cdot \ln(E[\exp(\alpha \cdot t_1)])$ instead of the original travel times.

We know that the deviations $t - \overline{t}$ are usually relatively small. Thus, to simplify the above expression, we can substitute $t = \overline{t} + (t - \overline{t})$ into the formula, expand the functions $\exp(z)$ and $\ln(z)$ into Taylor series and keep only the few first (major) terms in the expansion. Specifically, we have

$$\exp(\alpha \cdot t) = \exp(\alpha \cdot \overline{t}) \cdot \exp(\alpha \cdot (t - \overline{t})).$$

Here, the first factor does not depend on the random variable at all, so from the viewpoint of taking an expected value, it is simply a constant:

$$E[\exp(\alpha \cdot t)] = \exp(\alpha \cdot \overline{t}) \cdot E[\exp(\alpha \cdot (t - \overline{t}))].$$

We use the Taylor expansion of the exponential function:

$$\exp(z) = 1 + z + \frac{z^2}{2!} + \ldots = 1 + z + \frac{z^2}{2} + \ldots$$

Thus,

$$\exp(\alpha \cdot (t - \overline{t})) \approx 1 + \alpha \cdot (t - \overline{t}) + \frac{\alpha^2 \cdot (t - \overline{t})^2}{2},$$

and

$$E[\exp(\alpha \cdot (t - \overline{t}))] \approx 1 + \alpha \cdot E[t - \overline{t}] + \frac{\alpha^2 \cdot E[t - \overline{t})^2]}{2}.$$

By definition, $E[t - \overline{t}] = \overline{t} - \overline{t} = 0$, and $E[(t - \overline{t})^2]$ is the variance V. Thus, in our approximation,

$$E[\exp(\alpha \cdot t)] = \exp(\alpha \cdot \overline{t}) \cdot \left(1 + \frac{\alpha^2}{2} \cdot V\right).$$

So,

$$\frac{1}{\alpha} \cdot \ln(E[\exp(\alpha \cdot t)]) = \widetilde{t} + \frac{1}{\alpha} \cdot \ln\left(1 + \frac{\alpha^2}{2} \cdot V\right).$$

Using the Taylor expansion of the logarithm function $\ln(1 + z) = z + \ldots$, we conclude that

$$\frac{1}{\alpha} \cdot \ln(E[\exp(\alpha \cdot t)]) = \widetilde{t} + \frac{\alpha}{2} \cdot V.$$

Thus, minimizing the sum of these logarithmic expressions is equivalent to minimizing the sum of the expressions

$$\widetilde{t} = \overline{t} + \frac{\alpha}{2} \cdot V.$$

In other words, to make stochasticity into account, to each link's travel time, we add its variance (with an appropriate weight $\alpha/2$).

A seemingly natural description. In the case of the free flow traffic, there is no uncertainty; uncertainty occurs only if we have some volume on the road link – i.e., when the travel time t exceeds the free flow travel time t^f. Intuitively, the larger this excess $t - t^f$, the larger this uncertainty.

At first glance, it may seem natural to pick a proportion r_0 (e.g., 20%) and assume that for every link, the actual value $t - t_f$ can deviate by about $\pm 20\%$ (or whatever r is) from the average. In more precise terms, the standard deviation $\sigma \stackrel{\text{def}}{=} \sqrt{V}$ of the travel time is equal to $r_0 \cdot (\overline{t} - t^f)$.

Since $\sigma = \sqrt{V} = r_0 \cdot (\overline{t} - t^f)$, we conclude that $V = r_0^2 \cdot (\overline{t} - t_f)^2$.

Problem with seemingly natural assumption. Let us show that this seemingly natural assumption leads to counter-intuitive conclusions. Indeed, let us assume that we have two one-link

routes of equal quality leading from point A to point B, with the same free flow time t^f, same capacity c, and the same traffic volume v. In this case, for both links, we have the same expected travel time $\overline{t}$ and hence, the same variance – so, the values of the resulting minimized function are the same for both routes:

$$\widetilde{t}^{(1)} = \widetilde{t}^{(2)} = \overline{t} + \frac{\alpha}{2} \cdot r_0^2 \cdot (\overline{t} - t_f)^2.$$

Intuitively, if we subdivide one of the links into two equal sublinks of equal length (without changing anything of substance) we should end up with exactly the same selection. In reality, if we subdivide the first link, then for this link, we will have both t and t_f divided by 2: $\overline{t}_1 = \overline{t}_2 = \dfrac{\overline{t}}{2}$ and $t_1^f = t_2^f = \dfrac{t^f}{2}$. Hence, the variance V (proportional to $(t - t_f)^2$) will divide by 4. As a result, for each of these links, we get

$$\widetilde{t}_1 = \widetilde{t}_2 = \frac{\overline{t}}{2} + \frac{\alpha}{2} \cdot r_0^2 \cdot \frac{(\overline{t} - t_f)^2}{4}.$$

By adding these two values, we get the minimized value $\widetilde{t} = \widetilde{t}_1 + \widetilde{t}_2$ for the whole two-link route:

$$\widetilde{t} = \overline{t} + \frac{\alpha}{2} \cdot r_0^2 \cdot \frac{(\overline{t} - t_f)^2}{2}.$$

In this expression, the term proportional to the variance is twice smaller than for the second route, so this route will be selected.

Alternatively, if we keep the first route whole but subdivide the second route, we get a clear preference for the second route. Thus, the route selection depends on the exact subdivision into links – hence our seemingly natural assumption is really counter-intuitive.

Proposed solution. Our objective is to find a reasonable expression for the term

$$\widetilde{t} = \frac{1}{\alpha} \cdot \ln(E[\exp(\alpha \cdot t)]).$$

In general, this expression can depend on the free flow time t^f and on the average time $\overline{t}$.

As we have mentioned, in the absence of the traffic flow, when the travel time consists 100% of the free flow time t^f, there is no stochasticity. The larger the proportion of the excess time, i.e., the larger the ratio $r \stackrel{\text{def}}{=} \dfrac{\overline{t} - t^f}{t^f}$, the more stochasticity there is. Thus, it is reasonable to describe the desired expression for $\widetilde{t}$ in terms of t^f and r.

By definition of r, we have $\overline{t} - t^f = r \cdot t^f$ hence $\overline{f} = (1 + r) \cdot t^f$; so, once we know the dependence of $\widetilde{t}$ on t^f and $\overline{t}$, we can find its dependence on t^f and r as well. Thus, it is reasonable to claim that $\widetilde{t} = F(t^f, r)$ for some yet-to-be-determined function F.

The first desired property of the function F is that if the average time coincides with the free flow time, then there is no stochasticity, and $\widetilde{t} = t$. In other words, we must have $F(t, 0) = t$ for all t.

The second desired property is that when we subdivide a link into two sublinks, without changing the flow or capacity (and hence, without changing the ratio r), then the sum of the resulting values

$\widetilde{t}_1 + \widetilde{t}_2$ should be equal to the original value $\widetilde{t}$: $F(t_1^f, r) + F(t_2^f, r) = F(t_1^f + t_2^f, r)$. For each r, we get an equation $F'(a+b) = F'(a) + F'(b)$ for a monotonic function $F'(a) \stackrel{\text{def}}{=} F(a, r)$ hence (Aczel, 2006) $F'(a) = k \cdot a$ for some constant $k(r)$ which may depend on r. The fact that $F(t, 0) = t$ means that $k(0) = 1$.

In other words, we conclude that $\widetilde{t} = F(t^f, r) = t^f \cdot k(r)$. We know that $r = a \cdot \left(\frac{v}{c}\right)^\beta$, thus,

$$\widetilde{t} = t^f \cdot k\left(\left(\frac{v}{c}\right)^\beta\right).$$

Similarly to the above case, we can expand the dependence $k(r)$ into Taylor series and keep the first few terms in this expansion. Since $k(0) = r$, we conclude that $k(r) = 1 + a_1 \cdot r + a_2 \cdot r^2 + \ldots$, hence

$$\widetilde{t} = 1 + a_1 \cdot a \cdot \left(\frac{v}{c}\right)^\beta + a_2 \cdot a^2 \cdot \left(\frac{v}{c}\right)^{2\beta}.$$

Conclusion. The effect of stochasticity on the transportation problem can be described as follows:

- in the deterministic case, drivers select a route for which the overall travel time $\overline{t} = \overline{t}_1 + \ldots + \overline{t}_n$ is the smallest, where $\overline{t}_i = t_i^f \cdot \left[1 + a \cdot \left(\frac{v_i}{c_i}\right)^\beta\right]$;
- in the stochastic case, drivers select a route for which the expression $\widetilde{t} = \widetilde{t}_1 + \ldots + \widetilde{t}_n$ is the smallest, where $\widetilde{t}_i = t_i^f \cdot \left[1 + a_1 \cdot a \cdot \left(\frac{v_i}{c_i}\right)^\beta + b \cdot \left(\frac{v_i}{c_i}\right)^{2\beta}\right]$.

Thus, we can use the standard traffic assignment algorithms with a modified travel time function to find the corresponding traffic assignment.

Comment. Our experiments show that $a_1 \approx 1.4$ and $b \cdot 0$. So, to take the uncertainty into account, it is sufficient to replace the original value $a \approx 0.15$ in the BPR formula with the new value $a_1 \cdot a \approx 0.21$.

Acknowledgements

This work was supported in part by NSF grants HRD-0734825, EAR-0225670, and EIA-0080940, by Texas Department of Transportation grant No. 0-5453, by the Japan Advanced Institute of Science and Technology (JAIST) International Joint Research Grant 2006-08, and by the Max Planck Institut für Mathematik.

References

Aczel, J. *Lectures on Functional Equations and Their Applications*, Dover, New York, 2006.

Ahuja, R. K., T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

Appa, G. M. The transportation problem and its variants. *Oper. Res. Quarterly*, 24:79–99, 636–639, 1973.

Berinde, V. *Iterative approximation of fixed points*, Editura Efemeride, Baia Mare, 2002.

Braess, D. Über ein Paradox der Verkerhsplannung. *Unternehmenstorchung*, 12:258–268, 1968.

Charnes, A., and D. Klingman. The more-for-less paradox in the distribution model. *Cahiers du Centre d'Etudes de Recherche Operationelle*, 13:11–22, 1971.

Cheu, R., V. Kreinovich, Y.-C. Chiu, R. Pan, G. Xiang, S. Bhupathiraju, and S. R. Manduva. *Strategies for Improving Travel Time Reliability*, Texas Department of Transportation, Research Report 0-5453-R2, August 2007.

Chipman, J. The foundations of utility. *Econometrica*, 28:193–224, 1960.

Debreu, G. Review of R. D. Luce, "Individual Choice Behavior". *American Economic Review*, 50:186–188, 1960.

Jaynes, E. T., and G. L. Bretthorst (ed.), *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

Keeney, R. L., and H. Raiffa, *Decisions with Multiple Objectives*, John Wiley and Sons, New York, 1976.

Kohlenbach, U. Uniform asymptotic regularity for Mann iterates. *J. Math. Anal. Appl.*, 279(8):531–544, 2003.

Kohlenbach, U. Some computational aspects of metric fixed point theory. *Nonlinear Analysis*, 61(5):823–837, 2005.

Kohlenbach, U. Effective uniform bounds from proofs in abstract functional analysis", In: B. Cooper, B. Loewe, and A. Sorbi (eds.), *New Computational Paradigms: Changing Conceptions of What is Computable*, Springer Verlag, Berlin-Heidelberg-New York, 2007.

Kohlenbach, U.: *Applied Proof Theory: Proof Interpretations and their Use in Mathematics*. Springer Verlag, Berlin-Heidelberg, 2008.

Kohlenbach. U., and B. Lambov. Bounds on iterations of asymptotically quasi-nonexpansive mappings. In: J. G. Falset, E. L. Fuster, and B. Sims (eds.), *Proc. International Conference on Fixed Point Theory and Applications, Valencia 2003*, pages 143–172, Yokohama Publishers, 2004.

Luce, D. *Individual Choice Behavior*, John Wiley and Sons, New York, 1959.

Luce, D., and P. Suppes, Preference, utility, and subjective probability, In: D. Luce, R. Bush, and E. Galanter (eds.), *Handbook on Mathematical Psychology*, pages 249–410, John Wiley and Sons, New York, 1965.

McFadden, D. Conditional logit analysis of qualitative choice behavior, In: P. Zarembka (ed.), *Frontiers in Econometrics*, pages 105–142, Academic Press, New York, 1974.

McFadden, D. Economic choices, *American Economic Review*, 91:351–378, 2001.

Mirchandani, P., and H. Soroush, Generalized traffic equilibrium with probabilistic travle times and perceptions, *Transportation Science*, 21(3):133–152, 1987.

Noland, R. B. and K. A. Small. Travel time uncertainty, departure time choice, and the cost of morning commutes, *Transportation Researh Record*, 1493:150–158, 1995.

Noland, R. B., K. A. Small, P. M. Koseknoja, and X. Chu. Simulating travel reliability, *Regional Science and Urban Economics*, 28:535–564, 1998.

Pratt, J. W. Risk Aversion in the Small and in the Large. Econometrica, 32:122–136, 1964.

Raiffa, H. *Decision Analysis*, Addison-Wesley, Reading, Massachusetts, 1970.

Sheffi, Y. *Urban Transportation Networks*, Prentice Hall, Englewood Cliffs, NJ, 1985.

Su, Y., and X. Qin, Strong convergence theorems for asymptotically nonexpansive mappings and asymptotically nonexpansive semigroups. *Fixed Point Theory and Applications*, 2006, Article ID 96215, pp. 1–11.

Szwarc, W. The transportation paradox. *Naval Res. Logist. Quarterly*, 18:185–202, 1971.

Tatineni, M. *Solution properties of stochastic route choice models*, Ph.D. Dissertation, Department of Civil Engineering, University of Illinois at Chicago, 1996.

Tatineni, M., D. E. Boyce, and P. Mirchandani, Comparison of deterministic and stochastic traffic loading models, *Transportation Research Record*, 1607:16–23, 1997

Train, K. *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge, Massachusetts, 2003.

Wadsworth, H. M. (ed.), *Handbook of statistical methods for engineers and scientists*, McGraw-Hill Publishing Co., New York, 1990.

Appendix

A. Logit Discrete Choice Model: Towards a New Justification

Traditional approach to decision making. In decision making theory, it is proven that under certain reasonable assumption, a person's preferences are defined by his or her *utility function* $U(x)$ which assigns to each possible outcome x a real number $U(x)$ called *utility*; see, e.g., (Keeney and Raiffa, 1976; Raiffa, 1970). In many real-life situations, a person's choice s does not determine the outcome uniquely, we may have different outcomes $x_1, \ldots, x_n$ with probabilities, correspondingly, $p_1, \ldots, p_n$.

For example, drivers usually select the path with the shortest travel time. However, when a driver selects a path s, the travel time is often not uniquely determined: we may have different travel times $x_1, \ldots, x_n$ with corresponding probabilities $p_1, \ldots, p_n$.

For such a choice, we can describe the utility $U(s)$ associated with this choice as the expected value of the utility of outcomes: $U(s) = E[U(x)] = p_1 \cdot U(x_1) + \ldots + p_n \cdot U(x_n)$. Among several possible choices, a user selects the one for which the utility is the largest: a possible choice s is preferred to a possible choice s' (denoted $s > s'$) if and only if $U(s) > U(s')$.

It is important to mention that the utility function is not uniquely determined by the preference relation. Namely, for every two real numbers $a > 0$ and b, if we replace the original utility function $U(x)$ with the new one $U'(x) \stackrel{\text{def}}{=} a \cdot U(x) + b$, then for each choice s, we will have

$$U'(s) = E[a \cdot U(x) + b] = a \cdot E[U(x)] + b = a \cdot U(s) + b$$

and thus, $U'(s) > U'(s')$ if and only if $U(s) > U(s')$.

Situations in which we can only predict probabilities of different decision. One important application of decision making theory is predicting the user decisions. If we know the exact values $U(s)$ of the utilities, then we can predict the exact choice. For example, if the user has to choose between alternatives s and s', then the user chooses s if $U(s) \geq U(s')$ and s' if $U(s) \leq U(s')$.

In practice, we do not know the exact values $U(s)$ of the user's utility, we only know the approximate values $V(s) \approx U(s)$. Due to the difference between the observed (approximate) values $V(s)$ and the actual (unknown) values $U(s)$, we are no longer able to uniquely predict the user's behavior: e.g., even when $V(s) > V(s')$, we may still have $U(s) < U(s')$, and thus, it is possible that the user will prefer s.

If the differences $V(s) - U(s)$ and $V(s') - U(s')$ are small, then for $V(s) \gg V(s')$, we can be reasonably sure that $U(s) > U(s')$ and thus, that the user will select s. Similarly, if $V(s) \ll V(s')$, we can be reasonably sure that $U(s) < U(s')$ and thus, that the user will select s'. However, when the values $V(s)$ and $V(s')$ are close, then there is a certain probability that $U(s) > U(s')$ and thus, that the user will select s, and there is also a certain probability that $U(s) < U(s')$ and thus, that the user will select s'.

In this situation, based on the (approximate) utility values $V(s)$ and $V(s')$, we cannot exactly predict whether the user will prefer s or s' – because for the same values of $V(s)$ and $V(s')$, the user can prefer s and the user can also prefer s'. The best we can do in this situation is to predict the *probability* $P(s > s')$ of selecting s over s'.

Discrete choice: a formal description of the problem. Let us formulate the problem in precise terms. We have n different alternatives $s_1, \ldots, s_n$. For each of these alternative s_i, we know the (approximate) utility value $V_i \stackrel{\text{def}}{=} V(s_i)$. Based on these utility values $V(s_1), \ldots, V(s_n)$, we would like to predict the probability p_i that a user will select the alternative s_i.

Models used for such prediction are usually called *discrete choice models* (Train, 2003).

Invariance requirements in discrete choice models. As we have mentioned, the utility function is not uniquely determined by the preference relation. Namely, whenever the original utility function $U(s)$ describes the user's preference, then, for every $a > 0$ and b, the new function $U'(s) = a \cdot U(s) + b$ also describes the same preference. In other words, we can shift all the values of the utility function $u(s) \to U(s) + b$, and we can re-scale all the values $U(s) \to a \cdot u(s)$, and the resulting utility function will still describe the same preferences.

It is therefore reasonable to assume that if we shift the values of the approximate utility function, i.e., if we replace the original values $V(s_i)$ with the new values $V'(s_i) = V(s_i) + b$, then we should get the same preference probabilities:

$$p_i(V(s_1), V(s_2) \ldots, V(s_n)) = p_i(V(s_1) + b, V(s_2) + b, \ldots, V(s_n) + b).$$

In particular, if we take $b = -V(s_1)$, then we conclude that

$$p_i(V(s_1), V(s_2) \ldots, V(s_n)) = p_i(0, V(s_2) - V(s_1), \ldots, V(s_n) - V(s_1)),$$

i.e., that the probabilities depend only on the *differences* between the utility values – but not on the values themselves.

At first glance, it may seem reasonable to similarly require that the probability not change under re-scaling. However, in this case, re-scaling does not make intuitive sense, because we have a natural scale. For example, as a unit for such a scale, we can choose a standard deviation of the difference $U(s) - V(s)$ between the (unknown) actual utility $U(s)$ and the (known) approximate value of this utility $V(s)$.

In line with this analysis, in discrete choice models, it is usually assumed that the probabilities do not change with shift but it is *not* assumed that these probabilities are scale-invariant.

Logit: the most widely used discrete choice model. The most widely used discrete choice model is a *logit* model in which

$$p_i(V_1, \ldots, V_n) = \frac{e^{\beta \cdot V_i}}{\sum\limits_{j=1}^{n} e^{\beta \cdot V_j}} \tag{1}$$

for some parameter β. This model was first proposes in (Luce, 1959).

Logit: original justification. In (Luce, 1959), this model was justified based on the assumption of *independence of irrelevant alternatives*, according to which the relative probability of selecting s_1 or s_2 should not change if we add a third alternative s_3. In formal terms, this means that the probability of selecting s_1 out of two alternatives s_1 and s_2 should be equal to the conditional probability of selecting s_1 from three alternatives s_1, s_2, and s_3 under the condition that either s_1 or s_2 are selected.

It can be proven that under this assumption, the ratio p_i/p_j of the probabilities p_i and p_j should only depend on V_i and V_j; moreover, that we must have $p_i/p_j = f(V_i)/f(V_j)$ for some function $f(z)$. The requirement that this ratio be shift-invariant then leads to the conclusion that $f(z) = e^{\beta \cdot z}$ for some β – and thus, to the logit model.

Limitations of the original justification. At first glance, the above independence assumption sounds reasonable (and it is often reasonable). However, there are reasonable situations where this assumption is counter-intuitive; see, e.g., (Chipman, 1960; Debreu, 1960; Train, 2003).

For example, assume that in some cities, all the buses were originally blue. To get from point A to point B, a user can choose between taking a taxi (s_1) and taking a blue bus (s_2). A taxi is somewhat better to this user, so he selects a taxi with probability $p_1 = 0.6$ and a blue bus with the remaining probability $p_2 = 1 - 0.6 = 0.4$. In this case, the ratio p_1/p_2 is equal to 1.5.

Suppose now that the city decided to buy some new buses, and to paint them red. Let us also suppose that the comfort of the travel did not change, the buses are exactly the same. From the common sense viewpoint, it does not matter to the user whether buses are blue or red, so he should still select a taxi with probability $p_1 = 0.6$ and buses with probability 0.4. However, from the purely mathematical viewpoint, we now have *three* options: taking a taxi (s_1), taking a blue bus (s_2), and taking a red bus (s_3). Here, the probability of taking a bus is now $p_2 + p_3 = 0.4$. Hence, $p_2 < 0.4$ and so, the ratio p_1/p_2 is different from what we had before – contrary to the above independence assumption.

Current justification. An alternative justification for logit started with the unpublished result of Marley first cited in (Luce and Suppes, 1965). Marley has shown that if we assume that the approximation errors $\varepsilon(s) \stackrel{\text{def}}{=} U(s) - V(s)$ are independent and identically distributed, and if this distribution is the Gumbel distrubution, then the probability of selecting s_i indeed follows the logit formula.

Gumbel distribution can be characterized by the cumulative distribution function $F(\varepsilon) = e^{-e^{-\varepsilon}}$; it is a known distribution of extreme values.

In 1974, McFadden (McFadden, 2001) showed that, vice versa, if we assumed that the approximation errors $\varepsilon(s)$ are independent and identically distributed, and the choice probabilities are described by the logit formula, then the errors $\varepsilon(s)$ must follow the extreme value (Gumbel) distribution.

This justification was one of the main achievements for which D. McFadden received a Nobel prize in 2001 (McFadden, 2001).

Limitations of the current justification. The problem with this justification is that the logit model is known to work well even in the cases when different approximation errors are differently distributed; see, e.g., (Train, 2003).

For such situations, the only known alternative explanation is Luce's original one. The main limitation of this explanation was that it is based on the independence assumption. This is not so critical if we have three or more alternatives. Indeed, in this case, the empirical logit formula (that we are trying to explain) satisfies this assumption, so making this assumption in the situations when the logit formula holds makes sense.

This limitation, however, becomes crucial if we only consider the case of two alternatives. In this case, the independence assumption cannot even be formulated and therefore, Luce's justification does not apply. So, we arrive at the following problem.

Formulation of the problem. We need to come up with a new distribution-free justification for the logit formula, i.e., with a justification that does not depend on the assumption that approximation errors are independent and identically distributed. Such a justification is provided in this paper.

Preliminary analysis. In accordance with the above formulation of the problem, we are interested in the case of $n = 2$ alternatives s_1 and s_2. We know the approximate utility values V_1 and V_2, and we know that the probability p_1 of selecting the first alternative p_1 should only depend on the difference $V_1 - V_2$: $p_1 = F(V_1 - V_2)$ for some function $F(z)$. Our objective is to find this function $F(z)$. Let us first describe reasonable properties of this function $F(z)$.

When s_2 is fixed (hence V_2 is fixed) but the alternative s_1 is improving (i.e., V_1 is increasing), then the probability of selecting s_1 can only increase (or at least remain the same – e.g., if that probability was already equal to 1, it cannot further increase). In other words, as the difference $V_1 - V_2$ increases, the probability $p_1 = F(V_1 - V_2)$ should also increase (or at least remain the same). Thus, it is reasonable to require that the function $F(z)$ should be (non-strictly) increasing.

When s_2 and V_2 are fixed and s_1 becomes better and better, i.e., $V_1 \to +\infty$, then we should select s_1 with probability tending to 1. So, we must have $F(z) \to 1$ as $z \to +\infty$.

Similarly, s_2 and V_2 are fixed, and s_1 becomes worse and worse, i.e., $V_1 - V_2 \to -\infty$, then we should prefer s_2. So, we must have $F(z) \to 0$ as $z \to -\infty$.

Since we only have two alternatives, the probability $p_1 = F(V_1 - V_2)$ and the probability $p_2 = F(V_2 - V_1)$ must always add up to 1. Thus, we must have $F(z) + F(-z) = 1$ for all z.

So, we arrive at the following definition.

Definition 1. *By a* choice function*, we mean a function $F : R \to [0, 1]$ which is (non-strictly) increasing, and for which $F(z) \to 1$ as $z \to +\infty$, $F(z) \to 0$ as $z \to -\infty$, and $F(z) + F(-z) = 1$ for all z.*

Main idea. Our main idea is as follows. Up to now, we have discussed how to *describe* the user's behavior, but often, the ultimate objective is how to *modify* this behavior. For example, in transportation problems, the goal is often to use public transportation to relieve traffic congestion and related pollution. In this case, the problem is not just to estimate the probability of people using public transportation, but to find out how to increase this probability.

One way to increase this probability is to provide incentives. If we want to encourage people to prefer alternative s_1, then we can provide those who select this alternative with an additional benefit of value v_0. In this case, for alternatives $s_i \neq s_1$, the corresponding utility V_i remains the same, but for the alternative s_1, we have a new value of utility $V_1' = V_1 + v_0$.

After this addition, the original probability

$$p_1 = F(V_1 - V_2) \tag{2}$$

of selecting the alternative s_1 changes to a new value

$$p_1' = F(V_1' - V_2) = F(V_1 + v_0 - V_2). \tag{3}$$

These formulas can be simplified if we denote the difference $V_1 - V_2$ between the approximate utility values by ΔV. In these new notations, the original probability

$$p_1 = F(\Delta V) \tag{4}$$

is replaced by the new probability

$$p_1' = F(\Delta V + v_0). \tag{5}$$

This change of probability can be described in general terms: we receive new information – that there are now incentives. Based on this new information, we update our original probabilities p_i of selecting different alternatives s_i.

From the statistical viewpoint (see, e.g., (Jaynes and Bretthorst, 2003; Wadsworth, 1990)), when we receive new information, the correct way of updating probabilities is by using the Bayes formula. Specifically, if we have n incompatible hypotheses $H_1, \ldots, H_n$ with initial probabilities

$$P_0(H_1), \ldots, P_0(H_n), \tag{6}$$

then, after observations E, we update the initial probabilities to the new values:

$$P(H_i \,|\, E) = \frac{P(E \,|\, H_i) \cdot P_0(H_i)}{P(E \,|\, H_1) \cdot P_0(H_1) + \ldots + P(E \,|\, H_n) \cdot P_0(H_n)}. \tag{7}$$

Thus, we should require that the function $F(z)$ be such for which the transition from the old probability (4) to the new probability (5) can be described by the (fractionally linear) Bayes formula (7).

From the main idea to the exact formulas. Let us formalize the above requirement. In the case of two alternatives s_1 and s_2, we have two hypotheses: the hypothesis H_1 that the user will prefer s_1 and the opposite hypothesis H_2 that the user will prefer s_2. Initially, we did now know about any incentives, we only knew the approximate utility V_1 of the first alternative and the approximate utility V_2 of the second alternative. Based on the information that we initially had, we concluded that the probability of the hypothesis H_1 is equal to $p_1 = p(H_1) = F(\Delta V)$ (where $\Delta V = V_1 - V_2$), and the probability of the opposite hypothesis H_2 is equal to $p_2 = p(H_2) = 1 - p_1$.

Now, suppose that learn that there was no incentive to select alternative s_2 and an incentive of size v_0 to select alternative s_1. This new information E changes the probabilities of our hypotheses H_1 and H_2. Namely, according to Bayes formula, after the new information E, the probability p_1 should be updated to the following new value $p_1' = P(H_1 \,|\, E)$:

$$p_1' = \frac{P(E \,|\, H_1) \cdot P(H_1)}{P(E \,|\, H_1) \cdot p_1 + P(E \,|\, H_2) \cdot P(H_2)}. \tag{8}$$

The probability $P(E \,|\, H_1)$ is the conditional probability with which we can conclude that there was an incentive of size v_0 based on the fact that the user actually selected the alternative s_1. This conditional probability is, in general, different for different values v_0. To take this dependence into account, we will denote this conditional probability $P(E \,|\, H_1)$ by $A(v_0)$.

Similarly, the probability $P(E \,|\, H_2)$ is the conditional probability with which we can conclude that there was an incentive of size v_0 for alternative s_1 based on the fact that the user actually

selected the alternative s_2. This conditional probability is also, in general, different for different values v_0. To take this dependence into account, we will denote this conditional probability $P(E \,|\, H_2)$ by $B(v_0)$.

If we substitute the expressions $P(E \,|\, H_1) = A(v_0)$, $P(E \,|\, H_2) = B(v_0)$, $P(H_1) = F(\Delta V)$, and $P(H_2) = 1 - P(H_1) = 1 - F(\Delta V)$ into the above formula (8), then we conclude that

$$p_1' = \frac{A(v_0) \cdot F(\Delta V)}{A(v_0) \cdot F(\Delta V) + B(v_0) \cdot (1 - F(\Delta V))}. \tag{9}$$

On the other hand, once we know that there was an incentive v_0 to select the alternative s_1 and no incentive for the alternative s_2, then we have a better idea of the resulting utilities of the user: namely, the new value of the approximate utility is $V_1 + v_0$ for alternative s_1 and V_2 for the alternative s_2. In accordance with our expression for the choice probability based on the approximate utility values, the new probability of selecting s_1 should be equal to $F((V_1 + v_0) - V_2)$, i.e., to $F(\Delta V + v_0)$ (expression (4)).

If the probability update was done correctly, in full accordance with the Bayes formula, then this new value (4) must be equal to the value (9) that comes from using the Bayes formula. So, we arrive at the following definition:

Definition 2. *A choice function $F(z)$ is called* Bayes correct *if, for every v_0, there exist values $A(v_0)$ and $B(v_0)$ for which*

$$F(\Delta V + v_0) = \frac{A(v_0) \cdot F(\Delta V)}{A(v_0) \cdot F(\Delta V) + B(v_0) \cdot (1 - F(\Delta V))} \tag{10}$$

for all ΔV.

Comment. In other words, we require that the 2-parametric family of functions $F = \left\{ \dfrac{A \cdot F(\Delta V)}{A \cdot F(\Delta V) + B} \right\}$ corresponding to Bayesian updates be *shift-invariant* under a shift $\Delta V \to \Delta V + v_0$.

Theorem 1. *Every Bayes correct choice function $F(z)$ has the form*

$$F(\Delta V) = \frac{1}{1 + e^{-\beta \cdot \Delta V}} \tag{11}$$

for some real number β.

If we substitute $\Delta V = V_1 - V_2$ into this formula, and multiply the numerator and the denominator of the resulting formula by $e^{\beta \cdot V_1}$, then we conclude that for every Bayes correct choice function $F(z)$, we have

$$p_1 = F(V_1 - V_2) = \frac{e^{\beta \cdot V_1}}{e^{\beta \cdot V_1} + e^{\beta \cdot V_2}}. \tag{12}$$

Thus, for the desired case of two alternatives, we indeed provide a new distribution-free justification of the logit formula.

Proof. It is known that many formulas in probability theory can be simplified if instead of the probability p, we consider the corresponding odds

$$O = \frac{p}{1 - p}. \tag{13}$$

(If we know the odds O, then we can reconstruct the probability p as $p = O/(1+O)$.) The right-hand side of the formula (10) can be represented in terms of odds $O(\Delta V)$, if we divide both the numerator and the denominators by $1 - F(\Delta V)$. As a result, we get the following formula:

$$F(\Delta V + v_0) = \frac{A(v_0) \cdot O(\Delta V)}{A(v_0) \cdot O(\Delta V) + B(v_0)}. \tag{14}$$

Based on this formula, we can compute the corresponding odds $O(\Delta V + v_0)$: first, we compute the value

$$1 - F(\Delta V + v_0) = \frac{B(v_0)}{A(v_0) \cdot O(\Delta V) + B(v_0)}, \tag{15}$$

and then divide (14) by (15), resulting in:

$$O(\Delta V + v_0) = c(v_0) \cdot O(\Delta V), \tag{16}$$

where we denoted $c(v_0) \stackrel{\text{def}}{=} A(v_0)/B(v_0)$. It is known (see, e.g., (Aczel, 2006)) that all monotonic solutions of the functional equation (16) are of the form $O(\Delta V) = C \cdot e^{\beta \cdot \Delta V}$. Therefore, we can reconstruct the probability $F(\Delta V)$ as

$$F(\Delta V) = \frac{O(\Delta V)}{O(\Delta V) + 1} = \frac{C \cdot e^{\beta \cdot \Delta V}}{C \cdot e^{\beta \cdot \Delta V} + 1}. \tag{17}$$

The condition $F(z) + F(-z) = 1$ leads to $C = 1$. Dividing both the numerator and the denominator of the right-hand side by $e^{\beta \cdot \Delta V}$, we get the desired formula (11). Q.E.D.

B. Towards an Optimal Algorithm for Computing Fixed Points

Many practical situations eventually reach equilibrium. In many real-life situations, we have dynamical situations which eventually reach an equilibrium.

For example, in economics, when a situation changes, prices start changing (often fluctuating) until they reach an equilibrium between supply and demand.

In transportation, as we have mentioned, when a new road is built, some traffic moves to this road to avoid congestion on the other roads; this causes congestion on the new road, which, in its turn, leads drivers to go back to their previous routes, etc. (Sheffi, 1985).

It is often desirable to predict the corresponding equilibrium. For the purposes of the long-term planning, it is desirable to find the corresponding equilibrium. For example, for the purposes of economic planning, it is desirable to know how, in the long run, oil prices will change if we start exploring new oil fields in Alaska. For transportation planning, it is desirable to find out to what extent the introduction of a new road will relieve the traffic congestion, etc.

In order to describe how we can solve this practically important problem, let us describe this equilibrium prediction problem in precise terms.

Finding an equilibrium as a mathematical problem. To describe the problem of finding the *equilibrium* state(s), we must first be able to describe *all possible* states. In this paper, we assume that we already have such a description, i.e., that we know the set X of all possible states.

We must also be able to describe the fact that many states $x \in X$ are not equilibrium states. For example, if the price of some commodity (like oil) is set up too high, it will become profitable to explore difficult-to-extract oil fields; as a new result, the supply of oil will increase, and the prices will drop.

Similarly, as we have mentioned in the main text, if too many cars move to a new road, this road may become even more congested than the old roads initially were, and so the traffic situation will actually decrease – prompting people to abandon this new road.

To describe this instability, we must be able to describe how, due to this instability, the original state x gets transformed in the next moment of time. In other words, we assume that for every state $x \in X$, we know the corresponding state $f(x)$ at the next moment of time.

For non-equilibrium states x, the change is inevitable, so we have $f(x) \neq x$. The equilibrium state x is the state which does not change, i.e., for which $f(x) = x$. Thus, we arrive at the following problem: We are given a set X and a function $f : X \to X$; we need to find an element x for which $f(x) = x$.

In mathematical terms, an element x for which $f(x) = x$ is called a *fixed point* of the mapping f. So, there is a practical need to find fixed points.

The problem of computing fixed points. Since there is a practical need to compute the fixed points, let us give a brief description of the existing algorithms for computing these fixed points. Readers interested in more detailed description can look, e.g., in (Berinde, 2002).

Straightforward algorithm: Picard iterations. At first glance, the situation seems very simple and straightforward. We know that if we start with a state x at some moment of time, then in the next moment of time, we will get a state $f(x)$. We also know that eventually, we will get an equilibrium. So, a natural thing to do is to simulate how the actual equilibrium will be reached.

In other words, we start with an arbitrary (reasonable) state x_0. After we know the state x_k at the moment k, we predict the state x_{k+1} at the next moment of time as $x_{k+1} = f(x_k)$. This algorithm is called *Picard iterations* after a mathematician who started efficiently using it in the 19 century.

If the equilibrium is eventually achieved, i.e., if in real life the process converges to an equilibrium point x, then Picard's iterations are guaranteed to converge. Their convergence may be somewhat slow – since they simulate all the fluctuations of the actual convergence – but eventually, we get convergence.

Situations when Picard's iterations do not converge. In some important practical situations, Picard iterations do not converge.

The main reason is that in practice, we can have panicky fluctuations which prevent convergence. Of course, one expects fluctuations. For example, if the price of oil is high, then it will become profitable for companies to explore and exploit new oil fields. As a result, the supply of oil will drastically increase, and the price of oil will go down. Since this is all done in a unplanned way, with different companies making very rough predictions, it is highly probable that the resulting oil supply will exceed the demand. As a result, prices will go down, oil production in difficult-to-produce oil areas will become unprofitable, supply will go down, etc.

Such fluctuations have happened in economics in the past, and sometimes, not only they did not lead to an equilibrium, they actually led to deep economic crises.

As we have seen, similar situations happen in transportation as well.

How can we handle these situation: a natural practical solution. If the natural Picard iterations do not converge, this means that in practice, there is too much of a fluctuation. When at some moment k, the state x_k is not an equilibrium, then at the next moment of time, we have a state $x_{k+1} = f(x_k) \neq x_k$. However, this new state x_{k+1} is an not necessarily closer to the equilibrium: it "over-compensates" by going too far to the other side of the desired equilibrium.

For example, we started with a price x_k which was too high. At the next moment of time, instead of having a price which is closer to the equilibrium, we may get a new price x_{k+1} which is too low – and may even be further away from the equilibrium than the previous price.

In practical situations, such things do happen. In this case, to avoid such weird fluctuations and to guarantee that we eventually converge to the equilibrium point, a natural thing is to "dampen" these fluctuations: we know that a transition from x_k to x_{k+1} has gone too far, so we should only go "halfway" (or even smaller piece of the way) towards x_{k+1}.

How can we describe it in natural terms? In many practical situations, there is a reasonable linear structure on the set X on all the states, i.e., X is a linear space. In this case, going from x_k to $f(x_k)$ means adding, to the original state x_k, a displacement $f(x_k) - x_k$. Going halfway would then mean that we are only adding a half of this displacement, i.e., that we go from x_k to $x_{k+1} = x_k + \frac{1}{2} \cdot (f(x_k) - x_k)$, i.e., to

$$x_{k+1} = \frac{1}{2} \cdot x_k + \frac{1}{2} \cdot f(x_k).$$

The corresponding iteration process is called *Krasnoselskii iterations*. In general, we can use a different portions $\alpha \neq 1/2$, and we can also use different portions α_k on different moments of time. In general, we thus go from x_k to $x_{k+1} = x_k + \alpha_k \cdot (f(x_k) - x_k)$, i.e., to

$$x_{k+1} = (1 - \alpha_k) \cdot x_k + \alpha_k \cdot f(x_k).$$

These iterations are called *Krasnoselski-Mann iterations*.

Practical problem: the rate of convergence drastically depends on α_i. The above convergence results show that under certain conditions on the parameters α_i, there is a convergence. From the viewpoint of guaranteeing this convergence, we can select any sequence α_i which satisfies these conditions. However, in practice, different choice of α_i often result in drastically different rate of convergence.

To illustrate this difference, let us consider the simplest situation when already Picard iterations $x_{n+1} = f(x_n)$ converge, and converge monotonically. Then, in principle, we can have the same convergence if instead we use Krasnoselski-Mann iterations with $\alpha_n = 0.01$. Crudely speaking, this means that we replace each original step $x_n \to x_{n+1} = f(x_n)$, which bring x_n directly into x_{n+1}, by a hundred new smaller steps. Thus, while we still have convergence, we will need 100 times more iterations than before – and thus, we require a hundred times more computation time.

Since different values α_i lead to different rates of convergence, ranging from reasonably efficient to very inefficient, it is important to make sure that we select *optimal* values of the parameters α_i, values which guarantee the fastest convergence.

First idea: from the discrete iterations to the continuous dynamical system. In this section, we will describe the values α_i which are optimal in some reasonable sense. To describe this sense, let us go back to our description of the dynamical situation. In the above text, we considered observations made at discrete moments of time; this is why we talked about current moment of time, next moment of time, etc. In precise terms, we considered moments t_0, $t_1 = t_0 + \Delta t$, $t_2 = t_0 + 2\Delta t$, etc.

In principle, the selection of Δt is rather arbitrary. For example, in terms of prices, we can consider weekly prices (for which Δt is one week), monthly prices, yearly prices, etc. Similarly, for transportation, we can consider daily, hourly, etc. descriptions. The above discrete-time description is, in effect, a discrete approximation to an actual continuous-time system.

Similarly, Krasnoselski-Mann iterations $x_{k+1} - x_k = \alpha_k \cdot (f(x_k) - x_k)$ can be viewed as a discrete-time approximations to a continuous dynamical system which leads to the desired equilibrium. Specifically, the difference $x_{k+1} - x_k$ is a natural discrete analogue of the derivative $\frac{dx}{dt}$, the values α_k can be viewed as discretized values of an unknown function $\alpha(t)$, and so the corresponding continuous system takes the form

$$\frac{dx}{dt} = \alpha(t) \cdot (f(x) - x). \tag{18}$$

A discrete-time system is usually a good approximation to the corresponding continuous-time system. Thus, we can assume that, vice versa, the above continuous system is a good approximation for Krasnoselski-Mann iterations.

In view of this fact, in the following text, we will look for an appropriate (optimal) continuous-time system (18).

Scale invariance: natural requirement on a continuous-time system. In deriving the continuous system (18) from the formula for Krasnoselski-Mann iterations, we assumed that the original time interval Δt between the two consecutive iterations is 1. This means, in effect, that to measure time, we use a scale in which this interval Δt is a unit interval.

As we have mentioned earlier, the choice of the time interval Δt is rather arbitrary. If we make a different choice of this discretization time interval $\Delta t' \neq \Delta t$, then we would get a similar dynamical system, but described in a different time scale, with a different time interval $\Delta t'$ taken as a measuring unit. As a result of "de-discretizing" this new system, we would get a different continuous system of type (18) – a system which differs from the original one by a change in scale.

In the original scale, we identified the time interval Δt with 1. Thus, the time t in the original scale means physical time $T = t \cdot \Delta t$. In the new scale, this same physical time corresponds to the time $t' = \frac{T}{\Delta t'} = t \cdot \frac{\Delta t}{\Delta t'}$.

If we denote by $\lambda = \frac{\Delta t'}{\Delta t}$ the ratio of the corresponding units, then we conclude that the time t in the original scale corresponds to the time $t' = t/\lambda$ in the new scale. Let us describe the system (18) in terms of this new time coordinate t'. From the above formula, we conclude that $t = \lambda \cdot t'$; substituting $t = \lambda \cdot t'$ and $dt = \lambda \cdot dt'$ into the formula (18), we conclude that

$$\frac{1}{\lambda} \cdot \frac{dx}{dt'} = \alpha(\lambda \cdot t') \cdot (f(x) - x),$$

i.e., that

$$\frac{dx}{dt'} = (\lambda \cdot \alpha(\lambda \cdot t')) \cdot (f(x) - x). \tag{19}$$

It is reasonable to require that the optimal system of type (18) should not depend on what exactly time interval Δt we used for discretization.

Conclusion: optimal Krasnoselski-Mann iterations correspond to $\alpha_k = c/k$. Since a change of the time interval corresponds to re-scaling, this means the system (18) must be scale-invariant, i.e., to be more precise, the system (19) must have exactly the same form as the system (18) but with t' instead of t, i.e., the form

$$\frac{dx}{dt'} = \alpha(t') \cdot (f(x) - x). \tag{20}$$

By comparing the systems (19) and (20), we conclude that we must have

$$\lambda \cdot \alpha(\lambda \cdot t') = a(t')$$

for all t' and λ. In particular, if we take $\lambda = 1/t'$, then we get $\alpha(t') = \dfrac{\alpha(1)}{t'}$, i.e., $\alpha(t') = c/t'$ for some constant c $(= \alpha(1))$.

With respect to the corresponding discretized system, this means that we take $\alpha_k = \alpha(k) = c/k$.

Comment. The formula $\alpha_k = c/k$ is not exact: it comes form approximating the actual continuous dependence by a discrete one. This approximation makes asymptotic sense, but this formula cannot be applied for $k = 0$. To make this formula applicable, we must start with $k = 1$ – or, equivalently, start with $k = 0$ (since this is how most descriptions of iterations work), but use the expression $\alpha_k = c/(k + 1)$ instead.

Reasonable choice of the constant c and its interpretation. As we have mentioned, a reasonable idea is to use Picard iterations. This is not always a good idea, because we may get wild fluctuations. However, it makes some sense to start with the Picard iteration first, to get away from the initial state.

Picard iterations correspond to $\alpha_k = 1$; so, if we want $\alpha_0 = 1$, i.e., $c/(0 + 1) = 1$, we must take $c = 1$. The resulting iterations take the form

$$x_{k+1} = \left(1 - \frac{1}{k+1}\right) \cdot x_k + \frac{1}{k+1} \cdot f(x_k).$$

This formula (corresponding to $c = 1$) has a natural commonsense interpretation.

Namely, in Picard iterations, as a next iteration x_{k+1}, we take $f(x_k)$. When there are wild oscillations, these iterations do not converge. We expect, however, that these oscillations are going on around the equilibrium point. So, while the values x_i are oscillating and not converging at all, their averages

$$\frac{x_0 + \ldots + x_k}{k+1}$$

and the corresponding values

$$\frac{f(x_0) + \ldots + f(x_k)}{k+1}$$

will be getting closer and closer to the desired equilibrium. Thus, if we want to enhance convergence, then, instead of taking $f(x_k)$ as the next iteration, it makes sense to take an *average* of the previous values of $f(x_k)$:

$$x_{k+1} = \frac{f(x_0) + \ldots + f(x_{k-1}) + f(x_k)}{k+1}.$$

Let us show that this idea leads exactly to our choice $\alpha_k = 1/(k+1)$. Indeed, from $x_k = \frac{f(x_0) + \ldots + f(x_{k-1})}{k}$, we conclude that $f(x_0)+\ldots+f(x_{k-1}) = k \cdot x_k$, hence $f(x_0)+\ldots+f(x_{k-1})+f(x_k) = k \cdot x_k + f(x_k)$ and thus,

$$x_{k+1} = \frac{f(x_0) + \ldots + f(x_{k-1}) + f(x_k)}{k+1} = \frac{k \cdot x_k + f(x_k)}{k+1} = \left(1 - \frac{1}{k+1}\right) \cdot x_k + \frac{1}{k+1} \cdot f(x_k).$$

This selection seems to work well. The choice $a_k = 1/k$ have been successfully used and shown to be efficient. We have shown this on the example of our transportation problem. For other examples, see, e.g., (Su and Qin, 2006) and references therein.

C. Exponential Disutility Functions in Transportation Modeling: Justification

Stochastic approach, and the need to use utility or disutility functions. In real life, travel times are non-deterministic (*stochastic*): on each link, for the same capacity and flow, we may have somewhat different travel times (Sheffi, 1985).

In other words, for each link, the travel time t_i is no longer a uniquely determined real number, it is a *random variable* whose characteristics may depend on the capacity and flow along this link. As a result, the overall travel time t is also a random variable.

If we take this uncertainty into account, then it is no longer easy to predict which path will be selected: if we have two alternative paths, then it often happens that with some probability, the time along the first path is smaller, but with some other probability, the first path may turn out to be longer. How can we describe decision making under such uncertainty?

In decision making theory, it is proven that under certain reasonable assumption, a person's preferences are defined by his or her *utility function* $U(x)$ which assigns to each possible outcome x a real number $U(x)$ called *utility*; see, e.g., (Keeney and Raiffa, 1976; Raiffa, 1970). In many real-life situations, a person's choice s does not determine the outcome uniquely, we may have different outcomes $x_1, \ldots, x_n$ with probabilities, correspondingly, $p_1, \ldots, p_n$. For example, when a driver selects a path s, the travel time is often not uniquely determined: we may have different travel times $x_1, \ldots, x_n$ with corresponding probabilities $p_1, \ldots, p_n$. For such a choice, we can describe the utility $U(s)$ associated with this choice as the expected value of the utility of outcomes: $U(s) = E[U(x)] = p_1 \cdot U(x_1) + \ldots + p_n \cdot U(x_n)$. Among several possible choices, a user selects the one

for which the utility is the largest: a possible choice s is preferred to a possible choice s' (denoted $s > s'$) if and only if $U(s) > U(s')$.

For the applications presented in this paper, it is important to emphasize that the utility function is not uniquely determined by the preference relation. Namely, for every two real numbers $a > 0$ and b, if we replace the original utility function $U(x)$ with the new one $V(x) \stackrel{\text{def}}{=} a \cdot U(x) + b$, then for each choice s, we will have

$$V(s) = E[a \cdot U(x) + b] = a \cdot E[U(x)] + b = a \cdot U(s) + b$$

and thus, $V(s) > V(s')$ if and only if $U(s) > U(s')$.

In transportation, the main concern is travel time t, so the utility depends on time: $U = U(t)$. Of course, all else being equal, the longer it takes to travel, the less preferable the choice of a path; so, the utility function $U(t)$ must be strictly increasing: if $t < t'$, then $U(t) > U(t')$.

In general, decision making is formulated in terms of *maximizing* a utility function $U(x)$. In traditional (deterministic) transportation problems, however, decision making is formulated in terms of *minimization*: we select a route with the smallest possible travel time. Thus, when people apply decision making theory in transportation problems, they reformulate these problems in terms of a *disutility* function $u(x) \stackrel{\text{def}}{=} -U(x)$. Clearly, for every choice s, we have

$$u(s) \stackrel{\text{def}}{=} E[u(x)] = E[-U(x)] = -E[U(x)] = -U(s).$$

So, selecting the route with the *largest* value of expected utility $U(s)$ is equivalent to selecting the route with the *smallest* value of expected disutility $u(s)$. In line with this usage, in this paper, we will talk about disutility functions.

Since a disutility function $U(t)$ is strictly decreasing, the corresponding utility function $u(t) = -U(t)$ must be strictly increasing: if $t < t'$ then $u(t) < u(t')$.

Disutility functions traditionallly used in transportation: description and reasons. In transportation, traditionally, three types of disutility functions are used; see, e.g., (Mirchandani and Soroush, 1987; Tatineni, 1996; Tatineni et al., 1997).

First, we can use *linear* disutility functions $u(t) = a \cdot t + b$, with $a > 0$. As we have mentioned, multiplication by a constant $a > 0$ and addition of a constant b does not change the preferences, so we can safely assume that the utility function simply conincides with the travel time $u(t) = t$.

Second, we can use *risk-prone exponential* disutility functions

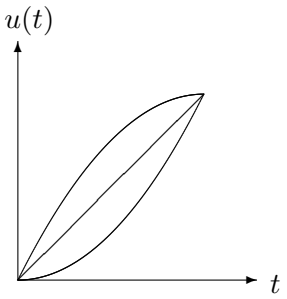
$$u(t) = -a \cdot \exp(-c \cdot t) + b$$

for some $a > 0$ and $c > 0$. This is equivalent to using $u(t) = -\exp(-c \cdot t)$.

Third, we can use *risk-averse exponential* disutility functions

$$u(t) = a \cdot \exp(c \cdot t) + b$$

for some $a > 0$ and $c > 0$. This is equivalent to using $u(t) = \exp(c \cdot t)$.



Several other possible disutility functions have been proposed, e.g., quadratic functions $u(t) = t + c \cdot t^2$; see, e.g., (Mirchandani and Soroush, 1987).

In practice, mostly linear and exponential functions are used. Actually, a linear function can be viewed as a limit of exponential functions:

$$t = \lim_{\alpha \to 0} \frac{1}{\alpha} \cdot (\exp(\alpha \cdot t) - 1),$$

so we can say that mostly exponential functions are used.

The main reason for using exponential disutility functions is that these functions are in accordance with common sense (Mirchandani and Soroush, 1987; Tatineni et al., 1997). Indeed:

- functions $-\exp(-c \cdot t)$ indeed lead to risk-prone behavior, i.e., crudely speaking, a behavior in which a person, when choosing between two paths, one with a deterministic time t_1 and another with a stochastic time t_2, prefers the second path if there is a large enough probability that $t_2 < t_1$ – even when the average time of the second path may be larger $\bar{t}_2 > t_1$;

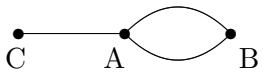
- functions $\exp(c \cdot t)$ indeed lead to risk-averse behavior, i.e., crudely speaking, a behavior in which a person, when choosing between two paths, one with a deterministic time t_1 and another with a stochastic time t_2, prefers the first path if there is a reasonable probability that $t_2 > t_1$ – even when the average time of the second path may be smaller: $\bar{t}_2 < t_1$.

This accordance, however, does not limit us to only exponential functions: e.g., quadratic functions are also in reasonably good accordance with common sense.

However, there is another common sense requirements that leads to linear or exponential functions.

A common sense assumption about the driver's preferences. Let us assume that we have several routes going from point A to point B, and a driver selected one of these routes as the best for him/her. For example, A may be a place at the entrance to the driver's department, and B is a similar department at another university located in a nearby town.

Let us now imagine a similar situation, in which the driver is also interested in reaching the point B, but this time, the driver starts at some prior point C. At this point C, there is only one possible way, and it leads to the point A; after A, we still have several possible routes. We can also assume that the time t_0 that it takes to get from C to A is deterministic. For example, C may be a place in the parking garage from where there is only one exit.

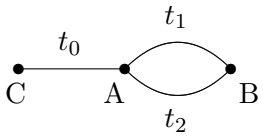


It is reasonable to assume that if the road conditions did not change, then, after getting to the point A, the driver will select the exact same route as last time, when this driver started at A.

Comment. Similarly, if two routes from A to B were equally preferable to the driver, then both routes should be equally preferable after we add a deterministic link from C to A to both routes.

In the deterministic case, this assumption is automatically satisfied. In the deterministic case, the travel time along each route is deterministic, and the driver selects a route with the shortest travel time.

Let us assume when going from A to B, the drive prefers the first route because its travel time t_1 is smaller than the travel time t_2 of the second route: $t_1 < t_2$. In this case, next time, when the travel starts from the point C, we have time $t_1 + t_0$ along the first route and $t_2 + t_0$ along the second route. Since we had $t_1 < t_2$, we thus have $t_1 + t_0 < t_2 + t_0$ – and therefore, the driver will still select the first route.



In the stochastic case, this assumption is not necessarily automatically satisfied. In the stochastic case, when going from A to B, the driver selects the first route if $E[u(t_1)] < E[u(t_2)]$, where $u(t)$ is the corresponding disutility function.

Next time, when the driver goes from C to B, the choice between the two routes depends on comparing different expected values: $E[u(t_1 + t_0)]$ and $E[u(t_2 + t_0)]$, where t_0 is the (deterministic) time of traveling from C to A. In principle, it may be possible that $E[u(t_1)] < E[u(t_2)]$ but

$$E[u(t_1 + t_0)] > E[u(t_2 + t_0)].$$

Let us describe a simple numerical example when this counter-intuitive phenomenon happens. In this example, we will use a simple non-linear disutility function: namely, the quadratic function $u(t) = t^2$. Let us assume that the first route from A to B is deterministic, with $t_1 = 7$, and the second route from A to B is highly stochastic: with equal probability 0.5, we may have $t_2 = 1$ and $t_2 = 10$. In this case, $E[u(t_1)] = t_1^2 = 49$ and

$$E[u(t_2)] = E[t_2^2] = \frac{1}{2} \cdot 1^2 + \frac{1}{2} \cdot 10^2 = 0.5 + 50 = 50.5.$$

Here, $E[u(t_1)] < E[u(t_2)]$, so the driver will prefer the first route.

However, if we add the same constant time $t_0 = 1$ for going from C to A to both routes, then in the first route, we will have $t_1 + t_0 = 7 + 1 = 8$, while in the second route, we will have $t_2 + t_0 = 1 + 1 = 2$ and $t_2 + t_0 = 10 + 1 = 11$ with equal probability 0.5. In this case,

$$E[u(t_1 + t_0)] = (t_1 + t_0)^2 = 8^2 = 64,$$

while

$$E[u(t_2+t_0)] = \frac{1}{2}\cdot 2^2 + \frac{1}{2}\cdot 11^2 = 2 + 60.5 = 62.5.$$

We see that here, $E[u(t_2+t_0)] < E[u(t_1+t_0)]$, i.e., the drive will select the second route instead of the first one.

This counter-intuitive phenomenon does not happen for linear or exponential disutility functions. Indeed, for a linear disutility function $u(t) = t$, we have $u(t_1+t_0) = t_1+t_0 = u(t_1)+t_0$; therefore, $E[u(t_1+t_0)] = E[u(t_1)]+t_0$ and similarly, $E[u(t_2+t_0)] = E[u(t_2)]+t_0$. Thus, if the driver selected the first route, i.e., if $E[u(t_1)] < E[u(t_2)]$, then by adding t_0 to both sides of this inequality, we can conclude that $E[u(t_1+t_0)] < E[u(t_2+t_0)]$ – i.e., that, in accordance with common sense, the same route will be selected if we start at the point C.

For the exponential disutility function $u(t) = \exp(\alpha\cdot t)$, we have $u(t_1+t_0) = \exp(\alpha\cdot(t_1+t_0)) = \exp(\alpha\cdot t_1)\cdot\exp(\alpha\cdot t_0)$ and therefore, $u(t_1+t_0) = u(t_1)\cdot\exp(\alpha\cdot t_0)$. Similarly, for the exponential disutility function $u(t) = -\exp(\alpha\cdot t)$, we have $u(t_1+t_0) = -\exp(\alpha\cdot(t_1+t_0)) = -\exp(\alpha\cdot t_1)\cdot\exp(\alpha\cdot t_0)$ and thus, $u(t_1+t_0) = u(t_1)\cdot\exp(\alpha\cdot t_0)$;

For both types of exponential disutility function, we have $E[u(t_1+t_0)] = \exp(\alpha\cdot t_0)\cdot E[u(t_1)]$ and similarly, $E[u(t_2+t_0)] = \exp(\alpha\cdot t_0)\cdot E[u(t_2)]$. Thus, if the driver selected the first route, i.e., if $E[u(t_1)] < E[u(t_2)]$, then by multiplying both sides of this inequality by the same constant $\exp(\alpha\cdot t_0)$, we can conclude that $E[u(t_1+t_0)] < E[u(t_2+t_0)]$ – i.e., that, in accordance with common sense, the same route will be selected if we start at the point C.

Resulting justification of exponential utility fuinctions. It turns out linear and exponential disutility functions are the only ones which are consistent with the above common sense requirement – for every other disutility function, a paradoxical counter-intuitive situation like the one described above is quite possible.

Let us describe this result in precise terms.

Definition 3. *By a* disutility function*, we mean a strictly increasing function $u(t)$ from non-negative real numbers to real numbers.*

Definition 4. *We say that two disutility functions $u(t)$ and $v(t)$ are* equivalent *if there exist real numbers $a > 0$ and b such that $v(t) = a\cdot u(t) + b$ for all t.*

Definition 5. *We say that a disutility function is* consistent with common sense *if it has the following property: let t_1 and t_2 be random variables with non-negative values, and let t_0 be an arbitrary (deterministic) non-negative real number; then,*

- *if $E[u(t_1)] < E[u(t_2)]$, then $E[u(t_1+t_0)] < E[u(t_2+t_0)]$;*
- *if $E[u(t_1)] = E[u(t_2)]$, then $E[u(t_1+t_0)] = E[u(t_2+t_0)]$.*

Theorem 2. *A disutility function is consistent with common sense if and only if it is equivalent to either the linear function $u(t) = t$, or to an exponential function $u(t) = \exp(c\cdot t)$ or $-\exp(-c\cdot t)$.*

Proof. Under an additional conditions of differentiability of the function $u(t)$, this result has been proven in (Pratt, 1964). For reader's convenience, we provide a new proof which does not require differentiability.

1°. We already know that linear and exponential disutility functions are consistent with common sense in the sense of Definition 5. It is therefore sufficient to prove that every disutility function $u(t)$ which is consistent with common sense is equivalent either to a linear one or to an exponential one.

2°. Let $u(t)$ be a disutility function which is consistent with common sense. By definition of computational simplicity, for every random variables t_1, once we know the values $u_1 = E[u(t_1)]$ and t_0, we can uniquely determine the value $E[u(t_1 + t_0)]$. Let us denote the value $E[u(t_1 + t_0)]$ corresponding to u_1 and t_0 by $F(u_1, t_0)$.

3°. Let t_1' be a non-negative number. For the case when $t_1 = t_1'$ with probability 1, we have $u_1' = E[u(t_1)] = u(t_1')$. In this case, $t_1 + t_0 = t_1' + t_0$ with probability 1, so $E[u(t_1 + t_0)] = u(t_1' + t_0)$. Thus, in this case, $u(t_1' + t_0) = F(u_1', t_0)$, where $u_1' = u(t_1')$.

4°. Let us now consider the case when t_1 is equal to t_1' with some probability $p_1' \in [0, 1]$, and to some smaller value $t_1'' < t_1'$ with the remaining probability $p_1'' = 1 - p_1'$. In this case,

$$u_1 = E[u(t_1)] = p_1' \cdot u(t_1') + (1 - p_1') \cdot u(t_1'').$$

We have already denoted $u(t_1')$ by u_1'; so, if we denote $u_1'' \stackrel{\text{def}}{=} u(t_1'')$, we can rewrite the above expression as

$$u_1 = p_1' \cdot u_1' + (1 - p_1') \cdot u_1''.$$

In this situation, $t_1 + t_0$ is equal to $t_1' + t_0$ with probability p_1' and to $t_1'' + t_0$ with probability $1 - p_1'$. Thus,

$$E[u(t_1 + t_0)] = p_1' \cdot u(t_1' + t_0) + (1 - p_1') \cdot u(t_1'' + t_0).$$

We already know that $u(t_1' + t_0) = F(u_1', t_0)$ and $u(t_1'' + t_0) = F(u_1'', t_0)$. So, we can conclude that

$$E[u(t_1 + t_0)] = p_1' \cdot F(u_1', t_0) + (1 - p_1') \cdot F(u_1'', t_0). \tag{21}$$

On the other hand, by the definition of the function F as $F(u_1, t_0) = E[u(t_1 + t_0)]$, we conclude that

$$E[u(t_1 + t_0)] = F(u_1, t_0),$$

i.e.,

$$E[u(t_1 + t_0)] = F(p_1' \cdot u_1' + (1 - p_1') \cdot u_1'', t_0). \tag{22}$$

Comparing the expressions (21) and (22) for $E[u(t_1 + t_0)]$, we conclude that

$$F(p_1' \cdot u_1' + (1 - p_1') \cdot u_1'', t_0) = p_1' \cdot F(u_1', t_0) + (1 - p_1') \cdot F(u_1'', t_0).$$

Let us analyze this formula. For every value $u_1 \in [u_1'', u_1']$, we can find the probability p_1' for which $u_1 = p_1' \cdot u_1' + (1 - p_1') \cdot u_1''$: namely, the desired equation means that $u_1 = p_1' \cdot u_1' + u_1'' - p_1' \cdot u_1''$; rearranging the terms, we get $u_1 - u_1'' = p_1' \cdot (u_1' - u_1'')$ and hence, the value $p_1' = \dfrac{u_1 - u_1''}{u_1' - u_1''}$. Substituting this expression into the above formula, we conclude that for a fixed t_0, the function $F(u_1, t_0)$ is a linear function of u_1:

$$F(u_1, t_0) = A(t_0) \cdot u_1 + B(t_0)$$

for some constants $A(t_0)$ and $B(t_0)$ which, in general, depend on t_0.

5°. We have already shown, in Part 3 of this proof, that $u(t'_1 + t_0) = F(u'_1, t_0)$. Thus, we conclude that for every $t'_1 \geq 0$ and $t_0 \geq 0$, we have

$$u(t'_1 + t_0) = A(t_0) \cdot u(t'_1) + B(t_0).$$

6°. For an arbitrary function $u(t)$, by introducing an appropriate constant $b = -u(0)$, we can always find an equivalent function $v(t)$ for which $v(0) = 0$. So, without losing generality, we can assume that $u(0) = 0$ for our original disutility function $u(t)$.

Since the disutility function is strictly increasing, we have $u(t) > 0$ for all $t > 0$.

For $t'_1 = 0$, the above formula takes the form $u(t_0) = B(t_0)$. Substituting this expression for $B(t_0)$ into the above formula, we conclude that

$$u(t'_1 + t_0) = A(t_0) \cdot u(t'_1) + u(t_0).$$

7°. The above property has to be true to arbitrary values of $t'_1 \geq 0$ and $t_0 \geq 0$. Swapping these values, we conclude that

$$u(t_0 + t'_1) = A(t'_1) \cdot u(t_0) + u(t'_1).$$

Since $t'_1 + t_0 = t_0 + t'_1$, we have $u(t'_1 + t_0) = u(t_0 + t'_1)$, hence

$$A(t_0) \cdot u(t'_1) + u(t_0) = A(t'_1) \cdot u(t_0) + u(t'_1).$$

Moving terms proportional to $u(t'_1)$ to the left hand side and terms proportional to $u(t_0)$ to the right hand side, we conclude that

$$(A(t_0) - 1) \cdot u(t'_1) = (A(t'_1) - 1) \cdot u(t_0). \tag{23}$$

In the following text, we will consider two possible situations:

- the first situation is when $A(t_0) = 1$ for some $t_0 > 0$;
- the second situation is when $A(t_0) \neq 1$ for all $t_0 > 0$.

In the first situation, $A(t_0) = 1$ for some $t_0 > 0$. For this t_0, the equation (23) takes the form $(A(t'_1) - 1) \cdot u(t_0) = 0$ for all t'_1. Since $u(t_0) > 0$ for $t_0 > 0$, we conclude that $A(t'_1) - 1 = 0$ for every real number $t'_1 \geq 0$, i.e., that the function $A(t)$ is identical to a constant function 1.

So, we have two possible situations:

- the first situation is when $A(t_0) = 1$ for some $t_0 > 0$; we have just shown that in this case, $A(t) = 1$ for all t; in the following text, we will show that in this situation, the disutility function $u(t)$ is linear;
- the second situation is when $A(t_0) \neq 1$ for all $t_0 > 0$; we will show that in this situation, the disutility function $u(t)$ is exponential.

8°. Let us first consider the situation in which $A(t)$ is always equal to 1. In this case, the above equation takes the form

$$u(t_0 + t_1') = u(t_0) + u(t_1').$$

In other words, in this case,

$$u(t_1 + t_2) = u(t_1) + u(t_2)$$

for all possible values $t_1 > 0$ and $t_2 > 0$.

In particular, for every $t_0 > 0$, we get:

- first, $u(2t_0) = u(t_0) + u(t_0) = 2u(t_0)$,
- then $u(3t_0) = u(2t_0) + u(t_0) = 2u(t_0) + u(t_0) = 3u(t_0)$, and,
- in general, $u(k \cdot t_0) = k \cdot u(t_0)$ for all integers k.

For every integer n and for $t_0 = 1/n$, we have $u(n \cdot t_0) = u(1) = n \cdot u(1/n)$, hence $u(1/n) = u(1)/n$. Then, for an arbitrary non-negative rational number k/n, we get

$$u(k/n) = u(k \cdot (1/n)) = k \cdot u(1/n) = k \cdot (1/n) \cdot u(1) = k/n \cdot u(1).$$

In other words, for every rational number $r = k/n$, we have $u(r) = r \cdot u(1)$.

Every real value t can be bounded, with arbitrary accuracy, by rational numbers k_n/n and $(k_n + 1)/n$: $k_n/n \le t \le (k_n + 1)/n$, where $k_n/n \to t$ and $(k_n + 1)/n \to t$ as $n \to \infty$. Since the disutility function $u(t)$ is strictly increasing, we conclude that $u(k_n/n) \le u(t) \le u((k_n + 1)/n)$. We already know that for rational values r, we have $u(r) = r \cdot u(1)$, so we have

$$k_n/n \cdot u(1) \le u(t) \le (k_n + 1)/n \cdot u(1).$$

In the limit $n \to \infty$, both sides of this inequality converge to $t \cdot u(1)$, hence $u(t) = t \cdot u(1)$.

So, in this case, we get a linear disutility function.

9°. Let us now analyze the case when $A(t) \ne 1$ for all $t > 0$. Since the values $u(t)$ are positive for all $t > 0$, we can divide both sides of the equality

$$(A(t_0) - 1) \cdot u(t_1') = (A(t_1') - 1) \cdot u(t_0)$$

by $u(t_0)$ and $u(t_1')$, and conclude that

$$\frac{A(t_0) - 1}{u(t_0)} = \frac{A(t_1') - 1}{u(t_1')}.$$

The ratio $\dfrac{A(t) - 1}{u(t)}$ has the same value for arbitrary two numbers $t = t_0$ and $t = t_1'$; thus, this ratio is a constant. Let us denote this constant by k; then, $A(t) - 1 = k \cdot u(t)$ for all $t > 0$. Since $A(t) \ne 1$, this constant k is different from 0.

Substituting the resulting expression $A(t) = 1 + k \cdot u(t)$ into the formula $u(t_1' + t_0) = A(t_0) \cdot u(t_1') + u(t_0)$, we conclude that

$$u(t_1' + t_0) = u(t_0) + u(t_1') + k \cdot u(t_0) \cdot u(t_1'),$$

i.e., that

$$u(t_1 + t_2) = u(t_1) + u(t_2) + k \cdot u(t_1) \cdot u(t_2)$$

for arbitrary numbers $t_1 > 0$ and $t_2 > 0$.

10°. Let us now consider a re-scaled function $v(t) \stackrel{\text{def}}{=} 1 + k \cdot u(t)$.

For this function $v(t)$, from the above formula, we conclude that

$$v(t_1 + t_2) = 1 + k \cdot u(t_1 + t_2) = 1 + k \cdot (u(t_1) + u(t_2)) + k^2 \cdot u(t_1) \cdot u(t_2).$$

On the other hand, we have

$$v(t_1) \cdot v(t_2) = (1 + k \cdot u(t_1)) \cdot (1 + k \cdot u(t_2)) =$$

$$1 + k \cdot (u(t_1) + u(t_2)) + k^2 \cdot u(t_1) \cdot u(t_2).$$

The expression for $v(t_1 + t_2)$ and for $v(t_1) \cdot v(t_2)$ coincide, so we conclude that

$$v(t_1 + t_2) = v(t_1) \cdot v(t_2)$$

for all possible values $t_1 > 0$ and $t_2 > 0$.

11°. When $k > 0$, then the new function $v(t)$ is an equivalent disutility function. We know that $u(0) = 0$ hence $v(0) = 1 + k \cdot 0 = 1$. Since $v(t)$ is a strictly increasing function, we thus conclude that $v(t) \geq v(0) > 0$ for all $t \geq 0$.

Thus, we can take a logarithm of all the values, and for the new function $w(t) \stackrel{\text{def}}{=} \ln(v(t))$, get an equation

$$w(t_1 + t_2) = \ln(v(t_1 + t_2)) = \ln(v(t_1) \cdot v(t_2)) = \ln(v(t_1)) + \ln(v(t_2)) = w(t_1) + w(t_2),$$

i.e., $w(t_1 + t_2) = w(t_1) + w(t_2)$ for all t_1 and t_2. The function $w(t)$ is increasing – as the logarithm of an increasing function. Thus, as we have already shown, $w(t) = c \cdot t$ for some $c > 0$.

From the logarithm $w(t) = \ln(v(t))$, we can reconstruct the original disutility function $v(t)$ as $v(t) = \exp(w(t))$. Since $w(t) = c \cdot t$, we conclude that the disuility function $v(t)$ has the desired risk-averse exponential form

$$v(t) = \exp(c \cdot t).$$

12°. When $k < 0$, the new function is strictly decreasing (and is thus not a disutility function; its opposite $-v(t)$ is a disutility function).

For the function $v(t)$, we cannot have $v(t_0) = 0$ for any t_0 – because otherwise we would have

$$v(t) = v(t_0 + (t - t_0)) = v(t_0) \cdot v(t - t_0) = 0$$

for all $t \geq t_0$ which contradicts to our conclusion that the function $v(t)$ should be strictly decreasing.

13°. For the function $v(t)$, we cannot have $v(t_0) < 0$ for any $t_0 > 0$ – because otherwise, we would have $v(2t_0) = v(t_0)^2 > 0$ hence $v(2t_0) > v(t_0)$ – which, since $2t_0 > t_0$, also contradicts to our conclusion that the function $v(t)$ should be strictly decreasing.

We thus conclude that $v(t) > 0$ for all t.

14°. Thus, we can take a logarithm of all the values, and for the new function $w(t) \stackrel{\text{def}}{=} \ln(v(t))$, get the equation $w(t_1 + t_2) = w(t_1) + w(t_2)$ for all t_1 and t_2. The function $w(t)$ is decreasing – as the logarithm of a decreasing function. Thus, $w(t) = -c \cdot t$ for some $c > 0$.

From the logarithm $w(t) = \ln(v(t))$, we can reconstruct the original function $v(t)$ as $v(t) = \exp(w(t)) = \exp(-c \cdot t)$, and the disutility function $u(t)$ as $-v(t) = -\exp(-c \cdot t)$.

So, we conclude that the disuility function $v(t)$ has the desired risk-prone exponential form $v(t) = -\exp(-c \cdot t)$.

The theorem is proven.

Extended precision with a rounding mode toward zero environment. Application on the CELL processor

Hong Diep Nguyen, Stef Graillat, and Jean-Luc Lamotte

CNRS, UMR 7606, LIP6, University Pierre et Marie Curie
4 place Jussieu, 75252 Paris cedex 05, France
email:hong.diep.nguyen@ens-lyon.fr, stef.graillat@lip6.fr, and Jean-Luc.Lamotte@lip6.fr

Abstract. In the field of scientific computing, the exactitude of the calculation is of prime importance. That leads to efforts made to increase the precision of the floating-point algorithms. One of them is to increase the precision of the floating-point number to double or quadruple the working precision. The building block of these efforts is the error-free transformations.

CELL processor is a microprocessor architecture jointly developed by a Sony, Toshiba, and IBM. Although its first major commercial application of Cell was in Sony's PlayStation 3 game console, it can provide a great potential for scientific computing with a peak single precision performance of 204.8 Gflop/s.

In this paper, we will do the study on how to implement the double working precision library, named single-single, on the SPEs (Synergistic Processing Element), the workhorse processors of the CELL. The methodology of this implementation is based on the paper of Yozo Hida, Xiaoye S. Li, and David H. Bailey, titled "Algorithms for quad-double precision floating point arithmetic".

To improve the performance, the FPU of the SPE supports only the truncation rounding. So all the floating point operations used in the implementation of the library can only use this rounding mode, which requires to make some modifications to the algorithms. That increases the complexity of the implementation. However by taking advantage of the characteristics of the SPE processor, among which the most important are the fully pipelined set of instructions in single precision and the FMA (Fused Multiplier-Add) function, we have managed to implement the error-free transformations very effectively, even more quickly than the ones used in the paper (Hida et al., 2001). With the SIMD characteristic, we can perform 4 operations at the same time. We also prove the exactitude of our modified error-free transformation, and the precision of our floating-point arithmetics by providing error bounds.

Even though the theoretical peak performance of the library is much less than the performance of the real double precision of the machine, which is about 2.7 Gflop/s in comparison with the 14.4 Gflop/s of the real double precision, the results of our test show that it is not such that bad. In the best case, the performance of our library and the performance of the real double are nearly equal. With the same approach, in the future, we will promote our work to the quad-single precision, which is very promising because the CELL processor does not support the quad precision.

Keywords: extended precision, rounding mode toward zero, CELL processor

1. Introduction

The CELL processor jointly developed by Sony, Toshiba, and IBM provides a great power of calculation with a peak performance in single precision of 204.8 Gflop/s. This performance is obtained with a set of SIMD processors which use single precision floating point numbers with rounding mode toward zero. The goal of our work is to develop extended precision libraries for this architecture.

In this paper we will study how to implement the double working precision library named single-single on the SPEs (Synergistic Processing Element) which are the workhorse processors of the CELL. Our approach is similar to those used in (Hida et al., 2001) for the quad-double precision arithmetic in the rounding mode to the nearest. The next CELL generation will provide powerfull computing power in double precision with a rounding toward zero. Our library will be easily fit into double-double library which will emulate the quad precision.

This paper begins with a brief introduction to the CELL processor, then we propose algorithms for the operators $(+, -, \times, /)$ of extended precision based on the error-free transformations for the rounding mode toward zero. The next section is devoted to the implementation of the single-single library on the SPE by taking into account the advantages of the SIMD characteristics, among which the most important are the fully pipelined single precision instructions set and the FMA (Fused Multiply-Add). Finally, the numerical experiments and the test results showing the library performance are presented.

2. Introduction to CELL processor

The CELL processor (Kahle et al., 2005; Williams et al., 2006) is composed of one "Power Processor Element" (PPE) and eight "Synergistic Processing Elements" (SPE). The PPE and SPEs are linked together by an internal high speed bus called "Element Interconnect Bus" (EIB).

The PPE is based on the Power Architecture. Despite its important computing power, in practical use, it only serves as a controller for the eight SPEs which perform most of the computational workload.

The SPE is composed of an "Synergistic Processing Unit" (SPU) and a "Memory Flow Controller" (MFC) which is devoted to the memory transfer via the DMA access. The SPE contains an SIMD processor for single and double precision (Jacobi et al., 2005; Gschwind et al., 2006), which can perform at the same time 4 operations in single precision or 2 operations in double precision. It supports all the 4 rounding modes for the double precision and only the rounding mode toward zero for the single precision.

The instruction set in single precision of the SPE is fully pipelined, one instruction can be issued for each clock cycle. It is based on the FMA function, which calculates the term $a * b + c$ in one operation and one rounding. So with a frequency of 3.2 GHz, each SPE can achieve the performance of $2 \times 4 \times 3.2 = 25.6$ GFLOPs on single precision numbers.

For the double precision, the instruction set is not fully pipelined. It is only possible to issue one instruction for each 7 cycles. So the peak performance of each SPE for the double precision is: $2 \times 2 \times 3.2/7 = 1.8$ GFLOPs.

Each SPE has a "Local Storage" (LS) of 256 KB for both data and code. In the opposite of the cache memory managment, there is no mechanism to load data in the LS. It is up to the programmer to explicit data transfer via DMA function call. The SPE possess a large set of registers (128 128-bits registers) which can be used directly by the program avoiding the load-and-store time.

3. Floating-point arithmetic and extended precision

In this section we briefly introduce the floating-point arithmetic and the methodology to extend the precision. In this paper, due to the specific environment of the CELL processor, we work only with the rounding mode toward zero.

In a computer, the set of floating-point numbers denoted $\mathbb{F}$ is the most frequently used to represent real numbers. A binary floating-point number is described as follows:

$$x = (\pm) \underbrace{1.x_1 \dots x_{p-1}}_{mantissa} \times 2^e, x_i \in \{0, 1\},$$

with p the precision and e the exponent of x. We use $\varepsilon = 2^{1-p}$ as the machine precision, and the value corresponding to the last bit of x is called *unit in the last place*, denoted $ulp(x)$ and $ulp(x) = 2^{e-p+1}$.

Let x, y be two floating-point numbers, $\circ$ be a floating-point operation ($\circ \in \{+, -, \times, /\}$). It is clear that $(x \circ y)$ is a real number but in most cases it is not representable by a floating-point number. Let $fl(x \circ y)$ be the representative floating-point number of $(x \circ y)$ obtained by a rounding. The difference between $(x \circ y)$ and $fl(x \circ y)$ corresponds to the rounding error denoted $err(x \circ y)$.

Given a specific machine precision, the precision of calculation can be increased by software. Instead of using a floating-point number, multiple floating-point numbers can be used to represent multiple parts of a real number. This is the idea of the extended precision. In our case, a single-single is defined as follows:

Definition 1. A single-single is a non-evaluated sum of 2 single precision floating-point numbers. The single-single represents the exact sum of these two floating-point numbers:

$$a = a_h + a_l.$$

There may be multiple couples of 2 floating-point numbers whose sums are equal. To ensure a unique representation, a_h and a_l should have the same sign and require to satisfy:

$$|a_l| < ulp(a_h). \tag{1}$$

To implement the extended precision we have to calculate the error produced by single precision operations by using the error-free transformations presented below.

3.1. The error-free transformations (EFT)

Let x, y be two floating-point numbers and $\circ$ be a floating-point operation. The error-free transformations are intended to calculate the rounding error caused by this operation. The EFTs transform $(x \circ y)$ into a couple of two floating-point numbers (r, e) so that:

$$r \approx x \circ y \quad \text{and} \quad r + e = x \circ y.$$

3.1.1. *Accurate sum*

There are two main algorithms for the accurate sum of two floating point numbers. For example for the rounding mode to nearest, there is the algorithm proposed by Knuth (Knuth, 1998) which uses 6 standard operations, or the algorithm proposed by Dekker (Dekker, 1971) which uses only 3 standard operations but with the assumption on the order between the absolute values of two input numbers.

In this paper, our work focuses only on the rounding mode toward zero. So, it is necessary to adapt these algorithms. Priest (Priest, 1992) has proposed an algorithm for an accurate sum using a rounding mode toward zero. To better use the pipelines, we proposed another algorithm.

Algorithm 1. Error-free transformation for the sum with rounding toward zero.

```
Two–Sum–toward–zero2 (a,b)
    if (|a| < |b|)
        swap(a,b)
    s = fl(a + b)
    d = fl(s - a)
    e = fl(b - d)
    if (|2 * b| < |d|)
        s = a, e = b
return (s,e)
```

The exactitude of the newly proposed algorithm is provided in the following theorem.

Theorem 1. Let a, b be two floating-point numbers. The result of `Two-Sum-toward-zero2` (s, e) in applying on a, b satisfies:

$$\begin{aligned} s + e &= a + b, \\ |e| &< ulp(s). \end{aligned}$$

The proof of all the theorems of this paper can be found in (Nguyen, 2007) (in french).

3.1.2. *Accurate product*

The calculation of the error-free transformation for the product is much more complicated than the sum (Dekker, 1971). But if the processor has a FMA (Fused Multiply-Add) which calculates the term $a * b + c$ in one operation then the classic algorithm for the product can be used.

Algorithm 2. The error-free transformation for the product of two floating-point numbers.

```
Two–Product–FMA (a,b)
    p = fl(a * b)
    e = fma(a,b,-p)
    return (p,e)
```

This algorithm is applicable for all the four rounding modes. The basic operation on the SIMD unit of the SPE being a FMA, our libray implements this algorithm.

4. Basic operations of single-single

4.1. Renormalisation

Using the EFTs toward zero, we can implement the basic operations for the single-single. Most of the algorithms described hereafter often produce an intermediate result of two overlapping floating point numbers. To respect the definition of the normalisation (1), it is necessary to apply a renormalisation step to transform these two floating-point numbers into a normalised single-single. The following function is proposed:

```
1  Renormalise2-toward-zero (a,b)
2      if (|a| < |b|)
3          swap(a,b)
4      s = fl(a + b)
5      d = fl(s - a)
6      e = fl(b - d)
7  return (s,e)
```

It is interesting to note that the renormalisation is the same for the rounding mode toward zero and to the nearest. But in the case of the rounding mode toward zero, it is not possible to give an exact result. The following theorem provides an error bound for this algorithm.

Theorem 2. Let a, b be two floating-point numbers. The result returned by `Renormalise2-toward-zero` is a couple of two floating-point numbers (s, e) which satisfies:

- s, e have the same sign and $|e| < ulp(s)$,
- $a + b = s + e + \delta$, where δ is error of normalisation and $|\delta| \leq \frac{1}{2}\varepsilon^2|a + b|$.

As we will see later, this error is much smaller than the errors produced by the following algorithms. To describe them, we use the notations in figure 4.1.

4.1.1. *Addition*

The figure 2 represents the algorithm for the addition of two single-singles a, b. The source code is as follows:

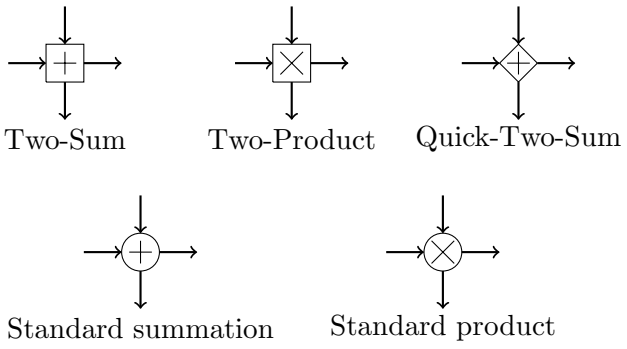


Figure 1. Notations

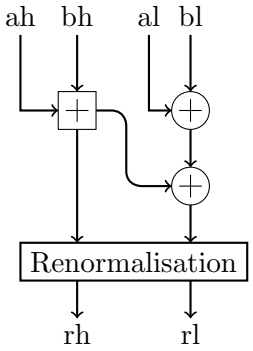


Figure 2. Algorithm for the addition of two single-singles

```
1       add_ds_ds (ah, al, bh, bl)
2          (th, tl) = Two-Sum-toward-zero (ah, bh)
3          tll = fl(al + bl)
4          tl = fl(tl + tll)
5          (rh, rl) = Renormalise2-toward-zero (th, tl)
6       return (rh, rl)
```

With two sums, a `Two-Sum-toward-zero` and a `Renormalise2-toward-zero`, the cost of the `add_ds_ds` algorithm is 11 FLOPs. The following theorem provides an error bound for this algorithm.

Theorem 3. Let $a_h + a_l$ and $b_h + b_l$ be two input single-singles and $r_h + r_l$ be the result of `add_ds_ds`. The error produced by this algorithm δ satisfies:

$$r_h + r_l = (a_h + a_l) + (b_h + b_l) + \delta,$$

$$|\delta| < max(2^{-23} * |a_l + b_l|, 2^{-43} * |a_h + a_l + b_h + b_l|) + 2^{-45} * |a_h + a_l + b_h + b_l|.$$

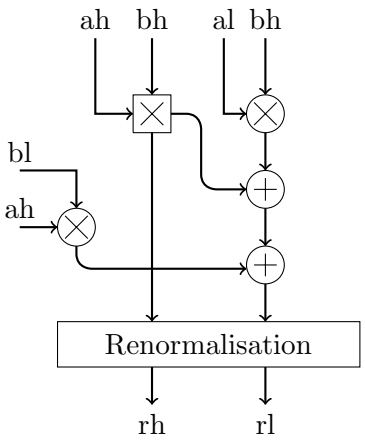


Figure 3. Algorithm for the product of two single-singles

4.1.2. *The subtraction*
The subtraction of two single-singles $a - b$ is implemented by a sum $a + (-b)$. To compute the opposite of a single-single, it is just necessary to get the opposite of the floating-point components. Therefore, the algorithms for the addition and the subtraction are similar.

4.1.3. *Product*
The product between two single-singles a and b can be considered as the product of two sum $a_h + a_l$ and $b_h + b_l$ so the exact product has 4 components:

$$\begin{aligned} p &= (a_h + a_l) \times (b_h + b_l) \\ &= a_h \times b_h + a_l \times b_h + a_h \times b_l + a_l \times b_l. \end{aligned}$$

Considering $a_h \times b_h$ as a term of order $\mathcal{O}(1)$, this product consists of 1 term $\mathcal{O}(1)$, 2 terms $\mathcal{O}(2)$ and 1 term $\mathcal{O}(3)$. To decrease the complexity of the algorithm the terms of order below $\mathcal{O}(2)$ will not be taken into account. Additionally, using the EFT for the product, $a_h \times b_h$ can be transformed exactly into 2 floating-point numbers of orders $\mathcal{O}(1)$ and $\mathcal{O}(2)$ respectively. So the product of two single-singles can be approximated by:

$$p \approx \underbrace{fl(a_h \times b_h)}_{\mathcal{O}(1)} + \underbrace{(err(a_h \times b_l) + a_l \times b_h + a_h \times b_l)}_{\mathcal{O}(2)}.$$

This approximation can be easily translated into the following algorithm:

```
1 mul_ds_ds (ah, al, bh, bl)
2       (th, tl) = Two-Product-FMA (ah, bh)
3       tll = fl(al * bh)
4       tll = fl(ah * bl + tll)
5       tl = fl(tl + tll)
```

```
6       (rh, rl) = Renormalise2-toward-zero (th, tl)
7       return (rh, rl)
```

This algorithm is described in figure 3.

The error bound of the algorithm `mul_ds_ds` is provided by the following theorem.

Theorem 4. Let $a_h + a_l$ and $b_h + b_l$ be two single-singles. Let $r_h + r_l$ be the result returned by the algorithm `mul_ds_ds` applying to $a_h + a_l$ and $b_h + b_l$. The error of this algorithm called δ satisfies:

$$|(r_h + r_l) - (a_h + a_l) \times (b_h + b_l)| < 2^{-43} \times |(a_h + a_l) \times (b_h + b_l)| + 9 \times 2^{-68} \times |(a_h + a_l) \times (b_h + b_l)|.$$

4.1.4. *The division*

The division of two single-singles is calculated by using the classic division algorithm.

Let $a = (a_h, a_l)$ and $b = (b_h, b_l)$ be two single-singles. To calculate the division of a by b, at first we calculate the approximate quotient by: $q_h = a_h/b_h$.

Then we calculate the residual $r = a - q_h \times b$, which allows to calculate the correction term by: $q_l = r/b_h$.

It can be written in detail as follows:

```
1  div_ds_ds(a,b)
2       ph = fl(ah / bh)
3       tmp1 = fl(ah - qh * bh)
4       tmp2 = fl(al - qh * bl)
5       r = fl(tmp1 + tmp2)
6       p1 = fl(r / bh)
7       (qh,ql) = Renormalise2-toward-zero(ph, pl)
8       return (qh,ql)
```

We also provide the following theorem to estimate the error of this algorithm.

Theorem 5. Let $a = (a_h, a_l)$ and $b = (b_h, b_l)$ be two single-singles, ε the machine precision, and ε_1 the error bound for the single precision division with $\mathcal{O}(\varepsilon_1) = \mathcal{O}(\varepsilon)$. The error of the algorithm `div_ds_ds` is bounded by:

$$|div\_ds\_ds(a,b)) - a/b| < [\varepsilon^2 \times (6.5 + 7 \times \varepsilon_1/\varepsilon + 2 \times (\varepsilon_1/\varepsilon)^2) + \mathcal{O}(\varepsilon^3))] \times |a/b|.$$

In most of cases we have $\varepsilon_1 = \varepsilon$. In this case, the error bound of this algorithm is:

$$|q - a/b| < [15.5 \times \varepsilon^2 + \mathcal{O}(\varepsilon^3)] \times |a/b|.$$

This inequality means that our division algorithm of two single-singles is accurate to 42 bits on a maximum of 48 bits. The accuracy of this algorithm can be increased by calculating another correction term q_2 but it has a great impact on the performance.

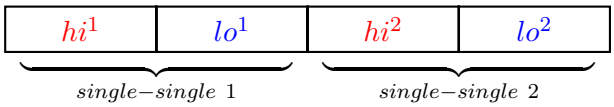


Figure 4. A vector of 2 single-singles

5. Implementation

The SPE (Synergistic Processor Element) of CELL processor contains a 32-bit 4-way SIMD processor together with a large set of 128 128-bits registers. It can perform the operations on the vectors of 16 `char` /`unsigned char`, 4 `int`/`unsigned int`, 4 `float` or 2 `double`.

The operations on scalars are implemented by using the vectorial operations. In this case, only one operation is performed on the preferred slot instead of 4 on vectors. For this reason, we implement only the vectorial operations for the single-singles.

5.1. Representation

A single-single is a couple of two floating-point numbers so each vector of 128 bits contains two single-singles (figure 4). So, the 128 bits register containing two single-single numbers could be seen as a vector of 4 floating points numbers.

5.2. Implementation of the error-free transformations

The EFT for the product is simply implemented by two instructions as follows:

```
1   Two–Prod–FMA (a,b)
2       p = spu_mul(a,b)
3       e = spu_msub(a, b, p)
4   return (p,e)
```

The algorithm of the EFT for the sum begins with a test and a swap. This test limits the possibility of parallelism. So, we have to first eliminate this test by the following procedure:

- evaluation of the condition. The result is a vector `comp` of type `unsigned int`, in which a value of zero means the condition holds and a value of `FFFFFFFF` for the opposite case.
- computation of the values of the two branches `val_1` (if the condition is satisfied) and `val_2` (if not).
- selection of the right value according to the vector of condition by using bit selection function:

$$d = spu\_sel(val\_2, val\_1, comp).$$

For each bit in the 128-bit vector *comp*, the corresponding bit from either vector val_2 or val_1 is selected. If the bit is 0, the bit from val_2 is selected; otherwise, the bit from val_1 is selected. The result is returned in vector d.

a	$a1$	$a2$	$a3$	$a4$
b	$b1 > a1$	$b2 < a2$	$b3 = a3$	$b4 > a4$
$comp = spu\_cmpabsgt(b, a)$				
	FFFFFFFF	00000000	00000000	FFFFFFFF
$hi = spu\_sel(a, b, comp)$				
	$b1$	$a2$	$a3$	$b4$
$lo = spu\_sel(b, a, comp)$				
	$a1$	$b2$	$b3$	$a4$

Figure 5. Example of the exchange of two vectors

For example, the test and the *swap* can be coded as follows:

```
1        comp = spu_cmpabsgt(b,a)
2        hi = spu_sel(a, b, comp)
3        lo = spu_sel(b, a, comp)
```

Figure 5 gives a concrete example of this exchange.

Each `spu_cmpabsgt` costs 2 clock cycles and the `spu_sel` too. Moreover, since the instructions of lines 2, 3 of this code are independent, they can be pipelined. So these 3 instructions cost only 5 clock cycles, which is less than a single precision operation (6 clock cycles for the FMA).

Applying the same procedure for the last conditional test of the algorithm `Two-Sum-toward-zero2`, this algorithm can be rewritten as follows:

```
1     Two-Sum-toward-zero2 (a,b)
2          comp = spu_cmpabsgt(b,a)
3          hi = spu_sel(a, b, comp)
4          lo = spu_sel(b, a, comp)
5          s = spu_add(a , b)
6          d = spu_sub(s , hi)
7          e = spu_sub(lo , d)
8          tmp = spu_mul(2 , lo)
9          comp = spu_cmpabsgt(d, tmp)
10         s = spu_sel(s, hi, comp)
11         e = spu_sel(e, lo, comp)
12    return (s,e)
```

Note that the addition of a and b does not change after the exchange. So we choose to use $a + b$ instead of $hi + lo$ to avoid the instructions dependencies. More precisely the 3 first instructions for the test and the exchange are independent of the instruction of line 5 which costs 6 cycles. So, they

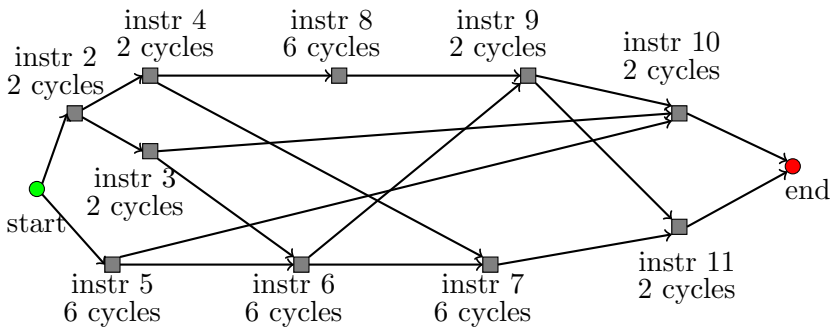


Figure 6. The dependencies between instructions of algorithm `Two-Sum-toward-zero`

can be executed in parallel[1]. The figure 6 emphasis the full independencies of instructions. This algorithm costs 20 clock cycles, which is a little bit more than the execution time of 3 consecutive double precision operations.

5.3. Renormalisation

The implementation of algorithm `Renormalise2-toward-zero` is similar to the `Two-Sum-toward-zero2` algorithm but without the conditional test and the exchange at the end.

```
1  Renormalise2-toward-zero (a,b)
2      s = spu_add(a , b)
3      comp = spu_cmpabsgt(b,a)
4      hi = spu_sel(a, b, comp)
5      lo = spu_sel(b, a, comp)
6      d = spu_sub(s , hi)
7      e = spu_sub(lo , d)
8      return (s,e)
```

With the same analysis as `Two-Sum-toward-zero2`, `Renormalise2-toward-zero` costs only 18 clock cycles. Now we will use these two functions to implement the arithmetic operators of single-singles.

5.4. Version 1

The natural version on single-single operations computes one operation on TWO single-singles. The SIMD processor allows us to manipulate simultaneously four 32 bits floating point numbers at the same time. When applying to vectors of single-singles, we can manipulate both the high and low components of these single-singles.

[1] On the SPE, there are 2 pipelines. The first one is devoted to numerical operations, the second one for control and logical operations. The two pipelines can be used in parallel.

Using the `Two-Sum-toward-zero2` presented above we calculate the sums and the rounding errors of two couples of high components and also of two couples of low components in the same time. Note that the rounding errors of these two couples of low components is computed, but not used by the algorithm.

Moreover, in the algorithms, it is necessary to compute operations between high and low components. This requires some extra operations to shuffle those components. So the first version does not take full advantage of the SIMD processor. We have implemented the first version for the sum and the product of single-singles which are `add_ds_ds_2`, `mul_ds_ds_2` and cost 50 cycles and 49 cycles respectively for two single-singles.

5.5. Version 2

The second version computes one operation on FOUR single-singles. It separates the high and the low components into two separated vectors (see figure 7) by using the function `spu_shuffle` of SPE which costs 4 clock cycles. This solution makes it possible a better optimisation of the pipelined instructions

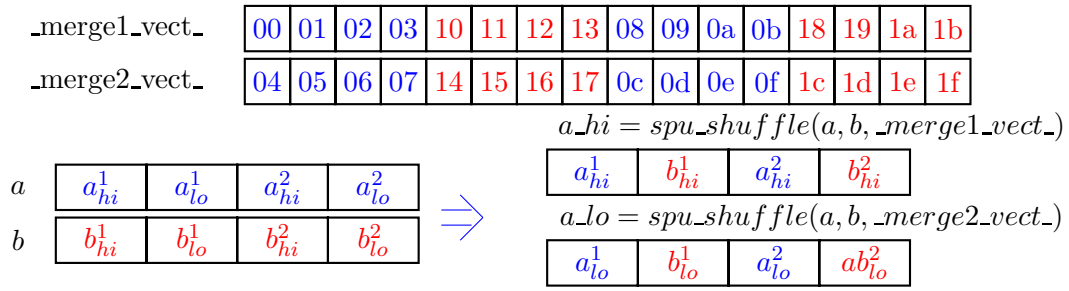


Figure 7. Merging of two vectors

Then, the operators can be implemented by applying directly the algorithms presented above on four operands separated in four vectors.

The intermediate result of these algorithms is also two vectors which contain respectively the four high parts and the four low parts of the result. At the end of the algorithm, the result vectors should be built by shuffling the high and the low components.

For example, the version 2 for the sum of single-singles is written as follows[2]

```
1  add_ds_ds_4 (vect_a1, vect_a2, vect_b1, vect_b2)
2       a_hi = spu_shuffle(vect_a1, vect_a2, _merge1_vect_)
3       a_lo = spu_shuffle(vect_a1, vect_a2, _merge2_vect_)
4       b_hi = spu_shuffle(vect_b1, vect_b2, _merge1_vect_)
5       b_lo = spu_shuffle(vect_b1, vect_b2, _merge2_vect_)
6       (s, e) = Two-Sum-toward-zero (a_hi, b_hi)
```

[2] The SIMD unit computes on 128-bits vectors. The 4 single-singles values of a and b are cut into two parts to keep the register organisation.

```
7           t1 = spu_add(a_lo , b_lo)
8           tmp = spu_add(t1 , e)
9           (hi, lo) = Renormalise2-toward-zero (s , tmp)
10          vect_c1 = spu_shuffle(hi, lo, _merge1_vect_)
11          vect_c2 = spu_shuffle(hi, lo, _merge2_vect_)
12      return (vect_c1, vect_c2)
```

Figure 8 shows the dependencies between instructions of this function. By using the tool `spu_timing` of IBM, the execution time of this function is **64 clock cycles** for four single-singles.

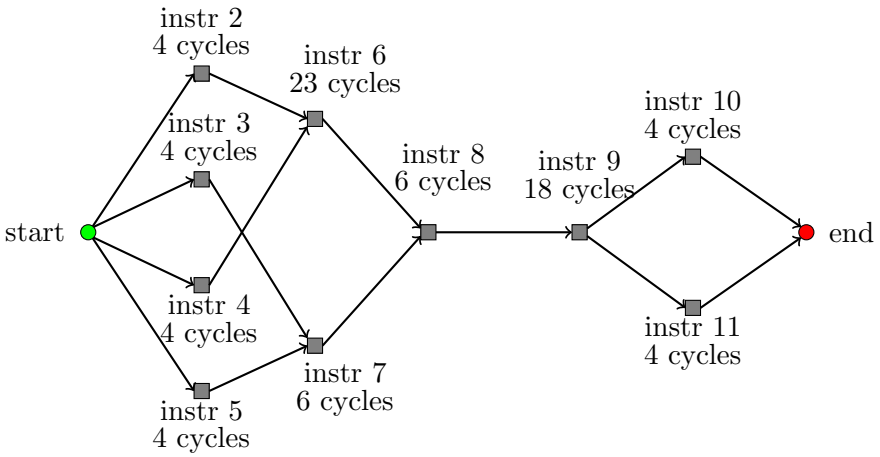


Figure 8. The dependencies between instructions of `add_ds_ds_4`

It is the same for the product of single-singles. We have successfully implemented the version 2 of the product of single-singles, called `mul_ds_ds_4` with an execution time of **60 clock cycles** for four single-singles.

The implementation of the division is more complicated. As described in the previous section, the division of single-singles `div_ds_ds` is based on the division in single precision, meanwhile the CELL processor does not support this kind of operation. It provides only a function to estimate the inverse of a floating-point number called `spu_re` which allows us to obtain a result precise up to 12 bits. So in order to implement the division of single-singles, we first have to implement the division in single precision.

The procedure to calculate the division of two 32 bits floating-point numbers a and b is as follows:

1. calculate the inverse of b,
2. multiply the inverse of b with a.

To improve the precision of the inversion we use the iterative Newton's method with the formula: $inv_{i+1} = inv_i + inv_i \times (1 - inv_i \times b)$. We also use the Newton's method for the multiplication with $a \times inverse(b)$ being the initial value. The division in single precision can be written as follows:

```
1        div (a, b)
2           tmp0 = spu_re(b)
3           rerr = spu_nmsub(tmp , b , 1)
4           inv = spu_madd(rerr , tmp0, tmp0)
5           rerr = spu_nmsub(inv , b , 1)
6           eerr = spu_mul(rerr , inv)
7           tmp = spu_mul(eerr , a)
8           q = spu_madd(a , inv , tmp)
9        return q
```

The precision of the algorithm `div` is provided by the following theorem.

Theorem 6. Let a, b be two floating-point numbers in single precision, ε being the machine precision. The relative error of the algorithm `div` is bounded by:

$$|div(a,b) - a/b| < [\varepsilon + \mathcal{O}(\varepsilon^2)] \times |a/b|.$$

Using the newly implemented single-precision division operator and the algorithm of division of single-singles presented above, we have implemented the function `div_ds_ds_4` which calculates fours single-single divisions at the same time. This function costs **111 clock cycles** for four single-singles.

5.6. Optimised algorithms

The versions 2 of the single-single operators performs four operations at the same time, and they have taken full advantage of the SIMD processor which provides an important performance of calculation. But using the `spu_timing` tool of IBM we recognized that there still left many non-used clock cycles in the process of calculation of each operator.

We can use these non-used clock cycles by increasing the number of operations executed at the same time.

With the restricted local storage (only 256 KB for both the code and data) we choose to implement operations on EIGHT single-singles. This third version is considered as the optimal version in our library. The third version of the sum, the product and the division are named `add_ds_ds_8`, `mul_ds_ds_8`, `div_ds_ds_8` and cost respectively **72 cycles**, **63 cycles** and **125 cycles** for eight single-singles. In comparison with the version 2 with only some supplementary clock cycles (for example 8 cycles for the sum and 3 cycles for the product) we can execute 8 single-single operations instead of 4. It means that we have achieved a coarse gain with the final version in terms of performance.

Almost every clock cycles being used, there would be no gain to deal with sixteen single-singles.

Table I. Theoretical results of the single-single library

Function	Number of operations	Execution time	Performance
add_ds_ds_2	2	50 cycles	0.128 GFLOPs
add_ds_ds_4	4	64 cycles	0.2 GFLOPs
add_ds_ds_8	8	72 cycles	0.355 GFLOPs
mul_ds_ds_2	2	49 cycles	0.130 GFLOPs
mul_ds_ds_4	4	60 cycles	0.213 GFLOPs
mul_ds_ds_8	8	63 cycles	0.406 GFLOPs
div_ds_ds_4	4	111 cycles	0.115 GFLOPs
div_ds_ds_8	8	125 cycles	0.2048 GFLOPs

5.7. Theoretical results

On a CELL processor with a frequency of 3.2 GHz, its theoretical performances (without memory access problem) of the single-single are presented in table I.

6. Numerical simulations

6.1. Experimental results

To test the performance of the single-single library, we created a program which performs the basic operators on two large vectors of single-single and also on two large double precision vectors of the same size. To achieve the peak performance of the library we use the third version of each operator. Double-buffering is used to hide data transfer time.

This program is executed on a IBM CELL Blade based at CINES, Montpellier, France. The CPU frequency is 3.2GHz. The results obtained are listed in the table II.

Figure 9 illustrates the performance of the addition on single-singles and on native double precision. Both have the same memory size. They are very close. It is interesting to note that the maximum performance with 64 bits floating point is not reached. In this case the program measures mainly the memory transfer time. The native double operation are completely hidden.

For the single-singles, the computing time of one operation is on the same order as the transfer memory necessary for one operation. This kind of program benefits for our library.

To have another comparison, another program is created which executes a large number of basic operators on a small number of data generated within the SPE without any data transfer. The execution time of the program is exactly the time of calculation. The results are presented

Table II. Real performances of the library single-single

Functions	Theoretical performance	Experimental performance
add_ds_ds_8	355 MFLOPs	250.4 MFLOPs
mul_ds_ds_8	406 MFLOPs	287.2 MFLOPs
div_ds_ds_8	204 MFLOPs	166.4 MFLOPs

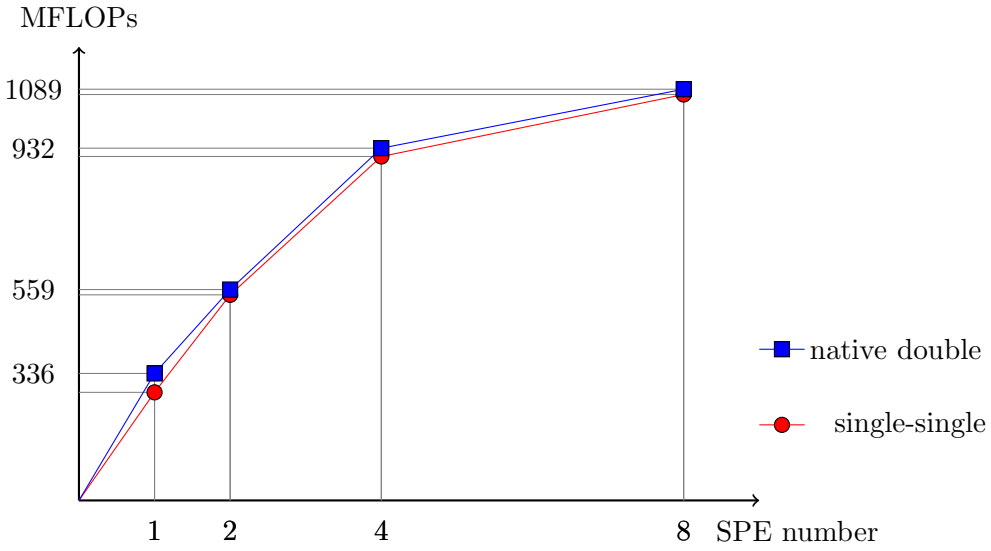


Figure 9. The performance of the library single-single: The addition

in table III. The peak performance for the multiplication on the CELL processor is achieved for native double precision.

With the single-singles numbers, it is not possible to achieve the same performance than with the native double precision. This is mainly due to two factors:

- the cost of the function call,
- the transfer from the local memory to the registers.

6.2. Exactitude

The exactitude of the library is tested by performing a large number of operations on random values of single-single and their corresponding double precision values. With 2^{24} comparisons, the results are summarized in table IV.

Table III. Performance of the single-single library and of the double precision of the machine, without data transfer

Functions	**Theoretical performance (1SPE)**	**Experimental performance (1 SPE)**	**Experimental performance (8 SPEs)**
add_ds_ds_8	355 MFLOPs	266 MFLOPs	2133 MFLOPs
mul_ds_ds_8	406 MFLOPs	320 MFLOPs	2560 MFLOPs
div_ds_ds_8	204 MFLOPs	172 MFLOPs	1383 MFLOPs
sum in double precision	914 MFLOPs	914 MFLOPs	7314 MFLOPs
product in double precision	914 MFLOPs	914 MFLOPs	7314 MFLOPs
division in double precision	(not supported)	86 MFLOPs	691 MFLOPs

Table IV. The exactitude of single-single library

Operation	**Max difference**	**Mean difference**
Sum	0.0e+00	0.00e+00
Product	2.964e-14	1.425e-16
Division	2.373e-14	1.758e-15

7. Conclusions and perspectives

This paper is based mostly on (Hida et al., 2001) with some adaptations to the rounding mode toward zero and to the implementation environment of CELL processor. First we propose an algorithm for the error-free transformation of the sum which is proved to be effectively implemented on the CELL processor. Then, we introduce the methodology to develop the extended precision of single-single with such basic operators that the sum, the product and the division. A large part of this paper is dedicated to the implementation of this library in exploiting the specific characteristics of CELL processor, among which the most important are the truncation rounding, the SIMD processor and the fully pipelined instruction set. The performance and the precision of the implemented library is tested by running test programs on a real CELL processor with a frequency of 3.2GHz.

In the future, this library could be completed by the treatment of numeric exceptions, by the binary operations, algebraic operations and transcendental operations.

Waiting for the next CELL generation, we are developing the quad-single precision library. With the next generation of CELL processor, we will be able to easily get:

– the quad precision implemented with double-double numbers with the methodology of the single-single library,

– the quad-double precision implemented with four double numbers with the methodology of the quad-single library.

These precisions are needed by more and more current applications.

Acknowledgements

The authors are very grateful to the CINES (Centre Informatique National de l'Enseignement Supérieur, Montpellier, France) for providing us access to their CELL blades.

References

Dekker, T. J.: 1971, 'A floating-point technique for extending the available precision'. *Numer. Math.* **18**, 224–242.

Gschwind, M., H. P. Hofstee, B. Flachs, M. Hopkins, Y. Watanabe, and T. Yamazaki: 2006, 'Synergistic Processing in Cell's Multicore Architecture'. *IEEE Micro* **26**(2), 10–24.

Hida, Y., X. S. Li, and D. H. Bailey: 2001, 'Algorithms for Quad-Double Precision Floating Point Arithmetic'. In: *Proc. 15th IEEE Symposium on Computer Arithmetic*. pp. 155–162, IEEE Computer Society Press, Los Alamitos, CA, USA.

Jacobi, C., H.-J. Oh, K. D. Tran, S. R. Cottier, B. W. Michael, H. Nishikawa, Y. Totsuka, T. Namatame, and N. Yano: 2005, 'The Vector Floating-Point Unit in a Synergistic Processor Element of a CELL Processor'. In: *ARITH '05: Proceedings of the 17th IEEE Symposium on Computer Arithmetic*. Washington, DC, USA, pp. 59–67, IEEE Computer Society.

Kahle, J. A., M. N. Day, H. P. Hofstee, C. R. Johns, T. R. Maeurer, and D. Shippy: 2005, 'Introduction to the cell multiprocessor'. *IBM J. Res. Dev.* **49**(4/5), 589–604.

Knuth, D. E.: 1998, *The Art of Computer Programming, Volume 2, Seminumerical Algorithms*. Reading, MA, USA: Addison-Wesley, third edition.

Priest, D. M.: 1992, 'On Properties of Floating Point Arithmetics: Numerical Stability and the Cost of Accurate Computations'. Ph.D. thesis, Mathematics Department, University of California, Berkeley, CA, USA. `ftp://ftp.icsi.berkeley.edu/pub/theory/priest-thesis.ps.Z`.

Nguyen, H. D.: 2007, 'Calcul précis et efficace sur le processeur CELL'. Master report, LIP6, UPMC (P. and M. Curie University), Paris, France `http://www-pequan.lip6.fr/~graillat/papers/rapport_Diep.pdf`.

Williams, S., J. Shalf, L. Oliker, S. Kamil, P. Husbands, and K. Yelick: 2006, 'The potential of the cell processor for scientific computing'. In: *CF '06: Proceedings of the 3rd conference on Computing frontiers*. New York, NY, USA, pp. 9–20, ACM Press.

Accurate Floating Point Product

Stef Graillat
Laboratoire LIP6, Département Calcul Scientifique
Université Pierre et Marie Curie (Paris 6)
4 place Jussieu, F-75252, Paris cedex 05, France
email:stef.graillat@lip6.fr

Abstract. Several different techniques and softwares intend to improve the accuracy of results computed in a fixed finite precision. Here we focus on a method to improve the accuracy of the product of floating point numbers. We show that the computed result is as accurate as if computed in twice the working precision. The algorithm is simple since it only requires addition, subtraction and multiplication of floating point numbers in the same working precision as the given data. Such an algorithm can be useful for example to compute the determinant of a triangular matrix and to evaluate a polynomial when represented by the root product form. It can also be used to compute the power of a floating point number.

Keywords: accurate product, exponentiation, finite precision, floating point arithmetic, faithful rounding, error-free transformations

AMS Subject Classification: 65-04, 65G20, 65G50

1. Introduction

In this paper, we present fast and accurate algorithms to compute the product of floating point numbers. Our aim is to increase the accuracy at a fixed precision. We show that the results have the same error estimates as if computed in twice the working precision and then rounded to working precision. Then we address the problem on how to compute a faithfully rounded result, that is to say one of the two adjacent floating point numbers of the exact result.

This paper was motived by papers (Ogita et al., 2005a; Rump et al., 2005; Graillat et al., 2005; Langlois and Louvet, 2007) and (Kornerup et al., 2007) where similar approaches are used to compute summation, dot product, polynomial evaluation and power.

The applications of our algorithms are multiple. One of the examples frequently used in Sterbenz's book (Sterbenz, 1974) is the computation of the product of some floating point numbers. Our algorithms can be used to compute the determinant of a triangle matrix. Another application is for evaluating a polynomial when represented by the root product form $p(x) = a_n \prod_{i=1}^{n}(x - x_i)$. We can also apply our algorithms to compute the power of a floating point number.

The rest of the paper is organized as follows. In Section 2, we recall notations and auxiliary results that will be needed in the sequel. We present the floating point arithmetic and the so-called error-free transformations. In Section 3, we present a classic algorithm to compute the product of floating point numbers. We give an error estimate as well as a validated error bound. We also

present a new compensated algorithm together with an error estimate and a validated error bound. We show that under mild assumptions, our algorithm gives a faithfully rounded result.

2. Notation and auxiliary results

2.1. Floating point arithmetic

Throughout the paper, we assume to work with a floating point arithmetic adhering to IEEE 754 floating point standard in rounding to nearest (IEEE Computer Society, 1985). We assume that no overflow nor underflow occur. The set of floating point numbers is denoted by $\mathbb{F}$, the relative rounding error by $\texttt{eps}$. For IEEE 754 double precision, we have $\texttt{eps} = 2^{-53}$ and for single precision $\texttt{eps} = 2^{-24}$.

We denote by $\mathrm{fl}(\cdot)$ the result of a floating point computation, where all operations inside parentheses are done in floating point working precision. Floating point

operations in IEEE 754 satisfy (Higham, 2002)

$$\mathrm{fl}(a \circ b) = (a \circ b)(1+\varepsilon_1) = (a \circ b)/(1+\varepsilon_2) \text{ for } \circ = \{+,-,\cdot,/\} \text{ and } |\varepsilon_\nu| \leq \texttt{eps}.$$

This implies that

$$|a \circ b - \mathrm{fl}(a \circ b)| \leq \texttt{eps}|a \circ b| \text{ and } |a \circ b - \mathrm{fl}(a \circ b)| \leq \texttt{eps}|\,\mathrm{fl}(a \circ b)| \text{ for } \circ = \{+,-,\cdot,/\}. \quad (1)$$

2.2. Error-free transformations

One can notice that $a \circ b \in \mathbb{R}$ and $\mathrm{fl}(a \circ b) \in \mathbb{F}$ but in general we do not have $a \circ b \in \mathbb{F}$. It is known that for the basic operations $+, -, \cdot$, the approximation error of a floating point operation is still a floating point number (see for example (Dekker, 1971)):

$$\begin{array}{lll} x = \mathrm{fl}(a \pm b) & \Rightarrow & a \pm b = x + y \quad \text{with } y \in \mathbb{F}, \\ x = \mathrm{fl}(a \cdot b) & \Rightarrow & a \cdot b = x + y \quad \text{with } y \in \mathbb{F}. \end{array} \quad (2)$$

These are *error-free* transformations of the pair (a, b) into the pair (x, y).

Fortunately, the quantities x and y in (2) can be computed exactly in floating point arithmetic. For the algorithms, we use Matlab-like notations. For addition, we can use the following algorithm by Knuth (Knuth, 1998, Thm B. p.236).

ALGORITHM 2.1 (Knuth (Knuth, 1998)). *Error-free transformation of the sum of two floating point numbers*

function $[x, y] = \texttt{TwoSum}(a, b)$
  $x = \mathrm{fl}(a + b)$
  $z = \mathrm{fl}(x - a)$
  $y = \mathrm{fl}((a - (x - z)) + (b - z))$

Another algorithm to compute an error-free transformation is the following algorithm from Dekker (Dekker, 1971). The drawback of this algorithm is that we have $x+y = a+b$ provided that $|a| \geq |b|$. Generally, on modern computers, a comparison followed by a branching and 3 operations costs more than 6 operations. As a consequence, `TwoSum` is generally more efficient than `FastTwoSum`.

ALGORITHM 2.2 (Dekker (Dekker, 1971)). *Error-free transformation of the sum of two floating point numbers with* $|a| \geq |b|$.

function $[x, y] = \texttt{FastTwoSum}(a, b)$
  $x = \mathrm{fl}(a + b)$
  $y = \mathrm{fl}((a - x) + b)$

For the error-free transformation of a product, we first need to split the input argument into two parts. Let p be given by $\texttt{eps} = 2^{-p}$ and define $s = \lceil p/2 \rceil$. For example, if the working precision is IEEE 754 double precision, then $p = 53$ and $s = 27$. The following algorithm by Dekker (Dekker, 1971) splits a floating point number $a \in \mathbb{F}$ into two parts x and y such that

$$a = x + y \quad \text{and} \quad x \text{ and } y \text{ nonoverlapping with } |y| \leq |x|.$$

ALGORITHM 2.3 (Dekker (Dekker, 1971)). *Error-free split of a floating point number into two parts*

function $[x, y] = \texttt{Split}(a, b)$
  $\texttt{factor} = \mathrm{fl}(2^s + 1)$
  $c = \mathrm{fl}(\texttt{factor} \cdot a)$
  $x = \mathrm{fl}(c - (c - a))$
  $y = \mathrm{fl}(a - x)$

With this function, an algorithm from Veltkamp (see (Dekker, 1971)) makes it possible to compute an error-free transformation for the product of two floating point numbers. This algorithm returns two floating point numbers x and y such that

$$a \cdot b = x + y \quad \text{with } x = \mathrm{fl}(a \cdot b).$$

ALGORITHM 2.4 (Veltkamp (Dekker, 1971)). *Error-free transformation of the product of two floating point numbers*

function $[x, y] = \texttt{TwoProduct}(a, b)$
  $x = \mathrm{fl}(a \cdot b)$
  $[a_1, a_2] = \texttt{Split}(a)$
  $[b_1, b_2] = \texttt{Split}(b)$
  $y = \mathrm{fl}(a_2 \cdot b_2 - (((x - a_1 \cdot b_1) - a_2 \cdot b_1) - a_1 \cdot b_2))$

The following theorem summarizes the properties of algorithms `TwoSum` and `TwoProduct`.

THEOREM 2.1 (Ogita, Rump and Oishi (Ogita et al., 2005a)). *Let $a, b \in \mathbb{F}$ and let $x, y \in \mathbb{F}$ such that $[x, y] = \texttt{TwoSum}(a, b)$ (Algorithm 2.1). Then,*

$$a + b = x + y, \quad x = \mathrm{fl}(a + b), \quad |y| \leq \texttt{eps}|x|, \quad |y| \leq \texttt{eps}|a + b|. \tag{3}$$

The algorithm `TwoSum` *requires* 6 *flops.*
Let $a, b \in \mathbb{F}$ and let $x, y \in \mathbb{F}$ such that $[x, y] = \texttt{TwoProduct}(a, b)$ (Algorithm 2.4). Then,

$$a \cdot b = x + y, \quad x = \mathrm{fl}(a \cdot b), \quad |y| \leq \texttt{eps}|x|, \quad |y| \leq \texttt{eps}|a \cdot b|. \tag{4}$$

The algorithm `TwoProduct` *requires* 17 *flops.*

3. Accurate floating point product

In this section, we present a new accurate algorithm to compute the product of floating point numbers. In Subsection 3.1, we recall the classic method and we give a theoretical error bound as well as a validated computable error bound. In Subsection 3.2, we present our new algorithm based on a compensated scheme together with a theoretical error bound. In Subsection 3.3, we give sufficient conditions on the number of floating point numbers so as to get a faithfully rounded result. Finally, in Subsection 3.4, we give a validated computable error bound for our new algorithm.

3.1. Classic method

The classic method for evaluating a product of n numbers $a = (a_1, a_2, \ldots, a_n)$

$$p = \prod_{i=1}^{n} a_i$$

is the following algorithm.

ALGORITHM 3.1. *Product evaluation*

function $\texttt{res} = \texttt{Prod}(a)$
  $p_1 = a_1$
  for $i = 2 : n$
    $p_i = \mathrm{fl}(p_{i-1} \cdot a_i)$
  end
  $\texttt{res} = p_n$

This algorithm requires $n - 1$ flops. Let us now analyse its accuracy.

We will use standard notations and standard results for the following error estimations (see (Higham, 2002)). The quantities γ_n are defined as usual (Higham, 2002) by

$$\gamma_n := \frac{n\texttt{eps}}{1 - n\texttt{eps}} \quad \text{for } n \in \mathbb{N}.$$

When using γ_n, we implicitly assume that $n\mathtt{eps} \leq 1$. A forward error bound is

$$|a_1a_2\cdots a_n - \mathtt{res}| = |a_1a_2\cdots a_n - \mathrm{fl}(a_1a_2\cdots a_n)| \leq \gamma_{n-1}|a_1a_2\cdots a_n|. \tag{5}$$

Indeed, by induction,

$$\mathtt{res} = \mathrm{fl}(a_1a_2\cdots a_n) = a_1a_2\cdots a_n(1+\varepsilon_2)(1+\varepsilon_3)\cdots(1+\varepsilon_n), \tag{6}$$

with $\varepsilon_i \leq \mathtt{eps}$ for $i = 2:n$. It follows from Lemma 3.1 of (Higham, 2002, p.63) that $(1+\varepsilon_2)(1+\varepsilon_3)\cdots(1+\varepsilon_n) = 1+\theta_n$ where $|\theta_{n-1}| \leq \gamma_{n-1}$.

A convenient device for keeping track of power of $1+\varepsilon$ term is described in (Higham, 2002, p.68). The relative error counter $\langle k\rangle$ denotes the product

$$\langle k\rangle = \prod_{i=1}^{k}(1+\varepsilon_i), \quad |\varepsilon_i| \leq \mathtt{eps}.$$

A useful rule for the counter is $\langle j\rangle\langle k\rangle = \langle j+k\rangle$. Using this notation, Equation (6) can be written $\mathtt{res} = \mathrm{fl}(a_1a_2\cdots a_n) = a_1a_2\cdots a_n\langle n-1\rangle$.

It is shown in (Ogita et al., 2005b) that for $a \in \mathbb{F}$, we have

$$(1+\mathtt{eps})^n \;\leq\; \frac{1}{(1-\mathtt{eps})^n} \leq \frac{1}{1-n\mathtt{eps}}, \tag{7}$$

$$\frac{|a|}{1-n\mathtt{eps}} \;\leq\; \mathrm{fl}\left(\frac{|a|}{1-(n+1)\mathtt{eps}}\right). \tag{8}$$

¿From Equation (6), it follows that

$$|a_1a_2\cdots a_n - \mathtt{res}| \leq (1+\mathtt{eps})^{n-1}\gamma_{n-1}|\mathtt{res}|.$$

If $m\mathtt{eps} \leq 1$ for $m \in \mathbb{N}$, $\mathrm{fl}(m\mathtt{eps}) = m\mathtt{eps}$ and $\mathrm{fl}(1-m\mathtt{eps}) = 1-m\mathtt{eps}$. Therefore,

$$\gamma_m \leq (1+\mathtt{eps})\,\mathrm{fl}(\gamma_m). \tag{9}$$

Hence,

$$\begin{aligned} |a_1a_2\cdots a_n - \mathtt{res}| \;&\leq\; (1+\mathtt{eps})^n\,\mathrm{fl}(\gamma_{n-1})|\mathtt{res}| \\ &\leq\; (1+\mathtt{eps})^{n+1}\,\mathrm{fl}(\gamma_{n-1}|\mathtt{res}|), \end{aligned}$$

and so

$$|a_1a_2\cdots a_n - \mathtt{res}| \leq \mathrm{fl}\left(\frac{\gamma_{n-1}|\mathtt{res}|}{1-(n+2)\mathtt{eps}}\right).$$

The previous inequality gives us a validated error bound that can be computed in pure floating point arithmetic in rounding to nearest.

3.2. Compensated method

We present hereafter a compensated scheme to evaluate the product of floating point numbers, i.e. the error of individual multiplication is somehow corrected. The technique used here is based on the paper (Ogita et al., 2005a).

ALGORITHM 3.2. *Product evaluation with a compensated scheme*

function res = CompProd(a)
  $p_1 = a_1$
  $e_1 = 0$
  for $i = 2 : n$
    $[p_i, \pi_i] = \texttt{TwoProduct}(p_{i-1}, a_i)$
    $e_i = \mathrm{fl}(e_{i-1}a_i + \pi_i)$
  end
  res = $\mathrm{fl}(p_n + e_n)$

This algorithm requires $19n - 18$ flops. For error analysis, we note that

$$p_n = \mathrm{fl}(a_1 a_2 \cdots a_n) \quad \text{and} \quad e_n = \mathrm{fl}\left(\sum_{i=2}^{n} \pi_i a_{i+1} \cdots a_n\right).$$

We also have

$$p = a_1 a_2 \ldots a_n = \mathrm{fl}(a_1 a_2 \ldots a_n) + \sum_{i=2}^{n} \pi_i a_{i+1} \cdots a_n = p_n + e, \tag{10}$$

where $e = \sum_{i=2}^{n} \pi_i a_{i+1} \cdots a_n$.

Before proving the main theorem, we will need two technical lemmas. The next lemma makes it possible to obtain a bound on the individual error of the multiplication namely π_i in function of the inital data a_i.

LEMMA 3.1. *Suppose floating point numbers $\pi_i \in \mathbb{F}$, $2 \le i \le n$ are computed by the following algorithm*

$p_1 = a_1$
for $i = 2 : n$
  $[p_i, \pi_i] = \texttt{TwoProduct}(p_{i-1}, a_i)$
end

Then,

$$|\pi_i| \le \texttt{eps}(1 + \gamma_{i-1})|a_1 \cdots a_i| \quad \textit{for } i = 2 : n.$$

Proof. From Equation (1), it follows that

$$|\pi_i| \le \texttt{eps}|p_i|.$$

Moreover, $p_i = \mathrm{fl}(a_1 \cdots a_i)$ so that from (5),

$$|p_i| \le (1 + \gamma_{i-1})|a_1 \cdots a_i|.$$

Hence, $|\pi_i| \le \texttt{eps}(1 + \gamma_{i-1})|a_1 \cdots a_i|$. □

The following lemma enables us to bound the rounding errors during the computation of the error during the full product.

LEMMA 3.2. *Suppose floating point numbers $e_i \in \mathbb{F}$, $1 \leq i \leq n$ are computed by the following algorithm*

$e_1 = 0$
for $i = 2 : n$
  $[p_i, \pi_i] = \texttt{TwoProduct}(p_{i-1}, a_i)$
  $e_i = \mathrm{fl}(e_{i-1}a_i + \pi_i)$
end

Then,

$$|e_n - \sum_{i=2}^{n} \pi_i a_{i+1} \cdots a_n| \leq \gamma_{n-1}\gamma_{2n}|a_1 a_2 \cdots a_n|.$$

Proof. First, one notices that $e_n = \mathrm{fl}(\sum_{i=2}^{n}(\pi_i a_{i+1} \cdots a_n))$. We will use the error counters described above. For n floating point numbers x_i, it is easy to see that (Higham, 2002, chap.4)

$$\mathrm{fl}(x_1 + x_2 + \cdots + x_n) = x_1\langle n-1\rangle + x_2\langle n-1\rangle + x_3\langle n-2\rangle + \cdots + x_n\langle 1\rangle.$$

This implies that

$$e_n = \mathrm{fl}(\sum_{i=2}^{n}(\pi_i a_{i+1} \cdots a_n)) = \mathrm{fl}(\pi_2 a_3 \cdots a_n)\langle n-2\rangle + \mathrm{fl}(\pi_3 a_4 \cdots a_n)\langle n-2\rangle + \cdots + \mathrm{fl}(\pi_n)\langle 1\rangle.$$

Furthermore, we have shown before that $\mathrm{fl}(a_1 a_2 \cdots a_n) = a_1 a_2 \cdots a_n\langle n-1\rangle$. Consequently,

$$e_n = \pi_2 a_3 \cdots a_n\langle n-2\rangle\langle n-1\rangle + \pi_3 a_4 \cdots a_n\langle n-3\rangle\langle n-1\rangle + \cdots + \pi_n\langle 1\rangle.$$

A straightforward computation yields

$$|e_n - \sum_{i=2}^{n} \pi_i a_{i+1} \cdots a_n| \leq \gamma_{2n-3} \sum_{i=2}^{n} |\pi_i a_{i+1} \cdots a_n|.$$

¿From Lemma 3.1, we have $|\pi_i| \leq \texttt{eps}(1 + \gamma_{i-1})|a_1 \cdots a_i|$ and hence

$$|e_n - \sum_{i=2}^{n} \pi_i a_{i+1} \cdots a_n| \leq (n-1)\texttt{eps}(1 + \gamma_{n-1})\gamma_{2n-3}|a_1 a_2 \cdots a_n|.$$

Since $\texttt{eps}(1 + \gamma_{n-1}) = \gamma_{n-1}/(n-1)$ and $\gamma_{2n-3} \leq \gamma_{2n}$, we obtain the desired result. □

One may notice that the computation of e_n is similar to the Horner scheme. One could have directly applied a result on the error of the Horner scheme (Higham, 2002, Eq.(5.3),p.95).

We can finally state the main theorem.

THEOREM 3.3. *Suppose Algorithm 3.2 is applied to floating point number $a_i \in \mathbb{F}$, $1 \leq i \leq n$, and set $p = \prod_{i=1}^{n} a_i$. Then,*

$$|\texttt{res} - p| \leq \texttt{eps}|p| + \gamma_n\gamma_{2n}|p|.$$

Proof. The fact that $\texttt{res} = \mathrm{fl}(p_n + e_n)$ implies that $\texttt{res} = (1+\varepsilon)(p_n + e_n)$ with $\varepsilon \leq \texttt{eps}$. So it follows

$$\begin{aligned}
|\texttt{res} - p| &= |\,\mathrm{fl}(p_n + e_n) - p| = |(1+\varepsilon)(p_n + e_n - p) + \varepsilon p| \\
&= |(1+\varepsilon)(p_n + \sum_{i=2}^{n} \pi_i a_{i+1} \cdots a_n - p) + (1+\varepsilon)(e_n - \sum_{i=2}^{n} \pi_i a_{i+1} \cdots a_n) + \varepsilon p| \\
&= |(1+\varepsilon)(e_n - \sum_{i=2}^{n} \pi_i a_{i+1} \cdots a_n) + \varepsilon p| \quad \text{by (10)} \\
&\leq \texttt{eps}|p| + (1+\texttt{eps})|e_n - \sum_{i=2}^{n} \pi_i a_{i+1} \cdots a_n| \\
&\leq \texttt{eps}|p| + (1+\texttt{eps})\gamma_{n-1}\gamma_{2n}|a_1 a_2 \cdots a_n|.
\end{aligned}$$

Since $(1+\texttt{eps})\gamma_{n-1} \leq \gamma_n$, it follows that $|\texttt{res} - p| \leq \texttt{eps}|p| + \gamma_n\gamma_{2n}|p|$. □

It may be interesting to study the condition number of the product evaluation. Ones defines

$$\mathrm{cond}(a) = \lim_{\varepsilon \to 0} \sup \left\{ \frac{|(a_1 + \Delta a_1)(a_2 + \Delta a_2) \cdots (a_n + \Delta a_n) - a_1 a_2 \cdots a_n|}{\varepsilon |a_1 a_2 \cdots a_n|} : |\Delta a_i| \leq \varepsilon |a_i| \right\}.$$

A standard computation yields

$$\mathrm{cond}(a) = n.$$

COROLLARY 3.4. *Suppose Algorithm 3.2 is applied to floating point number $a_i \in \mathbb{F}$, $1 \leq i \leq n$, and set $p = \prod_{i=1}^{n} a_i$. Then,*

$$\frac{|\texttt{res} - p|}{|p|} \leq \texttt{eps} + \frac{\gamma_n\gamma_{2n}}{n} \mathrm{cond}(a).$$

3.3. Faithful rounding

We define the floating point predecessor and successor of a real number r satisfying $\min\{f : f \in \mathbb{R}\} < r < \max\{f : f \in \mathbb{F}\}$ by

$$\mathrm{pred}(r) := \max\{f \in \mathbb{F} : f < r\} \quad \text{and} \quad \mathrm{succ}(r) := \min\{f \in \mathbb{F} : r < f\}.$$

DEFINITION 3.1. *A floating point number $f \in \mathbb{F}$ is called a* faithful rounding *of a real number $r \in \mathbb{R}$ if*

$$\mathrm{pred}(f) < r < \mathrm{succ}(f).$$

We denote this by $f \in \square(r)$. For $r \in \mathbb{F}$, this implies that $f = r$.

A faithful rounding is then one of the two adjacent floating point numbers of the exact result.

LEMMA 3.5 (Rump, Ogita and Oishi (Rump et al., 2005, lem. 2.5)). *Let $r, \delta \in \mathbb{R}$ and $\widetilde{r} := \mathrm{fl}(r)$. Suppose that $2|\delta| < \texttt{eps}|\widetilde{r}|$. Then $\widetilde{r} \in \square(r + \delta)$, that means $\widetilde{r}$ is a faithful rounding of $r + \delta$.*

Let res be the result of CompProd. Then we have $p = p_n + e$ and $\texttt{res} = \mathrm{fl}(p_n + e_n)$ with $e = \sum_{i=2}^{n} \pi_i a_{i+1} \cdots a_n$. It follows that $p = (p_n + e_n) + (e - e_n)$. This leads to the following lemma which gives a criterion to ensure that the result of CompProd is faithfully rounded.

LEMMA 3.6. *With the previous notations, if* $2|e - e_n| < \texttt{eps}|\texttt{res}|$ *then* res *is a faithful rounding of* p.

Since we have $|e - e_n| \leq \gamma_n\gamma_{2n}|p|$ and $(1 - \texttt{eps})|p| - \gamma_n\gamma_{2n}|p| \leq |\texttt{res}|$, a sufficient condition to ensure a faithful rounding is

$$2\gamma_n\gamma_{2n}|p| < \texttt{eps}((1 - \texttt{eps})|p| - \gamma_n\gamma_{2n}|p|)$$

that is

$$\gamma_n\gamma_{2n} < \frac{1 - \texttt{eps}}{2 + \texttt{eps}}\texttt{eps}.$$

Since $\gamma_n\gamma_{2n} \leq 2(n\texttt{eps})^2/(1 - 2n\texttt{eps})^2$, a sufficient condition is

$$2\frac{(n\texttt{eps})^2}{(1 - 2n\texttt{eps})^2} < \frac{1 - \texttt{eps}}{2 + \texttt{eps}}\texttt{eps}$$

which is equivalent to

$$\frac{n\texttt{eps}}{1 - 2n\texttt{eps}} < \sqrt{\frac{(1 - \texttt{eps})\texttt{eps}}{2(2 + \texttt{eps})}}$$

and then to

$$n < \frac{\sqrt{1 - \texttt{eps}}}{\sqrt{2}\sqrt{2 + \texttt{eps}} + 2\sqrt{(1 - \texttt{eps})\texttt{eps}}}\texttt{eps}^{-1/2}.$$

We have just shown that if $n < \alpha\texttt{eps}^{-1/2}$ where $\alpha \approx 1/2$ then the result is faithfully rounded. More precisely, in double precision where $\texttt{eps} = 2^{-53}$, if $n < 2^{25} \approx 5 \cdot 10^7$, we get a faithfully rounded result.

3.4. Validated error bound

We present here how to compute a valid error bound in pure floating point arithmetic in rounding to nearest. It holds that

$$\begin{aligned} |\texttt{res} - p| &= |\,\mathrm{fl}(p_n + e_n) - p| = |\,\mathrm{fl}(p_n + e_n) - (p_n + e_n) + (p_n + e_n) - p| \\ &\leq \texttt{eps}|\texttt{res}| + |p_n + e_n - p| \\ &\leq \texttt{eps}|\texttt{res}| + |e_n - e|. \end{aligned}$$

Since $|e_n - e| \leq \gamma_{n-1}\gamma_{2n}|p|$ and $|p| \leq (1 + \texttt{eps})^{n-1}\,\mathrm{fl}(|a_1a_2 \cdots a_n|)$ we obtain

$$\begin{aligned} |\texttt{res} - p| &\leq \texttt{eps}|\texttt{res}| + \gamma_{n-1}\gamma_{2n}|p| \\ &\leq \texttt{eps}|\texttt{res}| + \gamma_{n-1}\gamma_{2n}(1 + \texttt{eps})^{n-1}\,\mathrm{fl}(|a_1a_2 \cdots a_n|). \end{aligned}$$

Using (8) and (9), we get

$$\begin{aligned}
|\mathtt{res} - p| &\leq \mathrm{fl}(\mathtt{eps}|\mathtt{res}|) + (1+\mathtt{eps})^n \,\mathrm{fl}(\gamma_n)\,\mathrm{fl}(\gamma_{2n})\,\mathrm{fl}(|a_1a_2\cdots a_n|)\\
&\leq \mathrm{fl}(\mathtt{eps}|\mathtt{res}|) + (1+\mathtt{eps})^{n+2}\,\mathrm{fl}(\gamma_n\gamma_{2n}|a_1a_2\cdots a_n|)\\
&\leq \mathrm{fl}(\mathtt{eps}|\mathtt{res}|) + \mathrm{fl}\left(\frac{\gamma_n\gamma_{2n}|a_1a_2\cdots a_n|}{1-(n+3)\mathtt{eps}}\right)\\
&\leq (1+\mathtt{eps})\,\mathrm{fl}\left(\mathtt{eps}|\mathtt{res}| + \frac{\gamma_n\gamma_{2n}|a_1a_2\cdots a_n|}{1-(n+3)\mathtt{eps}}\right)\\
&\leq \mathrm{fl}\left(\left(\mathtt{eps}|\mathtt{res}| + \frac{\gamma_n\gamma_{2n}|a_1a_2\cdots a_n|}{1-(n+3)\mathtt{eps}}\right) / \,(1-2\mathtt{eps})\right).
\end{aligned}$$

We can summarize this as follows.

LEMMA 3.7. *Suppose Algorithm 3.2 is applied to floating point numbers $a_i \in \mathbb{F}$, $1 \leq i \leq n$ and set $p = \prod_{i=1}^{n} a_i$. Then, the absolute forward error affecting the product is bounded according to*

$$|\mathtt{res} - p| \leq \mathrm{fl}\left(\left(\mathtt{eps}|\mathtt{res}| + \frac{\gamma_n\gamma_{2n}|a_1a_2\cdots a_n|}{1-(n+3)\mathtt{eps}}\right) / \,(1-2\mathtt{eps})\right).$$

3.5. Validated error bound and faithful rounding

In the previous subsection, we have shown that

$$|e_n - e| \leq \mathrm{fl}\left(\frac{\gamma_n\gamma_{2n}|a_1a_2\cdots a_n|}{1-(n+3)\mathtt{eps}}\right). \tag{11}$$

Lemma 3.6 tells us that if $2|e - e_n| < \mathtt{eps}|\mathtt{res}|$ then `res` is a faithful rounding of p (where `res` is the result of `CompProd`).

As a consequence, if

$$\mathrm{fl}\left(2\frac{\gamma_n\gamma_{2n}|a_1a_2\cdots a_n|}{1-(n+3)\mathtt{eps}}\right) < \mathrm{fl}(\mathtt{eps}|\mathtt{res}|)$$

then we got a faitfully rounded result. This makes it possible to check *a posteriori* if the result is faithfully rounded.

4. Conclusion

In this paper, we provided an accurate algorithm for computing product of floating point numbers. We gave some sufficient conditions to obtain a faithfully rounded result as well as validated error bounds.

References

Dekker, T. J.: 1971, 'A floating-point technique for extending the available precision'. *Numer. Math.* **18**, 224–242.

Graillat, S., N. Louvet, and P. Langlois: 2005, 'Compensated Horner Scheme'. Research Report 04, Équipe de recherche DALI, Laboratoire LP2A, Université de Perpignan Via Domitia, France, 52 avenue Paul Alduy, 66860 Perpignan cedex, France.

Higham, N. J.: 2002, *Accuracy and stability of numerical algorithms.* Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), second edition.

IEEE Computer Society: 1985, *IEEE Standard for Binary Floating-Point Arithmetic, ANSI/IEEE Standard 754-1985.* New York: Institute of Electrical and Electronics Engineers. Reprinted in SIGPLAN Notices, 22(2):9–25, 1987.

Knuth, D. E.: 1998, *The Art of Computer Programming, Volume 2, Seminumerical Algorithms.* Reading, MA, USA: Addison-Wesley, third edition.

Kornerup, P., V. Lefevre, and J.-M. Muller: 2007, 'Computing Integer Powers in Floating-Point Arithmetic'. arXiv:0705.4369v1 [cs.NA].

Langlois, P. and N. Louvet: 2007, 'How to Ensure a Faithful Polynomial Evaluation with the Compensated Horner Algorithm'. In: *Proceedings of the 18th IEEE Symposium on Computer Arithmetic (ARITH '07), Montpellier, France.* pp. 141–149, IEEE Computer Society, Los Alamitos, CA, USA.

Ogita, T., S. M. Rump, and S. Oishi: 2005a, 'Accurate Sum And Dot Product'. *SIAM J. Sci. Comput.* **26**(6), 1955–1988.

Ogita, T., S. M. Rump, and S. Oishi: 2005b, 'Verified solution of linear systems without directed rounding'. Technical Report No. 2005-04, Advanced Research Institute for Science and Engineering, Waseda University.

Rump, S. M., T. Ogita, and S. Oishi: 2005, 'Accurate Floating-Point Summation'. Technical Report 05.12, Faculty for Information and Communication Sciences, Hamburg University of Technology.

Sterbenz, P. H.: 1974, *Floating-point computation.* Englewood Cliffs, N.J.: Prentice-Hall Inc. Prentice-Hall Series in Automatic Computation.

An interval based technique for FE model updating

Stefano Gabriele[1] and Claudio Valente[2]

[1]*Department of Structures, University of Rome "Roma Tre"*
gabriele@uniroma3.it
[2]*PRICOS Department, University of Chieti-Pescara "G. D'Annunzio"*
c.valente@unich.it

Abstract: Model updating techniques are largely used in civil and mechanical engineering to obtain reliable FE models. The model parameters are iteratively adjusted until the model response matches the measured structural response within a given tolerance. In this work it is assumed to know the response of a structure in terms of uncertain modal quantities. Accordingly, the model response is computed accounting for uncertainty by defining the model parameters as intervals.
The updating problem is formulated in the framework of interval analysis by exploiting the inclusion theorem. The solution is reached when the structural response is completely included by the FE model response and the parameters uncertainty is at a minimum. The presented method offers some advantages that are: each model parameter is included in a physical interval hence the solutions are guaranteed to be physical; the uncertainties of the measured response are naturally embodied into the problem. The method is discussed through a simple numerical example. The interval updating solution is then compared with conventional updating technique by applying it to a real case study.

Keywords: model updating, interval analysis, global optimization, FE model admissibility

1. Introduction

The work is framed in the field of finite element model (FEM) updating procedures (Friswell and Mottershead, 1995), that have the goal of calibrating the model parameters to get the best match between experimental and analytical modal data. In the case of civil engineering, model updating is a useful tool to know the actual state of the structures (diagnosis) and for the construction of predictive structural models (prognosis). The solution of the problem belongs to the field of inverse problems (Sorenson, 1980) and is classically faced using nonlinear programming algorithms. The objective function to be minimized is often chosen as the distance between measured and computed response quantities and the solution strategies can be distinguished by the algorithm used to search for the minimum. In structural dynamics, the updating is classically performed using modal data, that can be expressed either as the modal model or the response model of the structures (Ewins, 1984). In the case of civil engineering it is common practice to refer to modal shapes and related frequencies (Camillacci and Gabriele, 2005).
Two alternative formulations are usually followed: deterministic methods and statistical or Bayesian methods depending on the analyst preference and confidence on the uncertainty related to the problem. The first formulation makes use of deterministic or crisp parameters and measures, whereas the second

includes uncertainty through normal probability density functions or two crisp parameters and measures (average and standard deviation). In both cases the updating problem results to be ill-conditioned because of two fundamental aspects: the dependency of the problem on the ratio between the number of parameters and the number of independent measures (Gola et al., 2001) and the inherent uncertainties that characterize both the FE model (modelling errors) and the experimental data (measurement errors) (Capecchi and Vestroni, 1993).

In general no explicit uncertainty is associated to deterministic methods, for which the optimal parameters gauging is demanded to the search algorithm and to the model sensitivity. On the contrary, Bayesian approaches account for uncertainty by assigning appropriate values to the standard deviation to express confidence on the data. However, the statistical values, at least for the model data, are to be assigned a priori and strictly depend on the skill of the structural analyst (Collins et al., 1974).

The present work proposes an alternative way to treat the uncertainty in updating problems, that is based on the concepts of "interval analysis" (Moore, 1966). This methodology allows to represent uncertain quantities not by means of point values, but by bounding them inside possibility intervals. The interval width define the uncertainty level. In this respect, interval methods offer some advantages as compared to deterministic and stochastic methods in fact: they are capable to account explicitly for the uncertainties of the problem, do not need the introduction of distributions, as in the probabilistic case, let to define interval limits coherent with engineering bounds, do not require initial conditions to start the search algorithm.

Up to date various works have been issued concerning the computation of bounded eigenvalues and eigenvectors of mechanical structures (Shalaby, 2000), while it remains to deepen the possibility of using interval global optimization methods (Ratschek and Rokne, 1988, Hansen and Walster, 2004) to update parameters of FE models. The decision to use an interval approach also implies the use of interval finite element method (IFEM) for the development of numerical model to be updated. A formalized formulation of the IFEM can be found in the works of Muhanna and Mullen (1999) and Muhanna et al. (2006), also with applications in the static case. Previous applications of the static IFEM can be found in the works of Rao and Berke (1997), in Köyluoglu and Elishakoff (1998) where a comparison with probabilistic solutions is also presented. The interval FEM also finds applications in structural optimization procedures of truss structures (Pownuk, 1999). Some deepening in the dynamic case, that is of major interest for the treated arguments, can be found in Moens (2002).

The work is organized in three parts. In section 2 the basic concepts of interval analysis are given and the main properties of interval operations and functions are discussed in view of their subsequent use. In section 3 the interval model updating problem is discussed and applied to simple numerical example. In the last section the method is applied to a real test case, concerning the model updating of a simplified model of a building sub-structure, whose experimental modal data are made available by an independent experimental campaign. The method is discussed according to two possible cases: crisp or certain experimental measures and interval valued or uncertain measures.

2. Interval computations

In interval analysis (Moore 1966, Sunaga 1958) numbers are replaced by intervals in which they are contained, the larger the interval the larger the uncertainty in the evaluation of the number. An interval X could be denoted by infimum and supremum limits ($x_{\inf}$, $x_{\sup}$) or by the central notation, where the interval limits are obtained by respectively adding and subtracting the uncertainty radius Δx to the central value x_c , by way of the unit interval $e_\Delta = [-1,1]$ and by applying the interval addition rule.

$$X = \left\lfloor x_{\inf}, x_{\sup} \right\rfloor = x_c + \Delta x \cdot e_\Delta \tag{1}$$

The result of a generic interval operation "op" is the interval set of all the possible solutions when any operand varies independently in its own limits. From this definition follow the *inclusion property*, that is any possible result from the crisp operation "x op y" is included in the interval operation "X op Y", providing that $x \in X$ and $y \in Y$.
In the standard interval computations a result is generally overbounded. In this case the word "standard" means that any interval expression is evaluated according to the assumption of independency between operands. From this follows that the sharpest computed interval is evaluated from an expression that contains a minimum number of occurrences of the same operand. An example is given from the so called sub-distributivity property in equation (2).

$$X \cdot (Y + Z) \subseteq X \cdot Y + X \cdot Z \tag{2}$$

Let be $f(\mathbf{x})$ a real valued function that depends on the crisp parameters $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_n)$, there exist some ways to define its interval extension $F(\mathbf{X})$ (Moore, 1966). The interval functions considered in the paper are called natural extensions and are obtained by replacing every single occurrence x_i in the expression of f with the correspondent interval X_i in F. $F(\mathbf{X})$ maps $\mathbf{X} = (X_1,\ldots,X_n)$ into the real interval space and converges to f, i.e. $F(\mathbf{x}) = f(\mathbf{x})$, whenever $\mathbf{X}$ shrinks to crisp $\mathbf{x}$.
The inclusion property is settled for natural extensions by the inclusion theorem (Hansen and Walster, 2004). This theorem ensures, for various kind of interval extensions, that the inclusion range of $F(\mathbf{X})$ bounds all minima and maxima of $f(\mathbf{x})$ over $\mathbf{X}$. This theorem was firstly demonstrated for interval natural extensions that are also inclusion monotonic, i.e. $F(\cdot)$ is inclusion monotonic if taken $\{\mathbf{X} \subset \mathbf{Y} \mid X_i \subset Y_i, \forall\ i\}$, it follows that

$$F(\mathbf{X}) \subset F(\mathbf{Y}) \tag{3}$$

In this work it is of interest to discuss the interval analysis aspects related to inverse engineering problems. One of this is the convergence to crisp values of interval functions, and two different type of interval functions are considered. The first type, also called *thin* interval function, possesses interval variables $\mathbf{X}$ and crisp parameters p, $F(\mathbf{X},\mathbf{p})$. The thin attribute is referred to the kind of convergence of the

function as the radius $\Delta\mathbf{x}$ decrease and tends to zero. This is shown in the graphical example of Figure 1a, where the continuous line represents the crisp evaluated function $f(\mathbf{x},\mathbf{p})$, whereas the progressive decreasing monotonic boxes are the interval representation of its natural extension. From the figure is seen that as $\Delta x_i \to 0$, $\forall$ i, then $X_i \to x_{ci}$ and $F(\mathbf{X},\mathbf{p}) \to f(\mathbf{x}_c,\mathbf{p})$. The second type, also called *thick* interval function, possesses both interval variables $\mathbf{X}$ and parameters $\mathbf{P}$, $F(\mathbf{X},\mathbf{P})$. For thick functions only a relaxed type of convergence can be defined. In fact, if $F(\cdot,\mathbf{P})$ is evaluated on crisp $\mathbf{x}_c$ of the Figure 1b, then the best that can be obtained is that $F(\mathbf{x}_c,\mathbf{P}) \supset f(\mathbf{x}_c,\mathbf{p})$, $\forall\ \mathbf{p}\in\mathbf{P}$, but not equals it at $\mathbf{x}_c$. In this case the thick attribute refers to the impossibility of converging to crisp values of this type of functions, as a consequence of the presence of interval parameters $\mathbf{P}$ inside the function expression. From a geometrical point of view, as $\Delta x_i \to 0$, $\forall$ i, $F(\mathbf{X},\mathbf{P})$ converge to a segment, and covers a bundle of crisp functions.
This distinction between thin and thick function and their different kind of convergence are the main concepts embodied into the presented method together with the inclusion property. In fact, solutions to mechanical problems are considered physically plausible only if the method guarantees the inclusion of the experimental outcomes.

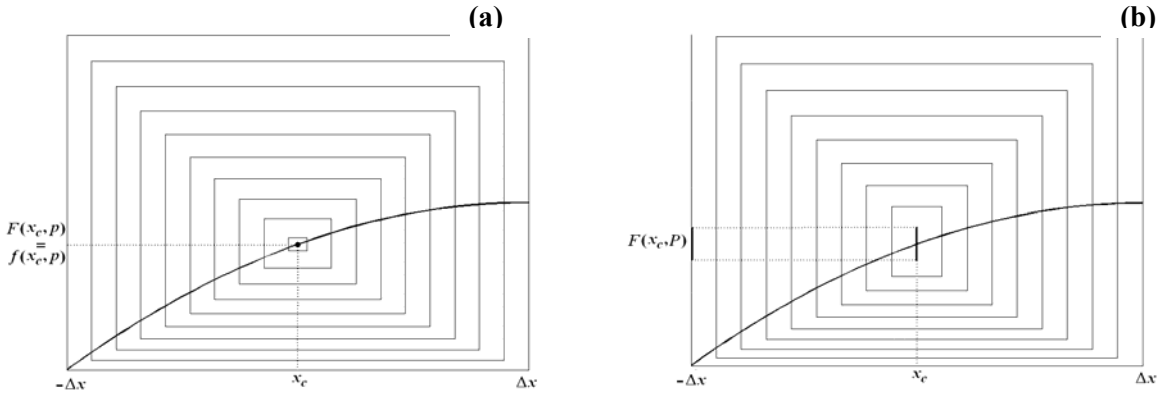


Figure 1 – Monotonic convergence of (a) thin function, (b) thick function

2.1 MODEL FUNCTION AND INTERVAL SOLUTION

The previously presented concepts about interval functions are now applied to define which kind of model are used in the present work, which kind of quantities are affected by uncertainty and which kind of interval solutions need to be computed.
One remembers that the purpose of the paper is to present an interval model updating procedure and the models to be updated are finite element (FE) representations of mechanical structures.

The formulation of the interval FE method can be found in Muhanna (1999) and Moens (2002), where uncertainties can appear in the mass and stiffness coefficients as well as in the geometry of the system. However, in the present context, only the constitutive parameters of the model are considered affected by uncertainty, therefore the stiffness matrix **K** is an interval matrix.
In view of model updating applications eigenvalues and eigenvectors of the system need to be computed and then compared with the experimental counterparts. In general these measured quantities are uncertain, due to experimental errors, and could be bounded in confidence intervals, that one wants to reproduce with the updated FE model.
For this purpose the model functions are defined as $\{\Lambda(\mathbf{K}), U(\mathbf{K})\}$, respectively the interval eigenvalues and eigenvectors, that depend on the interval stiffness variables that are present in the stiffness matrix **K**. $\Lambda(\mathbf{K})$ and $U(\mathbf{K})$ can also be considered as the interval extensions of the crisp eigenfunctions $\lambda(\mathbf{k})$ and $u(\mathbf{k})$.
According to the interval FE method, interval enclosure $\mathbf{\Lambda}=\Lambda(\mathbf{K})$ and $\mathbf{U}=U(\mathbf{K})$ can be computed as the solution of the generalized algebraic problem:

$$\mathbf{KU} = \mathbf{\Lambda m U} \tag{4}$$

To solve this problem one should find the inclusion set for the eigenvalues, defined as

$$\mathbf{\Gamma} = \left\{ \boldsymbol{\lambda} \in \mathrm{R}^{\mathrm{n}}, \mathbf{u} \in \mathrm{R}^{\mathrm{n}\times\mathrm{n}} \mid \mathbf{ku} = \boldsymbol{\lambda}\mathbf{mu}, \mathbf{u} \neq 0, \mathbf{k} \in \mathbf{K} \right\} \tag{5}$$

Unfortunately the methods proposed in the literature are capable to find the true solution $\mathbf{\Lambda} = \mathbf{\Gamma}$ only in limited cases (Shalaby, 2000) being the wider solution $\mathbf{\Lambda} \supset \mathbf{\Gamma}$ the only available solution for the general cases. Presently, a solution strategy similar to that developed by Qiu and Chen (1995) is followed. The choice comes from the observation that the interval computations can be replaced by crisp operations on the interval limits yet preserving the monotonic inclusion of the solution (Chiao, 1999). According to Qiu and Chen the problem (4) specialised for the *j*-th eigenvalue is written as:

$$\left(\mathbf{k}_{\mathrm{c}} + \Delta\mathbf{k} \cdot e_{\Delta}\right)\mathbf{U}_j = \Lambda_j \mathbf{m}\mathbf{U}_j \tag{6}$$

and the solution bounds are then computed by solving two crisp sub-problems obtained according to the following general interval property:

$$\begin{aligned} &\mathbf{K}\mathbf{u}_j = \mathbf{k}_c\mathbf{u}_j + \Delta\mathbf{k}\left|\mathbf{u}_j\right| \cdot e_{\Delta} = \mathbf{k}_c\mathbf{u}_j + \Delta\mathbf{k}\left(\mathbf{s}_j\mathbf{u}_j\right) \cdot e_{\Delta}; \\ &\text{with } \left|\mathbf{u}_j\right| = \mathbf{s}_j\mathbf{u}_j \text{ and } \mathbf{s}_j = diag\left(sign\left(\mathbf{u}_j\right)\right) \end{aligned} \tag{7}$$

that guarantees the monotonic inclusion in the working range. The infimum and supremum limits are obtained by the following expressions:

$$\begin{cases} \left(\mathbf{k}_c - \mathbf{s}_j^T \Delta\mathbf{k} \;\; \mathbf{s}_j\right)\mathbf{u}_{j,\inf} = \lambda_{j,\inf}\mathbf{m}\,\mathbf{u}_{j,\inf} \\ \left(\mathbf{k}_c + \mathbf{s}_j^T \Delta\mathbf{k} \;\; \mathbf{s}_j\right)\mathbf{u}_{j,\sup} = \lambda_{j,\sup}\mathbf{m}\,\mathbf{u}_{j,\sup} \end{cases} \tag{8}$$

Equations (8) gives $\mathbf{\Lambda} = [\boldsymbol{\lambda}_{\inf}, \boldsymbol{\lambda}_{\sup}] \supset \mathbf{\Gamma}$ where the over-bounding depends on the uncertainty radius $\Delta\mathbf{k}$. It has been shown in Moens (2002) that the above formulation ensures to include all the true solutions when the eigenvalues of the system are properly spaced.
The solution of the equations (8) is found under the hypothesis of sign invariance of the *j*-th eigenvector (Deif and Rhon, 1994). This restriction on the allowable eigenvectors is again necessary to guarantee the preservation of the inclusion property. In view of the updating problem it is also important to guarantee that the interval method used to calculate the function values $\{\mathbf{\Lambda}=\Lambda(\mathbf{K}),\ \mathbf{U}=U(\mathbf{K})\}$ is inclusion monotonic, in this case it is demonstrated (Hansen and Walster, 2004) that, for the defined extensions, the inclusion theorem holds. The authors are aware that exist many methods to compute interval eigenvalues and that the selected method is characterized by a great overbounding of the interval estimation. But the inclusion theorem validity it is, at author's judgment, more important for optimization problems applied to physical systems than the overbounding, at this work stage. This will be better explained in the following sections.

3. Interval model updating

The inverse model updating problem is classically formulated as the search for the minimum of a predefined objective function $l(\mathbf{x})$ that depends on the vector of the updating unknowns $\mathbf{x}$:

$$\min_{\substack{\mathbf{x}\in\mathbf{D} \\ \mathbf{D}\in R^n}} l(\mathbf{x}) \tag{9}$$

In those cases in which a matching between two sets of quantities is sought for, $l(\mathbf{x})$ is conveniently expressed as a measure of the distance between experimental and numerical quantities and a least squares formulation is followed (Camillacci and Gabriele, 2005). Therefore, in the present case, the objective function is specialized as the 2 norm distance:

$$l(\mathbf{k}) = \left\|\boldsymbol{\lambda}_s - \lambda(\mathbf{k})\right\|_2^2 \tag{10}$$

where the unknowns are stiffness variables, that are present in the matrix **k**, and where only the contribution of the eigenvalues is considered. A more general form than (10) would comprise the contribution of the eigenvectors as well (Gola et al. 2001), but this further sophistication is not within the purposes of the paper.
On the contrary, the interval model updating problem is here discussed.

3.1 INTERVAL GLOBAL MINIMIZATION

First of all the minimization of the error norm defined by equation (10), is a nonlinear programming problem and is possible to solve it in an interval space by interval global optimization algorithms (Hansen and Walster, 2004; Ratschek and Rockne, 1988). Such algorithms are generally comprised in the so called branch and bound methods (B&B), where an initial search domain is iteratively subdivided in smaller sub-domains and, for each created sub-domain, a first criterion (bounding step) is applied to verify if the sought solution could be contained in it or not. In the first case a sub-domain survives and a second criterion is applied to subdivide it again. In the second case the evaluated sub-domain is discarded from the search. The found solution is finally given by the surviving sub-domains. Interval B&B methods can be developed thanks to the existence of the interval inclusion theorem. In fact a non verified inclusion into generated sub-domains can be used as discarding criterion. Inclusions need to be verified by defining a proper interval extension of the crisp function to be minimized and a proper extension is that for which the inclusion theorem can be demonstrated, for example the class of inclusion monotonic extensions.
All the concepts briefly explained are now applied to the original updating problem defined by equation (10). The model updating problem is not only a mathematical programming problem, it is also a physical problem defined with some uncertainties, and is here important to underline the differences that arise with respect to a conventional setting. In fact, different cases should be accounted for, depending on which quantities are affected by the uncertainty.

1. The uncertainty source is only in the FE model stiffness parameters (**K**); in this case a natural extension for (10) is written as (11)

$$L(\mathbf{K}) = \left\| \boldsymbol{\lambda}_s - \Lambda(\mathbf{K}) \right\|_2^2 \tag{11}$$

where $\boldsymbol{\lambda}_s$ is the crisp vector of the experimental eigenvalues, $\Lambda(\mathbf{K})$ is the interval extension by which the FE model interval eigenvalues, $\boldsymbol{\Lambda}$, are calculated. If $\Lambda(\mathbf{K})$ is evaluated by the equations (8), it is inclusion monotonic and hence $L(\mathbf{K})$ is also inclusion monotonic.
$L(\mathbf{K})$ is a thin interval extension, because as $\Delta k_i \to 0$, $\forall$ i, then $K_i \to k_{ci}$ and $L(\mathbf{K}) \to f(\mathbf{k}_c)$. That means that the crisp updating problem (10) and the interval updating problem (11) have the same crisp solution, in the limit that the uncertainty approach to zero.

The solution in this case can be effectively obtained through standard interval B&B optimization techniques (Hansen and Walster, 2004; Jansson and Knüppel, 1995) that give good results even in the case of ill-conditioned problems and that, in the limit $\Delta\mathbf{k} \rightarrow 0$, converge to crisp solutions.

2. In a second case the uncertainty source is both in the experimental measures and in the model parameters; in this case it is required to match the interval vectors $\mathbf{\Lambda}_s$ and $\mathbf{\Lambda}=\mathit{\Lambda}(\mathbf{K})$.
   The natural extension of equation (10) is in this case given by

$$L(\mathbf{K}) = \left\| \mathbf{\Lambda}_s - \mathit{\Lambda}(\mathbf{K}) \right\|_2^2 \tag{12}$$

   and the above expression for $L(\mathbf{K})$ cannot be used unless a metric between intervals is introduced to replace the standard metric between crisp values (Moore, 1966).
   The equation (12) also defines an interval thick extension of (10), due to the fixed uncertainty in the experimental measures vector $\mathbf{\Lambda}_s$. In this case the objective of reducing the stiffness uncertainty, with the goal to converge around an optimal solution, is limited by this fact and numerical solutions can only be found with a final fixed uncertainty.

In the case 2., instead of introducing a metric between intervals, an approach coherent with the principles of interval analysis is proposed and named *Interval Intersection Method* (INTIM, Gabriele, 2004), where the basic branch and bound optimization technique present in Hansen (2004) is adopted.
One supposes to define the interval search domain $\mathbf{D}$, with $K_\mathrm{i} \in \mathbf{D}$, $\forall$ i. The branching step is left unchanged and its repeated application produces progressively smaller sub-domains, $\mathbf{D}_\mathrm{j} \subset \mathbf{D}$, in which the searched stiffness parameters are included. In the bounding step the basic operations between sets are applied to interval solutions $\mathbf{\Lambda}$, to verify the inclusion of the experimental vector $\mathbf{\Lambda}_s$. If the function $\mathit{\Lambda}(\mathbf{K})$ is inclusion monotonic and the inclusion theorem is verified, then the verified inclusion of $\mathbf{\Lambda}_s$ in $\mathbf{\Lambda}$ means that the FE model, endowed with the interval parameters $\mathbf{K}$, is capable to represent the experimental solution. This capability is here intended as FE *model admissibility*.
The degree of admissibility of a model, in the parameters domain, with respect to the known measured response is hence simply checked using the intersection operation to verify the inclusion:

$$\left\{ \mathbf{\Lambda}_s \cap \mathit{\Lambda}(\mathbf{K}) = \mathbf{\Lambda}_s, K_\mathrm{i} \in \mathbf{D}_\mathrm{j}, \forall \mathrm{i} \right\} \tag{13}$$

It can be assumed that if the total inclusion is not verified the model is not admissible to represent the real structure in the considered domain, in fact if the inclusion theorem holds no other eigensolution can be found outside the calculate interval ($\mathbf{\Lambda}$).
By inverting the previous statement, the equation (13) can be taken as exclusion criterion in the B&B search algorithm, for the sub-domains $\mathbf{D}_\mathrm{j}$, in the following pessimistic form:

$$\left\{ \mathbf{\Lambda}_s \cap \mathit{\Lambda}(\mathbf{K}) = \varnothing, K_\mathrm{i} \in \mathbf{D}_\mathrm{j}, \forall \mathrm{i} \right\} \tag{14}$$

In the interval updating algorithm the solution is iteratively found. Starting from the whole parameters space **D**, this is consecutively branched in sub-domains $\mathbf{D}_j$ that, in turn, are preserved or discarded according to (13) and (14). The procedure stops when for some $\mathbf{D}_j$ the criterion (13) holds and a pre-fixed radius of minimum uncertainty tolerance is reached, so they cannot be further branched.
It is important to note that the above procedure is indeed general and can be applied to case 1. as well, when $\boldsymbol{\lambda}_s$ is a crisp measures vector. Both cases and solution procedures are illustrated according to the numerical simulation discussed below.
In Figure 2 is depicted a graphical representation of the branching and the bounding steps, by thinking to apply admissibility criteria (13) and (14) for each generated sub-domain in the initial search box $\mathbf{K}_0$. In the figure the arrows represent the applications of the interval extension $\Lambda(\mathbf{K})$, in order to obtain the intervals $\boldsymbol{\Lambda}$ to be compared with $\boldsymbol{\Lambda}_s$ in the measures space.

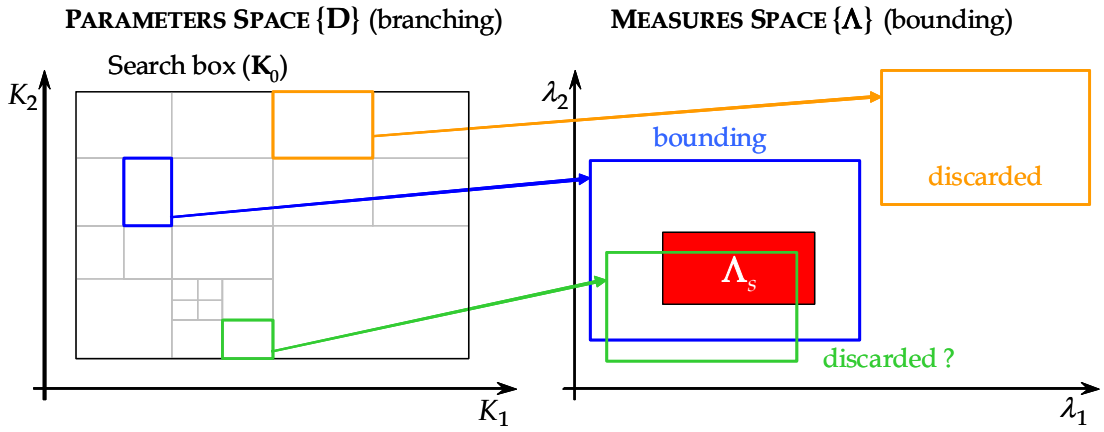


Figure 2 – Interval intersection method – branching and bounding step

3.2 NUMERICAL EXAMPLE

A simple mechanical system composed by a 2dofs mass-spring system is considered. The problem is again to find the parameters vector $\mathbf{k} = [k_1, k_2]$ that produces the best match between given experimental and numerical modal data. The mechanical system is used either to generate pseudo-experimental data or to compute the numerical frequencies according to the formulation given in section 2.1. It is initially assumed that the experimental frequencies are exactly identified (case 1.) and only the parameters are affected by uncertainty. Then, also the experimental frequencies are assumed to be identified within an interval (case 2.).

The uncertainty free pseudo-experimental frequencies are $\mathbf{f}_s$ = [0.082, 0.307] Hz, ($\lambda_s = 2\pi f_s^2$), and it corresponds to $\mathbf{k}_0$ = [1, 2]. In a full deterministic setting the 2-norm objective function (10) applies and conventional minimization schemes can be used. However, even in this simple situations the objective function can have more than one minimum as shown in Figure 3a where it is plotted in the form of a contour plot representation. In particular, the function has two global minima: one for the true vector of parameters $\mathbf{k}_0$ and the other for $\mathbf{k}$ = [3, 2/3]. Depending on the search domain, the solution algorithm and the initial value of the parameters the false minimum can be reached by the a crisp updating procedure.

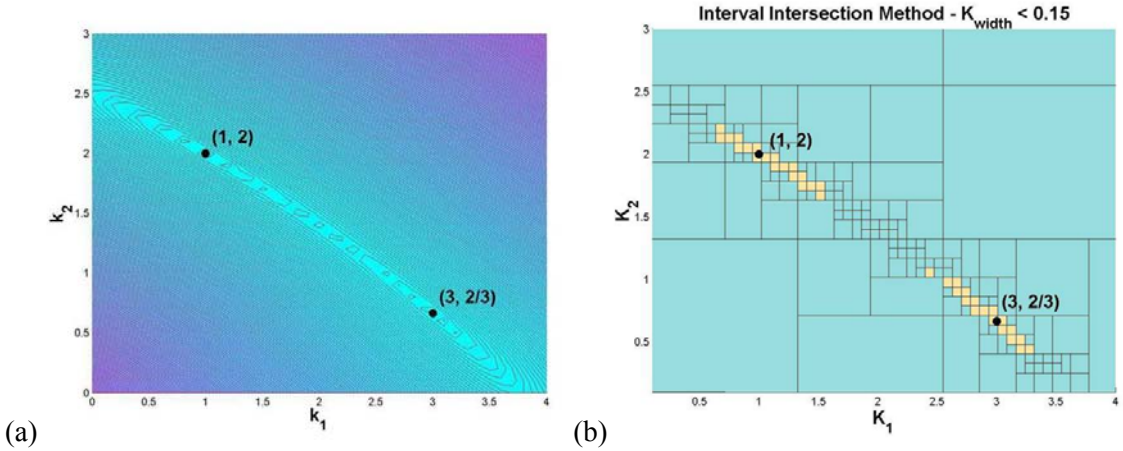


Figure 3 – (a) Crisp objective function, (b) INTIM solution for case 1.

In the case of crisp experimental (solution case 1.) INTIM technique is applied starting from the initial parameters domain $\mathbf{D} \supseteq \mathbf{K}_0$ = [[0,4], [0,3]] and the result is shown in Figure 3b, where the progressive partition in finer sub-domains tending to accumulate around the minima. The partitioning stops when for some $\mathbf{D}_j$ the criterion (13) holds and the radius of minimum uncertainty tolerance k_w, is reached, so that $\mathbf{D}_j$ cannot be further branched. In the figure the solution domain is given by the collection of the lightest boxes that are those for which $k_w < 0.15$. The uncertainty in the obtained solution is measured by the spread of the lightest boxes around the crisp minima.

In the case of interval valued experimental data the considerations done for solution case 2. are valid. Now the pseudo-experimental eigenvalues are collected in the interval vector $\mathbf{\Lambda}_s$ = [[0.08, 0.45], [3.55, 3.92]] and it corresponds to $\mathbf{K}^*$ = [[0.99, 1.01], [1.98, 2.02]].

The solution in the parameter space is given in Figure 4a. In the present case $\mathbf{k}_w$ = 0.31 and the solution is slightly more confined with a reduced number of branches, but with larger final boxes. Here again two distinct sub-domains solutions are detected.

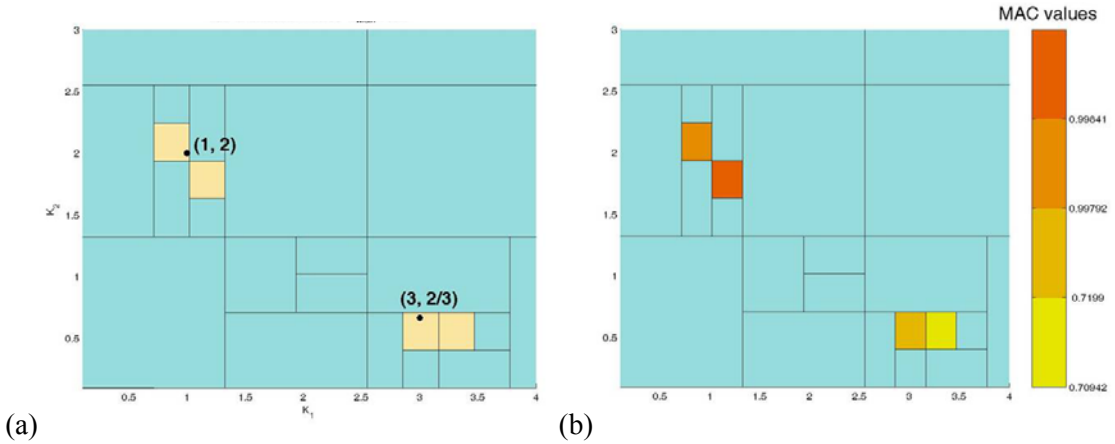


Figure 4 – INTIM solutions for case 2. (a) bounded solution, (b) solution choice

3.2.1. *Solution choice*

As shown by the example above, one of the main advantages of INTIM is the capability to find all the minima in the parameters space, in the form of a collection of boxes. Anyway, no further distinction between minima can be done to locate the box of the true parameters.

To make a choice among all the possible solutions an a posteriori processing of the solution is performed on the base of a choice criterion. If the experimental modal shapes are known the modal assurance criterion MAC can be used:

$$\mathrm{MAC}_{s,n} = \frac{\left(\mathbf{u}_s^T \cdot \mathbf{u}_n\right)^2}{\left(\mathbf{u}_s^T \cdot \mathbf{u}_s\right) \cdot \left(\mathbf{u}_n^T \cdot \mathbf{u}_n\right)} \tag{15}$$

where $\mathbf{u}_n$ and $\mathbf{u}_s$ stand for numerical and experimental central values of the eigenvectors. It is worth recalling that $0 \leq \mathrm{MAC} \leq 1$ and $\mathrm{MAC} = 1$ whenever $\mathbf{u}_n = \mathbf{u}_s$.

The MAC values have been computed for all the solution boxes in Figure 4a and have been reported in Figure 4b as a color scale superposed to the parameters domain. The darkest box is the parameters interval endowed with the highest MAC that is therefore chosen as the updating solution.

4. Real case study

The case study is taken from the ILVA-IDEM project in which one of the authors is involved (Mazzolani et al., 2004; Cardellicchio, Spina and Valente, 2004; Valente, Spina and Nicoletti, 2006). The experimental results were obtained during a large experimental campaign aimed at evaluating the

mechanical and strength characteristics of an existing reinforced concrete building that can be considered representative of many gravity-load designed reinforced concrete buildings located in the South of Italy, Figure 5a.

(a)  (b) 

Figure 5 – Case study: (a) Original building, (b) Selected structural module

	TR - Transient tests	SS - Steady state tests
1st longitudinal mode	[1.76, 1.87]	[1.74, 1.85]
2nd longitudinal mode	[5.65, 5.72]	[5.20, 5.65]

Table 1 – Identified experimental frequencies (Hz) of the longitudinal main frame.

	First floor	Second floor
Beams	1.75e10	1.75e10
Columns	[1.35, 1.90]e10	[0.89, 1.21]e10

Table 2 – Measured Young modulus (N/m^2) of the structural members.

Partition walls and external claddings were removed and, then, the original building was divided into six separate smaller structures (Figure 5b). Four of them, nominally identical each other, were subjected to dynamic testing aimed at identifying the modal model and at evaluating the scatter in the results. Different types of tests were performed by changing the excitation type. Impulse excitations were used to provide transient response (test TR) and harmonic excitations were used to provide steady state response (test SS). The analysis of the dynamic response was performed through well established methods (Ewins, 1984) and frequencies and modal shapes were identified for the first six modes. For the present purposes, only a small set of the whole available data are considered. They are referred to the structural module marked in Figure 5b. Further, for simplicity, only the modal behaviour in the plane of the main frames is considered Figure 6a. The frequencies of the first two longitudinal modes have been used in the updating procedure and their interval variation is shown in Table 1.

The mechanical properties of the concrete were measured in laboratory on core samples extracted from the structure and on site using NDT tests to check for the concrete uniformity. The results are given in Table 2, from which it is apparent that the uncertainty is limited to the columns.

4.1 RESULTS

The INTIM described in section 3 is applied to the 2D model of Figure 6b in order to update the stiffness of the columns. The Young modulus E is the parameter to update since it acts as a scale factor for the columns stiffness. It is assumed that the columns of a floor have all identical stiffness, therefore two interval values E_1 and E_2 are sought for, one per floor. A wide and identical intervals $E_1 = E_2 = [0.1, 3]\times10^{10}$ N/m^2 has been chosen to be the initial search domain **D**. A physical justification can be given to this choice in consideration of the large uncertainties related to the NTD tests, but it is unnecessary since it is the ability of the technique to work with box domains and its numerical efficiency that suggest to widen **D** in order to get a complete picture of the solution.

4.1.1. *Solution case 1.*

It is interesting to observe that if a crisp model updating procedure would be used, together with the initial conditions equal to the average values of the measured elastic moduli of Table 2 ($\mathbf{E}_s$), the following crisp values would be found: $e_1 = 2.27\times10^{10}$ N/m^2, $e_2 = 0.33\times10^{10}$ N/m^2 (Figure 7). They are very far from those listed in Table 2 that can be considered the physical solution range so that one can wonder if the adopted FE model is adequate to the problem or it should be revised.

In this case the interval solution calculated by INTIM and applied by choosing $\boldsymbol{\lambda}_s$ as the central values of the uncertain experimental measures (solution case 1.), is again far from the measured elastic moduli box ($\mathbf{E}_s$). But INTIM solution puts in evidence the presence of a second solution sub-domain that have a not null intersection with $\mathbf{E}_s$ only along the E_2 axis.

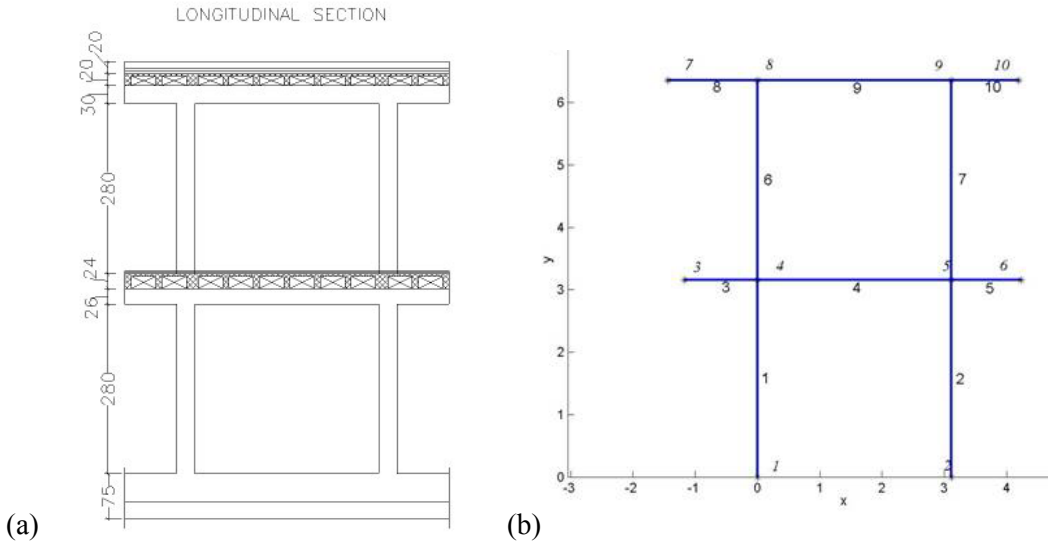


Figure 6 – Case study: (a) Longitudinal main frames, (b) 2D FE model

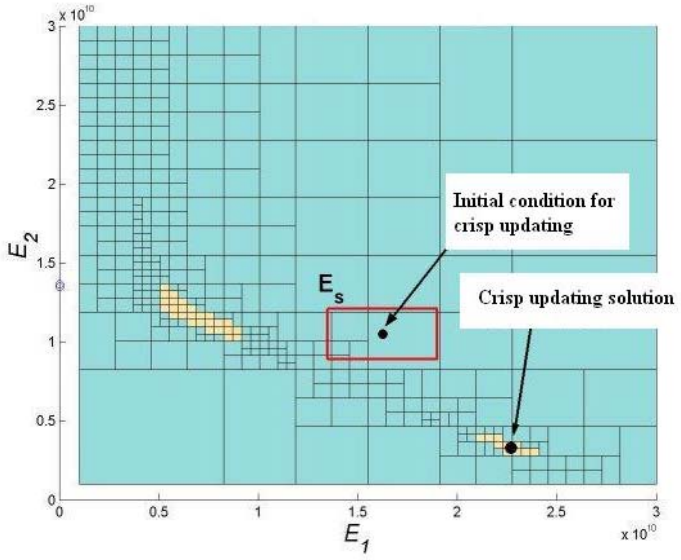


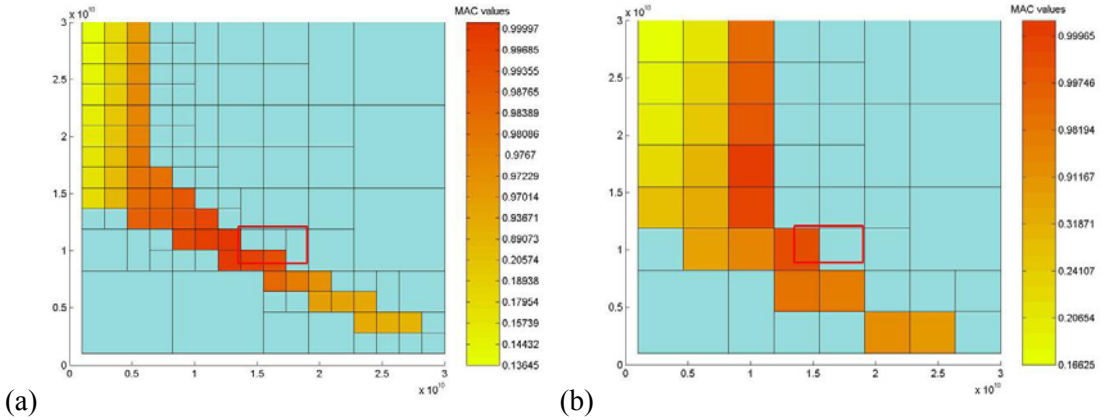
Figure 7 – Comparison between crisp updating solution and INTIM solution case 1.

4.1.2. *Solution case 2.*

The presented case study is really uncertain, as it clear from the values in Tables 1 and 2, and could be wrong to consider crisp values of the experimental frequencies as objective of the updating procedure.

If one considers the full measures uncertainty, firstly the interval updating technique can be used initially to check the admissibility of the FE model and then to find the solution. In the presence of experimental evaluations for the elastic moduli $\mathbf{E}_s$, admissible FE models are those for which the experimental eigensolution ($\mathbf{\Lambda}_s$) is completely included by the model response ($\mathbf{\Lambda}$), and parameters solution ($[E_1,E_2] \in \mathbf{D}_j$) has at least one non vanishing intersection with the experimental box $\mathbf{E}_s$.

The solution of the interval updating procedure is shown in the parameter space in Figure 8, where the results obtained from test TR and test SS are both reported. The empty rectangle shown in the figures is the box $\mathbf{E}_s$ of Table 2 and the most feasible solutions are in the color scale of the MAC values.



(a) (b)

Figure 8 – INTIM solution case 2. (a) test TR, (b) test SS

A finer partitioning in Figure 8a than in Figure 8b can be appreciated. In fact, for test TR and SS the stop tolerances are different, 1.83×10^9 N/m^2 and 3.67×10^9 N/m^2 respectively, because of the different level of uncertainties in the identified frequencies, see Table 1. Anyway, as expected, the qualitative behaviour of the solution is similar in the two cases.

It is worth to recall that all the boxes that are possible solutions have a full intersection in the frequency space (13). The MAC values are then used to discriminate among the possible solutions. From the figures, it can be observed that the darkest boxes are also those closest to the rectangle of the measured elastic moduli. Finally it deserves to underline that one could want to shrink the boxes radius to a point in order to have a crisp solution. However the convergence to a point value is not guaranteed in the parameters space, since in the frequency space it can happen that further radius reductions pushes the results of $\Lambda(\mathbf{K})$ to interval boxes for which the inclusion rule (13) does not apply, that is to say that in this case the objective function is an interval thick extension.

5. Conclusions

In this work the application of an interval updating technique is discussed and applied in both numerical and experimental cases. The technique is applied to uncertain interval FE models by taking distinct the case of certain (crisp) measures (solution case 1.) from the case of uncertain (interval) measures (solution case 2.). For this second case the updating approach, consistent with the principles of interval analysis and set theoretic comparisons, is presented and here called interval intersection method, INTIM. In the solution case 1. standard updating techniques, based on crisp objective function, and the new one are compared, by applying them to a simple 2dofs mechanical system. For the interval solution case 1. it is pointed out that all the admissible solutions can be found.
In the solution case 2. only the INTIM technique has been applied. In this case further developments are given for the choice of physical solutions in the FE model parameters space, based on MAC comparison of modal shapes. The interval intersection method is first numerically validated by applying it to the previous 2dofs system, is then applied for updating the column stiffness of a 2D model of an r/c experimented structure, by tacking its longitudinal modal behaviour. The obtained interval best results, in the parameters space, are found to be intersected with the interval of the equivalent measured mechanical properties.

Acknowledgements

The experimental results used in the work were obtained during a large experimental campaign aimed at evaluating the seismic performance of reinforced concrete structures trough tests up to collapse. The project was carried out by the ILVA-IDEM research group whose partners are Università “Federico II” di Napoli, Università della Basilicata, Università “G. d’Annunzio” di Chieti-Pescara and the Servizio Sismico Nazionale DPC.

References

Camillacci R. and S. Gabriele. Mechanical Identification and Model Validation for Shear-Type Frames, *Mechanical System and Signal Processing*, Vol. 19, Issue 3, pp. 597-614, 2005.

Capecchi D. and F. Vestroni. Identification of finite element models in structural dynamics, *Eng. Struct.* Vol. 15, No 1, pp. 21-30, 1993.

Cardellicchio S., Spina D. and C. Valente. Evaluation of the structural damage of r/c buildings through the use of modal parameters: the ILVA-IDEM project. In *Proceedings of XI Italian National Congress “L’ingegneria Sismica in Italia”*, Genoa, Italy, 2004.

Chiao K.P. Inclusion Monotonic Property of Courant–Fischer Minimax Characterization on Interval Eigenproblems for Symmetric Interval Matrices, *Tamsui Oxford Journal of Mathematical Sciences*, vol. 15 -pp.11-22, 1999

Collins J.D., Hart G. et. al. Statistical identification of structures, *AIAA Journal*, Vol. 12, No 2, pp. 185-190, 1974.

Deif A.S. and Rhon J. On the invariance of the sign pattern of matrix eigenvectors under perturbation. *Linera Algebra Appl*, 196, pp. 63-70, 1994.

Ewins D.J. *Modal Testing: Theory and Practice*, Research Studies, John Wiley and Sons, New York, 1984.

Friswell M.I. and J. E. Mottershead. *Finite Element Model Updating in Structural Dynamics*, Kluwer Academic Publishers, 1995.

Gabriele S. *FE Model Updating by Interval Analysis Techniques* (in Italian). PhD Thesis: University "Roma Tre", 2004.

Gola M.M., Somà A. and D. Botto. On theoretical limits of dynamic model updating using a sensitivity-based approach, Journal of Sound and Vibration, 244(4), pp. 583-595, 2001.

Köyluoglu H.U. and I. Elishakoff. A comparison of stochastic and interval finite elements applied to shear frames with uncertain stiffness properties. *Computer and Structures*, 67, pp. 91-98, 1998.

Hansen E.R. and G. Walster. *Global Optimization Using Interval Analysis*, *Second Ed*. Marcel Dekker, Inc., New York, 2004.

Jansson C. and O. Knüppel. A branch and bound algorithm for bound constrained optimization problems without derivatives. *Journal of Global Optimization*, 7, pp. 297-331, 1995.

Mazzolani F.M., Claderoni B., Spina D. and C. Valente. Structural identification of the existing building: the ILVA-IDEM project. In *Proceedings of XI Italian National Congress "L'ingegneria Sismica in Italia"*, Genoa, Italy, 2004.

Moens D., *A non-probabilistic finite element approach for structural dynamic analysis with uncertain parameters*. PhD Thesis: KU Leuven, 2002.

Moore R.E. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 1966.

Muhanna R.L. and R.L. Mullen. Formulation of fuzzy finite-element methods for solid mechanics problems. *Computer-Aided Civil and Infrastructure Engineering*, 14, pp. 107-117, 1999.

Muhanna L., Solin, Kreinovich V., Chessa J., Araiza R. and G. Xiang. Interval Finite Element Methods: New Directions. In *Proc. of the NSF Workshop on Reliable Engineering Computing, Modeling Errors and Uncertainty in Engineering Computations*, Savannah, Georgia Usa, 2006.

Pownuk A. Optimization of mechanical structures using interval analysis, *Computer Assisted Mechanics and Engineering Science*, 1-7, 1999.

Qiu Z., Chen S. and H. Jia . The Rayleigh Quotient Iteration Method for Computing Eigenvalue Bounds of Structures with Bounded Uncertain Parameters, *Computer and Structures*, Vol. 55 No. 2, pp. 221-227, 1995.

Rao S.S. and L. Berke. Analysis of uncertain structural systems using interval analysis, *AIAA Journal*, Vol. 35, No 4, pp. 727-735, 1997.

Ratschek H., Rokne J.. *New Computer Methods for Global Optimization*. Ellis Horwood Publ., 1988.

Shalaby A.M. The Interval Eigenvalue Problem: Review Article, In *Proceedings of ECCOMAS 2000*, Barcelona, 2000.

Sorenson H.W. *Parameter Estimation*, Marcel Dekker, Inc., New York, 1980.

Sunaga T. Theory of an interval algebra and its application to numerical analysis. *RAAG Memoirs*, 2, pp. 547-564,1958.

Valente C., Spina D. and M. Nicoletti M. Dynamic testing and modal identification. In *Seismic upgrading of r/c buildings by advanced techniques – The ILVA-IDEM project*. Monza/Italy: Polimetrica International Scientific Publisher; 2006.

Static Analysis of Uncertain Structures Using Interval Eigenvalue Decomposition

[1]Mehdi Modares and [2]Robert L. Mullen

[1]Department of Civil and Environmental Engineering
Tufts University
Medford, MA, 02155
email: mehdi.modares@tufts.edu

[2]Department of Civil Engineering
Case Western Reserve University
Cleveland, OH, 44106
email: rlm@case.edu

Abstract: Static analysis is an essential procedure to design a structure. Using static analysis, the structure's response to the applied external forces is obtained. This response includes internal forces/moments and internal stresses that is used in the design process. However, the mechanical characteristics of the structure possess uncertainties which alter the structure's response. One method to quantify the presence of these uncertainties is interval or unknown-but-bounded variables.

In this work a new method is developed to obtain the bounds on structure's static response using interval eigenvalue decomposition of the stiffness matrix. The bounds of eigenvalues are obtained using monotonic behavior of eigenvalues for a symmetric matrix subjected to non-negative definite perturbations. Moreover, the bounds of eigenvectors are obtained using perturbation of invariant subspaces for symmetric matrices. Comparisons with other interval finite element solution methods are presented. Using this method, it has shown that obtaining the bound on static response of an uncertain structure does not require a combinatorial or Monte-Carlo simulation procedure.

Keywords: Statics, Analysis, Interval, Uncertainty

1. Introduction

In design of structures, the performance of the structure must be guaranteed over its lifetime. Moreover, static analysis is a fundamental procedure for designing reliable structure that are subjected to static or quasi-static forces induced by various loading conditions and patterns.

However, in current procedures for static analysis of structural systems, the existence of uncertainty in either mechanical properties of the system or the characteristics of forcing function is generally not considered. These uncertainties can be attributed to physical imperfections, modeling inaccuracies and system complexities.

Although, in a design process, uncertainty is accounted for by a combination of load amplification and strength reduction factors that are based on probabilistic models of historic data, consideration of the effects of uncertainty has been removed from current static analysis of structural systems.

In this work, a new method is developed to perform static analysis of a structural system in the presence of uncertainty in the system's mechanical properties as well as uncertainty in the magnitude of loads. The presence of these uncertainties is quantified using interval or unknown-but-bounded variables.

This method obtains the bounds on structure's static response using interval eigenvalue decomposition of the stiffness matrix. The bounds of eigenvalues are obtained using the concept of monotonic behavior of eigenvalues for a symmetric matrix subjected to non-negative definite perturbations. Furthermore, the bounds of eigenvectors are obtained using perturbation of invariant subspaces for symmetric matrices. Using this method, it has shown that obtaining the bound on static response of an uncertain structure does not require a combinatorial or Monte-Carlo simulation procedure.

2. Deterministic Static Analysis

The equation of equilibrium for a multiple degree of freedom structure is defined as a linear system of equations as:

$$[K]\{U\} = \{P\} \tag{1}$$

where, $[K]$ is the stiffness matrix, $\{U\}$ is the vector of unknown nodal displacements, and $\{P\}$ is the vector of nodal forces. The solution to this system of equation is:

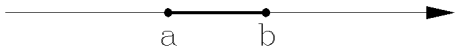
$$\{U\} = [K]^{-1}\{P\} \tag{2}$$

3. Interval Variables

The concept of interval numbers has been originally applied in the error analysis associated with digital computing. Quantification of the uncertainties introduced by truncation of real numbers in numerical methods was the primary application of interval methods (Moore 1966).

A real interval is a closed set defined by extreme values as (Figure 1):

$$\tilde{Z} = [z^l, z^u] = \{z \in \Re \mid z^l \le z \le z^u\} \tag{3}$$



$$\tilde{x} = [a, b]$$

Figure 1. An interval variable.

In this work, the symbol (~) represents an interval quantity. One interpretation of an interval number is a random variable whose probability density function is unknown but non-zero only in the range of interval.

Another interpretation of an interval number includes intervals of confidence for α-cuts of fuzzy sets. The interval representation transforms the point values in the deterministic system to inclusive set values in the system with bounded uncertainty.

3. Interval Static Analysis

Considering the presence of interval uncertainty in stiffness and force properties, the system of equilibrium equations, Eq.(1), is modified as an interval system of equilibrium equation as:

$$[\tilde{K}]\{U\} = \{\tilde{P}\} \tag{4}$$

where, $[\tilde{K}]$ is the interval stiffness matrix, $\{U\}$ is the vector of unknown nodal displacements, and $\{P\}$ is the vector of interval nodal forces. In development of interval stiffness matrix, the physical and mathematical characteristics of the stiffness matrix must be preserves.

This system of interval equations is mainly solved using computationally iterative procedures (Muhanna et al 2007) and (Neumaier and Pownuk 2007). The present method proposes a computationally efficient procedure with nearly sharp results using interval eigenvalue decomposition of stiffness matrix.

While the external force can also have uncertainties, in this work only problems with interval stiffness properties are addressed. However, for functional independent variations for both stiffness matrix and external force vector, the extension of the proposed work is straightforward.

3.1. DETERMINISTIC EIGENVALUE DECOMPOSITION

The deterministic symmetric stiffness matrix can be decomposed using matrix eigenvalue decomposition as:

$$[K] = [\Phi][\Lambda][\Phi]^T \tag{5}$$

where, $[\Phi]$ is the matrix of eigenvectors, and $[\Lambda]$ is the diagonal matrix of eigenvalues. Equivalently,

$$[K] = \sum_{i=1}^{N} \lambda_i \{\varphi_i\}\{\varphi_i\}^T \tag{6}$$

where, the values of λ_i is the eigenvalues and the vectors $\{\varphi_i\}$ are their corresponding eigenvectors. Therefore, the eigenvalue decomposition of the inverse of the stiffness matrix is:

$$[K]^{-1} = [\Phi][\Lambda]^{-1}[\Phi]^T \tag{7}$$

equivalently,

$$[K]^{-1} = \sum_{i=1}^{N} \frac{1}{\lambda_i} \{\varphi_i\}\{\varphi_i\}^T \tag{8}$$

Substituting Eq.(8) in the solution for the deterministic linear system of equation, Eq.(2), the solution for response is shown as:

$$\{U\} = (\sum_{i=1}^{N} \frac{1}{\lambda_i} \{\varphi_i\}\{\varphi_i\}^T)\{P\} \tag{9}$$

3.2. INTERVAL EIGENVALUE DECOMPOSITION

Similarly, the solution to interval system of equilibrium equations, Eq.(4), is:

$$\{\tilde{U}\} = (\sum_{i=1}^{N} \frac{1}{\tilde{\lambda}_i} \{\tilde{\varphi}_i\}\{\tilde{\varphi}_i\}^T)\{P\} \tag{10}$$

where, the values of $\tilde{\lambda}_i$ is the interval eigenvalues and, the vectors $\{\tilde{\varphi}_i\}$ are their corresponding interval eigenvectors that are to be determined.

4. Interval Eigenvalue Problem

4.1. BACKGROUND

The research in interval eigenvalue problem began to emerge as its applicability in science and engineering was realized. Hollot and Bartlett (1987) studied the spectra of eigenvalues of an interval matrix family which are found to depend on the spectrum of its extreme sets. Dief (1991) presented a method for computing interval eigenvalues of an interval matrix based on an assumption of invariance properties of eigenvectors.

In structural dynamics, Modares and Mullen (2004) have introduced a method for the solution of the interval eigenvalue problem which determines the exact bounds of the natural frequencies of a system using Interval Finite Element formulation.

4.2. DEFINITION

The eigenvalue problems for matrices containing interval values are known as the interval eigenvalue problems. If $[\tilde{A}]$ is an interval real matrix $(\tilde{A} \in \Re^{n \times n})$ and $[A]$ is a member of the interval matrix $([A] \in [\tilde{A}])$, the interval eigenvalue problem is shown as:

$$([A] - \lambda[I])\{x\} = 0, ([A] \in [\tilde{A}]) \tag{11}$$

4.2.1. *Solution for Eigenvalues*

The solution of interest to the real interval eigenvalue problem for bounds on each eigenvalue is defined as an inclusive set of real values $(\tilde{\lambda})$ such that for any member of the interval matrix, the eigenvalue solution to the problem is a member of the solution set. Therefore, the solution to the interval eigenvalue problem for each eigenvalue can be mathematically expressed as:

$$\{\lambda \in \tilde{\lambda} = [\lambda^l, \lambda^u] \mid \forall [A] \in [\tilde{A}] : ([A] - \lambda[I])\{x\} = 0\} \tag{12}$$

4.2.2. *Solution for Eigenvectors:*

The solution of interest to the real interval eigenvalue problem for bounds on each eigenvector is defined as an inclusive set of real values of vector $\{\tilde{x}\}$ such that for any member of the interval matrix, the eigenvector solution to the problem is a member of the solution set. Thus, the solution to the interval eigenvalue problem for each eigenvector is:

$$\{\{x\} \in \{\tilde{x}\} \mid \forall [A] \in [\tilde{A}], \lambda : ([A] - \lambda[I])\{x\} = 0\} \tag{13}$$

4.3. INTERVAL STIFFNESS MATRIX

The system's global stiffness can be viewed as a summation of the element contributions to the global stiffness matrix:

$$[K] = \sum_{i=1}^{n} [L_i][K_i][L_i]^T \tag{14}$$

where $[L_i]$ is the element Boolean connectivity matrix and $[K_i]$ is the element stiffness matrix in the global coordinate system. Considering the presence of uncertainty in the stiffness properties, the non-deterministic element elastic stiffness matrix is expressed as:

$$[\tilde{K}_i] = ([l_i, u_i])[K_i] \tag{15}$$

in which, $[l_i, u_i]$ is an interval number that pre-multiplies the deterministic element stiffness matrix. This procedure preserves the physical and mathematical characteristics of the stiffness matrix.

Therefore, the system's global stiffness matrix in the presence of any uncertainty is the linear summation of the contributions of non-deterministic interval element stiffness matrices:

$$[\tilde{K}] = \sum_{i=1}^{n} ([l_i, u_i])[L_i][K_i][L_i]^T = \sum_{i=1}^{n} ([l_i, u_i])[\overline{K}_i] \tag{16}$$

in which, $[\overline{K}_i]$ is the deterministic element elastic stiffness contribution to the global stiffness matrix.

4.4. INTERVAL EIGENVALUE PROBLEM FOR STATICS

The interval eigenvalue problem for a structure with stiffness properties expressed as interval values is:

$$[\tilde{K}]\{\tilde{\varphi}\} = (\tilde{\lambda})\{\tilde{\varphi}\} \tag{17}$$

Substituting Eq.(16) in Eq.(17):

$$(\sum_{i=1}^{n} ([l_i, u_i])[\overline{K}_i])\{\tilde{\varphi}\} = (\tilde{\lambda})\{\tilde{\varphi}\} \tag{18}$$

This interval eigenvalue problem can be transformed to a pseudo-deterministic eigenvalue problem subjected to a matrix perturbation. Introducing the central and radial (perturbation) stiffness matrices as:

$$[K_C] = \sum_{i=1}^{n} (\frac{l_i + u_i}{2})[\overline{K}_i] \tag{19}$$

$$[\tilde{K}_R] = \sum_{i=1}^{n} (\varepsilon_i)(\frac{u_i - l_i}{2})[\overline{K}_i] \quad , \quad \varepsilon_i = [-1,1] \tag{20}$$

Using Eqs. (19,20), the non-deterministic interval eigenpair problem, Eq.(18), becomes:

$$([K_C] + [\tilde{K}_R])\{\tilde{\varphi}\} = (\tilde{\lambda})\{\tilde{\varphi}\} \tag{21}$$

Hence, the determination of bounds on eigenvalues and bounds on eigenvectors of a stiffness matrix in the presence of uncertainty is mathematically interpreted as an eigenvalue problem on a central stiffness matrix ($[K_C]$) that is subjected to a radial perturbation stiffness matrix ($[\tilde{K}_R]$). This perturbation is in fact, a linear summation of non-negative definite deterministic element stiffness contribution matrices that are scaled with bounded real numbers (ε_i).

5. Solution

5.1. BOUNDS ON EIGENVALUES

The following concepts must be considered in order to bound the non-deterministic interval eigenvalue problem, Eq.(21). The classical linear eigenpair problem for a symmetric matrix is:

$$[A]\{x\} = \lambda\{x\} \tag{22}$$

with the solution of real eigenvalues ($\lambda_1 \le \lambda_2 \le ... \le \lambda_n$) and corresponding eigenvectors ($x_1, x_2, ..., x_n$). This equation can be transformed into a ratio of quadratics known as the *Rayleigh quotient*:

$$R(x) = \frac{\{x\}^T [A]\{x\}}{\{x\}^T \{x\}} \tag{23}$$

The Rayleigh quotient for a symmetric matrix is bounded between the smallest and the largest eigenvalues (Bellman 1960 and Strang 1976).

$$\lambda_1 \le R(x) = \frac{\{x\}^T [A]\{x\}}{\{x\}^T \{x\}} \le \lambda_n \tag{24}$$

Thus, the first eigenvalue (λ_1) can be obtained by performing an *unconstrained minimization* on the scalar-valued function of Rayleigh quotient:

$$\min_{x \in R^n} R(x) = \min_{x \in R^n} \left(\frac{\{x\}^T [A]\{x\}}{\{x\}^T \{x\}}\right) = \lambda_1 \tag{25}$$

For finding the next eigenvalues, the concept of maximin characterization can be used. This concept obtains the kth eigenvalue by imposing (*k-1*) constraints on the minimization of the Rayleigh quotient (Bellman 1960 and Strang 1976):

$$\lambda_k = \max[\min\ R(x)]$$
$$\text{(subject to constrains } (x^T z_i = 0), i = 1, ...k-1, k \ge 2) \tag{26}$$

5.1.1. *Bounding the Eigenvalues for Statics*

Using the concepts of minimum and maximin characterizations of eigenvalues for symmetric matrices, the solution to the interval eigenvalue problem for the eigenvalues of a system with uncertainty in the stiffness characteristics (Eq.(21)) for the first eigenvalue can be shown as:

$$\tilde{\lambda}_1 = \min_{x \in R^n}\left(\frac{\{x\}^T[\tilde{K}]\{x\}}{\{x\}^T\{x\}}\right) = \min_{x \in R^n}\left(\frac{\{x\}^T\left(\sum_{i=1}^{n}([l_i, u_i])[\overline{K}_i]\right)\{x\}}{\{x\}^T\{x\}}\right) \tag{27}$$

for the next eigenvalues:

$$\tilde{\lambda}_k = \max\left[\min_{x.z_i=0, i=1,\ldots,k-1} \frac{\{x\}^T[\tilde{K}]\{x\}}{\{x\}^T\{x\}}\right] = \max\left[\min_{x.z_i=0, i=1,\ldots,k-1}\left(\frac{\{x\}^T([K_C]+[\tilde{K}_R])\{x\}}{\{x\}^T[M]\{x\}}\right)\right] \tag{28}$$

5.1.2. *Deterministic Eigenvalue Problems for Bounding Eigenvalues in Statics*

Substituting and expanding the right-hand side terms of Eqs. (27,28):

$$\left(\frac{\{x\}^T[K_C]\{x\}}{\{x\}^T\{x\}} + \frac{\{x\}^T[\tilde{K}_R]\{x\}}{\{x\}^T\{x\}}\right) = \sum_{i=1}^{n}\left(\frac{l_i + u_i}{2}\right)\left(\frac{\{x\}^T[\overline{K}_i]\{x\}}{\{x\}^T\{x\}}\right) + \sum_{i=1}^{n}(\tilde{\varepsilon}_i)\left(\frac{u_i - l_i}{2}\right)\left(\frac{\{x\}^T[\overline{K}_i]\{x\}}{\{x\}^T\{x\}}\right) \tag{29}$$

Since the matrix $[\overline{K}_i]$ is non-negative definite, the term $\left(\frac{\{x\}^T[\overline{K}_i]\{x\}}{\{x\}^T\{x\}}\right)$ is non-negative.

Therefore, using the monotonic behavior of eigenvalues for symmetric matrices, the upper bounds on the eigenvalues in Eqs.(19,20) are obtained by considering maximum values of interval coefficients of uncertainty $(\tilde{\varepsilon}_i = [-1,1])$, $((\varepsilon_i)_{\max} = 1)$, for all elements in the radial perturbation matrix.

Similarly, the lower bounds on the eigenvalues are obtained by considering minimum values of those coefficients, $((\varepsilon_i)_{\min} = -1)$, for all elements in the radial perturbation matrix. Also, it can be observed that any other element stiffness selected from the interval set will yield eigenvalues between the upper

and lower bounds. This imonotonic behavior of eigenvalues can also be used for parameterization purposes.

Using these concepts, the deterministic eigenvalue problems corresponding to the maximum and minimum eigenvalues are obtained (Modares and Mullen 2004) as:

$$(\sum_{i=1}^{n}(u_i)[\overline{K}_i])\{\varphi\} = (\lambda_{\max})\{\varphi\} \tag{30}$$

$$(\sum_{i=1}^{n}(l_i)[\overline{K}_i])\{\varphi\} = (\lambda_{\min})\{\varphi\} \tag{31}$$

5.2. BOUNDS ON EIGENVECTORS

5.2.1. *Invariant Subspace*

The subspace χ is defined to be an *invariant subspace* of matrix $[A]$ if:

$$A\chi \subset \chi \tag{32}$$

Equivalently, if χ is an invariant subspace of $[A]_{n\times n}$ and also, columns of $[X_1]_{n\times m}$ form a basis for χ, then there is a unique matrix $[L_1]_{m\times m}$ such that:

$$[A][X_1] = [X_1][L_1] \tag{33}$$

The matrix $[L_1]$ is the representation of $[A]$ on χ with respect to the basis $[X_1]$ and the eigenvalues of $[L_1]$ are a subset of eigenvalues of $[A]$. Therefore, for the invariant subspace, $(\{v\},\lambda)$ is an eigenpair of $[L_1]$ if and only if $(\{[X_1]\{v\}\},\lambda)$ is an eigenpair of $[A]$.

5.2.2. *Theorem of Invariant Subspaces*

For a real symmetric matrix $[A]$, considering the subspace χ with the linearly independent columns of $[X_1]$ forming a basis for χ and the linearly independent columns of $[X_2]$ spanning the complementary subspace $\chi^{\perp}$, then, χ is an invariant subspace of $[A]$ iff:

$$[X_2]^T[A][X_1]=[0] \tag{34}$$

Therefore, invoking this condition and postulating the definition of invariant subspaces, the symmetric matrix $[A]$ can be reduced to a diagonalized form using a unitary similarity transformation as:

$$[X_1X_2]^T[A][X_1X_2]=\begin{bmatrix}[X_1]^T[A][X_1] & [X_1]^T[A][X_2]\\ [X_2]^T[A][X_1] & [X_2]^T[A][X_2]\end{bmatrix}=\begin{bmatrix}[L_1] & [0]\\ [0] & [L_2]\end{bmatrix} \tag{35}$$

where $[L_i]=[X_i]^T[A][X_i],\ i=1,2.$

5.2.3. *Simple Invariant Subspace*

An invariant subspace is *simple* if the eigenvalues of its representation $[L_1]$ are distinct from other eigenvalues of $[A]$. Thus, using the reduced form of $[A]$ with respect to the unitary matrix $[[X_1][X_2]]$, χ is a *simple* invariant subspace if the eigenvalues of $[L_1]$ and $[L_2]$ are distinct:

$$\lambda([L_1])\bigcap\lambda([L_2])=\varnothing \tag{36}$$

5.2.4. *Perturbed Eigenvector*

Considering the column spaces of $[X_1]$ and $[X_2]$ to span two complementary simple invariant subspaces, the perturbed orthogonal subspaces are defined as:

$$[\hat{X}_1]=[X_1]+[X_2][P] \tag{37}$$

$$[\hat{X}_2]=[X_2]-[X_1][P]^T \tag{38}$$

in which $[P]$ is a matrix to be determined.

Thus, each perturbed subspace is defined as a summation of the exact subspace and the contribution of the complementary subspace. Considering a symmetric perturbation $[E]$, the perturbed matrix is defined as:

$$[\hat{A}]=[A]+[E] \tag{39}$$

Applying the theorem of invariant subspaces for perturbed matrix and perturbed subspaces, and linearizing due to a small perturbation compared to the unperturbed matrix, Eq.(34) is rewritten as:

$$[P][L_1]-[L_2][P]=[X_2]^T[E][X_1] \tag{40}$$

This perturbation problem is an equation for unknown $[P]$ in the form of a Sylvester's equation in which, the uniqueness of the solution is guaranteed by the existence of simple perturbed invariant subspaces.

Finally, specializing the result for one eigenvector and solving the above equation, the perturbed eigenvector is (Stewart and Sun 1990):

$$\{\hat{x}_1\}=\{x_1\}+[X_2](\lambda_1[I]-[L_2])^{-1}[X_2]^T[E]\{x_1\} \tag{41}$$

5.2.5 *Bounding Eigenvectors for Statics*

For the perturbed eigenvalue problem for statics, Eq.(21), the error matrix is:

$$[E]=[\tilde{K}_R]=(\sum_{i=1}^{n}(\varepsilon_i)(\frac{u_i-l_i}{2})[\overline{K}_i]) \tag{42}$$

Using the error matrix in eigenvector perturbation equation for the first eigenvector, Eq.(33) the perturbed eigenvector is:

$$\{\tilde{\varphi}_1\}=\{\varphi_1\}+([\Phi_2](\lambda_1[I]-[\Lambda_2])^{-1}[\Phi_2]^T(\sum_{i=1}^{n}(\varepsilon_i)(\frac{u_i-l_i}{2})[\overline{K}_i]))\{\varphi_1\} \tag{43}$$

in which, $\{\varphi_1\}$ is the first eigenvector, (λ_1) is the first eigenvalue, $[\Phi_2]$ is the matrix of remaining eigenvectors and $[\Lambda_2]$ is the diagonal matrix of remaining eigenvalues obtained from the deterministic eigenvalue problem. Eq.(30,31 and 43) is used to calculate the bounds on interval eigenvalues and interval eigenvectors in the response equation, Eq.(9).
In order to attain sharper results, the functional dependency of intervals in direct interval multiplications in Eq.(9) is considered. Also, input intervals are subdivided and the union of responses of subset results is obtained.

6. Numerical Example Problem

The bounds on the static response for a 2-D statically indeterminate truss with interval uncertainty present in the modulus of elasticity of each element are determined (Figure 2). The cross-sectional area A, the

length for horizontal and vertical members L, the Young's moduli E for all elements are $\tilde{E} = ([0.99,1.01])E$.

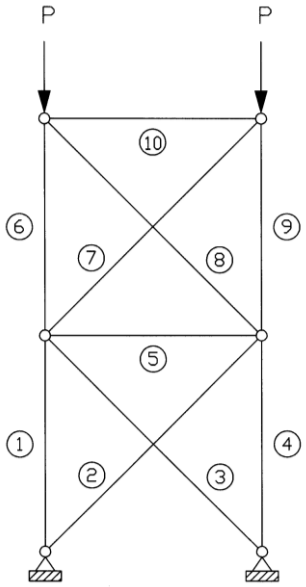


Figure 2. The structure of 2-D truss

The problem is solved using the method presented in this work. The functional dependency of intervals in the response equation is considered. A hundred-segment subdivision of input intervals is performed and the union of responses is obtained. For comparison, an exact combinatorial analysis has performed which considers lower and upper values of uncertainty for each element i.e. solving ($2^n = 2^{10} = 1024$) deterministic problems.

The static analysis results obtained by the present method and the brute force combination solution for the vertical displacement of the top nodes in are summarized Table (1).

	Lower Bound *Present Method*	Lower Bound *Combination Method*	Upper Bound *Combination Method*	Upper Bound *Present Method*	Error *%*
$\frac{U}{\left(\frac{PL}{AE}\right)}$	-1.6265	-1.6244	-1.5859	-1.5838	% 0.12

Table1. Bounds on Vertical Displacement of Top Nodes

The results show that the proposed robust method yields nearly sharp results in a computationally efficient manner as well as preserving the system's physics.

4. Conclusions

A finite-element based method for static analysis of structural systems with interval uncertainty in mechanical properties is presented.

This method proposes an interval eigenvalue decomposition of stiffness matrix. By obtaining the exact bounds on the eigenvalues and nearly sharp bounds on the eigenvectors, the proposed method is capable to obtain the nearly sharp bounds on the structure's static response.

Some conservative overestimation in response occurs that can be attributed to the linearization in formation of bounds of eigenvectors and also, the functional dependency of intervals in the dynamic response formulation.

This method is computationally feasible and it shows that the bounds on the static response can be obtained without combinatorial or Monte-Carlo simulation procedures.

This computational efficiency of the proposed method makes it attractive to introduce uncertainty into structural static analysis and design. While this methodology is shown for structural systems, its extension to various mechanics problems is straightforward.

References

Bellman, R. Introduction to Matrix Analysis, McGraw-Hill, New York 1960.

Dief, A., Advanced Matrix theory for Scientists and Engineers, pp.262-281. Abacus Press 1991.

Hollot, C. and A. Bartlett. On the eigenvalues of interval matrices, *Technical Report*, Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 1987.

Modares, M. and R. L. Mullen. Free Vibration of Structures with Interval Uncertainty. *9th ASCE Specialty Conference on Probabilistic Mechanics and Structural Reliability* 2004.

Moore, R. E. Interval Analysis. Prentice Hall, Englewood, NJ 1966.

Muhanna, R. L. and R. L. Mullen. Uncertainty in Mechanics Problems-Interval-Based Approach. *Journal of Engineering Mechanics* June-2001, pp.557-566 2001.

Muhanna, R. L., Zhang H. and R. L. Mullen. Interval Finite Element as a Basis for Generalized Models of Uncertainty in Engineering Mechanics, *Reliable Computing*, Vol. 13, pp. 173-194, 2007.

Neumaier, A. Interval Methods for Systems of Equations. Cambridge University Press, Cambridge 1990.

Neumaier, A. and A. Pownuk. Linear Systems with Large Uncertainties, with Applications to Truss Structures, *Reliable Computing*, Vol. 13, pp. 149-172, 2007.

Strang, G. Linear Algebra and its Applications, Massachusetts Institute of Technology, 1976.

Stewart, G.W. and J. Sun. Matrix perturbation theory, Chapter 5. Academic Press, Boston, MA 1990.

General Interval FEM Program Based on Sensitivity Analysis

Andrzej Pownuk
The University of Texas at El Paso
Department of Mathematical Sciences
email:ampownuk@utep.edu

Abstract. Today there are many methods for solution of equation with interval parameters (Moens and Vandepitte, 2005). Unfortunately there are very few efficient methods which can be directly applied for solution of complex engineering problems. Sensitivity analysis method (Pownuk, 2004) gives very good inner approximation of the exact solution set. This method was implemented in C++ language by the author and the program can be recompiled on Windows, Linux and Solaris operating systems. The program is able to solve 1D, 2D and 3D linear problems of electrostatics with interval parameters.

Additionally it is possible to solve problems with uncertain functional parameters (Pownuk, 2006). In order to do that it is necessary to create special finite elements. It is possible to consider also uncertain shapes The program is very universal and can be applied to the solution of complex engineering problem. The program is a part web application, which is written in php language and can be run on the web page http://andrzej.pownuk.com.

Keywords: interval stresses, stress distribution, sensitivity analysis, functional parameters

1. Design of structures with the interval parameters

One of the simplest method of modelling of uncertain parameters is based on the intervals (Moore, 1966). In that case in order to describe values of the parameter p it is necessary to know only two numbers i.e. upper $\overline{p}$ and lower bound $\underline{p}$.
In civil and mechanical engineering one of the most popular method of mathematical modeling of engineering structures is the finite element method (Zienkiewicz and Taylor, 2000). The FEM method leads to the following system of parameter dependent system of linear or nonlinear equations

$$K(p)u(p) = Q(p) \tag{1}$$

where K is the stiffness matrix, Q is the right hand side and p is the vector of uncertain parameters

$$p = [p_1, ..., p_m]^T. \tag{2}$$

In this paper the following notation for the interval parameters and the interval functions will be applied. If we have the function $f(p)$ then

$$f(\mathbf{p}) = \{f(p) : p \in \mathbf{p}\} \tag{3}$$

$$\mathbf{f}(\mathbf{p}) = \square f(\mathbf{p}) = \square\{f(p) : p \in \mathbf{p}\} \tag{4}$$

where $\mathbf{p}$ is the interval or a vector of the interval parameters. Function f can be real valued or vector valued. $\mathbf{p}$ can be an interval in R (i.e. $\mathbf{p} = [\underline{p}_i, \overline{p}_i] \subset R$) or in R^m (i.e. $\mathbf{p} = [\underline{p}_1, \overline{p}_1] \times ... \times [\underline{p}_m, \overline{p}_m]$). If the parameters p_i belong to some know intervals $p_i \in [\underline{p}_i, \overline{p}_i]$, then the solution can be defined as the smallest interval which contain the exact solution set.

$$u(\mathbf{p}) = \{u : K(p)u(p) = Q(p), p \in \mathbf{p}\} \tag{5}$$

$$\mathbf{u}(\mathbf{p}) = \square u(\mathbf{p}) = \square\{u : K(p)u(p) = Q(p), p \in \mathbf{p}\} \tag{6}$$

2. Sensitivity analysis method

There are different methods of calculation of the set (6) (Moens and Vandepitte, 2005; Neumaier, 1990). One of the simplest and most efficient method of solution of system of equations with the interval parameters is the sensitivity analysis method (Pownuk, 2004).

Sensitivity analysis method for general explicit function $u_i = u_i(p)$.

1. Calculate the mid point solution $u(p_0)$ from the following system of equations

$$u_0 = u(p_0) \tag{7}$$

   where p_0=mid($\mathbf{p}$).

2. Calculate the sensitivity $\frac{\partial u(p_0)}{\partial p_i}$ at the mid point p_0.

3. Find the combination of parameters which corresponds to the extreme values of the solution.

$$If\ \frac{\partial u_i(p_0)}{\partial p_j} \geqslant 0\ then\ p_{i,j}^{max} = \overline{p}_j, p_{i,j}^{min} = \underline{p}_j, \tag{8}$$

$$if\ \frac{\partial u_i(p_0)}{\partial p_j} < 0\ then\ p_{i,j}^{max} = \underline{p}_j, p_{i,j}^{min} = \overline{p}_j. \tag{9}$$

   Combination of endpoints which correspond to the extreme value of function $u_i = u_i(p)$ will be denoted in the following way

$$p_i^{min} = (p_{i,1}^{min}, p_{i,2}^{min}, ..., p_{i,m}^{min}), \tag{10}$$

$$p_i^{max} = (p_{i,1}^{max}, p_{i,2}^{max}, ..., p_{i,m}^{max}). \tag{11}$$

4. Create a list L of all critical endpoints combinations.

$$L = \{p_1^{min}, p_1^{max}, p_2^{min}, p_2^{max}, ..., p_m^{min}, p_m^{max}\} = \tag{12}$$

$$= \{p_1, p_2, ..., p_{2m}\} \tag{13}$$

5. Now it is possible to create a new list L^*, which contain only different endpoints

$$L^* = \{p_1^*, p_1^*, ..., p_{n_*}^*\}. \tag{14}$$

6. For all elements in the list L^* calculate a value of the vector u

$$u_{i,j}^* = u_i(p_j^*), \quad for\ j = 1, ..., n^*. \tag{15}$$

7. Calculate the extreme values of the solution

$$\underline{u}_i = \min\{u_i(p_0), u_{i,1}^*, u_{i,2}^*, ..., u_{i,n^*}^*\}, \quad \overline{u}_i = \max\{u_i(p_0), u_{i,1}^*, u_{i,2}^*, ..., u_{i,n^*}^*\}. \tag{16}$$

The results are exact if the sign of the derivative is constant.

3. Interval functional parameters

3.1. Equations with interval functional parameters

In order to get reliable results it is possible to approximate the values of the unknown function p by using some upper and lower bounds

$$p(x) \in [\underline{p}, \overline{p}] = \mathbf{p} \tag{17}$$

Better approximation can be obtained using functional intervals

$$p(x) \in [\underline{p}(x), \overline{p}(x)] = \mathbf{p}(x) \tag{18}$$

Lets assume that the behaviour of the structure with interval parameters is described by the following equation

$$F(x, u, p) = 0 \tag{19}$$

where u is a vector of the solutions and p is a vector of parameters. The solution of the equation (19) can be defined in the following way (Neumaier, 1990)

$$u(x, \mathbf{p}) = \{u : F(x, u, p) = 0, p(x) \in \mathbf{p}(x)\}, \quad x \in \Omega. \tag{20}$$

The set $u(x, \mathbf{p})$ is in general very complicated (Neumaier, 1990), because of that in applications it is easier to use the smallest interval which contain the exact solution set.

$$\mathbf{u}(x, \mathbf{p}) = \Box u(x, \mathbf{p}) = \Box\{u : F(x, u, p) = 0, p(x) \in \mathbf{p}(x)\}, \quad x \in \Omega. \tag{21}$$

If the equation is not directly dependent on x then the solution set is the following

$$u(\mathbf{p}) = \{u : F(u, p) = 0, p(x) \in \mathbf{p}(x)\}, \tag{22}$$

$$\mathbf{u}(\mathbf{p}) = \Box u(\mathbf{p}) = \Box\{u : F(u, p) = 0, p(x) \in \mathbf{p}(x)\}. \tag{23}$$

3.2. General concept of monotonicity

A map $T : X \to Y$ is monotone if $(X, \geqslant)$ is a partially ordered set and $x, y \in X$, $x \geqslant y \Rightarrow T(x) \geqslant T(y)$. Typically, X will be a subset of a Banach space Y with a cone Y_+ of positive elements and $x \leqslant y$ is equivalent to $y - x \in Y_+$ (Hirsch and Smith, 2005).

3.3. Solution of the equations with the interval functional parameters - general case

In general it is very hard to get the solution set (23) or (21). Fortunately in many applications it is possible to apply the method which is based on sensitivity analysis, Taylor expansion and/or functional derivative (Pownuk, 2006). These methods allow us to get very actuate solution and have low computational complexity.
Let us consider a function $u = u(p)$ where $p : R^n \supset \Omega \to p(x) \in R$, X is a functional space which contain the functions p, u is the function form the space X to the space R i.e. $u : X \ni p \to u(p) \in R$. Lets consider only positive variation of the function p i.e.

$$\delta p(x) = p_1(x) - p_0(x) > 0 \tag{24}$$

where $p_1, p_2 \in X$. If one add positive variation to the function p_0 then the results (i.e. $p_0 + \delta p$) is bigger than the function p_0 i.e.

$$p_0 + \delta p(x) > p_0(x) \tag{25}$$

If the difference $u(p + \delta p) - p(p_0)$ has constant sign the the function u is monotone.
If the function u is differentiable then finite increment of the functions u can be approximated by the differential

$$u(p_0 + \delta p) - u(p_0) = \delta u(p_0, \delta p) + R(p_0, \delta p) \tag{26}$$

where

$$\lim_{\|\delta p\| \to 0} \frac{|R(p_0, \delta p)|}{\|\delta p\|} = 0, \tag{27}$$

and for small variations δp we can write

$$u(p_0 + \delta p) - u(p_0) \approx \delta u(p_0, \delta p) \tag{28}$$

If the differential $\delta u(p_0, \delta p)$ is positive then the function $u = u(p)$ is monotone around the point p_0 (Hirsch and Smith, 2005).

Theorem 1
If the function $u : X \to R$ is differentiable and $\delta u(p_0, \delta p) \geqslant 0$ for all $p \in [\underline{p}, \overline{p}] \subset X$ and some δp, then $u = u(p)$ is monotone in the interval $[\underline{p}, \overline{p}]$.

Proof

$$u(p_0 + \delta p) - u(p_0) = \int_0^1 \delta u(p_0 + t\delta p, \delta p) dt \tag{29}$$

if $\delta u(p_0 + t\delta p, \delta p) \geqslant 0$ then $\int\limits_0^1 \delta u(p_0 + t\delta p, \delta p)dt \geqslant 0$ and then

$$u(p_0 + \delta p) \geqslant u(p_0) \tag{30}$$

i.e. the function u is monotone. Now it is possible to calculate extreme values of the function $u = u(p)$ for $p \in \mathbf{p}$ if the sign of the differential is constant.

General sensitivity analysis with functional parameters

1. **if** $\delta u(p, \delta p) \geqslant 0$ **then** $p^{min} = \underline{p}$, $p^{max} = \overline{p}$.
2. **if** $\delta u(p, \delta p) < 0$ **then** $p^{min} = \overline{p}$, $p^{max} = \underline{p}$.
3. $\underline{u} = u(p^{min})$, $\overline{u} = u(p^{max})$.

The algorithm is not very practical because in general it is hard to verify the sign of the differential $\delta u(p_0, \delta p)$. In order to make that method a little more practical it is necessary to consider some special cases.

3.4. Extreme values of the integral in the form $u(p) = \int\limits_\Omega L(x, p(x))dx$

Differential of the function $u(p) = \int\limits_\Omega L(x, p(x))dx$ has the following form

$$\delta u(p_0, \delta p) = \int\limits_\Omega \frac{\partial L(x, p(x))}{\partial p(x)} \delta p(x)dx = \left\langle \frac{\delta u}{\delta p}, \delta p \right\rangle \tag{31}$$

where

$$\frac{\delta u}{\delta p(x)} = \frac{\partial L(x, p(x))}{\partial p(x)} \tag{32}$$

is the functional derivative of the function $u = u(p)$ and $\langle ., . \rangle$ is the scalar product.

Theorem 2
If $\frac{\delta u}{\delta p(x)} \geqslant 0$ for $p \in [\underline{p}, \overline{p}] \subset X$, then the function $u = u(p)$ is monotone in the interval $\mathbf{p}$.
Proof
If $\frac{\delta u}{\delta p(x)} \geqslant 0$ and $\delta p(x) \geqslant 0$ then $\delta u(p_0, \delta p) = \left\langle \frac{\delta u}{\delta p}, \delta p \right\rangle \geqslant 0$ and according to the theorem 1 the function $u = u(p)$ is monotone.

Now it is possible to use more efficient version of the algorithm

Sensitivity analysis based on functional derivative

1. **if** $\frac{\delta u}{\delta p(x)} \geqslant 0$ **then** $p^{min} = \underline{p}$, $p^{max} = \overline{p}$.
   if $\frac{\delta u}{\delta p(x)} < 0$ **then** $p^{min} = \overline{p}$, $p^{max} = \underline{p}$.
   $\underline{u} = u(p^{min})$, $\overline{u} = u(p^{max})$.

If the sign of the functional derivative is not constant, then it is possible to apply approximate method for finding extreme values of the solutions. According to the equation (28) the finite increment of the functions can be approximated by the differential. If the differential is positive (i.e. $\delta u(p_0, \delta p) \geqslant 0$) then for very small variations δp we can assume that $u(p + \delta p) \geqslant u(p)$. The product $\frac{\delta u}{\delta p(x)}\delta p(x)$ is nonnegative if $\frac{\delta u}{\delta p(x)} \geqslant 0$ and $\delta p(x) \geqslant 0$ or $\frac{\delta u}{\delta p(x)} \leqslant 0$ and $\delta p(x) \leqslant 0$. If we have the function $p_0 \in [\underline{p}, \overline{p}]$ and the value of functional derivative $\frac{\delta u(p_0)}{\delta p(x)}$ is not constant, then it is possible to change the sign of the variation δp is such a way which make the differential positive. It is possible to define the small variations in the following way

$$\delta p^u(x) = \lambda(x)\frac{\delta u(p_0)}{\delta p(x)}, \quad \delta p^l(x) = -\lambda(x)\frac{\delta u(p_0)}{\delta p(x)} \tag{33}$$

where $\lambda(x)$ is an arbitrary positive function. If the variations $\delta p^l, \delta p^u$ are small enough then $\delta u(p_0, \delta p^u) \geqslant 0$, $\delta u(p_0, \delta p^l) \leqslant 0$ and according to the relation (28) we can write

$$u(p_0 + \delta p^u) \geqslant u(p_0) \tag{34}$$

$$u(p_0 + \delta p^l) \leqslant u(p_0) \tag{35}$$

Above described properties can be applied to the creation of approximate algorithm for finding upper and lower bound of the function $u = u(p)$.

Calculation of upper bound $\overline{u}$

1. $p(x) = p_0(x)$
2. choose the function $\lambda(x)$
3. $\delta p^u(x) = \lambda(x)\frac{\delta u(p)}{\delta p(x)}$
4. $p_{old}(x) = p(x)$
5. $p(x) := p(x) + \delta p^u(x)$
6. **if** $p(x) > \overline{p}(x)$ **then** $p(x) = \overline{p}(x)$
7. **if** $p(x) < \underline{p}(x)$ **then** $p(x = \underline{p}(x)$
8. **if** $\|p_{old} - p\| > \varepsilon$ **then** goto step 2
9. $\overline{u} = u(p)$
10. **stop**

The lower bound can be calculated in the similar way.

3.5. Extreme values of the functions and the integrals

In more complicated cases the function $u = u(p)$ is a superposition of algebraic function f and the integrals in the form $\int\limits_{\Omega} L_i(x, p(x))dx$

$$\begin{gathered} u(p) = f\left(y_1, ..., y_q\right) = f(y) \\ y_1 = I_1(p) = \int\limits_{\Omega} L_1(x, p(x))dx, ..., y_q = I_q(p) = \int\limits_{\Omega} L_q(x, p(x))dx \end{gathered} \tag{36}$$

Differential in this case is equal to:

$$\delta u(p, \delta p) = \sum_i \frac{\partial f(y)}{\partial y_i} I_i(p, \delta p) = \sum_i \frac{\partial f(y)}{\partial y_i} \int\limits_{\Omega} \frac{\partial L_i}{\partial p(x)} \delta p(x) dx \tag{37}$$

Functional derivative can be defined in this case in the following way

$$\frac{\delta u}{\delta p(x)} = \sum_i \frac{\partial f(y)}{\partial y_i} \frac{\delta I_i(p)}{\delta p(x)} = \sum_i \frac{\partial f(y)}{\partial y_i} \frac{\partial L_i(x, p(x))}{\partial p(x)} \tag{38}$$

In matrix notation

$$\frac{\delta u(p)}{\delta p} = \left[\frac{\partial f(y)}{\partial y_1}, ..., \frac{\partial f(y)}{\partial y_p} \right] \begin{bmatrix} \frac{\partial L_1}{\partial p(x)} \\ ... \\ \frac{\partial L_p}{\partial p(x)} \end{bmatrix} \tag{39}$$

If the sign of the functional derivative is constant, then the sign of the differential is constant (for very small perturbations δp) and according to the theorem **??** the function $u = u(p)$ is monotone. In order to calculate the extreme values of the solutions by using the algorithm 3.4. If the sign of the derivative is not constant then it is possible to apply algorithm **??** and **??**.
It is also interesting to study the function u in the case when it depend on many functions p_i i.e.

$$\begin{gathered} u(p) = f\left(y_1, ..., y_q\right) = f(y) \\ y_1 = I_1(p) = \int\limits_{\Omega} L_1(x, p(x))dx, ..., y_q = I_q(p) = \int\limits_{\Omega} L_q(x, p(x))dx \end{gathered} \tag{40}$$

where $p = (p_1, ..., p_m)$. The differential is equal to

$$\delta u(p, \delta p) = \sum_i \frac{\partial f(y)}{\partial y_i} I_i(p, \delta p) = \sum_i \frac{\partial f(y)}{\partial y_i} \int\limits_{\Omega} \left(\sum_j \frac{\partial L_i}{\partial p_j(x)} \delta p_j(x) \right) dx \tag{41}$$

Now it is possible to calculate the functional derivative which is in this case a vector with the following components

$$\frac{\delta u(p)}{\delta p} = \left[\sum_i \frac{\partial f(y)}{\partial y_i} \frac{\partial L_i}{\partial p_1(x)}, ..., \sum_i \frac{\partial f(y)}{\partial y_i} \frac{\partial L_i}{\partial p_m(x)} \right] \tag{42}$$

In matrix notation

$$\frac{\delta u(p)}{\delta p} = \left[\frac{\partial f(y)}{\partial y_1}, ..., \frac{\partial f(y)}{\partial y_p}\right] \left[\begin{array}{ccc} \frac{\partial L_1}{\partial p_1(x)} & \cdots & \frac{\partial L_1}{\partial p_m(x)} \\ \cdots & \cdots & \cdots \\ \frac{\partial L_p}{\partial p_1(x)} & \cdots & \frac{\partial L_p}{\partial p_m(x)} \end{array}\right] \tag{43}$$

The differential is positive if the variations δp_j have the same sign as $\sum\limits_i \frac{\partial f(y)}{\partial y_i}\frac{\partial L_i}{\partial p_j(x)}$. It is also possible to create discrete version of these methods.

4. Sensitivity with respect to changes of the region of integration

4.1. Introduction

Lets consider a function $u = u(\Omega)$ where Ω is a domain of integration.

$$u(\Omega) = \int\limits_{\Omega} L(x)dx \tag{44}$$

Lets consider the following increment

$$u(\Omega + \Delta\Omega) - u(\Omega) = \int\limits_{\Omega+\Delta\Omega} L(x)dx - \int\limits_{\Omega} L(x)dx = \int\limits_{\Delta\Omega} L(x)dx \tag{45}$$

The operation "$\Omega + \Delta\Omega$" is a sum of two set i.e. "$\Omega \cup \Delta\Omega$". If the set is convex then from main value theorem

$$\int\limits_{\Delta\Omega} L(x)dx = |\Delta\Omega| L(x^*) \tag{46}$$

where $x^* \in \Delta\Omega$.

$$\frac{u(\Omega + \Delta\Omega) - u(\Omega)}{|\Delta\Omega|} = L(x^*) \tag{47}$$

In the limit case

$$\frac{\delta u}{\delta\Omega(x)} = \lim_{|\Delta\Omega(x)|\to 0} \frac{u(\Omega + \Delta\Omega(x)) - u(\Omega)}{|\Delta\Omega(x)|} = L(x). \tag{48}$$

If $\underline{\Omega} \subset \Omega \subset \overline{\Omega}$ then extreme values of the function $u = u(p)$ by using sensitivity analysis method.

The inclusion $\subset$ can be treat as the partial order relation $\geqslant$. Because of that it is possible to take into account "set intervals"

$$[\underline{\Omega}, \overline{\Omega}] = \{\Omega : \Omega \subset \Omega \subset \overline{\Omega}\}. \tag{49}$$

If the sign of the derivative $\frac{\delta u}{\delta\Omega(x)}$ is not constant then it is possible to create the sets Ω^{max} and Ω^{min} in the following way.

$$\Omega^{max} = \underline{\Omega} \cup \left\{x : \frac{\delta u}{\delta\Omega(x)} \geqslant 0, x \in \overline{\Omega} - \underline{\Omega}\right\} \tag{50}$$

$$\Omega^{min} = \underline{\Omega} \cup \left\{ x : \frac{\delta u}{\delta \Omega(x)} < 0, x \in \overline{\Omega} - \underline{\Omega} \right\} \tag{51}$$

Extreme values of the function $u = u(\Omega)$ are equal $\underline{u} = u(\Omega^{min})$, $\overline{u} = u(\Omega^{max})$.
The function $u = u(\Omega)$ may be a superposition of algebraic function and the integral.

$$u(\Omega) = f(y), \quad y = \int_{\Omega} L(x)dx \tag{52}$$

$$\frac{\delta u}{\delta \Omega(x)} = \frac{df(p)}{dy} \frac{\delta}{\delta \Omega(x)} \int_{\Omega} L(x)dx = \frac{df(p)}{dy} L(x) \tag{53}$$

The function u can be dependent on many integrals.

$$u(\Omega) = f(y), \quad y_1 = \int_{\Omega} L_1(x)dx, ..., y_n = \int_{\Omega} L_p(x)dx \tag{54}$$

$$\frac{\delta u}{\delta \Omega(x)} = \sum_i \frac{df(p)}{dy_i} \frac{\delta}{\delta \Omega(x)} \int_{\Omega} L_i(x)dx = \sum_i \frac{df(p)}{dy_i} L_i(x) \tag{55}$$

4.2. Moment of inertia of cross-section with uncertain shape

Polar moment of inertia

$$I_0(\Omega) = \int_{\Omega} r^2 d\Omega = \iint_{\Omega} (x^2 + y^2)dxdy \tag{56}$$

Because the limit is positive in the set $\overline{\Omega} - \underline{\Omega}$

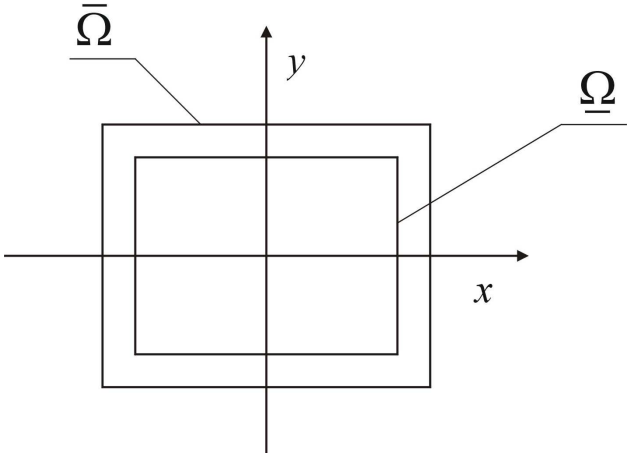


Figure 1. Uncertain shape of cross-section

$$\frac{\delta I_0}{\delta \Omega(x,y)} = x^2 + y^2 \geqslant 0 \tag{57}$$

then

$$\underline{I}_0 = I_0(\underline{\Omega}) = \iint\limits_{\underline{\Omega}} (x^2+y^2)dxdy, \quad \overline{I}_0 = I_0(\overline{\Omega}) = \iint\limits_{\overline{\Omega}} (x^2+y^2)dxdy \tag{58}$$

In the case of product moment of inertia

$$I_{xy}(\Omega) = \iint\limits_{\Omega} xydxdy \tag{59}$$

the limit

$$\frac{\delta I_{xy}}{\delta \Omega(x,y)} = xy \tag{60}$$

is sometimes positive and sometimes negative. From the picture 2 we can see that $xy \geqslant 0$ in the

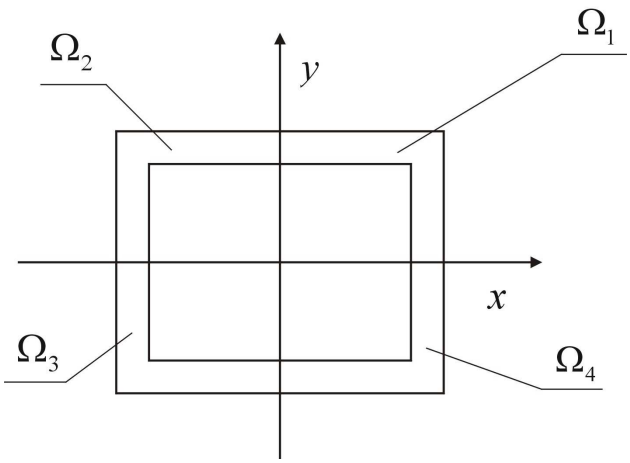


Figure 2. Uncertain shape

sets Ω_1 and Ω_3. $xy \leqslant 0$ in the set Ω_2 and Ω_4. Because of that

$$\underline{I}_{xy} = I_{xy}(\underline{\Omega} \cup \Omega_2 \cup \Omega_4) = \iint\limits_{\underline{\Omega}\cup\Omega_2\cup\Omega_4} (x^2+y^2)dxdy \tag{61}$$

$$\overline{I}_{xy} = I_{xy}(\underline{\Omega} \cup \Omega_1 \cup \Omega_3) = \iint\limits_{\underline{\Omega}\cup\Omega_1\cup\Omega_3} (x^2+y^2)dxdy \tag{62}$$

5. General case

In general it is possible to consider a functional which is dependent on parameters $h_i \in \mathbf{p}$, functional parameters $p_i(x) \in \mathbf{p}_i(x)$ and integrals which are dependent on the sets Ω_i

$$u = F(h_1, ..., h_m, y_1, ..., y_p, z_1, ..., z_q) \tag{63}$$

$$y_1 = \int_{\Gamma_1} L_1(x, p_1(x), ..., p_k(x))dx \tag{64}$$

$$y_2 = \int_{\Gamma_2} L_2(x, p_1(x), ..., p_k(x))dx \tag{65}$$

$$... \tag{66}$$

$$y_q = \int_{\Gamma_q} L_q(x, p_1(x), ..., p_k(x))dx \tag{67}$$

$$z_1 = \int_{\Omega_1} \Psi_1(x, p_1(x), ..., p_k(x))dx \tag{68}$$

$$z_2 = \int_{\Omega_2} \Psi_2(x, p_1(x), ..., p_k(x))dx \tag{69}$$

$$... \tag{70}$$

$$z_q = \int_{\Omega_q} \Psi_q(x, p_1(x), ..., p_k(x))dx \tag{71}$$

If the sign of each derivative is constant then it is possible to apply sensitivity analysis to each uncertain parameters separately.

6. Direct method of calculation of sensitivity from differential equation

6.1. Sensitivity with respect to real valued parameters

Lets us consider tension-compression differential equation

$$\frac{d}{dx}\left(EA\frac{du}{dx}\right) + n = 0 \tag{72}$$

with the following boundary condition

$$EA\frac{du(0)}{dx} = P, \quad u(0) = 0 \tag{73}$$

After integration we will get

$$EA\frac{du}{dx} + \int\limits_0^x n dx = EA\frac{du(0)}{dx} \tag{74}$$

$$EA\frac{du}{dx} + \int\limits_0^x n dx = P \tag{75}$$

$$\frac{du}{dx} = \frac{P}{EA} - \frac{1}{EA}\int\limits_0^x n dx \tag{76}$$

$$u(x) = u(0) + \int\limits_0^x \frac{P}{EA} d\eta - \int\limits_0^x \left(\frac{1}{EA} \int\limits_0^\eta n d\xi \right) d\eta \tag{77}$$

$$u(x) = \int\limits_0^x \frac{P}{EA} d\eta - \int\limits_0^x \left(\frac{1}{EA} \int\limits_0^\eta n d\xi \right) d\eta \tag{78}$$

For constant values of E, A and n we will get

$$u(x) = \frac{Px}{EA} - \frac{nx^2}{2EA} \tag{79}$$

Partial derivative of the solution is equal to

$$\frac{\partial u(x)}{\partial E(y)} = -\frac{Px}{E^2 A} + \frac{nx^2}{2E^2 A} \tag{80}$$

Functional derivative of the differential equation with respect to the uncertain parameter p_i

$$\frac{\partial}{\partial p_i}\left[\frac{d}{dx}\left(EA\frac{du}{dx} \right) + n \right] = 0 \tag{81}$$

$$\frac{d}{dx}\left(\frac{\partial (EA)}{\partial p_i}\frac{du}{dx} \right) + \frac{d}{dx}\left(EA\frac{d}{dx}\left(\frac{\partial u}{\partial p_i} \right) \right) + \frac{\partial n}{\partial p_i} = 0 \tag{82}$$

For example if $p_i = E$ then

$$\frac{d}{dx}\left(\frac{\partial (EA)}{\partial E}\frac{du}{dx} \right) + \frac{d}{dx}\left(EA\frac{d}{dx}\left(\frac{\partial u}{\partial E} \right) \right) + \frac{\partial n}{\partial E} = 0 \tag{83}$$

$$\frac{d}{dx}\left(A\frac{du}{dx} \right) + \frac{d}{dx}\left(EA\frac{d}{dx}\left(\frac{\partial u}{\partial E} \right) \right) = 0 \tag{84}$$

After integration

$$A\frac{du}{dx} + EA\frac{d}{dx}\left(\frac{\partial u}{\partial E} \right) = C \tag{85}$$

Derivative of boundary conditions

$$\frac{d}{dx}\left(\frac{\partial u(0)}{\partial E}\right) = -\frac{P}{E^2 A}, \quad \frac{\partial u(0)}{\partial E} = 0 \tag{86}$$

then

$$\frac{P}{E} - EA\frac{P}{E^2 A} = C \;\Rightarrow\; 0 = C \tag{87}$$

From boundary conditions we will get

$$A\frac{du}{dx} + EA\frac{d}{dx}\left(\frac{\partial u}{\partial E}\right) = 0 \tag{88}$$

$$\frac{d}{dx}\left(\frac{\partial u}{\partial E}\right) = -\frac{1}{E}\frac{du}{dx} \tag{89}$$

$$\frac{d}{dx}\left(\frac{\partial u}{\partial E}\right) = -\frac{1}{E}\left(\frac{P}{EA} - \frac{1}{EA}\int_0^x n d\eta\right) \tag{90}$$

$$\frac{d}{dx}\left(\frac{\partial u}{\partial E}\right) = -\frac{P}{E^2 A} + \frac{1}{E^2 A}\int_0^x n d\eta \tag{91}$$

After integration

$$\frac{\partial u}{\partial E} = \frac{\partial u(0)}{\partial E} - \int_0^x \frac{P}{E^2 A} d\xi + \int_0^x \left(\frac{1}{E^2 A}\int_0^\xi n d\eta\right) d\xi \tag{92}$$

$$\frac{\partial u}{\partial E} = -\int_0^x \frac{P}{E^2 A} d\xi + \int_0^x \left(\frac{1}{E^2 A}\int_0^\xi n d\eta\right) d\xi \tag{93}$$

For constant values

$$\frac{\partial u}{\partial E} = -\frac{Px}{E^2 A} + \frac{nx^2}{2E^2 A} \tag{94}$$

Using this method it is possible to avoid approximation errors.

6.2. Sensitivity with respect to functional parameters

The solution of the equation (72) is given by the formula (78). The functional derivative with the respect the the values of Young modulus $E(y)$ is equal to

$$\frac{\delta u(x)}{\delta E(y)} = \frac{\delta}{\delta E(y)}\int_0^x \frac{P}{EA} d\eta - \frac{\delta}{\delta E(y)}\int_0^x \left(\frac{1}{EA}\int_0^\eta n d\xi\right) d\eta \tag{95}$$

$$\frac{\delta u(x)}{\delta E(y)} = -\frac{P}{E^2(y)A(y)} + \frac{1}{E^2(y)A(y)} \int\limits_0^y n d\xi \tag{96}$$

It is possible to calculate the functional derivative of the solution of the equation (72) with respect of the functional parameter $E = E(y)$

$$\frac{d}{dx}\left(E(x)A(x)\frac{du(x,E)}{dx}\right) + n(x) = 0 \tag{97}$$

$$\frac{d}{dx}\left((E(x)+\delta E(x))A(x)\frac{du(x,E+\delta E)}{dx}\right) + n(x) = 0 \tag{98}$$

The last equation for a small perturbation can be written in the following way

$$u(x, E+\delta E) \approx u(x,E) + \delta u_E(x, \delta E) \tag{99}$$

After neglecting quadratic terms we will get

$$\begin{array}{c} \frac{d}{dx}\left(E(x)A(x)\frac{du(x,E)}{dx}\right) + \frac{d}{dx}\left(\delta E(x)A(x)\frac{du(x,E)}{dx}\right) + \\ + \frac{d}{dx}\left(E(x)A(x)\frac{d}{dx}\delta u_E(x,\delta E)\right) + n(x) = 0 \end{array} \tag{100}$$

If we subtract the equations (97) and (100) the result is

$$\frac{d}{dx}\left(\delta E(x)A(x)\frac{du(x,E)}{dx}\right) + \frac{d}{dx}\left(E(x)A(x)\frac{d}{dx}\delta u_E(x,\delta E)\right) = 0 \tag{101}$$

After integration we will get

$$\delta E(x)A(x)\frac{du(x,E)}{dx} + E(x)A(x)\frac{d}{dx}\delta u_E(x,\delta E) = C \tag{102}$$

The functional derivative of the boundary conditions is given by the following formulas

$$u(0,E) = 0, \tag{103}$$

$$u(0, E+\delta E) = 0, \tag{104}$$

$$u(0,E) + \delta u_E(0,\delta E) = 0 \tag{105}$$

then

$$\delta u(0, \delta E) = 0 \tag{106}$$

$$\frac{d}{dx}u(0,E) = \frac{P}{E(0)A(0)}, \tag{107}$$

$$\frac{d}{dx}u(0,E+\delta E) = \frac{P}{(E(0)+\delta E(0))A(0)}, \tag{108}$$

$$\frac{d}{dx}u(0,E) + \frac{d}{dx}\delta u_E(0,\delta E) = \frac{P}{E(0)A(0)} - \frac{P\delta E(0)}{E^2(0)A(0)} \tag{109}$$

then

$$\frac{d}{dx}\delta u_E(0,\delta E) = -\frac{P\delta E(0)}{E^2(0)A(0)} \tag{110}$$

For $x = 0$ we have

$$\delta E(0)A(0)\frac{du(0,E)}{dx} + E(0)A(0)\frac{d}{dx}\delta u_E(0,\delta E) = C \tag{111}$$

From boundary conditions

$$\delta E(0)A(0)\frac{P}{E(0)A(0)} - E(0)A(0)\frac{P\delta E(0)}{E^2(0)A(0)} = C \tag{112}$$

$$0 = C \tag{113}$$

Now the equation has the following form

$$\delta E(x)A(x)\frac{du(x,E)}{dx} + E(x)A(x)\frac{d}{dx}\delta u_E(x,\delta E) = 0 \tag{114}$$

$$\frac{d}{dx}\delta u_E(x,\delta E) = -\frac{\delta E(x)}{E(x)}\frac{du(x,E)}{dx} \tag{115}$$

From the equation (76)

$$\frac{d}{dx}\delta u_E(x,\delta E) = -\frac{\delta E(x)}{E(x)}\left(\frac{P}{E(x)A(x)} - \frac{1}{E(x)A(x)}\int_0^x n(x)dx\right) \tag{116}$$

$$\frac{d}{dx}\delta u_E(x,\delta E) = -\frac{P\delta E(x)}{E^2(x)A(x)} + \frac{\delta E(x)}{E^2(x)A(x)}\int_0^x n(\eta)d\eta \tag{117}$$

After integration

$$\begin{aligned}\delta u_E(x,\delta E) = \delta u_E(0,\delta E) - \int_0^x \tfrac{P\delta E(\xi)}{E^2(\xi)A(\xi)}d\xi + \\ + \int_0^x \left(\tfrac{\delta E(\xi)}{E^2(\xi)A(\xi)}\int_0^\xi n(\eta)d\eta\right)d\xi\end{aligned} \tag{118}$$

for $x = 0$ we know that $\delta u(0,\delta E) = 0$, then

$$\delta u_E(x,\delta E) = -\int_0^x \tfrac{P\delta E(\xi)}{E^2(\xi)A(\xi)}d\xi + \int_0^x \left(\tfrac{\delta E(\xi)}{E^2(\xi)A(\xi)}\int_0^\xi n(\eta)d\eta\right)d\xi \tag{119}$$

$$\delta u_E(x,\delta E) = \int_0^x \left(\tfrac{-P}{E^2(\xi)A(\xi)} + \tfrac{1}{E^2(\xi)A(\xi)}\int_0^\xi n(\eta)d\eta\right)\delta E(\xi)d\xi \tag{120}$$

then

$$\frac{\delta u(x)}{\delta E(\xi)} = \frac{-P}{E^2(\xi)A(\xi)} + \frac{1}{E^2(\xi)A(\xi)}\int_0^\xi n(\eta)d\eta \tag{121}$$

7. FEM with uncertain functional parameters

Finite element method lead to the following parameter dependent system of equations (Zienkiewicz and Taylor, 2000)

$$K(p)u(p) = Q(p) \tag{122}$$

where K is the stiffness matrix, Q is the load vector and u is the vector of the solutions. The functional derivative $\frac{\delta u(p)}{\delta p_i(x)}$ of the solution can be calculated from the following equation

$$K(p)\frac{\delta u(p)}{\delta p_i(x)} = \frac{\delta Q(p)}{\delta p_i(x)} - \frac{\delta K(p)}{\delta p_i(x)}u(p). \tag{123}$$

The solution $\frac{\delta u(p)}{\delta p_i(x)}$ can be applied in the algorithms, which are described in the previous sections. It is not possible to calculate the functional derivative $\frac{\delta u(p)}{\delta p_i(x)}$ in all points $x \in \Omega$. Because of that functional derivative should be calculated in as many grid points as possible x_k. The sign of the functional derivative $\frac{\delta u(p)}{\delta p_i(x)}$ is calculated by using the nearest grid points x_k i.e. $\frac{\delta u(p)}{\delta p_i(x_k)}$.

8. Postprocessing of the interval solution based on sensitivity analysis

8.1. 3D elasticity

In structural mechanics solution of the system of equations (122) is used for calculations of other mechanical quantities like for example stress and strain. In linear elasticity the relation between the strain tensor ε and displacement vector u is the following

$$\varepsilon(x) = \frac{1}{2}\left(\nabla^T u(x) + \nabla u(x)\right) \tag{124}$$

$$\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \tag{125}$$

In cartesian coordinate system it is possible to approximate the displacement field $u_i(x)$ (i is a direction of the displacement i.e. $x, y, z, \varphi_x, \varphi_y, \varphi_z$) in the element Ω_e using shape functions $N_{ek}(x)$ (e is a number of element, k is a number of node) and the values of the function u_i in the nodal points x_{ek} (usually $u_{eki} = u_i(x_{ek})$, however u_{eki} can be also defined using derivatives of the function u_i, e is a number of element, k is a number of node, i is a direction of the displacement)

$$u_{ei}(x) \approx \sum_k N_{ek}(x)u_{eki} \tag{126}$$

From the equation (124) and (126) we have

$$\varepsilon_{eij} = \frac{1}{2}\left(\frac{\partial}{\partial x_j}\left(\sum_k N_{ek}(x)u_{eki}\right) + \frac{\partial}{\partial x_i}\left(\sum_k N_{ek}(x)u_{ekj}\right)\right) = \tag{127}$$

$$= \frac{1}{2}\left(\sum_k \frac{\partial N_{ek}(x)}{\partial x_j} u_{eki} + \sum_k \frac{\partial N_{ek}(x)}{\partial x_i} u_{eki}\right) = \tag{128}$$

$$= \frac{1}{2}\sum_{k,p}\left(\frac{\partial N_{ek}(x)}{\partial x_j}\delta_{pi} + \frac{\partial N_{ek}(x)}{\partial x_i}\delta_{pj}\right) u_{ekp} = \tag{129}$$

$$= \sum_{k,p} B_{eijkp} u_{ekp} \tag{130}$$

then

$$B_{eijkp} = \frac{1}{2}\sum_{k,p}\left(\frac{\partial N_{ek}(x)}{\partial x_j}\delta_{pi} + \frac{\partial N_{ek}(x)}{\partial x_i}\delta_{pj}\right) \tag{131}$$

Relation between the global solution vector u_q (q is a number of degree of freedom in the solution vector) and the vector of local solution of the elements u_{ekp} (e is a number of element, k is a number of node in the element e, p is a direction of the displacement) is the following

$$u_{ekp} = \sum_q U_{ekpq} u_q \tag{132}$$

Sensitivity of the displacements

$$\frac{\partial u_{ekp}}{\partial p_j} = \sum_q U_{ekpq} \frac{\partial u_q}{\partial p_j} \tag{133}$$

In the case of linear elastic materials the relation between the stress σ_{ij} and strain ε_{ij} is the following

$$\sigma_{emn} = \sum_{i,j} C_{emnij}\varepsilon_{eij} = \sum_{i,j,k,p} C_{emnij} B_{eijkp} u_{ekp} \tag{134}$$

The sensitivity of the strain field can be calculated as a derivative

$$\frac{\partial \varepsilon_{eij}}{\partial p_l} = \frac{\partial}{\partial p_l}\left(\sum_{k,p} B_{eijkp} u_{ekp}\right) = \tag{135}$$

$$= \sum_{k,p}\left(\frac{\partial B_{eijkp}}{\partial p_l} u_{ekp} + B_{eijkp}\frac{\partial u_{ekp}}{\partial p_l}\right) \tag{136}$$

or in the case of functional parameters

$$\frac{\delta \varepsilon_{eij}}{\delta p_l(x)} = \frac{\delta}{\delta p_l(x)}\left(\sum_{k,p} B_{eijkp} u_{ekp}\right) = \tag{137}$$

$$= \sum_{k,p}\left(\frac{\delta B_{eijkp}}{\delta p_l(x)} u_{ekp} + B_{eijkp}\frac{\delta u_{ekp}}{\delta p_l(x)}\right) \tag{138}$$

The sensitivity of the stress field can be calculated from the equation (134)

$$\frac{\partial \sigma_{emn}}{\partial p_l} = \frac{\partial}{\partial p_l}\left(\sum_{i,j,k,p} C_{emnij} B_{eijkp} u_{ekp}\right) = \tag{139}$$

$$= \sum_{i,j,k,p}\left(\frac{\partial (C_{emnij} B_{eijkp})}{\partial p_l} u_{ekp} + B_{eijkp} C_{emnij} \frac{\partial u_{ekp}}{\partial p_l}\right) \tag{140}$$

or in the case of functional parameters

$$\frac{\delta \sigma_{emn}}{\delta p_l(x)} = \frac{\delta}{\delta p_l(x)}\left(\sum_{i,j,k,p} C_{emnij} B_{eijkp} u_{ekp}\right) \tag{141}$$

$$= \sum_{i,j,k,p}\left(\frac{\delta (C_{emnij} B_{eijkp})}{\delta p_l(x)} u_{ekp} + B_{eijkp} C_{emnij} \frac{\delta u_{ekp}}{\delta p_l(x)}\right) \tag{142}$$

If we know the derivatives of the strain and stress field then it is possible to calculate the extreme values of the solution using the methods which are described in the previous sections.
Potential energy can be calculated as

$$V = \sum_{e,m,n,i,j} \int\limits_{\Omega} C_{emnij} \varepsilon_{eij} \varepsilon_{emn} d\Omega - \sum_{e,i} \int\limits_{\Omega} f_{ei} u_{ei} d\Omega \tag{143}$$

where f_{ei} are the loads. The local stiffness matrix can be calculated from the following formula

$$K_{ekplq} = \sum_{e,m,n,i,j} \int\limits_{\Omega} C_{emnij} B_{emnkp} B_{eijlp} d\Omega \tag{144}$$

Global stiffness matrix

$$K_{\alpha\beta} = \sum_{e} K_{e\alpha\beta} \tag{145}$$

where

$$K_{e\alpha\beta} = \sum_{k,p,l,q} K_{ekplq} U_{ekp\alpha} U_{ekp\beta} \tag{146}$$

Above relation is linear that is way it is possible to calculate sensitivity of global stiffness matrix using linear relation

$$\frac{\partial K_{\alpha\beta}}{\partial p_\gamma} = \sum_{e} \frac{\partial K_{e\alpha\beta}}{\partial p_\gamma} \tag{147}$$

$$\frac{\partial K_{e\alpha\beta}}{\partial p_\gamma} = \sum_{k,p,l,q} \frac{\partial K_{ekplq}}{\partial p_\gamma} U_{ekp\alpha} U_{ekp\beta} \tag{148}$$

Local load vector can be calculated using shape functions $N_{ek}(x)$ and load vector t_{ei}

$$Q_{eki} = \int\limits_{\Omega_e} t_{ei} N_{ek} d\Omega \tag{149}$$

Global load vector Q_p can be assembled from the local load vectors Q_{eki}

$$Q_p = \sum_{eki} U_{ekip} Q_{eki} \tag{150}$$

then the sensitivity of the global load vector can be calculated from the sensitivity of the local load vectors

$$\frac{\partial Q_p}{\partial p_l} = \sum_{eki} U_{ekip} \frac{\partial Q_{eki}}{\partial p_l} \tag{151}$$

8.2. Tension-compression problem

The displacement field u in the case of tension-compression problem is described by second order differential equation

$$\frac{d}{dx}\left(EJ\frac{du}{dx}\right) + n = 0 \tag{152}$$

where E is Young modulus, J is a moment of inertia, n is a vector of continuous loads and u is a displacement. After discretization in the case of constant E, A, L we will get the following stiffness matrix

$$K_e = \begin{bmatrix} k_{e11} & k_{e12} \\ k_{e21} & k_{e22} \end{bmatrix} = \begin{bmatrix} \frac{E_e A_e}{L_e} & -\frac{E_e A_e}{L_e} \\ -\frac{E_e A_e}{L_e} & \frac{E_e A_e}{L_e} \end{bmatrix} \tag{153}$$

Sensitivity with respect to the variation of Young modulus

$$\frac{\partial K_e}{\partial E_p} = \begin{bmatrix} \frac{\delta_{ep} A_e}{L_e} & -\frac{\delta_{ep} A_e}{L_e} \\ -\frac{\delta_{ep} A_e}{L_e} & \frac{\delta_{ep} A_e}{L_e} \end{bmatrix} \tag{154}$$

in similar way it is possible to calculate sensitivity with the respect of other parameters. Global stiffness matrix can be calculated in by using the connectivity matrix.

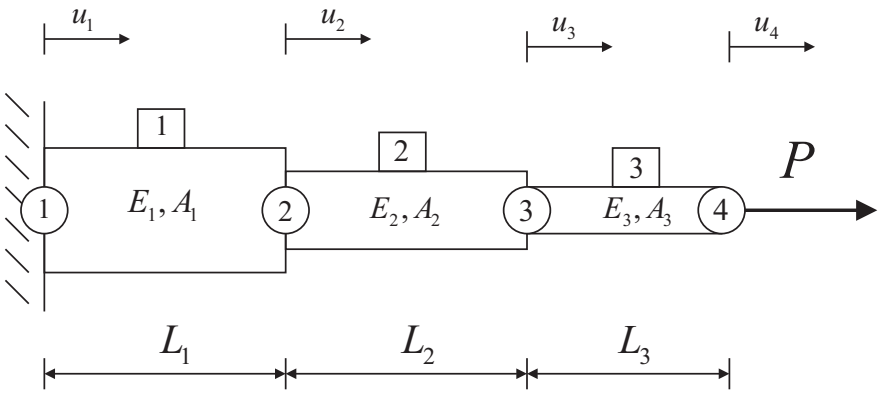


Figure 3. Tension problem

Global stiffness matrix can be calculated in the following way

$$K_1 = \begin{bmatrix} \frac{E_1 A_1}{L_1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{155}$$

$$K_2 = \begin{bmatrix} \frac{E_2 A_2}{L_2} & -\frac{E_2 A_2}{L_2} & 0 \\ -\frac{E_2 A_2}{L_2} & \frac{E_2 A_2}{L_2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{156}$$

$$K_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{E_3 A_3}{L_3} & -\frac{E_3 A_3}{L_3} \\ 0 & -\frac{E_3 A_3}{L_3} & \frac{E_3 A_3}{L_3} \end{bmatrix} \tag{157}$$

$$K = K_1 + K_2 + K_3 = \begin{bmatrix} \frac{E_1 A_1}{L_1} + \frac{E_2 A_2}{L_2} & -\frac{E_2 A_2}{L_2} & 0 \\ -\frac{E_2 A_2}{L_2} & \frac{E_2 A_2}{L_2} + \frac{E_3 A_3}{L_3} & -\frac{E_3 A_3}{L_3} \\ 0 & -\frac{E_3 A_3}{L_3} & \frac{E_3 A_3}{L_3} \end{bmatrix} \tag{158}$$

Global load vector after applying boundary conditions

$$Q = \begin{bmatrix} 0 \\ 0 \\ P \end{bmatrix} \tag{159}$$

Mid point solution is a solution of the following system of equation

$$Ku = Q \tag{160}$$

where

$$u = \begin{bmatrix} u_2 \\ u_3 \\ u_4 \end{bmatrix} \tag{161}$$

Sensitivity of the displacement u with respect of value of Young modulus E_2 can be calculated from the following system of equation

$$K\frac{\partial u}{\partial E_2} = \frac{\partial Q}{\partial E_2} - \frac{\partial K}{\partial E_2}u \tag{162}$$

where

$$\frac{\partial K}{\partial E_2} = \frac{\partial K_1}{\partial E_2} + \frac{\partial K_2}{\partial E_2} + \frac{\partial K_3}{\partial E_2} = \begin{bmatrix} \frac{A_2}{L_2} & -\frac{A_2}{L_2} & 0 \\ -\frac{A_2}{L_2} & \frac{A_2}{L_2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{163}$$

$$\frac{\partial Q}{\partial E_2} = \frac{\partial}{\partial E_2}\begin{bmatrix} 0 \\ 0 \\ P \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{164}$$

Displacements in the first element

$$u_1(x) = [N_{11}(x), N_{12}(x)] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{165}$$

Sensitivity of the displacements in the first element

$$\frac{\partial u_1(x)}{\partial E_2} = \left[1 - \frac{x}{L_1}, \frac{x}{L_1}\right] \begin{bmatrix} \frac{\partial u_1}{\partial E_2} \\ \frac{\partial u_2}{\partial E_2} \end{bmatrix} \tag{166}$$

Displacements in the second element

$$u_2(x) = [N_{21}(x), N_{22}(x)] \begin{bmatrix} u_2 \\ u_3 \end{bmatrix} \tag{167}$$

Sensitivity of the displacements in the first element

$$\frac{\partial u_2(x)}{\partial E_2} = \left[1 - \frac{x}{L_2}, \frac{x}{L_2}\right] \begin{bmatrix} \frac{\partial u_2}{\partial E_2} \\ \frac{\partial u_3}{\partial E_2} \end{bmatrix} \tag{168}$$

Displacements in the third element

$$u_3(x) = [N_{31}(x), N_{32}(x)] \begin{bmatrix} u_3 \\ u_4 \end{bmatrix} \tag{169}$$

Sensitivity of the displacements in the third element

$$\frac{\partial u_3(x)}{\partial E_2} = \left[1 - \frac{x}{L_3}, \frac{x}{L_3}\right] \begin{bmatrix} \frac{\partial u_3}{\partial E_2} \\ \frac{\partial u_4}{\partial E_2} \end{bmatrix} \tag{170}$$

Now it is possible to calculate the strain in all elements

$$\varepsilon_1 = \frac{du}{dx} = \left[-\frac{1}{L_1}, \frac{1}{L_1}\right] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{171}$$

$$\varepsilon_2 = \frac{du}{dx} = \left[-\frac{1}{L_2}, \frac{1}{L_2}\right] \begin{bmatrix} u_2 \\ u_3 \end{bmatrix} \tag{172}$$

$$\varepsilon_2 = \frac{du}{dx} = \left[-\frac{1}{L_3}, \frac{1}{L_3}\right] \begin{bmatrix} u_3 \\ u_4 \end{bmatrix} \tag{173}$$

Sensitivity of the strain field

$$\frac{\partial \varepsilon_1}{\partial E_2} = \left[-\frac{1}{L_1}, \frac{1}{L_1}\right] \begin{bmatrix} \frac{\partial u_1}{\partial E_2} \\ \frac{\partial u_2}{\partial E_2} \end{bmatrix} \tag{174}$$

$$\frac{\partial \varepsilon_2}{\partial E_2} = \left[-\frac{1}{L_2}, \frac{1}{L_2}\right] \begin{bmatrix} \frac{\partial u_2}{\partial E_2} \\ \frac{\partial u_3}{\partial E_2} \end{bmatrix} \tag{175}$$

$$\frac{\partial \varepsilon_3}{\partial E_2} = \left[-\frac{1}{L_3}, \frac{1}{L_3} \right] \begin{bmatrix} \frac{\partial u_3}{\partial E_2} \\ \frac{\partial u_4}{\partial E_2} \end{bmatrix} \tag{176}$$

The stress in elements

$$\sigma_1 = E_1 \varepsilon_1 \tag{177}$$

$$\sigma_2 = E_2 \varepsilon_2 \tag{178}$$

$$\sigma_3 = E_3 \varepsilon_3 \tag{179}$$

The sensitivity of the stress filed

$$\frac{\partial \sigma_1}{\partial E_2} = E_1 \frac{\partial \varepsilon_1}{\partial E_2} \tag{180}$$

$$\frac{\partial \sigma_2}{\partial E_2} = 1 \cdot \varepsilon_2 + E_2 \frac{\partial \varepsilon_2}{\partial E_2} \tag{181}$$

$$\frac{\partial \sigma_3}{\partial E_2} = E_3 \frac{\partial \varepsilon_3}{\partial E_2} \tag{182}$$

Above described formulas are true only if the Young modulus $E = E(x)$ and the area of cross-section $A = A(x)$ is constant inside each element. If these functions are not constant then the stiffness matrix have to be calculated by using the integration.

$$K_1 = \begin{bmatrix} \int\limits_0^L E_1(x) A_1(x) \frac{dN_{12}}{dx} \frac{dN_{12}}{dx} dx & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{183}$$

$$K_2 = \begin{bmatrix} \int\limits_0^L E_2(x) A_2(x) \frac{dN_{21}}{dx} \frac{dN_{21}}{dx} dx & \int\limits_0^L E_2(x) A_2(x) \frac{dN_{21}}{dx} \frac{dN_{12}}{dx} dx & 0 \\ \int\limits_0^L E_2(x) A_2(x) \frac{dN_{22}}{dx} \frac{dN_{21}}{dx} dx & \int\limits_0^L E_2(x) A_2(x) \frac{dN_{22}}{dx} \frac{dN_{22}}{dx} dx & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{184}$$

$$K_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \int\limits_0^L E_3(x) A_3(x) \frac{dN_{31}}{dx} \frac{dN_{31}}{dx} dx & \int\limits_0^L E_3(x) A_3(x) \frac{dN_{31}}{dx} \frac{dN_{32}}{dx} dx \\ 0 & \int\limits_0^L E_3(x) A_3(x) \frac{dN_{32}}{dx} \frac{dN_{31}}{dx} dx & \int\limits_0^L E_3(x) A_3(x) \frac{dN_{32}}{dx} \frac{dN_{33}}{dx} dx \end{bmatrix} \tag{185}$$

The functional derivative can be calculated without differentiation.

$$\frac{\delta K_1}{\delta E_2(x)} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{186}$$

$$\frac{\delta K_2}{\delta E_2(x)} = \begin{bmatrix} A_2(x) \frac{dN_{21}}{dx} \frac{dN_{21}}{dx} & A_2(x) \frac{dN_{21}}{dx} \frac{dN_{12}}{dx} & 0 \\ A_2(x) \frac{dN_{22}}{dx} \frac{dN_{21}}{dx} & A_2(x) \frac{dN_{22}}{dx} \frac{dN_{22}}{dx} & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{187}$$

$$\frac{\delta K_3}{\delta E_2(x)} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{188}$$

$$\frac{\delta K}{\delta E_2(x)} = \frac{\delta K_1}{\delta E_2(x)} + \frac{\delta K_2}{\delta E_2(x)} + \frac{\delta K_3}{\delta E_2(x)} \tag{189}$$

$$\frac{\delta Q}{\delta E_2(x)} = \frac{\delta}{\delta E_2(x)} \begin{bmatrix} 0 \\ 0 \\ P \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{190}$$

Functional derivative of the displacements can be calculated from the following system of equations

$$K \frac{\delta u}{\delta E_2(x)} = \frac{\delta Q}{\delta E_2(x)} - \frac{\delta K}{\delta E_2(x)} u \tag{191}$$

In the same way it is possible to calculate the functional derivative of the displacements, stress and strain fields.

8.3. Truss structures

Using the sensitivity analysis method it is possible to calculate the interval displacements in the truss structures with the interval Young modulus and the area of cross-section (Fig. 4). The struc-

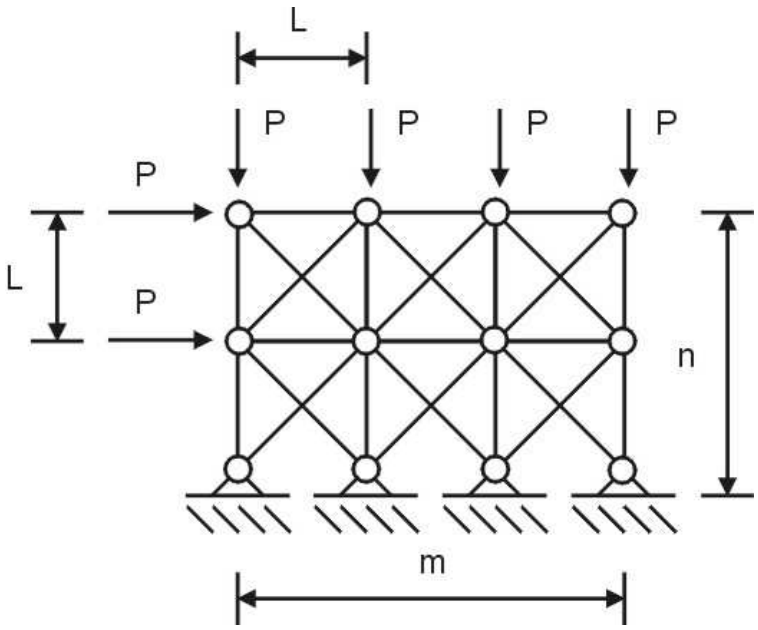


Figure 4. Plain stress-strain problem

ture can be described in two steps. In the first step the truss structure is described by using ANSYS FEM (http://www.ansys.com) program internal scripting language. In this case the uncertainty of

Table I. Interval displacements of truss structure

ID	Interval displacement $[m]$	node	dof
u[0]	[2.54206368e-05, 2.88890319e-05]	5	1
u[1]	[-2.41613231e-06, -5.5936232e-07]	5	2
u[2]	[1.89488493e-05, 2.18240888e-05]	6	1
u[3]	[-1.18336781e-05, -9.68801242e-06]	6	2
u[4]	[1.74375666e-05, 2.00368684e-05]	7	1
u[5]	[-1.53016570e-05, -1.28438219e-05]	7	2
u[6]	[2.23883755e-05, 2.55322229e-05]	8	1
u[7]	[-2.43184098e-05, -2.13175562e-05]	8	2
u[8]	[4.47984203e-05, 5.07482294e-05]	9	1
u[9]	[-1.25873042e-05, -9.13828295e-06]	9	2
u[10]	[3.58319463e-05, 4.09641151e-05]	10	1
u[11]	[-2.03184368e-05, -1.75638790e-05]	10	2
u[12]	[3.30408793e-05, 3.79901925e-05]	11	1
u[13]	[-2.87524495e-05, -2.54594638e-05]	11	2
u[14]	[3.51831538e-05, 4.042328624e-05]	12	1
u[15]	[-4.18322390e-05, -3.7394527e-05]	12	2

the Young modulus is 5% (MP, EX, 1, 5) and the uncertainty of the area of cross-section is also 5% (R, 1, 5). The interval displacements are shown in the Table I.
This example shows that the sensitivity analysis can be use as an extension of existing FEM code.

8.4. Plain stress

Let us consider a 2D structure which is shown on Fig. 5.

In calculation linear-elastic plain stress-strain mathematical model was used. Young nodulus was uncertain and equal to $E \in [210 \cdot 10^9, 212 \cdot 10^9] \frac{N}{m^2}$, Poisson number $\nu \in [0.2, 0.4]$, thickness $h = 0.1m$, width $L = 1m$, height $h = 1m$ surface load $t_y \in [3, 2]kN$. Numerical results are shown in the Table II. The results are show in the following format
$u[number] = [lower\ bound,\ midpoint\ solution,\ upper\ bound]$.

8.5. Interval stress in 3D elastic body

Using described theory it is possible to calculate the interval stress using the 3D brick elements (Fig. 6). Let us consider 6 finite elements with continuous loads $q \in [1, 3] \frac{kN}{m}$, Young modulus $E \in [210, 212] 10^9 \frac{N}{m^2}$, Poisson number $\nu \in [0.2, 0.4]$ which are shown in the Fig. 7.
In each element there are 27 Gauss points. Results of calculations are shown in the table below.

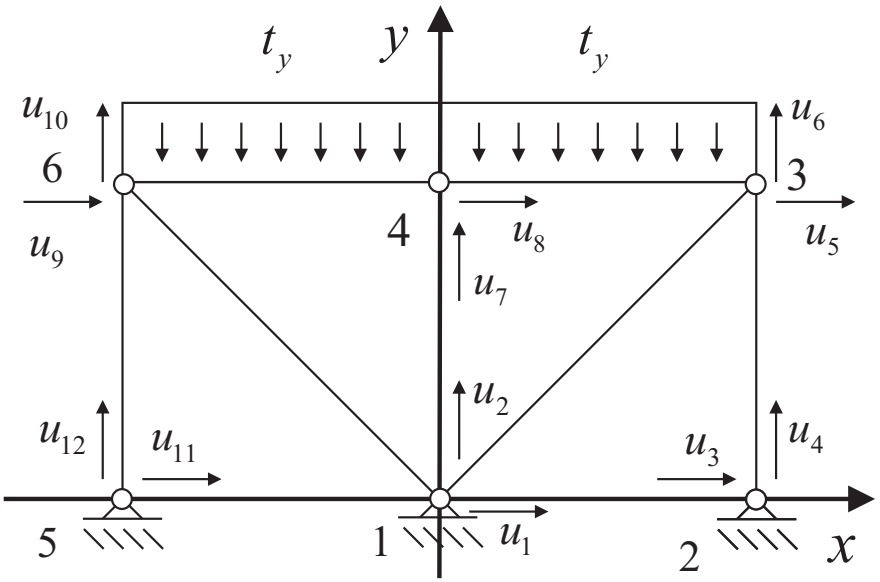


Figure 5. Plain stress-strain problem

Table II. Interval displacements in the truss structure

ID	Interval displacement [m]
u[5]	[7.010160e-08, 9.384325e-08, 1.175510e-07]
u[6]	[-4.461538e-07, -3.479902e-07, -2.587601e-07]
u[7]	[-4.600000e-07, -3.619668e-07, -2.716981e-07]
u[9]	[-1.175510e-07, -9.384325e-08, -7.010160e-08]
u[10]	[-4.461538e-07, -3.479902e-07, -2.587601e-07]

The program can be run from the web page http://andrzej.pownuk.com. The structure is described using some easy to understand scripting language.

9. The computer program

Sensitivity analysis is implemented in object oriented C++ computer program. In the program there is 11 finite elements. The program allow to use the following analysis types

1. Liner static analysis (classical FEM solution)
2. Liner static analysis with interval combinatoric
3. Liner static analysis with sensitivity analysis

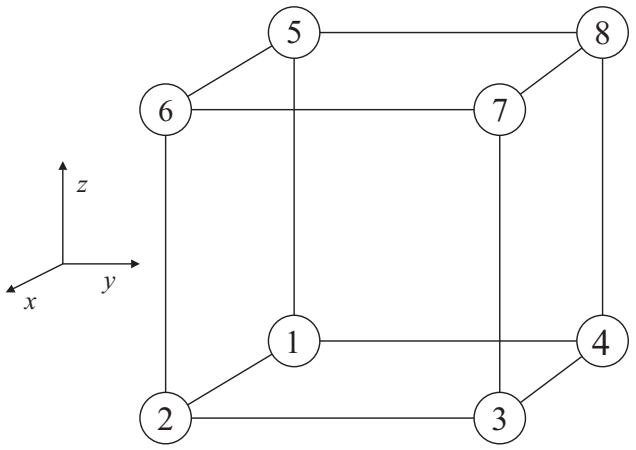


Figure 6. 3D brick element

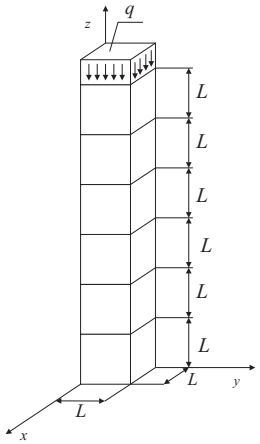


Figure 7. 3D brick elements with surface load

4. Liner static analysis with Taylor expansion method

5. Liner static analysis with functional derivative method

6. Liner static analysis with combination of functional derivative and sensitivity analysis method

The program can be run on-line from the web page http://andrzej.pownuk.com. In the first box there is a description of the problem Fig. 8.
After clicking the button "calculate" the results will appear in the second box.
In order to see all steps of the calculations "debug" command can be apply (e.g. debug_interval_solution).
In order to see the intermediate results commands "print" can be applied (e.g. print_global_stiffness_matrix).

Table III. Interval stress in the element 1

Number of Gauss point	interval stress $\sigma_{zz}\left[\frac{N}{m^2}\right]$
1	[-3.000000e+03, -1.000000e+03]
2	[-3.000000e+03, -1.000000e+03]
3	[-3.000000e+03, -1.000000e+03]
4	[-3.000000e+03, -1.000000e+03]
5	[-3.000000e+03, -1.000000e+03]
6	[-3.000000e+03, -1.000000e+03]
etc.	etc.

Table IV. Interval von Mises stress in the element 1

Number of Gauss point	interval von Mises stress $\sigma_M\left[\frac{N}{m^2}\right]$
1	[1.000000e+03, 3.000000e+03]
2	[1.000000e+03, 3.000000e+03]
3	[1.000000e+03, 3.000000e+03]
4	[1.000000e+03, 3.000000e+03]
5	[1.000000e+03, 3.000000e+03]
6	[1.000000e+03, 3.000000e+03]
etc.	etc.

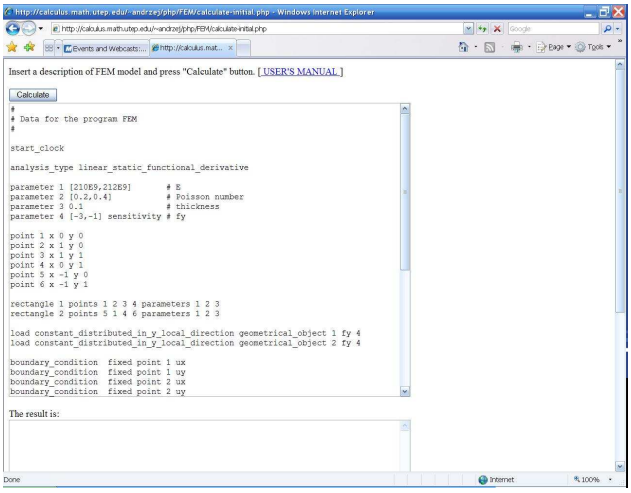


Figure 8. Web application

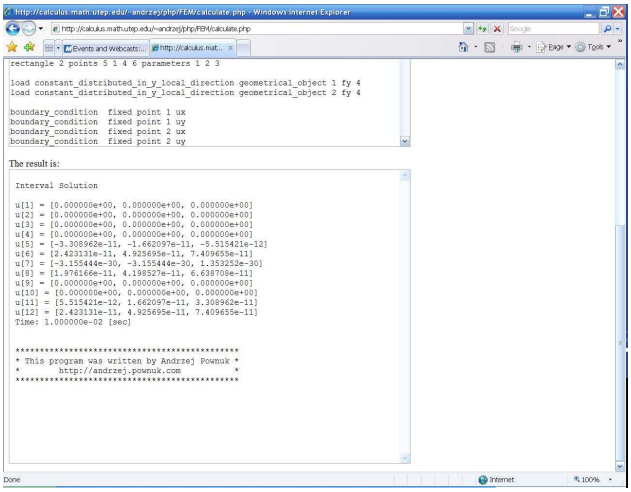


Figure 9. Results of the calculations

10. Interval eigenvalue

10.1. Sensitivity of the eigenvalues

In the dynamical problems of structural mechanics the finite element method lead to the following system of differential equations

$$M\ddot{u} + Ku = 0 \tag{192}$$

If we assume that the solution is in the following form

$$u = u_0 \sin(\omega t + \phi) \tag{193}$$

then

$$\dot{u} = \omega u_0 \cos(\omega t + \phi), \quad \ddot{u} = -\omega^2 u_0 \sin(\omega t + \phi) \tag{194}$$

and from the equation (192) we have

$$-M\omega^2 u_0 \sin(\omega t + \phi) + K u_0 \sin(\omega t + \phi) = 0 \tag{195}$$

$$(K - \omega_j^2 M) u_j = 0 \tag{196}$$

Eigenvectors $u_1, ..., u_n$ are M-orthogonal

$$u_i^T M u_j = \delta_{ij} \tag{197}$$

then from the equation (196)

$$u_i^T K u_j = \omega_j^2 \delta_{ij} \tag{198}$$

Sensitivity with the respect to the parameter p

$$\left(\frac{\partial K}{\partial p} - \frac{\partial \omega_j^2}{\partial p} M - \omega_j^2 \frac{\partial M}{\partial p} \right) u_j + (K - \omega_j^2 M) \frac{\partial u_j}{\partial p} = 0 \tag{199}$$

Lets multiply above equation by u_i^T

$$u_i^T \left(\frac{\partial K}{\partial p} - \frac{\partial \omega_j^2}{\partial p} M - \omega_j^2 \frac{\partial M}{\partial p} \right) u_j = 0 \tag{200}$$

$$u_i^T \left(\frac{\partial K}{\partial p} - \omega_j^2 \frac{\partial M}{\partial p} \right) u_j = \frac{\partial \omega_j^2}{\partial p} u_i^T M u_j \tag{201}$$

$$u_i^T \left(\frac{\partial K}{\partial p} - \omega_j^2 \frac{\partial M}{\partial p} \right) u_j = \frac{\partial \omega_j^2}{\partial p} \delta_{ij} \tag{202}$$

Then sensitivity of the frequency of vibration ω_j^2 can be calculated from the following formula (Lund, 1994; Hilbert and Courant, 1953)

$$\frac{\partial \omega_j^2}{\partial p} = u_j^T \left(\frac{\partial K}{\partial p} - \omega_j^2 \frac{\partial M}{\partial p} \right) u_j \tag{203}$$

The interval frequency of vibration can be calculated using sensitivity analysis and derivative $\frac{\partial \omega_j}{\partial p}$.

$$\frac{\partial \omega_j^2}{\partial p} = 2\omega_j \frac{\partial \omega_j}{\partial p} \tag{204}$$

$$\frac{\partial \omega_j}{\partial p} = \frac{1}{2\omega_j} \frac{\partial \omega_j^2}{\partial p} = \frac{1}{2\omega_j} u_j^T \left(\frac{\partial K}{\partial p} - \omega_j^2 \frac{\partial M}{\partial p} \right) u_j \tag{205}$$

If the derivative of stiffness matrix $\frac{\partial K}{\partial p}$ and the mass matrix $\frac{\partial M}{\partial p}$ are constant then the sign of the derivative $\frac{\partial \omega_j^2}{\partial p}$ is constant and extreme values of ω^2 can be calculated by using sensitivity analysis. Let us consider the system of first order differential equation in the matrix form

$$\dot{x} = Ax \tag{206}$$

If we assume that the solution has the following form $x = x_0 e^{\lambda t}$, $x = \lambda x_0 e^{\lambda t}$ then

$$\lambda x_0 e^{\lambda t} = A x_0 e^{\lambda t}, \quad \Rightarrow \quad (A - \lambda I) x_0 = 0 \tag{207}$$

Then we have the standard eigenvalue problem. Derivative with respect of parameter p is equal to the following

$$\left(\frac{\partial A}{\partial p} - \frac{\partial \lambda_j}{\partial p} I \right) x_j + (A - \lambda_j I) \frac{\partial x_j}{\partial p} = 0 \tag{208}$$

$$x_i^T \left(\frac{\partial A}{\partial p} - \frac{\partial \lambda_j}{\partial p} I \right) x_j = 0, \quad \Rightarrow \quad \frac{\partial \lambda_j}{\partial p} x_i^T x_j = x_i^T \frac{\partial A}{\partial p} x_j \tag{209}$$

Finally derivative of the eigenvalue can be calculated from the following formula

$$\frac{\partial \lambda_j}{\partial p} = x_j^T \frac{\partial A}{\partial p} x_j \tag{210}$$

Now it is possible to apply sensitivity analysis method in order to calculate upper and lower bound of the eigenvalue λ_j.
If the derivative of the matrix A i.e. $\frac{\partial A}{\partial p}$ is constant then the sign of the derivative $\frac{\partial \lambda}{\partial p}$ is constant and extreme values of λ can be calculated by using sensitivity analysis.
Different method which is based on perturbation of positive definite matrices is described in the paper (Modares, Mullen and Muhanna, 2006).

10.2. Vibration of multibody system

Dynamics of the mechanical system, which is shown in the Fig. 10 is described by the following system of differential equation

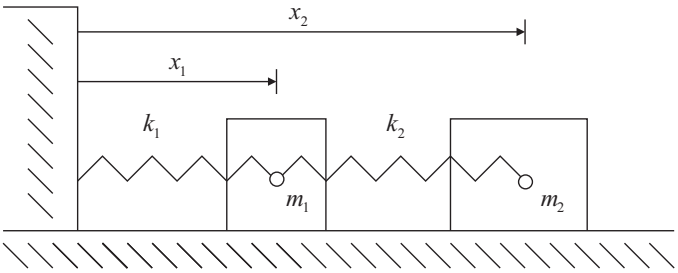


Figure 10. Multibody system

$$\begin{bmatrix} m & 0 \\ 0 & m \end{bmatrix} \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{bmatrix} + \begin{bmatrix} 2k & -k \\ -k & k \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{211}$$

or shortly

$$M\ddot{x} + Kx = 0 \tag{212}$$

where $k = k_1 = k2$ and $m = m_1 = m_2$. The eigenvalue problem

$$det\left(K - \omega^2 M\right) = 0 \tag{213}$$

has the following solution

$$\omega_1 = \sqrt{(3-\sqrt{5})\frac{k}{2m}}, \quad \omega_2 = \sqrt{(3+\sqrt{5})\frac{k}{2m}} \tag{214}$$

The eigenvectors x_1, x_2 satisfy the following system of linear equations

$$(K - \omega_1^2 M)x_1 = 0 \tag{215}$$

$$x_1 = \begin{bmatrix} \frac{\sqrt{5}-1}{\sqrt{2(5-\sqrt{5})m}} \\ \frac{\sqrt{2}}{\sqrt{(5-\sqrt{5})m}} \end{bmatrix} \tag{216}$$

$$(K - \omega_2^2 M)x_2 = 0 \tag{217}$$

$$x_2 = \begin{bmatrix} -\frac{\sqrt{5}+1}{\sqrt{2(5+\sqrt{5})m}} \\ \frac{\sqrt{2}}{\sqrt{(5+\sqrt{5})m}} \end{bmatrix} \tag{218}$$

$$\frac{\partial \omega_1^2}{\partial m} = x_1^T \left(\frac{\partial K}{\partial m} - \omega_1^2 \frac{\partial M}{\partial m} \right) x_1 = -(3-\sqrt{5})\frac{k}{2m^2} < 0 \tag{219}$$

$$\frac{\partial \omega_2^2}{\partial m} = x_2^T \left(\frac{\partial K}{\partial m} - \omega_2^2 \frac{\partial M}{\partial m} \right) x_2 = -(3+\sqrt{5})\frac{k}{2m^2} < 0 \tag{220}$$

$$\frac{\partial \omega_1^2}{\partial k} = x_1^T \left(\frac{\partial K}{\partial k} - \omega_1^2 \frac{\partial M}{\partial k} \right) x_1 = (3-\sqrt{5})\frac{1}{2m} > 0 \tag{221}$$

$$\frac{\partial \omega_2^2}{\partial k} = x_2^T \left(\frac{\partial K}{\partial k} - \omega_2^2 \frac{\partial M}{\partial k} \right) x_2 = (3+\sqrt{5})\frac{1}{2m} > 0 \tag{222}$$

If we assume that the sign of the eigenvalue is constant, then extreme values of the eigenvalues can be calculated in the following way

$$\underline{\omega}_1 = \omega_1(\overline{m}, \underline{k}), \quad \overline{\omega}_1 = \omega_1(\underline{m}, \overline{k}) \tag{223}$$

$$\underline{\omega}_2 = \omega_2(\overline{m}, \underline{k}), \quad \overline{\omega}_2 = \omega_2(\underline{m}, \overline{k}) \tag{224}$$

where $m \in [\underline{m}, \overline{m}]$, $k \in [\underline{k}, \overline{k}]$.

11. Conclusions

Using functional derivative it is possible to check monotonicity of the function with uncertain functional parameters. If the function $u = u(p)$ is monotone then extreme values of the results can be calculated by using upper and lower bound of the functional intervals and sensitivity analysis (Neumaier and Pownuk, 2004; Pownuk, 2004). Sensitivity can be use as an extension of the existing FEM programs. Using quasi analytical method it is possible to avoid approximation errors. Functional derivative can be sometimes calculated without integration. This property may increase accuracy of the solution. Using the sensitivity analysis method it is possible to calculate the interval eigenvalues. Interval eigenvalues can be calculated also in the case of structures with uncertain shape and uncertain functional parameters.

Presented sensitivity analysis method can be applied to the solution of any problem with functional parameters in which it is possible to calculate the functional derivative and verify monotonicity. For non-monotone problems it is possible to apply an extension of the algorithm, which gives only inner bounds.

The approach presented can be applied together with any numerical method for the solution of the underlying problem, including techniques for partial differential equations e.g. FEM, FDM, BEM, FVM etc. Extended version of this paper was published as a research report at the web page of the University of Texas at El Paso (Pownuk, 2007).

References

M.W. Hirsch and H. Smith. Monotone maps: a review. *Journal of Difference Equations and Applications*, 11(4-5):379–398, 2005.

E. Lund. *Finite Element Based Design Sensitivity Analysis and Optimization*. Ph.D. Dissertation, Aalborg University, Aalborg, 1994.

M. Modares, R. L. Mullen, and R.L. Muhanna. Natural frequencies of a structure with bounded uncertainty. *Journal of Engineering Mechanics, ASCE*, 132, 2006.

D. Moens and D. Vandepitte. A survey of non-probabilistic uncertainty treatment in finite element analysis. *Computer Methods in Applied Mechanics and Engineering*, 194(14-16), 2005.

R.E. Moore. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, 1966.

A. Neumaier. *Interval Methods for Systems of Equations*. Cambridge Univ. Press, Cambridge, 1990.

A. Neumaier and A. Pownuk. Linear systems with large uncertainties with applications to truss structures. *Reliable Computing*, 13(2):149–172, 2007.

A. Pownuk. Numerical solutions of fuzzy partial differential equation and its application in computational mechanics. *Fuzzy Partial Differential Equations and Relational Equations: Reservoir Characterization and Modeling, (M. Nikravesh, L. Zadeh and V. Korotkikh, eds.), Studies in Fuzziness and Soft Computing, Physica-Verlag*, pages 308–347, 2004.

A. Pownuk. Numerical solution of fem equations with uncertain functional parameters. *FEMTEC 2006, The University of Texas at El Paso, December 11 - 15, El Paso, USA*, 2006.

A. Pownuk, General Interval FEM Program Based on Sensitivity Analysis *The University of Texas at El Paso,* Department of Mathematical Sciences Research Reports Series Texas Research Report No. 2007-06, El Paso, Texas, USA, http://www.math.utep.edu/preprints/2007/2007-06.pdf

D. Hilbert R. Courant. *Methods of Mathematical Physics, Vol. 1*. Interscience Publishers, New York, 1953.

O.C. Zienkiewicz and R.L. Taylor. *The Finite Element Method. Fifth edition. Volume 1: Basis*, volume 1. Butterworth Heinemann, London, 5 edition, 2000.

Worst Case Bounds on the Point-wise Discretization Error in Boundary Element Method for Elasticity Problem

B. F. Zalewski and R. L. Mullen
Department of Civil Engineering
Case Western Reserve University
Cleveland, OH 44106
e-mail: bxz10@case.edu; rlm@case.edu

Abstract: In this work, local discretization error is bounded via interval approach for the elasticity problem using the interval boundary element formulation. The formulation allows for computation of the worst case bounds on the errors in the solution of elasticity problem. From these bounds the worst case bounds on the discretization error of any point in the domain of the boundary can be computed. Examples are presented to demonstrate the effectiveness of the treatment of local discretization error in elasticity problem via interval methods.

Keywords: interval boundary element method, interval analysis, discretization error, elastostatics

1. Introduction

Most of the problems in engineering mechanics are governed by partial differential equations, to which solutions, in general, cannot be obtained exactly due to complexities in the geometry of the system for which the applied boundary conditions must be satisfied. Therefore, numerical methods have been developed to approximate the true solution by a polynomial interpolation between discrete values. The foremost method is the finite element method (FEM), in which the domain of the system is discretized into elements consisting of polynomial interpolation functions between discrete values which are to be computed. Another numerical method used to approximate the solutions to partial differential equations is the boundary element method (BEM). In boundary element analysis (BEA), the domain variables are transformed to the boundary variables, thus decreasing the dimension of the problem by one. This allows, in general, decreasing the time necessary for mesh generation or mesh refinement. The domain transformation is performed by the use of fundamental solutions to the linear partial differential equations, thus restricting classical BEM to problems for which the fundamental solution is known. The boundary integral equations, resulting from weighted residual formulation, are solved using point collocation methods, in which the residual is set to zero in the domain and exists only
on the boundary of the system. To achieve such residual, the weighted residual function in a weak formulation of the partial differential equation, takes the form of the fundamental solution. The transformed boundary integral equations are then solved by approximating the true solution over discrete

boundaries, thus introducing the discretization error. Although discretization error estimates have been made for BEM (Rencis and Jong 1989) the worst case bounds on the local discretization error have been computed only for the Laplace problem (Zalewski and Mullen 2007).

In this work the point-wise discretization error is studied for the elasticity problem. The boundary integral equations are bounded by interval boundary integral equations, eventually resulting in interval linear system of equations. A parametric solver is reviewed that enables the computation of non-naive bounds. Example problems are presented to illustrate the behavior of the discretization error bounds.

2. Boundary Element Analysis of Elasticity Problem

2.1. BEA FORMULATION FOR ELASTICITY PROBLEM

The boundary element formulation for the behavior of an isotropic and homogeneous body is discussed in the literature (Brebbia 1992, Hartmann 1889, Pilkey and Wunderlich 1994). The following section reviews the two dimensional boundary element formulation for the elasticity problem. The elasticity problem is:

$$\left.\begin{array}{c} \sigma_{ij,j} + b_i = 0 \quad in\ \Omega \\ u_i = \hat{u}_i \quad on\ \Gamma_1, \quad t_i = \hat{t}_i \quad on\ \Gamma_2 \\ \bigcup_{i=1}^{2} \Gamma_i = \Gamma \quad and \quad \bigcap_{i=1}^{2} \Gamma_i = 0 \end{array}\right\} \tag{1}$$

where Ω is the domain of the system, Γ is the boundary of the system, σ_{ij} is the stress tensor, b_i is the vector of body force, u_i is the displacement vector with a forced boundary condition $\hat{u}_i$ on Γ_1, and t_i is the traction vector with a natural boundary condition $\hat{t}_i$ on Γ_2. The first step in approximating the solution to Eq. (1) is to express it in a weighted residual form or a weak form:

$$\int_{\Omega} \left(\sigma_{ij,j} + b_i\right) u_i^* d\Omega = \int_{\Gamma_2} (t_i - \hat{t}_i) u_i^* d\Gamma_2 - \int_{\Gamma_1} (u_i - \hat{u}_i) t_i^* d\Gamma_1 \tag{2}$$

where u_i^* and t_i^* are the weighted residual functions. In the following steps Betti's reciprocal theorem is reviewed and used to formulate boundary integral equations. Expanding the left side of Eq. (2) results in:

$$\int_\Omega \left(\sigma_{ij,j} + b_i\right) u_i^* d\Omega = \int_\Omega \sigma_{ij,j} u_i^* d\Omega + \int_\Omega b_i u_i^* d\Omega = 0 \tag{3}$$

Applying the chain rule to the first integral on the right side of Eq. (3) yields:

$$\int_\Omega \sigma_{ij,j} u_i^* d\Omega = \int_\Omega \left(\sigma_{ij} u_i^*\right)_{,j} d\Omega - \int_\Omega \sigma_{ij} u_{i,j}^* d\Omega \tag{4}$$

Substituting $\varepsilon_{ij}^* = u_{i,j}^*$ in Eq. (4) results in:

$$\int_\Omega \sigma_{ij,j} u_i^* d\Omega = \int_\Omega \left(\sigma_{ij} u_i^*\right)_{,j} d\Omega - \int_\Omega \sigma_{ij} \varepsilon_{ij}^* d\Omega \tag{5}$$

where ε_{ij} is the linear strain tensor. Applying Gauss integral theorem to the first integral on the right side of Eq. (5):

$$\int_\Omega \left(\sigma_{ij} u_i^*\right)_{,j} d\Omega = \int_\Gamma \sigma_{ij} u_i^* n_j d\Gamma = \int_\Gamma \sigma_{ij} n_j u_i^* d\Gamma = \int_\Gamma t_i u_i^* d\Gamma \tag{6}$$

Substituting the result of Eq. (6) into Eq. (5) and rearranging terms yields:

$$\int_\Omega \sigma_{ij} \varepsilon_{ij}^* d\Omega + \int_\Omega \sigma_{ij,j} u_i^* d\Omega = \int_\Gamma t_i u_i^* d\Gamma \tag{7}$$

The equilibrium condition, $\sigma_{ij,j} = -b_i$, is substituted into Eq. (7) to obtain:

$$\int_\Omega \sigma_{ij} \varepsilon_{ij}^* d\Omega - \int_\Gamma b_i u_i^* d\Gamma = \int_\Gamma t_i u_i^* d\Gamma \tag{8}$$

Following the same procedure, Eq. (3) through Eq. (8), the following equation can be obtained:

$$\int_\Omega \sigma_{ij}^* \varepsilon_{ij} d\Omega - \int_\Gamma b_i^* u_i d\Gamma = \int_\Gamma t_i^* u_i d\Gamma \tag{9}$$

It is then considered that the body follows the linear elastic constitutive model:

$$\sigma_{ij} = E_{ijkl}\varepsilon_{kl} \tag{10}$$

where E_{ijkl} is the fourth order linear elasticity tensor. Eq. (10) can also be written as:

$$\sigma_{ij} = \frac{E}{1+\nu}\varepsilon_{ij} + \frac{\nu E}{(1+\nu)(1-2\nu)}\delta_{ij}\varepsilon_{kk} \tag{11}$$

Also by expansion of σ_{ij} tensor and symmetry of E_{ijkl} tensor with respect to i, j and k, l indices:

$$\sigma_{ij}\varepsilon^*_{ij} = E_{ijkl}\varepsilon_{kl}\varepsilon^*_{ij} = E_{klij}\varepsilon_{ij}\varepsilon^*_{kl} = E_{klij}\varepsilon^*_{kl}\varepsilon_{ij} = E_{ijkl}\varepsilon^*_{kl}\varepsilon_{ij} = \sigma^*_{ij}\varepsilon_{ij} \tag{12}$$

By equating the first integral terms in Eq. (8) and Eq. (9) due to Eq. (12), Betti's reciprocal theorem can be obtained:

$$\int_\Gamma t_i u^*_i d\Gamma + \int_\Gamma b_i u^*_i d\Gamma = \int_\Gamma t^*_i u_i d\Gamma + \int_\Gamma b^*_i u_i d\Gamma \tag{13}$$

Eq. (13) is the starting point of the boundary element formulation for the elasticity problem. Equilibrium equation $\sigma^*_{ij,j} = -b^*_i$ is substituted into Eq. (13) resulting in:

$$-\int_\Gamma \sigma^*_{ij,j} u_i d\Gamma + \int_\Gamma t^*_i u_i d\Gamma = \int_\Gamma u^*_i b_i d\Gamma + \int_\Gamma u^*_i t_i d\Gamma \tag{14}$$

In order to decrease the dimension of the integral equation, Eq. (14), the weighted residual function is set to be the Green's function, which is obtained by applying a point load in direction a_i. This can be expressed as:

$$\sigma^*_{ij,j} = -\delta(x-\xi)a_i \tag{15}$$

where ξ is a source point at which a concentrated force is applied, x is a field point at which a response to the concentrated force is observed, and $\delta(x-\xi)$ is the Dirac delta function. The resulting fundamental solution is:

$$u^*_i = u^*_{ji}a_j \tag{16}$$

$$t_i^* = t_{ji}^* a_j \tag{17}$$

where u_{ji}^* and t_{ji}^* are i components of the displacements and tractions, respectively, due to a concentrated force in the j direction, and a_j is a unit vector in the direction of the applied concentrated force. The kernel functions u_{ji}^* and t_{ji}^* are given as:

$$u_{ij}^* = \frac{1}{8\pi(1-\nu)G}\left[(4\nu-3)\ln(r)\delta_{ij} + \frac{(\vec{x}-\vec{\xi})\bullet\vec{i}}{r}\cdot\frac{(\vec{x}-\vec{\xi})\bullet\vec{j}}{r}\right] \tag{18}$$

$$q_{ij}^* = \frac{-1}{4\pi(1-\nu)r}\left\{\begin{array}{c}\left[(1-2\nu)\delta_{ij} + 2\frac{(\vec{x}-\vec{\xi})\bullet\vec{i}}{r}\cdot\frac{(\vec{x}-\vec{\xi})\bullet\vec{j}}{r}\right]\cdot\frac{(\vec{x}-\vec{\xi})\bullet\vec{n}}{r} \\ -(1-2\nu)\left[\frac{(\vec{x}-\vec{\xi})\bullet\vec{i}}{r}n_y - \frac{(\vec{x}-\vec{\xi})\bullet\vec{j}}{r}n_x\right]\end{array}\right\} \tag{19}$$

Substituting Eq. (15), Eq. (16), and Eq. (17) into Eq. (14) yields:

$$u_i(\xi)a_i + \int_\Gamma t_{ji}^* a_j u_i d\Gamma = \int_\Gamma u_{ji}^* a_j b_i d\Gamma + \int_\Gamma u_{ji}^* a_j t_i d\Gamma, \quad \xi \in \Omega \tag{20}$$

The indices are exchanged in all the integral terms in Eq. (20) as:

$$u_i(\xi)a_i + \int_\Gamma t_{ij}^* a_i u_j d\Gamma = \int_\Gamma u_{ij}^* a_i b_j d\Gamma + \int_\Gamma u_{ij}^* a_i t_j d\Gamma, \quad \xi \in \Omega \tag{21}$$

The a_i coefficients are constant and can be canceled out from Eq. (21):

$$u_i(\xi) + \int_\Gamma t_{ij}^* u_j d\Gamma = \int_\Gamma u_{ij}^* b_j d\Gamma + \int_\Gamma u_{ij}^* t_j d\Gamma, \quad \xi \in \Omega \tag{22}$$

Assuming that the body force is zero, Eq. (22) can be simplified to:

$$u_i(\xi) + \int_\Gamma t_{ij}^* u_j d\Gamma = \int_\Gamma u_{ij}^* t_j d\Gamma, \quad \xi \in \Omega \tag{23}$$

Eq. (23) is integrated such that the source point, ξ, is included on the circular boundary of radius ε, as $\varepsilon \to 0$. This results in the right side integral vanishing. For constant elements the left side integral results in $-1/2u_i(\xi)$. Thus on the boundary of the system, Eq. (23) can be rewritten as:

$$\frac{1}{2}u_i(\xi) + \int_\Gamma t_{ij}^*(x,\xi)u_j(x)d\Gamma = \int_\Gamma u_{ij}^*(x,\xi)t_j(x)d\Gamma, \quad \xi \in \Gamma \tag{24}$$

In most cases, the exact solution to Eq. (24) cannot be found. Therefore Eq. (24) can be approximately solved using numerical methods such as BEM.

2.2. BOUNDARY DISCRETIZATION USING CONSTANT ELEMENT

In general, boundary integral equations, such as Eq. (24), cannot be solved analytically. To obtain approximate solutions, the boundary integral equation is discretized into boundary elements for which the true solution is approximated by a polynomial interpolation between known values of either u or t. In this work, only boundary elements with constant shape functions are used to generate significant discretization errors. Higher order approximation is assumed to approximate the true solutions better thus decreasing the discretization error. Constant elements contain one node per element, leading to the following discretization:

$$u(x) = \Phi u_i \tag{25}$$

$$t(x) = \Phi t_i \tag{26}$$

where u_i and t_i are the vectors of nodal values of u or t, respectively, at node i, and Φ is the vector of constant shape functions. The discretized Eq. (24) can be written as:

$$\frac{1}{2}u_i + \sum_{Elements}\int_{\Gamma_x} t_{ij}^*(x,\xi)\Phi d\Gamma_x u_j = \sum_{Elements}\int_{\Gamma_x} u_{ij}^*(x,\xi)\Phi d\Gamma_x t_j \tag{27}$$

Eq. (27) can be written in a matrix form:

$$Hu = Gt \tag{28}$$

where matrix H is singular and therefore satisfies the rigid body motion. To obtain a unique solution to Eq. (28) at least one boundary condition in each direction of the problem must be specified for the displacement. Eq. (28) is then rearranged according to the appropriate boundary conditions and solved as a linear algebra problem:

$$Ax = f \tag{29}$$

The terms of H and G matrices can either be determined explicitly or are computed numerically using numerical integration schemes. The effects of the integration error and truncation error have been studied (Zalewski et al. 2007) and can be implemented to enclose the true solution of Eq. (29). In this work the impact of the discretization error on the solution to Eq. (24) is studied, following the boundary element formulation, using interval methods.

3. Interval Analysis

In this work, the discretization error in BEM is treated using an interval approach. The following is a review of interval analysis (Moore 1966, Neumaier 1990). An interval number $\tilde{x} = [a, b]$ is a set of real numbers such that:

$$[a, b] = \{x \mid a \le x \le b\} \tag{30}$$

where $(a, b) \in \Re$. Interval variables $\tilde{x} = [a, b]$ and $\tilde{y} = [c, d]$ behave according to the following operations:

Addition:

$$\tilde{x} + \tilde{y} = [a + c, b + d] \tag{31}$$

Subtraction:

$$\tilde{x} - \tilde{y} = [a - d, b - c] \tag{32}$$

Multiplication:

$$\tilde{x} \cdot \tilde{y} = [\min\{ac, ad, bc, bd\}, \max\{ac, ad, bc, bd\}] \tag{33}$$

Division:

$$\frac{\tilde{x}}{\tilde{y}} = [a,b]\cdot\left[\frac{1}{d},\frac{1}{c}\right],\ 0 \notin \tilde{y} \tag{34}$$

Integration of interval-valued function $f(x,\tilde{\xi})$, which is a class of all possible functions bounded by a given interval is performed as:

$$\int_{\Gamma} f(x,\tilde{\xi})d\Gamma = \left[\int_{\Gamma} \underline{f}(x,\tilde{\xi})d\Gamma, \int_{\Gamma} \bar{f}(x,\tilde{\xi})d\Gamma\right],\ \ \xi \in [\underline{\xi},\bar{\xi}] \tag{35}$$

Subdistributive property:

$$\tilde{x}\cdot(\tilde{y}+\tilde{z}) \subseteq \tilde{x}\cdot\tilde{y} + \tilde{x}\cdot\tilde{z} \tag{36}$$

One of the major sources of overestimation or underestimation in interval solutions is the subdistributive property of interval numbers. Great emphasis should be made to the correct order of operations in interval analysis. If the correct representation is given by the left term in Eq. (36), expressing the operation by the right term may cause overestimation. If the correct representation is expressed as the right term in Eq. (36), expressing it as the left term may result in inner bounds and the enclosure of the solution may not be guaranteed. This issue will be farther referred to in considering interval kernel functions.

Another source of overestimation occurs due to the dependency of interval numbers, either linear or nonlinear. Linear dependency of interval numbers for $\tilde{x} = [-1,1]$ and $\tilde{y} = [-1,1]$ can be illustrated as:

$$\tilde{x}\cdot\tilde{y} = [-1,1] \tag{37}$$

$$\tilde{x}\cdot\tilde{x} = [0,1] \tag{38}$$

Eq. (37) considers the two sets to be independent; therefore, the operation must enclose all possible values. Eq. (38) takes into account that the same set is multiplied by itself; therefore, every number in set $\tilde{x}$ is multiplied by itself. For engineering problems interval dependency occurs mostly due to the physics of the problem and needs to be considered for sharp solutions. Naive interval application may results in wide and unrealistic bounds. Considering an example:

$\tilde{y} = 6\cdot\tilde{x}\cdot\tilde{x} + 3\cdot\tilde{x},\ \ \tilde{x} = [-1,1]$

direct interval operation results in naive bounds for the solution, $\tilde{y} = [-9,9]$. However, considering interval dependency, the bounds on the solution result in exact bounds, $\tilde{y} = [-0.375,9]$.

Another source of overestimation is the order of operations in interval linear algebra. To obtain sharp results, interval operations should be performed last to reduce the overestimation due to the dependency in interval matrix coefficients. The following example demonstrates this consideration.

$\tilde{y}_1 = A \cdot (B \cdot \tilde{x}),\ \ \tilde{y}_2 = (A \cdot B) \cdot \tilde{x}$, where

$$A = \begin{bmatrix} \mathrm{a}_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},\ B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix},\ \tilde{x} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix}.$$

$$\tilde{y}_1 = \begin{bmatrix} a_{11}(b_{11}\tilde{x}_1 + b_{12}\tilde{x}_2) + a_{12}(b_{21}\tilde{x}_1 + b_{22}\tilde{x}_2) \\ a_{21}(b_{11}\tilde{x}_1 + b_{12}\tilde{x}_2) + a_{22}(b_{21}\tilde{x}_1 + b_{22}\tilde{x}_2) \end{bmatrix},\ \tilde{y}_2 = \begin{bmatrix} (a_{11}b_{11} + a_{12}b_{21})\tilde{x}_1 + (a_{11}b_{12} + a_{12}b_{22})\tilde{x}_2 \\ (a_{21}b_{11} + a_{22}b_{21})\tilde{x}_1 + (a_{21}b_{12} + a_{22}b_{22})\tilde{x}_2 \end{bmatrix}$$

It can be clearly seen that $\tilde{y}_2$ is sharper then $\tilde{y}_1$ due to the considered dependency of $\tilde{x}_1$ and $\tilde{x}_2$ throughout the rows of $\tilde{y}_2$. Therefore special care should be given to the order of interval operations to obtain sharp bounds on the solution.

4. Interval Linear System of Equations

The interval linear system of equations of the form of Eq. (29) is solved using Krawczyk iteration (Krawczyk 1969) based on Brouwer's fixed point theorem (Mullen and Muhanna 1999, Muhanna and Mullen 2001, Muhanna et al. 2005). One approach of self-validating (SV) methods to find the zero of the function $f(x) = 0,\ \Re^n \to \Re^n$ is to consider a fixed point function $g(x) = x$. The transformation between $f(x)$ and $g(x)$ for a non-singular preconditioning matrix C is:

$$f(x) = 0 \Leftrightarrow g(x) = x \tag{39}$$

$$g(x) = x - C \cdot f(x) \tag{40}$$

where the function $g(x)$ is considered as a Newton operator. From Brouwer's fixed point theorem and from:

$$g(\tilde{x}) \subseteq \tilde{x} \quad for\ some\ \tilde{x} \in \Re^n \tag{41}$$

the following is true:

$$\exists\ x \in \tilde{x}: \quad f(x) = 0 \tag{42}$$

This method is used to solve linear system of equations of the form of Eq. (29). The preconditioning matrix C is chosen as $C = A^{-1}$. From Eq. (40) and Eq. (41) it follows that:

$$Cb + (I - CA)\tilde{x} \subseteq \tilde{x} \tag{43}$$

The left hand side of Eq. (43) is the Krawczyk operator (Krawczyk 1969). For the iteration to provide finite solution, the preconditioning matrix needs to be proven regular (Neumaier 1990, Rump 2001). The following proves this condition.

Theorem 1. (Rump 2001) *Let* $A, C \in \Re^{n\times n}, b \in \Re^n,$ *and* $\tilde{x} \in \Re^n$ *be given*. If

$$Cb + (I - CA)\tilde{x} \subseteq \text{int}(\tilde{x}) \tag{44}$$

then C *and* A *are regular and the unique solution of* $Ax = b$ *satisfies* $A^{-1}b \in \tilde{x}$.

$\text{int}(\tilde{x})$ refers to the interior of $\tilde{x}$. However, all terms in Eq. (29) can be interval terms, thus the following is a proof for the guarantee of the solution for the equation of this form.

Theorem 2. (Rump 2001) *Let* $\tilde{A} \in \Re^{n\times n}, C \in \Re^{n\times n}, \tilde{b} \in \Re^n,$ *and* $\tilde{x} \in \Re^n$ *be given*. If

$$C\tilde{b} + (I - C\tilde{A})\tilde{x} \subseteq \text{int}(\tilde{x}) \tag{45}$$

then C *and every matrix* $A \in \tilde{A}$ *is regular and*

$$\sum(\tilde{A}, \tilde{b}) = \{x \in \Re^n \mid \exists\ A \in \tilde{A}\ \exists\ b \in \tilde{b} : Ax = b\} \subseteq \tilde{x} \tag{46}$$

Eq. (46) guarantees the solution to the interval linear system of equations of the form of Eq. (29). The residual form of Eq. (46) is (Neumaier 1990):

$$C\tilde{b} - C\tilde{A}x_0 + (I - C\tilde{A})\tilde{\delta} \subseteq \text{int}(\tilde{\delta}) \tag{47}$$

where $\tilde{x} = x_0 + \tilde{\delta}$. A good initial guess is $x_0 = C\hat{b}$, where $C = \hat{A}^{-1}$, $\hat{A}$ is the midpoint matrix of A, and $\hat{b}$ is the midpoint vector of b. The following sections describe the treatment of point-wise discretization error via interval methods.

5. Discretization Error Bounds for Boundary Element Method

The discretization error in the solutions to integral equations results from considering a finite number of collocation points for which these solutions are computed. In general, the true solutions to integral equations are functions, not discrete values, and therefore the space of the approximate solutions does not cover the space of the true solutions. The boundary integral equations can be obtained by the use of collocation methods resulting in equation of the form of Eq. (24). The boundary integral equations are satisfied exactly only if all the locations of the source point ξ on the boundary are considered. However, to obtain a linear system of equations, a finite number of source points are considered. Moreover, the location of the source points is unique and the solution is considered as a polynomial interpolation between discrete values, whose location corresponds to the location of the source point. This allows for the solution of the linear system of equations to be unique and thus the system can be solved for the unknown boundary values. It should be noted that if al non countable source points are considered, the boundary values at all points can be computed, resulting in the true solution. The boundary integral equation can also be evaluated over n sub-domains as expressed by Eq. (27). The unique location of the source point and its correspondence to the point at which the approximate solution is computed must be satisfied for all sub-domains. Eq. (27) is satisfied exactly only if all the locations of the source point are considered. Thus the discretization error is introduced in the same manner as in Eq. (24).

In the analysis of the discretization error, all the locations of the source point, $\tilde{\xi}$, in the continuous boundary integral equation:

$$\frac{1}{2}u_i(\xi) + \int_\Gamma t_{ij}^*(x,\xi)u_j(x)d\Gamma = \int_\Gamma u_{ij}^*(x,\xi)t_j(x)d\Gamma, \quad \xi \in \Gamma \tag{48}$$

are treated via interval approach. Considering interval bounds $\tilde{\xi}$ on all the possible locations of the source points ξ allows obtaining an interval solution which bounds the true solution. From the interval bounds on the boundary values, the bounds on the true solution for any point in the domain can be computed. Eq. (48) is bounded by an interval boundary integral equation in which the terms $u_{ij}^{*}(x,\xi)$ and $t_{ij}^{*}(x,\xi)$ are known interval-valued functions. The unknown functions $u_j(x)$ and $t_j(x)$ in Eq. (48) are then bounded by interval values enclosing the true solution.

The integral over the domain can be expressed as the sum of the integrals over the elements and thus the boundary integral equation must be bounded on each element for all the locations of the source points. Hence, for the boundary Γ subdivided into n boundary elements, for each element k the interval values $\tilde{u}$ and $\tilde{t}$ that bound the functions $u(x)$ and $t(x)$ are found (Figure 1).

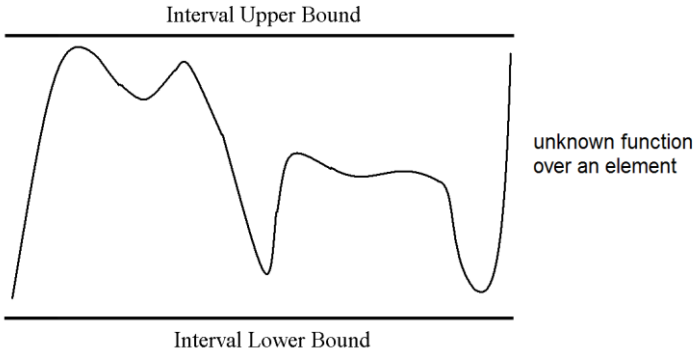


Figure 1. Constant interval bounds on a function.

For higher order elements the interval valued function, of the order of the polynomial approximation, encloses the true solution. The bounding of the function using linear elements is shown (Figure 2).

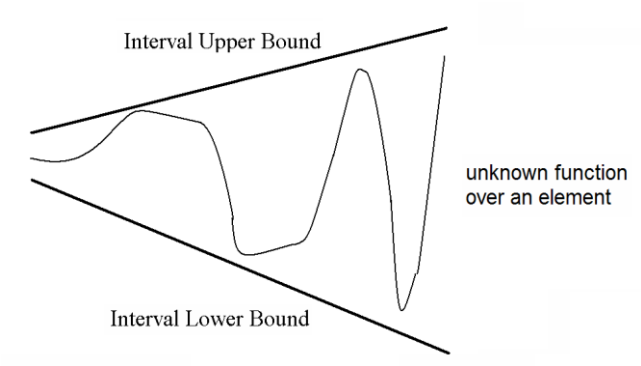


Figure 2. Linear interval bounds on a function.

It is assumed that on all other elements, except for the element in consideration, the bounds on all boundary values are known. Also either the bounds on the Dirichlet or the Neumann boundary condition bounds are known for the element in consideration. Then the remaining boundary value for the single element in consideration is bounded. The process is repeated for the second element with the assumed bounds for all the other elements, a computed bound for the previously considered element, and either the Dirichlet or the Neumann boundary condition bounds for the second element in consideration. This procedure, known as the interval Gauss-Seidel iteration (Neumaier 1990), is performed for all elements until the true solution is enclosed. Mathematically the above statement can be expressed as:

$\forall k \in \{1,2,...,n\}$ *Assume* $\underline{u_m} \le u_m \le \overline{u_m}$, $\underline{t_m} \le t_m \le \overline{t_m}$ *is known* $\forall m \neq k$.

Also known $\underline{t_k} \le t_k \le \overline{t_k}$. *Find* $\underline{u_k} \le u_k \le \overline{u_k}$

$$\forall \xi_k \left| \begin{array}{l} \frac{1}{2}u_{ik}(\xi_k) + \int\limits_{\Gamma_k} t^*_{ij}(x,\xi_k)u_{jk}(x)d\Gamma_k = \\ \sum\limits_{m=1}^{n} \int\limits_{\Gamma_m} u^*_{ij}(x,\xi_k)t_{jm}(x)d\Gamma_m + \int\limits_{\Gamma_k} u^*_{ij}(x,\xi_k)t_{jk}(x)d\Gamma_k - \sum\limits_{m=1}^{n} \int\limits_{\Gamma_m} t^*_{ij}(x,\xi_k)u_{jm}(x)d\Gamma_m \end{array} \right.$$

Or

$\forall k \in \{1,2,...,n\}$ *Assume* $\underline{u_m} \le u_m \le \overline{u_m}$, $\underline{t_m} \le t_m \le \overline{t_m}$ *is known* $\forall m \neq k$.

Also known $\underline{u_k} \le u_k \le \overline{u_k}$. *Find* $\underline{t_k} \le t_k \le \overline{t_k}$

$$\forall \xi_k \left| \begin{array}{l} \int\limits_{\Gamma_k} u^*_{ij}(x,\xi_k)t_{jk}(x)d\Gamma_k = \frac{1}{2}u_{ik}(\xi_k) + \int\limits_{\Gamma_k} t^*_{ij}(x,\xi_k)u_{jk}(x)d\Gamma_k + \\ \sum\limits_{m=1}^{n} \int\limits_{\Gamma_m} t^*_{ij}(x,\xi_k)u_{jm}(x)d\Gamma_m - \sum\limits_{m=1}^{n} \int\limits_{\Gamma_m} u^*_{ij}(x,\xi_k)t_{jm}(x)d\Gamma_m \end{array} \right. \quad (49)$$

Each term of the summation in Eq. (49) is represented graphically (Figure 3).

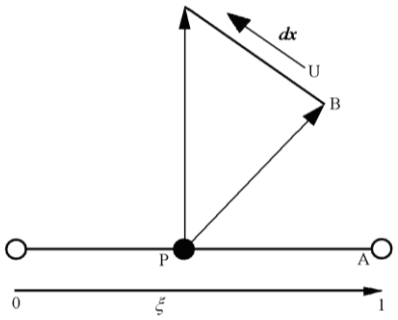


Figure 3. Integration from element B from point P on element A.

If u or q are specified boundary conditions, the interval integration can be performed explicitly as described in section 3, Eq. (35). In this work, for computational efficiency purposes, the underlying system of interval equations is solved using Krawczyk iteration (Krawczyk 1969), rather than using the interval Gauss-Seidel iteration (Neumaier 1990). This substitution of the method for bounding the unknown boundary values can be made since both of these methods are iterative methods for solving

interval linear systems of equations and both obtain guaranteed bounds for the solution. Hence, the interval boundary element method (IBEM) formulation is performed such that the resulting interval linear system of equations is of the form of Eq. (29).

6. Interval Kernel Splitting Technique

The analysis of the discretization error requires that the boundary integral equations for each element be bounded for all the locations of the source point ξ. The integral equation in the boundary element formulation has the form of the Fredholm equation of the first kind. Kernel splitting techniques have been used to bound the interval Fredholm equation of the first kind in which the right side is deterministic (Dobner 2002) as:

$$\int_{\Gamma} \tilde{a}(x,\xi)u(x)d\Gamma = b(\xi) \tag{50}$$

However, the interval boundary integral equations considered herein have an interval right side, due to the interval valued location of the source point $\tilde{\xi}$, therefore a new Interval Kernel Splitting Technique (IKST) is developed. The integral of the product of two functions is bounded considering interval bounds on the unknown value as:

$$\int_{\Gamma} a(x,\tilde{\xi})\tilde{u}d\Gamma \supseteq \int_{\Gamma} a(x,\tilde{\xi})u(x)d\Gamma = b(\tilde{\xi}) \tag{51}$$

To separate the kernels such that the unknown $\tilde{u}$ can be taken out of the integral on Γ, the left side integral from Eq. (51) is expressed as a sum of the integrals:

$$\int_{\Gamma} a(x,\tilde{\xi})\tilde{u}d\Gamma = \int_{\Gamma_1} a(x,\tilde{\xi})\tilde{u}d\Gamma_1 + \int_{\Gamma_2} a(x,\tilde{\xi})\tilde{u}d\Gamma_2 \tag{52}$$

where $\Gamma_1 \bigcup \Gamma_2 = \Gamma,\ \Gamma_1 \bigcap \Gamma_2 = 0$ and:

$$a(x,\tilde{\xi}) > 0 \ \ or \ \ a(x,\tilde{\xi}) < 0 \ \ on \ \ \Gamma_1 \tag{53}$$

$$a(x,\tilde{\xi}) \in 0 \ \ on \ \ \Gamma_2 \tag{54}$$

The interval kernel is of the same sign on Γ_1, thus $\tilde{u}$ can be directly taken out of the integral on Γ_1 as:

$$\int_{\Gamma_1} a(x,\tilde{\xi})\tilde{u}d\Gamma_1 = \int_{\Gamma_1} a(x,\tilde{\xi})d\Gamma_1\tilde{u} \tag{55}$$

Due to the subdistributive property of interval numbers, Eq. (36), $\tilde{u}$ cannot be taken out of the integral on Γ_2. The direct application of the subdistributive property may result in inner bounds on the interval integral as:

$$\int_{\Gamma_2} a(x,\tilde{\xi})d\Gamma_2\tilde{u} \subseteq \int_{\Gamma_2} a(x,\tilde{\xi})\tilde{u}d\Gamma_2 \tag{56}$$

Hence the interval kernel is bounded by its limits on Γ_2:

$$\int_{\Gamma_2} \tilde{a}\tilde{u}d\Gamma_2 \supseteq \int_{\Gamma_2} a(x,\tilde{\xi})\tilde{u}d\Gamma_2 \tag{57}$$

where $\tilde{a}$ is defined as:

$$\tilde{a} = [\min\{a(x+\tilde{\varepsilon},\tilde{\xi})\}, \max\{a(x+\tilde{\varepsilon},\tilde{\xi})\}] \tag{58}$$

$$\tilde{\varepsilon} = [-\varepsilon,\varepsilon] \tag{59}$$

ε is the tolerance level of the nonlinear solver used to find the zero location of $a(x,\tilde{\xi})$. To show that by bounding the kernel on Γ_2 allows $\tilde{u}$ to be taken out from the integral on Γ_2, the integral on Γ_2 is expressed as an infinite sum:

$$\int_{\Gamma_2} \tilde{a}\tilde{u}d\Gamma_2 = \lim_{\Delta\to 0}\sum_{i=1}^{n}(\Delta\tilde{a}\tilde{u})\Big|_{\Gamma_2} = \lim_{\Delta\to 0}(n\Delta\tilde{a}\tilde{u})\big|_{\Gamma_2} = \lim_{\Delta\to 0}(n\Delta\tilde{a})\tilde{u}\big|_{\Gamma_2} = \lim_{\Delta\to 0}\sum_{i=1}^{n}(\Delta\tilde{a})\Big|_{\Gamma_2}\tilde{u} = \int_{\Gamma_2}\tilde{a}d\Gamma_2\tilde{u} \tag{60}$$

where Δ is a small part of Γ_2. Thus $\tilde{u}$ can be taken out of both integrals on Γ_1 and on Γ_2 and the split interval boundary integral equation becomes:

$$\int_{\Gamma_1} a(x,\tilde{\xi})d\Gamma_1\tilde{u} + \int_{\Gamma_2} \tilde{a}d\Gamma_2\tilde{u} \supseteq \int_{\Gamma} a(x,\tilde{\xi})\tilde{u}d\Gamma \supseteq \int_{\Gamma} a(x,\tilde{\xi})u(x)d\Gamma = b(\tilde{\xi}) \tag{61}$$

The kernels are bounded for all the elements resulting in interval linear system of equations:

$$\tilde{A}_1\tilde{u} + \tilde{A}_2\tilde{u} \supseteq \tilde{b} \tag{62}$$

IKST bounds the continuous boundary integral equation for all the locations of the source point ξ and Eq. (48) is guaranteed to be satisfied for all the weighting functions. The solution to Eq. (62) is described in the following sections.

7. Iterative Solver for the Interval Linear System of Equations

The bounding of the original boundary integral equation using IKST results in the interval linear system of equations different from that of Eq. (29). Hence, the algorithm to solve the interval linear system of equations, Eq. (62), must be developed. This section describes the transformation of Eq. (62) to obtain it in the form of Eq. (29). Then, Krawczyk iteration (Krawczyk 1969) is performed to obtain the guaranteed bounds on the solution. Considering the linear system of equations:

$$\tilde{A}_{1e}\tilde{x}_e + \tilde{A}_{2e}\tilde{x}_e = \tilde{b}_e \tag{63}$$

where $\tilde{A}_{1e} \in \tilde{A}_1$, $\tilde{A}_{2e} \in \tilde{A}_2$, $\tilde{b}_e \in \tilde{b}$, $\tilde{x}_e \in \tilde{x}$ and A_{1e} is regular $\forall A_{1e} \big| A_{1e} \in \tilde{A}_{1e}$. Eq. (63) is pre-multiplied by $\tilde{A}_{1e}^{-1}$ as:

$$\tilde{A}_{1e}^{-1}\tilde{A}_{1e}\tilde{x}_e + \tilde{A}_{1e}^{-1}\tilde{A}_{2e}\tilde{x}_e = \tilde{A}_{1e}^{-1}\tilde{b}_e \tag{64}$$

By substituting $\tilde{A}_{1e}^{-1}\tilde{A}_{1e} = I$, $\tilde{A}_{1e}^{-1}\tilde{A}_{2e} = \tilde{A}_{3e}$ and $\tilde{A}_{1e}^{-1}\tilde{b}_e = \tilde{b}_{1e}$, Eq. (64) can be rewritten as:

$$\tilde{x}_e + \tilde{A}_{3e}\tilde{x}_e = \tilde{b}_{1e} \tag{65}$$

Since the first term in Eq. (65) is a deterministic identity matrix pre-multiplying $\tilde{x}_e$, the following substitution can be made directly. Letting $I + \tilde{A}_{3e} = \tilde{A}_e$ results in:

$$\tilde{A}_e \tilde{x}_e = \tilde{b}_{1e} \tag{66}$$

The transformed system of equations is subjected to Krawczyk iteration (Krawczyk 1969) as described in the previous section.

8. Discretization Error in Interval Boundary Element Method

In the preceding formulation, the bounds on the unknown boundary values are found using iterative techniques. The obtained bounds, however, are greatly overestimated since the dependency of interval values was not considered. One reason for this overestimation is that the interval kernels are bounded such that the source point ξ is allowed to vary along the entire element. Thus, for two adjacent elements, two source points are allowed to be connecting point between the elements and have the same location, resulting in the reduction of the rank of the system of equations. The unique location of a single source point is also not considered throughout the rows of H and G matrices, which are in $R^{n\times n}$. Thus, the parameterization of the interval location of the source point, $\tilde{\xi}$, in the $\tilde{H}$ and $\tilde{G}$ matrices must be considered in the solver to obtain n independent interval equations and to reduce the overestimation which results from a non-unique location of the source point on any individual element. For convenience, the system is parameterized such that $\tilde{\xi} = [0, 1]$ is the location scaled by a length of an element. In performing interval matrix products, the value of $\tilde{\xi}$ is decomposed into sub-intervals such that:

$$\bigcup_{i=1}^{n} \tilde{\xi}_i = \tilde{\xi} \;\; and \;\; \bigcap_{i=1}^{n} \tilde{\xi}_i = 0 \tag{67}$$

The parameterized boundary integral equation is bounded by IKST for each subinterval $\tilde{\xi}_i$, resulting in the linear system of equations:

$$H_1(\tilde{\xi}_i)\tilde{u} + H_2(\tilde{\xi}_i)\tilde{u} = G_1(\tilde{\xi}_i)\tilde{t} + G_2(\tilde{\xi}_i)\tilde{t} \tag{68}$$

where the kernel is of the same sign for $H_1(\tilde{\xi}_i)$ and $G_1(\tilde{\xi}_i)$ and contains zero for $H_2(\tilde{\xi}_i)$ and $G_2(\tilde{\xi}_i)$. The system of equations is rearranged according to the boundary conditions as:

$$A_1(\tilde{\xi}_i)\tilde{x} + A_2(\tilde{\xi}_i)\tilde{x} = b(\tilde{\xi}_i) \tag{69}$$

Steps described in the previous section lead to the equation of the form:

$$\tilde{A}(\tilde{\xi}_i)\tilde{x} = b_1(\tilde{\xi}_i) \tag{70}$$

The initial interval guess is then considered as:

$$\tilde{x}_0 = A^{-1}\bigcup_{i=1}^{n} b_1(\tilde{\xi}_i) \tag{71}$$

where A is computed for $\xi = 1/2$. The difference between I and the preconditioning matrix A^{-1} post-multiplied by the interval matrix $\tilde{A}(\tilde{\xi}_i)$ is computed as:

$$\tilde{I}_d = I - \bigcup_{i=1}^{n} A^{-1}\tilde{A}(\tilde{\xi}_i) \tag{72}$$

The difference between the solution and the initial guess is computed for each $\tilde{\xi}_i$ pre-multiplied by the preconditioning matrix I, which numerically gave the sharpest results:

$$\tilde{\delta} = \bigcup_{i=1}^{n}\left(\tilde{b}(\tilde{\xi}_i) - \tilde{A}_1(\tilde{\xi}_i)\tilde{x}_0 - \tilde{A}_2(\tilde{\xi}_i)\tilde{x}_0\right) \tag{73}$$

Also:

$$\tilde{\delta}_1 = \tilde{\delta} \tag{74}$$

The iteration is performed as:

$$\tilde{del} = \tilde{\delta}_1 \tag{75}$$

$$\tilde{\delta}_1 = \tilde{\delta} + \tilde{I}_d\tilde{del} \tag{76}$$

$$\text{If } \tilde{del} \supset \tilde{\delta}_1 \tag{77}$$

$$\tilde{x} = \tilde{x}_0 + \tilde{\delta}_1 \tag{78}$$

For any point n on element k the bounds on the discretization error are found as:

$$\tilde{E}_{nk}^{discretization} = \tilde{x}_k - x_n \tag{79}$$

where $\tilde{x}_k$ are the solution bounds over an element k and x_n is the solution from a conventional boundary element analysis for point n.

9. Examples

The first example demonstrates the IBEM considering discretization error for the elasticity problem. A unit square domain of the problem as well as the boundary element mesh is shown (Figure 4). The body has a unit elastic modulus and a zero Poisson ratio. The left and right sides have a zero traction boundary condition; the bottom boundary has a zero displacement boundary condition, while the top boundary has a zero traction condition in the x direction and a unit displacement in the y direction.

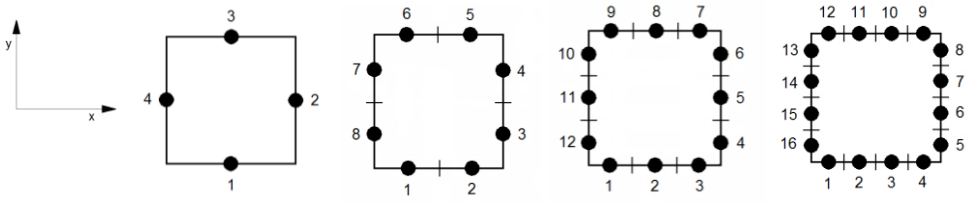


Figure 4. Boundary discretization using constant boundary elements.

The behavior of the y displacement bounds such as solution width, effectivity index, and solution bounds is depicted (Figure 5-7) for nodes 2, 3, 4, and 5 on the four respective meshes. The interval bounds, depicted by a solid line enclosing the dashed true solution, for the right edge displacement in the y direction are shown (Figure 8). The effect of the parameterization for the traction in the x direction for element 1 for the 4 and 8 element meshes is also shown (Figure 9, Figure 10).

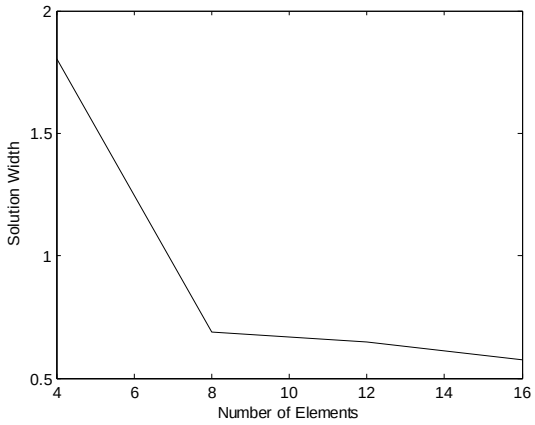


Figure 5. Behavior of the width of the interval solution with problem size.

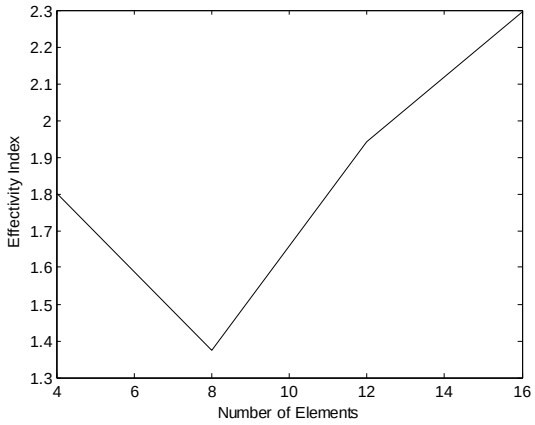


Figure 6. Behavior of the effectivity index with problem size.

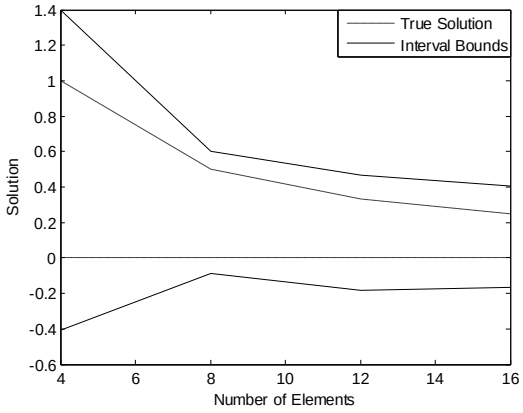


Figure 7. Behavior of the interval bounds with problem size.

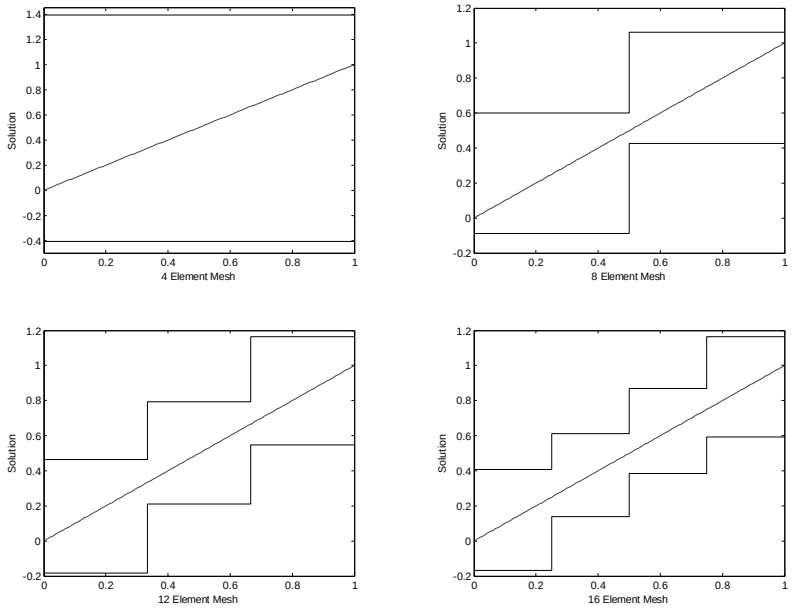


Figure 8. Behavior of the interval bounds for the different meshes.

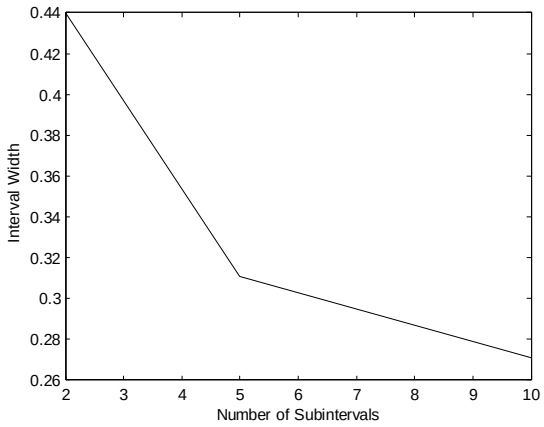


Figure 9. Behavior of the width of the interval solution with parameterization for a 4 element mesh.

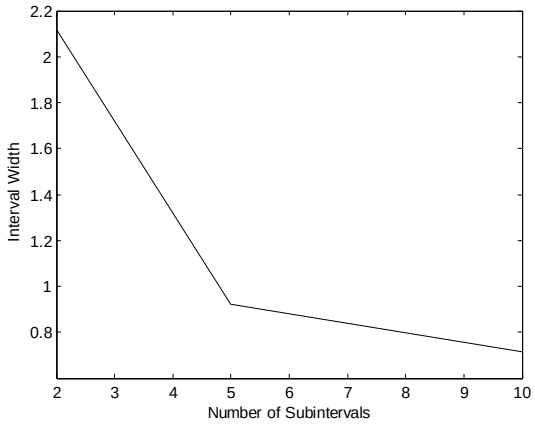


Figure 10. Behavior of the width of the interval solution with parameterization for an 8 element mesh.

The second example obtains bounds on the solution, considering the discretization error, to a hexagonal plate subjected to a unit displacement in the y direction at the top and a unit displacement in the $-y$ direction on the bottom (Figure 11). The body has a unit elastic modulus and a zero Poisson ratio.

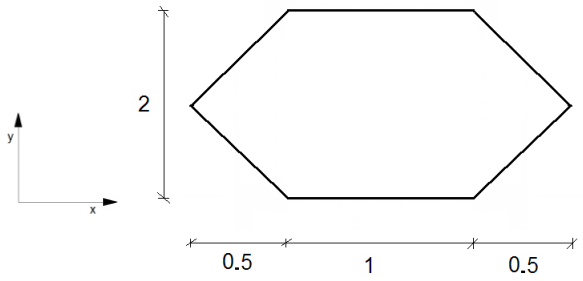


Figure 11. Hexagonal plate subjected to a unit displacement.

A symmetry model is considered, to decrease the computational time, with a unit displacement at the top and is uniformly discretized using constant boundary elements (Figure 12, Figure 13).

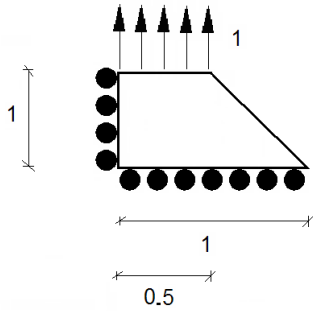


Figure 12. Symmetry model.

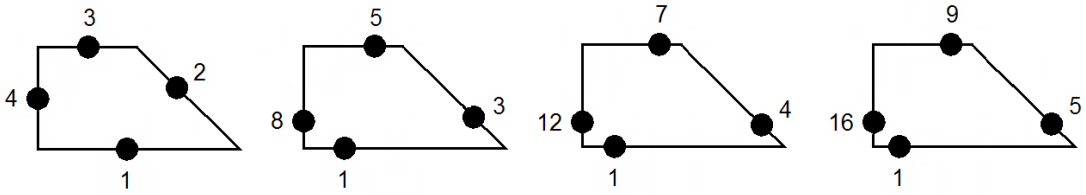


Figure 13. Uniform boundary discretization using constant boundary elements.

The behavior of the solution width, effectivity index, and solution bounds is depicted (Figure 14-16) for the displacement in the y direction for nodes 4, 8, 12, and 16 on the four respective meshes shown above. The interval bounds, depicted by a solid line enclosing the dashed true solution, for the left edge displacement in the y direction are shown (Figure 17).

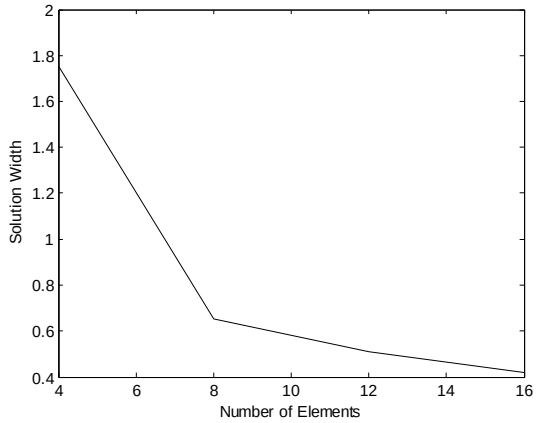


Figure 14. Behavior of the width of the interval solution with problem size.

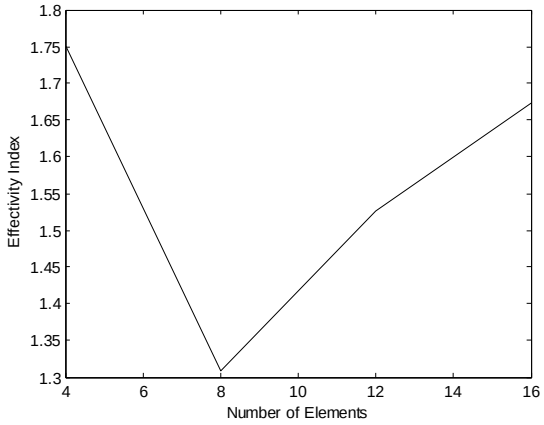


Figure 15. Behavior of the effectivity index with problem size.

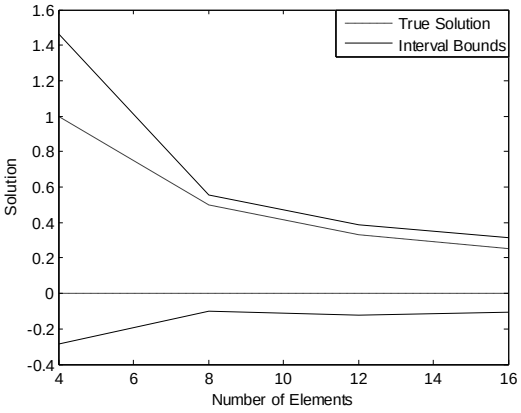


Figure 16. Behavior of the interval bounds with problem size.

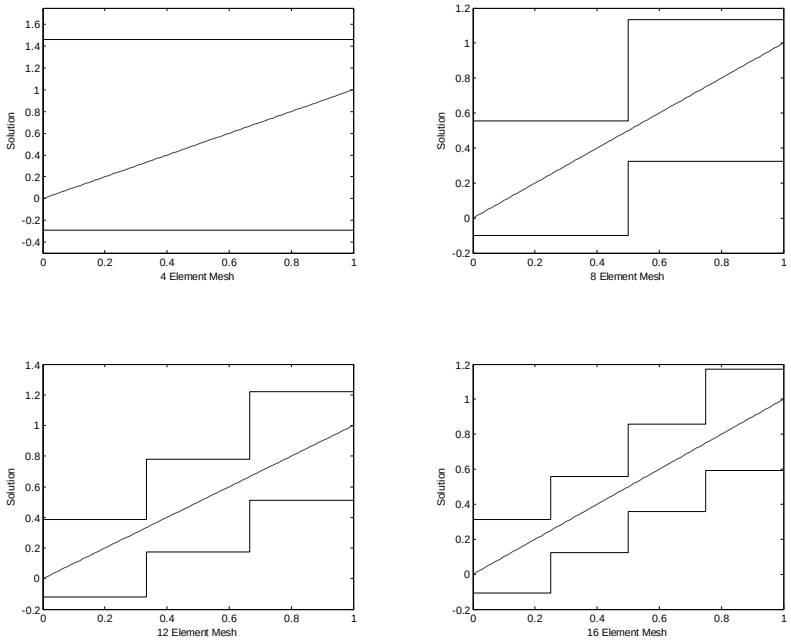


Figure 17. Behavior of the interval bounds for the different meshes.

10. Conclusion

In this work the discretization error for the elasticity problem is bounded using interval boundary element method. The interval bounds on the true solution are shown to converge for the meshes considered despite the increase in the effectivity index. The increase in the effectivity index is attributed to the slower convergence of the interval bounds than the true solution. The overestimation in the interval bounds is due to the overestimation of the terms in the interval boundary integral equation using IKST, imperfect parameterization of the location of the source point throughout the rows of the matrices H and G, and the overestimation in the iterative interval solver. There are two sources of overestimation in the iterative scheme solving the interval system of linear equations. The first one is due to the inherent overestimation when Krawczyk iteration is used to solve interval linear system of equations. This source of overestimation occurs due to the orthogonal multidimensional interval bounds enclosing a true solution which may not be, and in most cases is not, orthogonal and/or oriented in the same direction as the interval bounds (Figure 18).

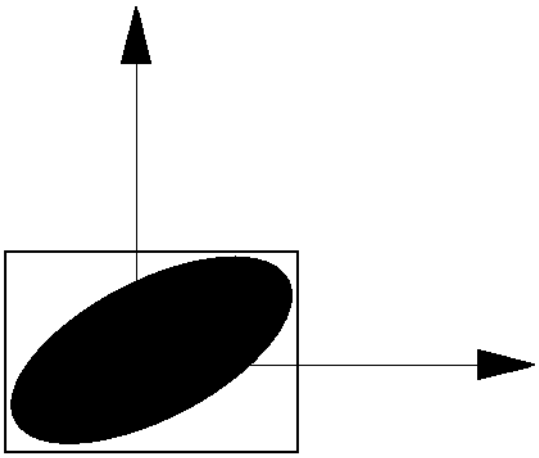


Figure 18. Interval bounds on the solution.

The second source of overestimation on the interval solver comes from incomplete consideration of the interval parameterization in Eq. (76). Each term in Eq. (76) is parameterized; however, each of these terms must be dealt with in its entirety when operated with. The solution of the linear system of equations must be satisfied for the entire system and thus the residual has to be calculated for the entire interval width, not for the length of the subinterval. If the residual is computed for the portion of the interval, for instance an interval width corresponding to a subinterval such that a complete interval parameterization can be utilized in Eq. (76), the enclosure in no longer guaranteed.

References

Alefeld, G. and J. Herzberger. *Introduction to Interval Computations*, Academic Press, New York, NY, 1983.

Brebbia, C.A. and J. Dominguez. *Boundary Elements: An Introductory Course*, Computational Mechanics, New York, McGraw-Hill, 1992.

Dobner, H. Kernel-Splitting Technique for Enclosing the Solution of Fredholm Equations of the First Kind, *Reliable Computing*, Vol. 8, pp. 469-179, 2002.

Friedman, A. *Partial differential equations***,** R. E. Krieger Pub. Co., Huntington, N.Y., 1976.

Gay, D. M. Solving Interval Linear Equations, *SIAM Journal on Numerical Analysis*, Vol. 19, 4, pp. 858-870, 1982.

Hansen, E. *Interval arithmetic in matrix computation*, J. S. I. A. M., series B, Numerical Analysis, part I, 2, 308-320, 1965.

Hartmann, F. *Introduction to boundary elements : theory and applications*, New York, Springer-Verlag, 1989.

Jansson, C. Interval Linear System with Symmetric Matrices, Skew-Symmetric Matrices, and Dependencies in the Right Hand Side, *Computing*, Vol. 46, pp. 265-274, 1991.

Krawczyk, R. Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken, *Computing*, Vol. 4, pp. 187-201, 1969.

MATLAB 6.5.1. *Using Matlab Version 6 OEM Manual the Language of Technical Computing*, Mathworks, 2002.

Moore, R. E. *Interval Analysis*, Prentice-Hall, Inc., Englewood Cliffs, N. J, 1966.

Muhanna, R. L. and R. L. Mullen. Uncertainty in Mechanics Problems - Interval-Based Approach, *Journal of Engineering Mechanics, ASCE*, Vol. 127, No. 6, pp. 557-566, 2001.

Muhanna et al. Penalty-Based Solution for the Interval Finite-Element Methods. *Journal of Engineering Mechanics, ASCE*, Vol. 131, 10, pp.1102-1111, 2005.

Mullen, R. L. and R. L. Muhanna. Interval-Based Finite Element Methods. *Reliable Computing 5* pp. 97-100, 1999.

Mullen, R. L. and R. L. Muhanna. Bounds of Structural Response for All Possible Loadings, "*Journal of Structural Engineering, ASCE*, Vol. 125, No. 1, pp 98-106, 1999.

Neumaier, A. Overestimation in Linear Interval Equations, *SIAM Journal on Numerical Analysis*, Vol. 24, 1, pp. 207-214, 1987.

Neumaier, A. Rigorous Sensitivity Analysis for Parameter-Dependent Systems of Equations, *Journal of Mathematical Analysis and Applications*, Vol. 144, pp. 16-25, 1989.

Neumaier, A. *Interval methods for systems of equations*, Cambridge University Press, 1990.

Pilkey, W. D. and W. Wunderlich. *Mechanics of Structures, Variational and Computational Methods*, CRS Press, London, 1994.

Rencis, J. J. and K-Y. Jong. An Error Estimator for Boundary Element Computations, *ASCE Journal of Engineering Mechanics*, Vol. 115 (9), pp. 1993-2010, 1989.

Rump, S. M. *Kleine Fehleschranken bei Matrixproblem*, Dissertation, Universitat Karlsruhe, 1980.

Rump, S. M. Rigorous Sensitivity Analysis for Systems of Linear and Nonlinear Equations, *Mathematics of Computations*, Vol. 54, 190, pp. 721-736, 1990.

Rump, S. M. Self-validating Methods, *Linear Algebra and its Applications*, Vol. 324 pp. 3-13, 2001.

Sunaga, T. Theory of interval algebra and its application to numerical analysis, *RAAG Memoirs 3*, pp 29-46, 1958.

Taylor B. *Methodus incrementorum directa & inversa*, Londini, typis Pearsnianis, 1715.

Zalewski B. F. et al. Interval Boundary Element Method in the Presence of Uncertain Boundary Conditions, Integration Errors, and Truncation Errors, *Engineering Analysis with Boundary Elements*, 2007 (in review).

Zalewski B. F. and R. L. Mullen. Local Discretization Errors in Boundary Element Analysis, *Journal of Computational and Applied Mathematics*, 2007 (in review).

Stress Analysis of a Singly Reinforced Concrete Beam with Uncertain Structural Parameters

M.V.Rama Rao[1], A. Pownuk[2] and I. Skalna[3]
[1]Department of Civil Engineering,
Vasavi College of Engineering, Hyderabad-500 031, India
email: dr.mvrr@gmail.com
[2]Department of Mathematical Sciences, University of Texas at El Paso
500 West University Avenue El Paso, Texas 79968, USA
email: andrzej@pownuk.com
[3]Department of Applied Computer Science
University of Science and Technology AGH, ul. Gramatyka 10, Cracow, Poland
email: skalna@agh.edu.pl

Abstract. This paper presents the efforts by the authors to introduce interval uncertainty in the stress analysis of reinforced concrete flexural members. A singly reinforced concrete beam with interval values of steel reinforcement and corresponding Young's modulus and subjected to an interval bending moment is taken up for analysis. Using extension principle, the internal moment of resistance of the beam is expressed as a function of interval values of stresses in concrete and steel. The stress distribution model for the cross section of the beam given by IS 456-2000 (Indian standard code of practice for plain and reinforced concrete) is modified for this purpose. The internal moment of resistance is then equated to the external bending moment due to interval loads acting on the beam. The stresses in concrete and steel are obtained as interval values for various combinations of interval values of structural parameters. The interval stresses and strains in concrete and steel obtained using combinatorial solution; search-based algorithm and sensitivity analysis are found to be in excellent agreement.

Keywords: interval stresses; stress distribution; sensitivity analysis; search-based algorithm

1. Introduction

Analysis of rectangular beams of reinforced concrete is based on nonlinear and/or discontinuous stress-strain relationships and such analyses are difficult to perform. Provided the nature of loading, the beam dimensions, the materials used and the quantity of reinforcement are known, the theory of reinforced concrete permits the analysis of stresses, strains, deflections, crack spacing and width and also the collapse load. Further, the aim of analyzing the beam is to locate the neutral axis depth, find out the stresses in compression concrete and tensile reinforcement and also compute the moment of resistance. The aim of the designer of reinforced concrete beams is to predict the entire spectrum of behavior in mathematical terms, identify the parameters which influence this behavior, and obtain the cracking, deflection and collapse limit loads. There are usually innumerable answers to a design problem. Thus the

design is followed by analysis and a final selection is obtained by a process of iteration. Thus the design process becomes clear only when the process of analysis is learnt thoroughly.

In the traditional (deterministic) methods of analysis, all the parameters of the system are taken to be precisely known. In practice, however, there is always some degree of uncertainty associated with the actual values for structural parameters. As a consequence of this, the structural system will always exhibit some degree of uncertainty. A reliable approach to handle uncertainty in a structural system is the use of interval algebra. In this approach, uncertainties in structural parameters will be introduced as interval values i.e., the values are known to lie between two limits, but the exact values are unknown. Thus, the problem is of determining conservative intervals for the structural response. Though interval arithmetic was introduced by Moore (Moore, 1966), the application of interval concepts to structural analysis is more recent. Modeling with intervals provides a link between design and analysis where uncertainty may be represented by bounded sets of parameters. Interval computation has become a significant computing tool with the software packages developed in the past decade. In the present work, a singly-reinforced concrete beam with interval area of steel reinforcement and corresponding interval Young's modulus and subjected to an interval moment is taken up for analysis. Interval algebra is used to establish the bounds for the stresses and strains in steel and concrete.

2. Literature Survey

In the literature there are several methods for solution of equations with interval parameters. In the year 1966, Moore (1966) discussed the problem of solution of system of linear interval equations. Neumaier (1990) discussed several methods of solution of linear interval equations in his book. Ben-Haim and Elishakoff (1990) introduced ellipsoid uncertainty. System of linear interval equation with dependent parameters and symmetric matrix was discussed by Jansson (1991). In their work Köylüoglu, Cakmak, Nielsen (1995) applied the concept of interval matrix to solution of FEM equations with uncertain parameters. Rao and Chen (1998) developed a new search-based algorithm to solve a system of linear interval equations to account for uncertainties in engineering problems. The algorithm performs search operations with an accelerated step size in order to locate the optimal setting of the hull of the solution.

McWilliam (2000) described several method of solution of interval equations. Akpan *et. al* (2001) used response surface method in order to approximate fuzzy solution. Vertex solution methodology that was based on α-cut representation was used for the fuzzy analysis. Muhanna and Mullen (2001) handled uncertainty in mechanics problems on using an interval-based approach. Muhanna's algorithm is modified by Rama Rao (2006) to study the cumulative effect of multiple uncertainties on the structural response. Neumaier and Pownuk (2007) explored properties of positive definite interval matrices. Their algorithm works even for very large uncertainty in parameters. Skalna, Rama Rao and Pownuk (2007) investigated the solution of systems of fuzzy equations in structural mechanics.

Several models were proposed to describe the stress distribution in the cross section of a concrete beam subjected to pure flexure. Initially, the parabolic model was proposed by Hognestad (1955) in 1951. This was followed by an exponential model proposed by Smith and Young (1955) and Desai and Krishnan model (1964). These models are applicable to concretes with strength below 40 MPa. The Indian standard code of practice for plain and reinforced concrete IS 456-2000 (2000) allows the assumption of any suitable relationship between the compressive stress distribution in concrete and the strain in concrete i.e. rectangle, trapezoid, parabola or any other shape which results in prediction of strength in substantial agreement with the results of test.

Rama Rao and Pownuk (2007) made the initial efforts to introduce uncertainty in the stress analysis of reinforced concrete flexural members. A singly reinforced concrete beam subjected to an interval load is taken up for analysis. Using extension principle, the internal moment of resistance of the beam is expressed as a function of interval values of stresses in concrete and steel. The stress distribution model for the cross section of the beam given by IS 456-2000 is modified for this purpose. The internal moment of resistance is then equated to the external bending moment due to interval loads acting on the beam. The stresses and strains in concrete and steel are obtained as interval values. The sensitivity of stresses in steel and concrete to corresponding variation of interval values of load about its mean values is explored.

A study of the effect of multiple uncertainties on the stress distribution across the cross section of a singly-reinforced concrete member is taken up by the authors in the present work. The stress distribution model suggested by the Indian code IS 456-2000 is followed in the present study (Figure 1). Post cracking behavior up to Limit State of Serviceability is considered in the present work (Purushottaman, 1986).

3. Stress analysis of a singly reinforced concrete section

3.1 STRESS DISTRIBUTION DUE TO A CRISP MOMENT

A singly reinforced concrete section shown in Figure 1 with is taken up for analysis of stresses and strains in concrete and steel. The beam has a width of b and an effective depth of d. The beam is subjected to a maximum external moment M. Strain-distribution is linear and ε_{cc} is the strain in concrete at the extreme compression fiber and ε_s is the strain in steel. Let x be the neutral axis depth from the extreme compression fiber. The aim of analyzing the beam is to locate this neutral axis depth, find out the stresses in compression concrete and the tensile reinforcement and also compute the moment of resistance.

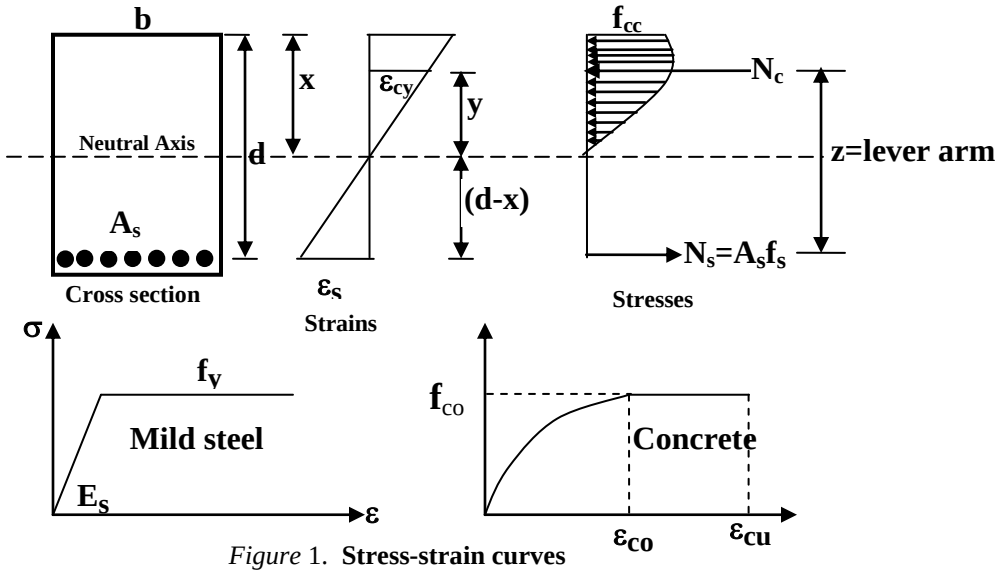


Figure 1. **Stress-strain curves**

The stress-distribution is concrete is parabolic and concrete in tension is neglected. The strain ε_{cy} at any level y below the neutral axis ($y \le x$) is

$$\varepsilon_{cy} = \left(\frac{y}{x}\right)\varepsilon_{cc} \tag{1}$$

The corresponding stress f_{cy} is

$$f_{cy} = f_{co}\left[2\left(\frac{\varepsilon_{cy}}{\varepsilon_{co}}\right) - \left(\frac{\varepsilon_{cy}}{\varepsilon_{co}}\right)^2\right] \quad \text{for} \quad \varepsilon_{cy} \le \varepsilon_{co} \text{ and } f_{cy} = f_{co} \text{ for } \varepsilon_{cy} = \varepsilon_{co} \tag{2}$$

Total compressive force in concrete N_c is given by

$$N_c = \int_{y=0}^{y=x} f_{cy} b dy = \left[C_1\varepsilon_{cc} - C_1\varepsilon_{cc}^2\right]x \tag{3}$$

where

$$C_1 = \left(\frac{bf_{co}}{\varepsilon_{co}}\right) \quad \text{and} \quad C_2 = \left(\frac{bf_{co}}{3\varepsilon_{co}^2}\right) \tag{4}$$

Tensile stress in steel

$$N_s = (A_s E_s)\varepsilon_{cc}\left(\frac{d-x}{x}\right) \tag{5}$$

If there are no external loads, the equation of longitudinal equilibrium, $N_s = N_c$ leads to the quadratic equation

$$\left[C_1 - C_2\varepsilon_{cc}\right]x^2 + A_s E_s x - A_s E_s d = 0 \tag{6}$$

Depth of resultant compressive force from the neutral axis $\bar{y}$ is given by

$$\bar{y} = \frac{\int_{y=0}^{y=x} bf_{cy}\,y\,dy}{\int_{y=0}^{y=x} bf_{cy}\,dy} = \frac{\left[\left(\frac{2C_1}{3}\right) - \left(\frac{3C_2}{4}\right)\varepsilon_{cc}\right]}{\left[C_1 - C_2\varepsilon_{cc}\right]}x \tag{7}$$

Internal resisting moment M_R is given by

$$M_R = N_c \times z = N_c \times (\bar{y} + d - x) \tag{8}$$

For equilibrium the external moment M is equated to the internal moment of resistance M_R as

$$M \le M_R \tag{9}$$

The neutral axis depth x can determined by solving equation (6) only when ε_{cc} is known. Thus a trial and error procedure is adopted where in ε_{cc} is assumed and the corresponding values of N_C , $\bar{y}$ and internal resisting moment M_R are obtained using equation (3), equation (7) and equation (8) such that equation (9) is satisfied.

Strain in steel $\varepsilon_s = \left(\frac{d-x}{x}\right)\varepsilon_{cc}$ (10)

Stress in steel $f_s = E_s \varepsilon_s = E_s \left(\frac{d-x}{x} \right) \varepsilon_{cc} \leq 0.87 f_y$ (11)

Total tensile force in steel reinforcement = $N_s = A_s f_s$ (12)

4. Stress Distribution due to uncertain interval parameters

Consider the case of a singly reinforced concrete beam with interval values of area of steel reinforcement $\mathbf{A_s}$ with corresponding interval Young's modulus $\mathbf{E_s}$ and subjected to an interval external bending moment $\mathbf{M} = \left[\underline{M}, \overline{M} \right]$. The uncertainty in external moment arises out of uncertainty of loads acting on the beam. Correspondingly the resulting stresses and strains in concrete and steel are also uncertain and are modeled using interval numbers.

Using extension principle (Zadeh, 1965) all the equations developed in the previous section can be extended and made applicable to the interval case. The objective of the present study is to determine distribution of stresses and strain across the cross section of the beam. Two new approaches have been proposed for this purpose: a search based algorithm and a procedure based on Pownuk's sensitivity analysis (Pownuk, 2004).These methods are outlined as follows:

4.1 SEARCH-BASED ALGORITHM (SBA)

A search based algorithm (SBA) is developed to perform search operations with an accelerated step size in order to compute the optimal setting for the interval value of strain in concrete is $\boldsymbol{\varepsilon}_{cc} = \left[\underline{\varepsilon_{cc}}, \overline{\varepsilon}_{cc} \right]$. The algorithm is outlined below:

4.1.1 *Algorithm -1 (Search-based algorithm)*

a) The mid-value M of the given interval moment $\mathbf{M}$ is computed as $M = \frac{\underline{M} + \overline{M}}{2}$

b) The mid-value A_s of interval area of steel reinforcement $\mathbf{A_s}$ is computed as $A_s = \frac{\underline{A_s} + \overline{A}_s}{2}$

c) The mid-value E_s of interval Young's modulus of steel $\mathbf{E_s}$ is computed as $E_s = \frac{\underline{E_s} + \overline{E}_s}{2}$

d) Now the interval form of quadratic equation (6) given below is solved using the procedure outlined by Hansen and Walster (2002)

$$\left[C_1 - C_2\boldsymbol{\varepsilon}_{cc}\right]\mathbf{x}^2 + \mathbf{A_s E_s x} - \mathbf{A_s E_s} d = 0 \tag{13}$$

Various values of $\boldsymbol{\varepsilon}_{cc}$ are assumed and the neutral axis depth $\mathbf{x}$ and the corresponding values of $\mathbf{N_c}, \bar{\mathbf{y}}$ and $\mathbf{M_R}$ are determined by using a trial and error procedure outlined in the previous section.

e) The interval strain in concrete $\boldsymbol{\varepsilon}_{cc}$ is initially approximated as the point interval $\left[\varepsilon_{cc}, \varepsilon_{cc}\right]$.

f) The lower and upper bounds of $\boldsymbol{\varepsilon}_{cc}$ are obtained as $\boldsymbol{\varepsilon}_{cc} = \left[\varepsilon_{cc} - \lambda_1 d\underline{\varepsilon}, \varepsilon_{cc} + \lambda_2 d\bar{\varepsilon}\right]$ where $d\underline{\varepsilon}$ and $d\bar{\varepsilon}$ are the step sizes in strain to obtain the lower and upper bounds, λ_1 and λ_2 being the corresponding multipliers. Initially λ_1 and λ_2 are taken as unity.

g) While both λ_1, λ_2 are non-zero, $d\underline{\varepsilon}$ and $d\bar{\varepsilon}$ are incremented and $\boldsymbol{\varepsilon}_{cc}$ is computed. The procedure is continued iteratively till the interval form of (9) i.e. $\mathbf{M} \le \mathbf{M_R}$ is satisfied. The computations performed are outlined as follows:

1) The interval values of $\mathbf{x}, \bar{\mathbf{y}}, \mathbf{z}, \mathbf{N_c}$ and the interval internal resisting moment $\mathbf{M_R} = [\underline{M}_R, \bar{M}_R]$ are computed. If η is a very small number

2) λ_1 is set to zero if $\left|\dfrac{\underline{M}_R - \underline{M}}{\underline{M}_R}\right| \le \eta$ (14)

3) λ_2 is set to zero if $\left|\dfrac{\bar{M}_R - \bar{M}}{\bar{M}_R}\right| \le \eta$ (15)

4) The search is discontinued when $\lambda_1 = \lambda_2 = 0$.

4.2 SENSITIVITY ANALYSIS METHOD

4.2.1 *Extreme values of ε_{cc} and x.*

Unknown variables ε_{cc} and x can be found from the system of equation (8) and equilibrium equation $N_s = N_c$. Let us introduce a new notation

$$\begin{cases} F_1 = F_1(\varepsilon_{cc}, x, p_1, ..., p_m) = M_R - N_c \cdot (\bar{y} + d - x) = 0 \\ F_2 = F_2(\varepsilon_{cc}, x, p_1, ..., p_m) = N_s - N_c = 0 \end{cases} \tag{16}$$

where $p_1 = M$, $p_2 = f_{co}$, $p_3 = A_S$, $p_4 = \varepsilon_{co}$, $p_5 = E_S$, $p_6 = b$, $p_7 = d$.

Because the problem is relatively simple and the intervals $\left[\underline{p}_i, \overline{p}_i \right]$ are usually narrow, then it is possible to solve the problem using sensitivity analysis method (Pownuk, 2004). Let us calculate sensitivity of the solution with respect to the parameter p_i.

$$\frac{\partial}{\partial p_i} F_1 = \frac{\partial F_1}{\partial \varepsilon_{cc}} \frac{\partial \varepsilon_{cc}}{\partial p_i} + \frac{\partial F_1}{\partial x} \frac{\partial x}{\partial p_i} + \frac{\partial F_1}{\partial p_i} = 0 \tag{17}$$

$$\frac{\partial}{\partial p_i} F_2 = \frac{\partial F_2}{\partial \varepsilon_{cc}} \frac{\partial \varepsilon_{cc}}{\partial p_i} + \frac{\partial F_2}{\partial x} \frac{\partial x}{\partial p_i} + \frac{\partial F_2}{\partial p_i} = 0 \tag{18}$$

In matrix form

$$\begin{bmatrix} \frac{\partial F_1}{\partial \varepsilon_{cc}} & \frac{\partial F_1}{\partial x} \\ \frac{\partial F_2}{\partial \varepsilon_{cc}} & \frac{\partial F_2}{\partial x} \end{bmatrix} \begin{bmatrix} \frac{\partial \varepsilon_{cc}}{\partial p_i} \\ \frac{\partial x}{\partial p_i} \end{bmatrix} = \begin{bmatrix} -\frac{\partial F_1}{\partial p_i} \\ -\frac{\partial F_2}{\partial p_i} \end{bmatrix} \tag{19}$$

Using Cramer's rule the solution is given by the following formulas

$$\frac{\partial \varepsilon_{cc}}{\partial p_i} = -\frac{\begin{vmatrix} \frac{\partial F_1}{\partial p_i} & \frac{\partial F_1}{\partial x} \\ \frac{\partial F_2}{\partial p_i} & \frac{\partial F_2}{\partial x} \end{vmatrix}}{\begin{vmatrix} \frac{\partial F_1}{\partial \varepsilon_{cc}} & \frac{\partial F_1}{\partial x} \\ \frac{\partial F_2}{\partial \varepsilon_{cc}} & \frac{\partial F_2}{\partial x} \end{vmatrix}} = -\frac{\frac{\partial (F_1, F_2)}{\partial (p_i, x)}}{\frac{\partial (F_1, F_2)}{\partial (\varepsilon_{cc}, x)}}, \quad \frac{\partial x}{\partial p_i} = -\frac{\begin{vmatrix} \frac{\partial F_1}{\partial \varepsilon_{cc}} & \frac{\partial F_1}{\partial p_i} \\ \frac{\partial F_2}{\partial \varepsilon_{cc}} & \frac{\partial F_2}{\partial p_i} \end{vmatrix}}{\begin{vmatrix} \frac{\partial F_1}{\partial \varepsilon_{cc}} & \frac{\partial F_1}{\partial x} \\ \frac{\partial F_2}{\partial \varepsilon_{cc}} & \frac{\partial F_2}{\partial x} \end{vmatrix}} = -\frac{\frac{\partial (F_1, F_2)}{\partial (\varepsilon_{cc}, p_i)}}{\frac{\partial (F_1, F_2)}{\partial (\varepsilon_{cc}, x)}} \tag{20}$$

If all Jacobians

$$\frac{\partial (F_1, F_2)}{\partial (\varepsilon_{cc}, x)}, \frac{\partial (F_1, F_2)}{\partial (p_i, x)}, \frac{\partial (F_1, F_2)}{\partial (\varepsilon_{cc}, p_i)} \tag{21}$$

are regular then the derivatives have constant sign and the relations $\varepsilon_{cc} = \varepsilon_{cc}(p_i), x = x(p_i)$ are monotone. All variables p_i belong to known intervals $p_i \in \left[\underline{p}_i, \overline{p}_i\right]$ because of that sign of the Jacobians can be checked using interval global optimization method (Pownuk, 2004).

If

$$0 \neq \min_{\varepsilon_{cc}\in\left[\underline{\varepsilon}_{cc},\bar{\varepsilon}_{cc}\right],x\in\left[\underline{x},\bar{x}\right],p_1\in\left[\underline{p}_1,\bar{p}_1\right],\ldots,p_m\in\left[\underline{p}_m,\bar{p}_m\right]} \left|\Delta(\varepsilon_{cc}, x, p_1, \ldots, p_m)\right| \tag{22}$$

then the Jacobian Δ is regular, where $\Delta = \frac{\partial(F_1, F_2)}{\partial(\varepsilon_{cc}, x)}$, $\Delta = \frac{\partial(F_1, F_2)}{\partial(p_i, x)}$ or $\Delta = \frac{\partial(F_1, F_2)}{\partial(\varepsilon_{cc}, p_i)}$. If the sign of the derivatives is constant then extreme values of the solution can be calculated using endpoints of the intervals $\left[\underline{p}_i, \overline{p}_i\right]$ and sensitivity analysis method (Pownuk, 2004). The whole algorithm of calculation is the following:

4.2.2 *Algorithm-2 (Sensitivity Analysis)*

1) Calculate mid point of the intervals $p_{i0} = \frac{\underline{p}_i + \overline{p}_i}{2}$.
2) Solve the system of equation (16) and calculate ε_{cc0}, x_0.
3) Calculate sensitivity of the solution $\frac{\partial \varepsilon_{cc}}{\partial p_i}$, $\frac{\partial x}{\partial p_i}$ from the system of equation (19).
4) If $\frac{\partial \varepsilon_{cc}}{\partial p_i} \geq 0$ then $p_i^{\min,\varepsilon_{cc}} = \underline{p}_i$, $p_i^{\max,\varepsilon_{cc}} = \overline{p}_i$, if $\frac{\partial \varepsilon_{cc}}{\partial p_i} < 0$ then $p_i^{\min,\varepsilon_{cc}} = \overline{p}_i$, $p_i^{\max,\varepsilon_{cc}} = \underline{p}_i$.
5) If $\frac{\partial x}{\partial p_i} \geq 0$ then $p_i^{\min,x} = \underline{p}_i$, $p_i^{\max,x} = \overline{p}_i$, if $\frac{\partial x}{\partial p_i} < 0$ then $p_i^{\min,x} = \overline{p}_i$, $p_i^{\max,x} = \underline{p}_i$.
6) Extreme values of ε_{cc}, x can be calculated as a solution of the following system of equations.
7) Verification of the results. If the derivatives have the same sign at the endpoints $p_i^{\min,x}, p_i^{\max,x}, p_i^{\min,\varepsilon_{cc}}, p_i^{\max,\varepsilon_{cc}}$ and in the midpoint then the solution is very reliable.

4.2.3 *Interval stress in extreme concrete fiber*

Sensitivity of stress in extreme concrete fiber f_{cc} can be calculated in the following way

$$\frac{\partial}{\partial p_i} f_{cc} = \frac{\partial f_{cc}}{\partial \varepsilon_{cc}} \frac{\partial \varepsilon_{cc}}{\partial p_i} + \frac{\partial f_{cc}}{\partial x} \frac{\partial x}{\partial p_i} + \frac{\partial f_{cc}}{\partial p_i} \tag{23}$$

where $\frac{\partial \varepsilon_{cc}}{\partial p_i}$ and $\frac{\partial x}{\partial p_i}$ are solution of the equation (19).

If $\frac{\partial f_{cc}}{\partial p_i} \geq 0$ then $p_i^{\min, f_{cc}} = \underline{p}_i$, $p_i^{\max, f_{cc}} = \overline{p}_i$, if $\frac{\partial f_{cc}}{\partial p_i} < 0$ then $p_i^{\min, f_{cc}} = \overline{p}_i$, $p_i^{\max, f_{cc}} = \underline{p}_i$.

$$\underline{f}_{cc} = f_{cc}\left(\varepsilon_{cc}^{\min, f_{cc}}, x^{\min, f_{cc}}, p_1^{\min, f_{cc}}, ..., p_m^{\min, f_{cc}}\right), \tag{24}$$

$$\overline{f}_{cc} = f_{cc}\left(\varepsilon_{cc}^{\max, f_{cc}}, x^{\max, f_{cc}}, p_1^{\max, f_{cc}}, ..., p_m^{\max, f_{cc}}\right) \tag{25}$$

In the midpoint sensitivity is equal to

$$\frac{\partial}{\partial M} f_{cc} = \frac{\partial f_{cc}}{\partial \varepsilon_{cc}} \frac{\partial \varepsilon_{cc}}{\partial M} + \frac{\partial f_{cc}}{\partial x} \frac{\partial x}{\partial M} + \frac{\partial f_{cc}}{\partial M} \tag{26}$$

Extreme values of stress in extreme concrete fiber calculated form the equations (24) and (25).

4.2.4 *Interval stress in steel*

Sensitivity of stress in steel f_s can be calculated in the following way

$$\frac{\partial}{\partial p_i} f_s = \frac{\partial f_s}{\partial \varepsilon_{cc}} \frac{\partial \varepsilon_{cc}}{\partial p_i} + \frac{\partial f_s}{\partial x} \frac{\partial x}{\partial p_i} + \frac{\partial f_s}{\partial p_i} \tag{27}$$

where $\frac{\partial \varepsilon_{cc}}{\partial p_i}$ and $\frac{\partial x}{\partial p_i}$ are solution of the equation (19).

If $\frac{\partial f_s}{\partial p_i} \geq 0$ then $p_i^{\min, f_s} = \underline{p}_i$, $p_i^{\max, f_s} = \overline{p}_i$, if $\frac{\partial f_s}{\partial p_i} < 0$ then $p_i^{\min, f_s} = \overline{p}_i$, $p_i^{\max, f_s} = \underline{p}_i$.

$$\underline{f}_s = f_s\left(\varepsilon_{cc}^{\min, f_s}, x^{\min, f_s}, p_1^{\min, f_s}, ..., p_m^{\min, f_s}\right), \tag{28}$$

$$\overline{f}_s = f_s\left(\varepsilon_{cc}^{\max, f_s}, x^{\max, f_s}, p_1^{\max, f_s}, ..., p_m^{\max, f_s}\right). \tag{29}$$

Sensitivity at the mid point is computed as

$$\frac{\partial}{\partial M} f_s = \frac{\partial f_s}{\partial \varepsilon_{cc}} \frac{\partial \varepsilon_{cc}}{\partial M} + \frac{\partial f_s}{\partial x} \frac{\partial x}{\partial M} + \frac{\partial f_s}{\partial M} \tag{30}$$

4.3 COMBINATORIAL SOLUTION

Combinatorial solution is obtained by considering the upper and lower bounds of the external interval moment and computing the corresponding deterministic values of ε_{cc}, x, $\bar{y}$, N_c and M_R are determined. The lower and upper values taken by these quantities are utilized to obtain the corresponding interval values of $\mathbf{x}$, $\bar{\mathbf{y}}$, $\mathbf{z}$, $\mathbf{N_c}$ and $\mathbf{M_R}$.

5. Example Problem

A singly reinforced beam with rectangular cross section is taken up to illustrate the validity of the above methods. The beam has the dimensions b = 300 mm and D = 550 mm and effective depth d = 500 mm. The beam is reinforced with 6 numbers of Tor50 bars of 25 mm diameter ($A_S = 6 \times 491\ mm^2 = 2946 mm^2$).The bending moment acting on the beam is $\mathrm{M} = 100\mathrm{kNm}$. Allowable compressive stress in concrete f_{co} *is* $13.4\ N/mm^2$ and allowable strain in concrete ε_{co} is 0.002. Young's modulus of steel E_S *is* $200GPa$.The stress-strain curves for concrete and steel as detailed in IS 456-2000 are adopted (Figure 1)

Interval uncertainty is considered in the bending moment, area of steel reinforcement and Young's modulus of steel reinforcement. The corresponding membership functions are shown in Figure 2, Figure 3 and Figure 4 respectively. Interval values of bending moment, area of steel reinforcement and corresponding Young's modulus can be extracted from these figures using α-cut approach at any desired level of uncertainty for use in the stress analysis. For example, corresponding to α = 0.8, the interval values considered are $\mathbf{M}$ = [98,102] kNm, $\mathbf{A_s}$ = [2917, 2975] mm^2 and $\mathrm{E_s} = [198, 202]\mathrm{GPa}$. The corresponding interval values of neutral axis depth, strain and stress in concrete and stress in steel reinforcement are computed at various levels of uncertainty and membership functions are plotted.

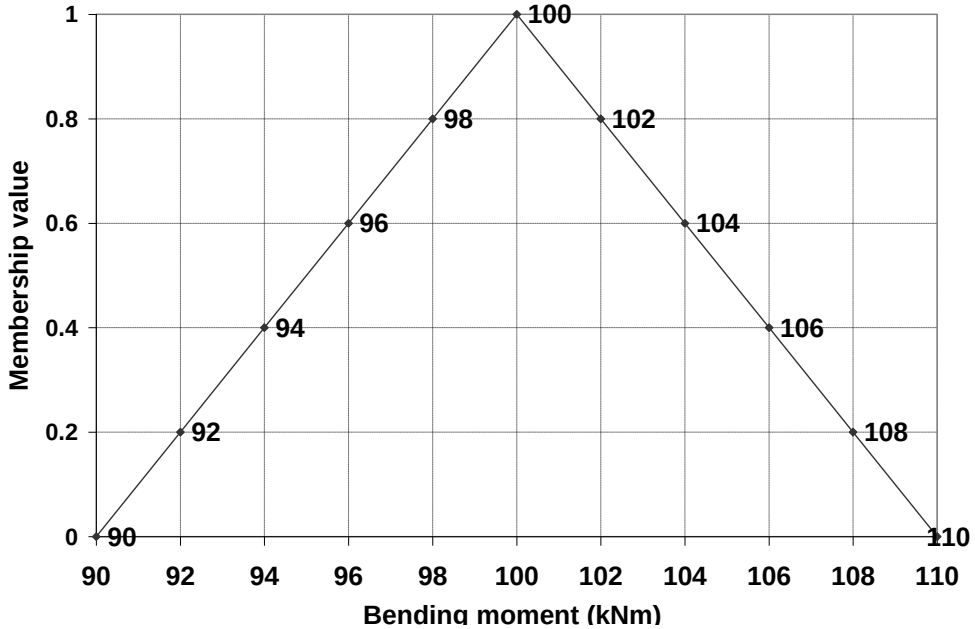


Figure 2. Membership function for bending moment

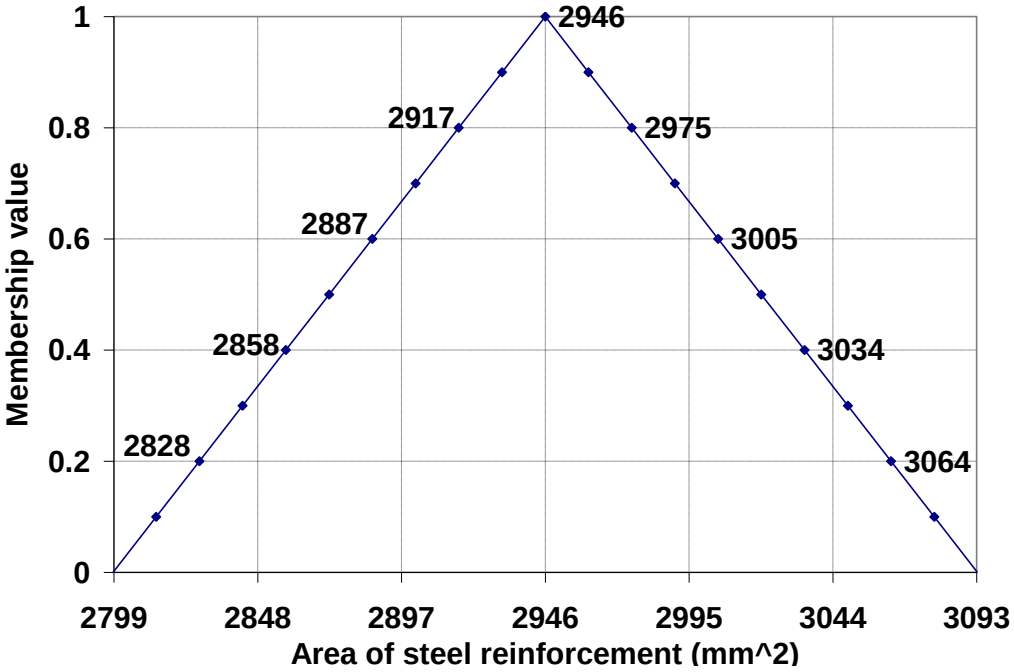
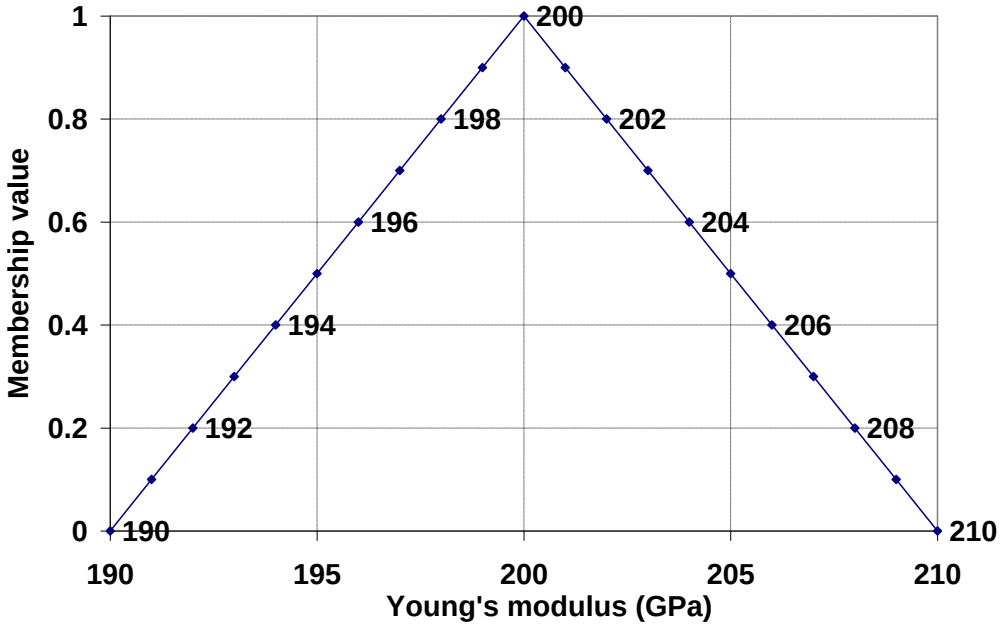


Figure 3. Membership function for area of steel reinforcement



Fiaure 4. Membership function for Young's modulus of steel reinforcement

In the present study, interval values of neutral axis depth $\mathbf{x}$, strain $\boldsymbol{\varepsilon}_{cc}$ and stress $\mathbf{f}_{cc}$ in extreme compression fiber of concrete and stress in steel f_s computed using search-based algorithm (SBA) and sensitivity analysis (SA) approach for the following cases:

a) Case 1 :
External interval moment $\mathbf{M}$ = [96,104] kNm
Area of Steel reinforcement A_S = 2946 mm^2
Young's modulus of Steel reinforcement E_S = 2×10^5 N/mm^2

b) Case 2:
External interval moment $\mathbf{M}$ = [90,110] kNm
Interval area of Steel reinforcement $\mathbf{A_s}$ = [0.9,1.1]*2946 mm^2
Young's modulus of Steel reinforcement E_S = 2×10^5 N/mm^2

c) Case 3:
External interval moment $\mathbf{M}$ = [80,120] kNm
Area of Steel reinforcement A_S = 2946 mm^2

Interval Young's modulus of Steel reinforcement $\mathbf{E_s}$ = [0.98, 1.02]*2×10^5 N/mm^2

d) Case 4:
External interval moment $\mathbf{M}$ = [90,110] kNm
Interval area of Steel reinforcement $\mathbf{A_s}$ = [0.98, 1.02]*2946 mm^2
Interval Young's modulus of Steel reinforcement $\mathbf{E_s}$ = [0.98, 1.02]*2×10^5 N/mm^2

6. Results and Discussion

A web application is developed by the authors and is posted at the URL http://calculus.math.utep.edu/~andrzej/php/concrete-beam . Computations are performed using this web application. The screen capture of the web application is shown in Figure 5. The screen capture of the results obtained using this web application is shown in Figure 6.

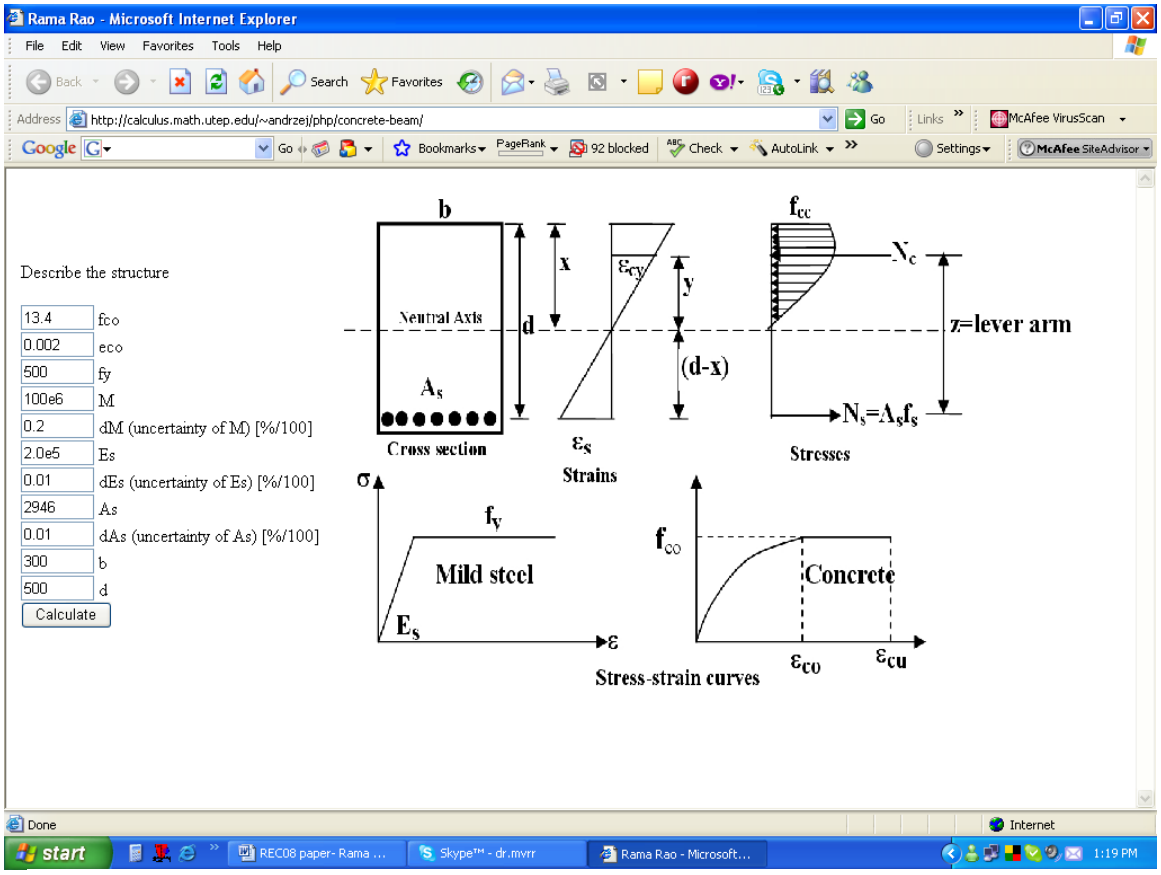


Figure 5. Web Application

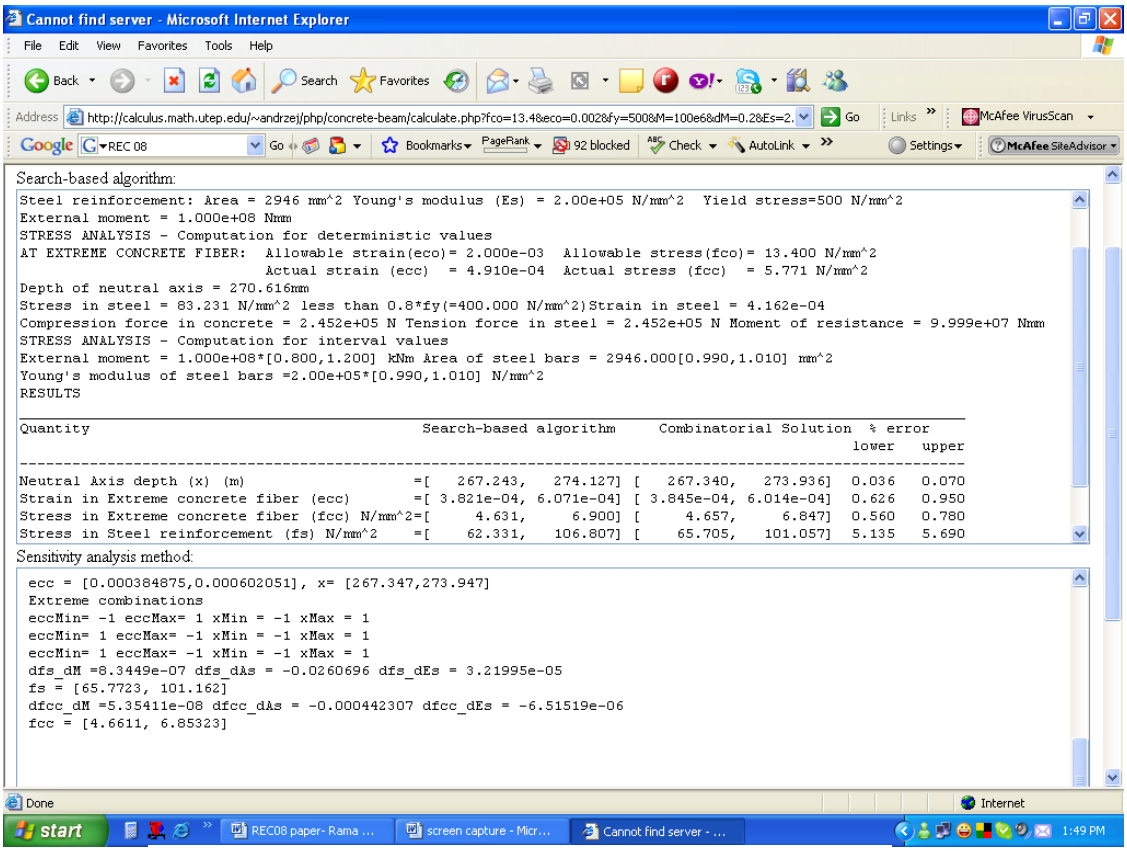


Figure 6. Results obtained using web application

Tables 1 *through* 4 present the results obtained using these three approaches viz. search-based algorithm, sensitivity analysis and combinatorial solution. Relative difference is computed for results obtained using search-based algorithm and sensitivity analysis in comparison with results obtained using combinatorial approach.

Table 1 presents the results obtained for Case 1. It is observed that the relative difference computed for search-based algorithm range from 0.003 percent to 0.416 percent while the corresponding relative difference computed using sensitivity analysis ranges from 0.0 percent to 0.005 percent. Table 2 presents the results obtained for Case 2. It is observed that the relative difference computed for search-based algorithm range from 0.179 percent to 2.2 percent while the corresponding relative difference computed using sensitivity analysis is almost zero. Table 3 presents the results obtained for Case 3. It is observed that the relative difference computed for search-based algorithm range from 0.039 percent to 7.567 percent while the corresponding relative difference computed using sensitivity analysis while the

corresponding relative difference computed using sensitivity analysis is almost zero. Table 4 presents the results obtained for Case 4. It is observed that the relative difference computed for search-based algorithm range from 0.061 percent to 4.701 percent while the corresponding relative difference computed using sensitivity analysis is almost zero. Thus it is observed that the relative difference is very small. Thus these methods agree very well with the combinatorial solution.

Table. 1 Comparison of results obtained using the three approaches for M = [96,104]kNm

	$\varepsilon_{cc} \times 10^{-4}$		f_{cc} (N/mm^2)		**x** (mm)		f_s (N/mm^2)	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Combinatorial	4.699	5.123	5.557	5.985	270.291	270.945	79.870	86.612
Search based approach	4.704	5.117	5.562	5.980	270.299	270.937	79.543	86.972
% difference	0.106	0.117	0.090	0.084	0.003	0.003	0.409	0.416
Sensitivity Analysis	4.699	5.122	5.557	5.985	270.291	270.946	79.871	86.614
% difference	0.002	0.005	0.001	0.011	0.000	0.000	0.001	0.005

Table. 2 Comparison of results obtained using the three approaches for M = [90,110] kNm , A_s = [0.9,1.1]*2946mm^2

	$\varepsilon_{cc} \times 10^{-4}$		f_{cc} (N/mm^2)		**x** (mm)		f_s (N/mm^2)	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Combinatorial	4.279	5.601	5.121	6.454	261.07	279.374	68.467	101.149
Search based approach	4.261	5.631	5.102	6.483	260.693	279.874	67.029	103.374
% difference	0.421	0.536	0.371	0.449	0.144	0.179	2.100	2.200
Sensitivity Analysis	4.279	5.601	5.121	6.454	261.07	279.374	68.467	101.149
% difference	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table. 3 Comparison of results obtained using the three approaches for M = [80,120] kNm and E_s = [0.98,1.02]*2946mm^2

	$\varepsilon_{cc} \times 10^{-4}$		f_{cc} (N/mm^2)		x (mm)		f_s (N/mm^2)	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Combinatorial	3.848	6.021	4.661	6.853	267.338	273.939	66.339	100.286
Search based approach	3.821	6.071	4.631	6.901	267.235	274.119	61.707	107.875
% difference	0.702	0.830	0.644	0.700	0.039	0.066	6.982	7.567
Sensitivity Analysis	3.848	6.021	4.661	6.853	267.338	273.939	66.339	100.286
% difference	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table. 4 Comparison of results obtained using the three approaches for M = [96,104] kNm, A_s = [0.98,1.02]*2946mm^2 E_s = [0.98,1.02]*2.0e5kN/m^2

	$\varepsilon_{cc} \times 10^{-4}$		f_{cc} (N/mm^2)		x (mm)		f_s (N/mm^2)	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Combinatorial	4.651	5.178	5.507	6.040	266.935	274.239	78.303	88.379
Search based approach	4.643	5.188	5.499	6.051	266.771	274.423	74.797	92.534
% difference	0.172	0.193	0.145	0.182	0.061	0.067	4.477	4.701
Sensitivity Analysis	4.651	5.178	5.507	6.040	266.935	274.239	78.303	88.379
% difference	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

6.1 COMBINED MEMBERSHIP FUNCTION FOR STRESSES AND STRAINS

Combined membership functions are plotted for stresses and strains in concrete and steel as well as neutral axis depth using the three approaches viz. search-based algorithm, sensitivity analysis and combinatorial approach. These membership functions are obtained using the procedure suggested by Moens and Vandepitte (2005). Figure 7 shows the plots of membership function for the depth of neutral axis. Combined membership functions for the strain and stress in extreme concrete fiber are presented in Figure 8 and Figure 9 respectively. Combined membership function for the stress in steel reinforcement is shown in Figure 10. It is observed that all these membership functions are triangular with linear variation

of the response about the corresponding mean value. The plots of combined membership functions obtained using search-based approach and sensitivity analysis agree well with the membership functions plotted using combinatorial approach.

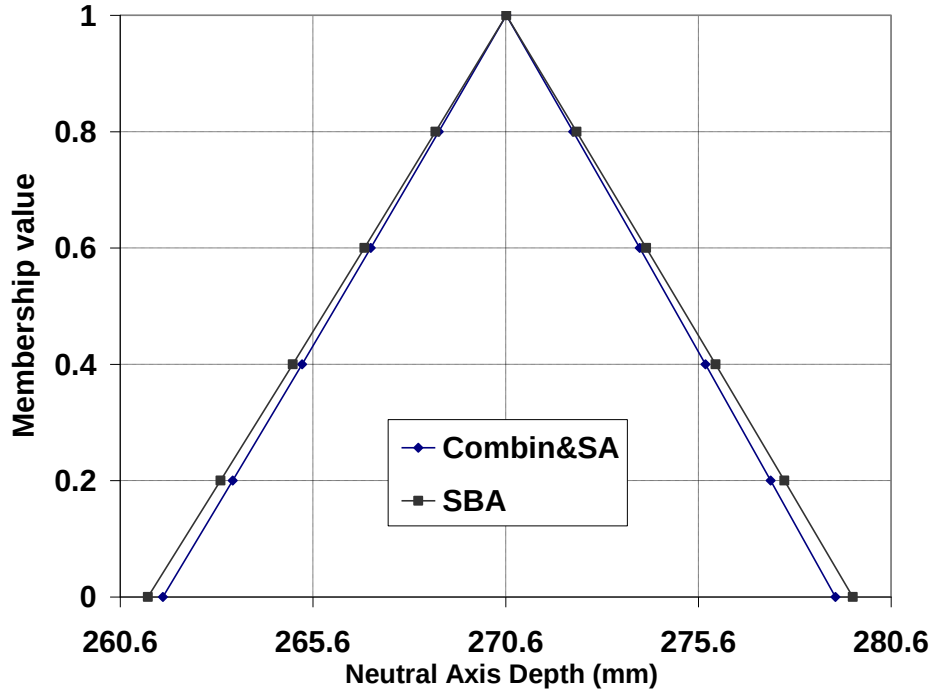


Figure .7 Combined membership function for neutral axis depth(x)

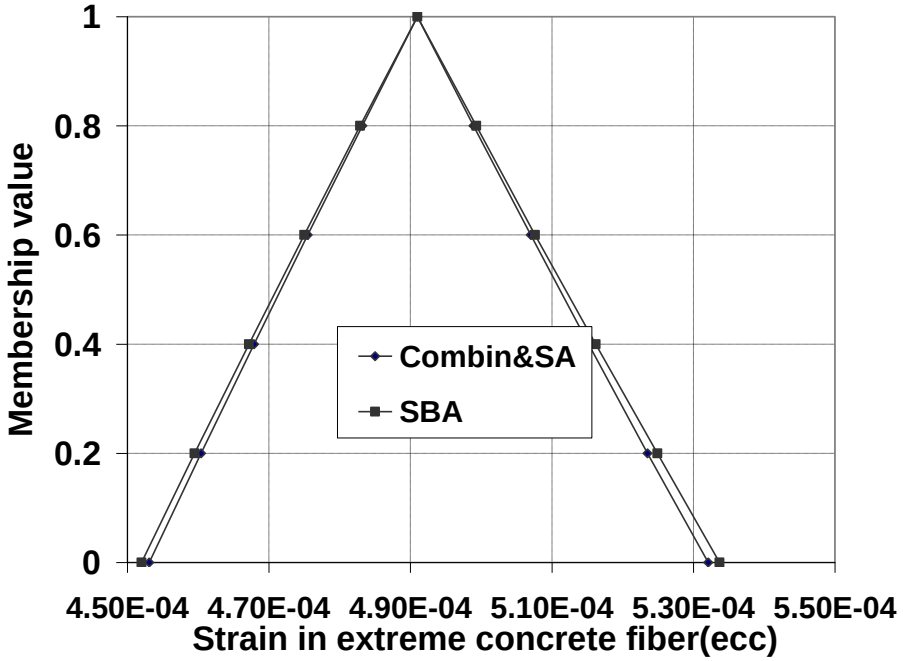


*Figure .*8 Combined membership function for strain (ε_{cc}) in extreme concrete fiber

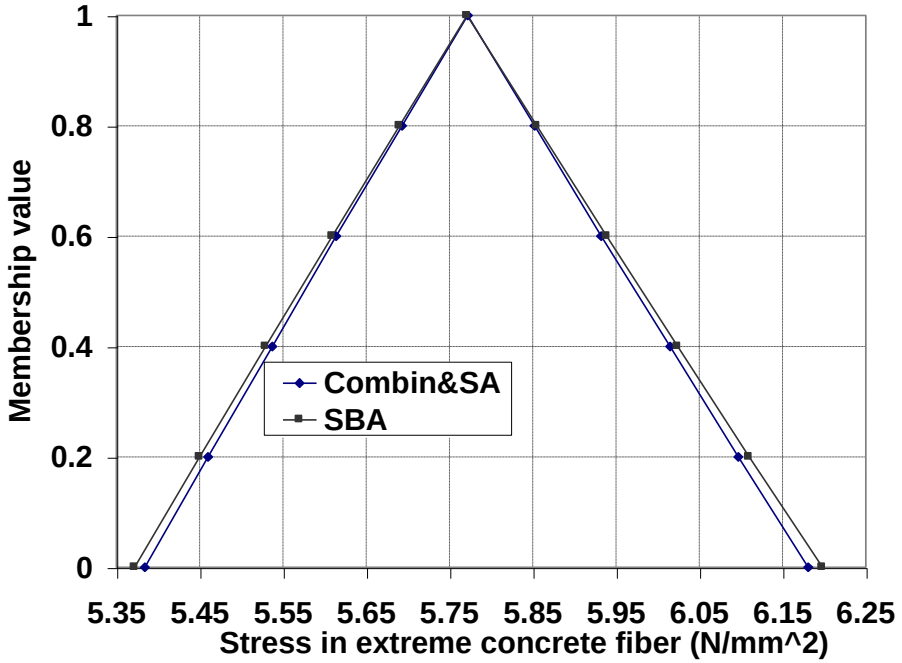


Figure 9. Combined membership function for stress (f_{cc}) in extreme concrete fiber

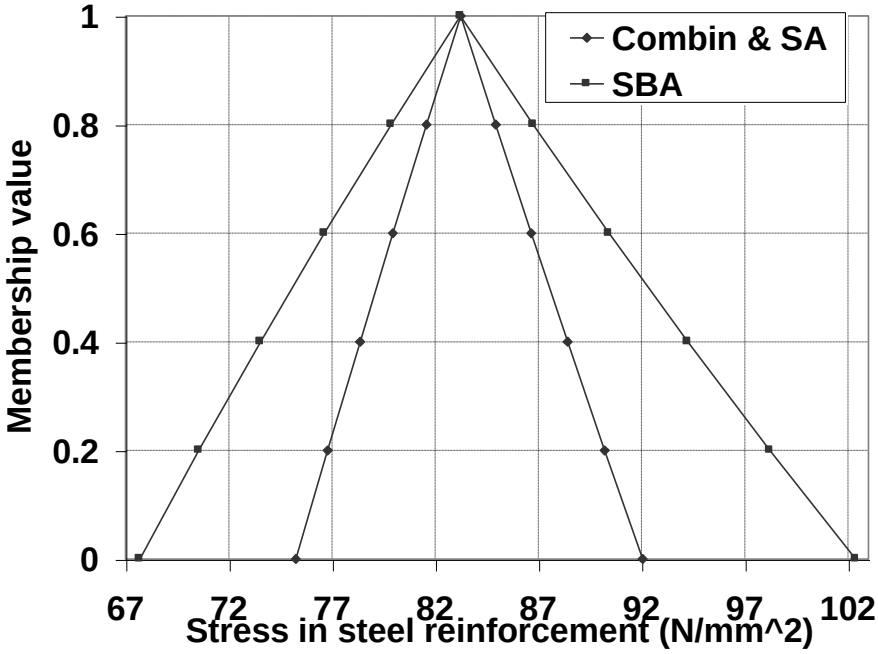


Figure .10 Combined membership function for stress (f_{cc}) in steel reinforcement

7. Conclusions

In the present paper, analysis of stresses in the cross section of a singly reinforced beam with interval values of area of steel reinforcement with corresponding interval Young's modulus and subjected to an interval external bending moment is taken up. The stress analysis is performed by three approaches viz. a search based algorithm and sensitivity analysis and combinatorial approach. It is observed that the results obtained are in excellent agreement. These approaches allow the designer to have a detailed knowledge about the effect of uncertainty on the stress distribution of the beam. The combined membership functions are plotted for neutral axis depth and stresses in concrete and steel and are found to be triangular.

Interval stresses and strains are also calculated using sensitivity analysis. The sign of the derivatives at the mid point and also at the endpoints is found to be same thus establishing the validity of the solution. More accurate monotonicity tests based on second and higher order derivatives (Pownuk, 2004) can also be used to establish sharp bounds on interval solution. Results with guaranteed accuracy can also be calculated using interval global optimization (Hansen, 1992 and Neumaier, 1990). Initial efforts by the authors in this direction gave encouraging results. Extended version of this paper will be published on the web page of the Department of Mathematical Science at the University of Texas at El Paso (http://www.math.utep.edu/preprints/).

8. References

Ben-Haim,Y., Elishakoff, I. Convex *Models of Uncertainty in Applied Mechanics*. Elsevier Science Publishers, New York, 1990.

Desai, P. and Krishnan, S. (1964), "Equation for stress strain curves of concrete", ACI J., 61(3), 345-350.

Hansen, E. *Global Optimization Using Interval Analysis*. Dekker, New York, 1992.

Hansen,E. and. Walster,G.W. Sharp Bounds on Interval Polynomial Roots. *Reliable Computing*, 8: 115–122, Kluwer Academic Publishers, Netherlands, 2002.

Hognestad, E., Hanson, N. W. and McHenry, D. Concrete stress distribution in ultimate strength design. *ACI Journal*, 52(6): 455-479, 1955.

*Indian Standard Code for Plain and Reinforced Concrete (*IS 456-2000*)*, Bureau of Indian standards, India, Fourth Revision, 2000.

Jansson, C. Interval linear systems with symmetric matrices, skew- symmetric matrices and dependencies in the right hand side. *Computing*, 46(3):265–274, 1991.

Köylüoglu, H.U., Cakmak, A. and Nielsen, S.R.K. Interval mapping in structural mechanics, In*: Spanos, ed. Computational Stochastic Mechanics*, Balkema, 125–133, 1995.

McWilliam, S. Anti-optimization of uncertain structures using interval analysis. *Computers and Structures*, 79:421–430, 2000.

Moens, D. and Vandepitte, D. A fuzzy finite element procedure for the calculation of uncertain frequency response functions of damped structures. *Journal. of Sound and Vibration*, 288 : 431-462,2005.

Moore, R.E. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, New York, 1966.

Muhanna,R. and Mullen, R.L. Uncertainty in mechanics problems - interval-based approach. *Journal of Engineering Mechanics*, ASCE, 127(6):557–566, 2001.

Neumaier A., *Interval Methods for Systems of Equations. Encyclopedia of Mathematics and its Applications*, 37, Cambridge Univ. Press, Cambridge, 1990.

Neumaier A., *Global Optimization*, http://www.mat.univie.ac.at/~neum/glopt.html

Neumaier A. and Pownuk A. Linear systems with large uncertainties with applications to truss structures. *Reliable Computing*, 13(1):149–172, 2007.

Orisamolu I.R., Gallant ,B.K., Akpan U.O., Koko.T.S. Practical fuzzy finite element analysis of structures. *Finite Elements in Analysis and Design*, 38:93-111,2000.

Pownuk, A. *Numerical solutions of fuzzy partial differential equation and its application in computational mechanics, Fuzzy Partial Differential Equations and Relational Equations: Reservoir Characterization and Modeling* (M. Nikravesh, L. Zadeh and V. Korotkikh, eds.). *Studies in Fuzziness and Soft Computing*, Physica-Verlag, , pp. 308-347,2004.

Purushottaman, P. *Reinforced Concrete Structural Elements-Behaviour, Analysis and Design*. Tata McGraw-Hill Publishing Company Limited, New Delhi, India, 1986.

Rama Rao, M.V. and Ramesh Reddy, R. Fuzzy finite element analysis of structures with uncertainty in load and material properties. *Journal of Structural Engineering*, Vol. 33(2): 129-137, 2006.

Rama Rao, M.V., Pownuk, A. Stress distribution in a reinforced concrete flexural member with uncertain structural parameters. *The University of Texas at El Paso, Department of Mathematical Sciences Research Reports Series, Texas Research Report No. 2007-05*, El Paso, Texas, USA,2007.

Rao, S.S. and Chen, L. Numerical solution of fuzzy linear equations in engineering analysis. *International Journal for Numerical Methods in Engineering*, 43:391–408, 1998.

Skalna ,I., Rama Rao, M.V, Pownuk, A, "Systems of fuzzy equations in structural mechanics", *The University of Texas at El Paso, Department of Mathematical Sciences Research Reports Series, Texas Research Report No. 2007-01*, El Paso, Texas, USA,2007.

Smith G. M. and Young, L. E. "Ultimate flexural analysis based on stress strain curves of cylinders", *ACI Journal*, 53(6): 597-609, 1955.

Zadeh, L.A. Fuzzy sets, *Information and Control*, 8: 338-353, 1965.

Evaluation of Inconsistent Engineering data

Michael Beer
Department of Civil Engineering
National University of Singapore
BLK E1A #07-03, 1 Engineering Drive 2, Singapore 117576
email: cvebm@nus.edu.sg

Abstract. In this paper options for a realistic evaluation of engineering data characterized by inconsistency regarding uncertainty and imprecision are discussed. The proposed methods are linked to the generalized uncertainty model fuzzy randomness. This enables a quantification of uncertainty and imprecision simultaneously with a smooth transition between fuzziness and randomness. Statistical information is exploited with traditional statistical methods, whereas imprecision is dealt with using fuzzy methods. Statistical uncertainty and imprecision are considered within the same model but not mixed with one another. In this manner, both components are reflected separately in the computational results from a subsequent structural or safety analysis. Quantification techniques are elucidated for three typical engineering cases of inconsistent information; (i) small sample size and expert knowledge, (ii) imprecise sample elements, and (iii) inconsistent environmental conditions and expert knowledge. The usefulness of the proposed quantification methods for a subsequent structural analysis and safety assessment is demonstrated by way of engineering examples.

Keywords: Inconsistent data; Imprecise data; Fuzzy methods; Fuzzy probabilities; Uncertain structural analysis; Safety assessment.

1. Introduction

The usefulness of the results from an engineering analysis depends significantly on the realistic modeling of the input parameters. Shortcomings, in this regard, may lead to biased computational results, wrong decisions, and serious consequences [18]. This applies, in particular, if the data are characterized by uncertainty and imprecision. A variety of mathematical models have been formulated to take account of the available information as realistically as possible [3, 6, 7, 10, 13, 14, 15, 17, 23, 24, 28, 29]. The usefulness and capabilities of these models have already been demonstrated in the solution of practical problems, for example, in civil/mechanical engineering [1, 4, 5, 8, 9, 11, 12, 14, 16, 19, 21, 22, 25].

In engineering practice the available information frequently appears as partly stochastic and partly imprecise – in a mixed stochastic/non-stochastic form. In those cases the model fuzzy randomness [19] provides a proper basis to utilize traditional statistical methods together with quantification methods from fuzzy set theory. In this manner, a broad spectrum of typical engineering cases can be covered; and the introduction of unwarranted information is avoided. This is demonstrated in the sequel with proposals of quantification techniques for three typical engineering situations. First, the quantification of data from a small sample together with expert knowledge is considered. This is associated with the problem of weak statistical information from estimations and tests. A solution is obtained by utilizing the statistical imprecision in the specification of fuzzy parameters and fuzzy distribution types of a fuzzy random quantity.

Second, samples with imprecise elements are evaluated, which requires the application of statistics with fuzzy quantities. For this purpose, fuzzy arithmetic is implemented in statistical estimations and tests. Third, inconsistent environmental conditions are dealt with together with expert knowledge. This leads to critical conditions for statistical estimations and tests. For solution, a separation of fuzziness and randomness is applied in the quantification procedure by constructing groups of consistent data.

In all three cases fuzzy random quantities are obtained which reflect the stochastic uncertainty and the imprecision of the underlying information simultaneously and separately. The fuzzy probability distributions are described as a bunch of distributions that cover all possible stochastic models within the range of imprecision. Bunch parameters are fuzzy quantities $\tilde{p}_t$, which include distribution parameters as well as parameters for the specification of the distribution type. Then, each crisp point from the $\tilde{p}_t$ specifies one real-valued random quantity associated with a certain membership degree according to fuzzy set theory. For a detailed description see [19]. This enables the utilization and combination of sophisticated and numerically efficient methods from stochastic mechanics [26, 27] and from interval [22] and fuzzy structural analysis [20] in subsequent engineering computations. The respective algorithms of fuzzy stochastic structural analysis and safety assessment are discussed in [19].

2. Small Sample size and Expert Knowledge

Assume that a concrete sample of small size is available. The sample elements are random realizations. The available information on the sample is insufficient, however, to describe a real-valued random variable free of doubt. The type of the distribution function and the parameters cannot be determined uniquely; additional uncertainty exists. Expert knowledge and experience are available from similar cases in the past. This uncertainty is rather non-stochastic and may be accounted for with the aid of fuzzy set theory [2, 30]. Statistical methods may be used as a basis for quantification, which are supplemented by fuzzy methods to finalize the modeling. Depending on the available information it is possible to formulate an imprecise parametric or nonparametric estimation problem. On this basis, the type and the parameters of the sought distribution are determined in as imprecise quantities, namely, as fuzzy quantities. These fuzzy quantities are, subsequently, lumped together as fuzzy parameters $\tilde{p}_t(\tilde{X})$, in which $\tilde{X}$ represents a fuzzy random quantity – for convenience, limited to the one-dimensional case. The $\tilde{p}_t(\tilde{X})$ may be determined from imprecise empirical statistical information extracted from the sample together with expert knowledge.

If, for example, the type of distribution is known with sufficient certainty, this implies an imprecise, parametric estimation problem. The sample functions applied in statistical methods yield more or less acceptable estimation values for the parameters of a distribution. In order to take account of the imprecision of the estimator, confidence intervals may be determined for the estimator in question. The probabilistic propositions for confidence intervals applied in statistical methods may then serve as additional information for the specification of the membership functions $\mu(p_t(X))$ of the $\tilde{p}_t(\tilde{X})$ in the present case. Expert knowledge is brought in with regard to

- the specification of the distribution type,
- the choice of the estimator,
- the construction of confidence intervals (type and levels),

– the assignment of membership degrees to the selected confidence levels, and

– the subsequent modification of the initial draft of the membership functions $\mu(p_t(X))$.

Table I. Sample of the cylinder compressive strength f_c of a concrete

Number i of realization	Compressive strength $x_i = f_{ci}[N/mm^2]$	Number i of realization	Compressive strength $x_i = f_{ci}[N/mm^2]$
1	28.3	11	26.8
2	31.5	12	35.3
3	35.2	13	26.3
4	29.8	14	23.1
5	27.6	15	20.2
6	30.7	16	29.2
7	25.2	17	25.7
8	34.6	18	34.2
9	28.9	19	24.8
10	19.2	20	22.8

Table II. Statistical estimation and assignment of membership values for $\tilde{m}_x$ and $\tilde{\sigma}_x$

Estimation	Confidence level	m_X	σ_X	α-level
Point	- –	27.97	4.75	1.00
Interval	0.50	[27.24, 28.70]	[4.35, 5.43]	0.75
	0.75	[26.71, 29.23]	[4.05, 5.92]	0.50
	0.90	[26.13, 29.81]	[3.77, 6.52]	0.25
	0.99	[24.93, 31.01]	[3.34, 7.92]	0.00

Suppose that a sample of size 20 is available for the cylinder compressive strength f_c of a concrete according to Table 2. A normal distribution is assumed based on expert knowledge, and the parameters m_x and σ_x are determined as fuzzy values $\tilde{m}_x$ and $\tilde{\sigma}_x$. For this purpose interval estimations are applied. From the 20 measured values of the compressive strength the central confidence intervals for the confidence levels 0.50, 0.75, 0.90, and 0.99 are determined. Dependencies between the parameters are not taken into account. Additionally, common point estimations are used to specify crisp values for the expected value (as the mean value of the sample) and the standard deviation (based on the sample variance). The results (Table 2) are then taken as a basis for the specification of the parameters as fuzzy quantities. Membership values are assigned to the estimation results by subjective assessment. That is, the confidence intervals are interpreted as being α-level sets of the fuzzy values $\tilde{m}_x$ and $\tilde{\sigma}_x$; see Table 2. The mean values of the fuzzy numbers are taken

from the point estimations. Eventually, the fuzzy quantities $\tilde{m}_x$ and $\tilde{\sigma}_x$ are obtained according to Fig. **??**. As dependencies between the parameters in the interval estimations are neglected, interaction between $\tilde{m}_x$ and $\tilde{\sigma}_x$ is not obtained.

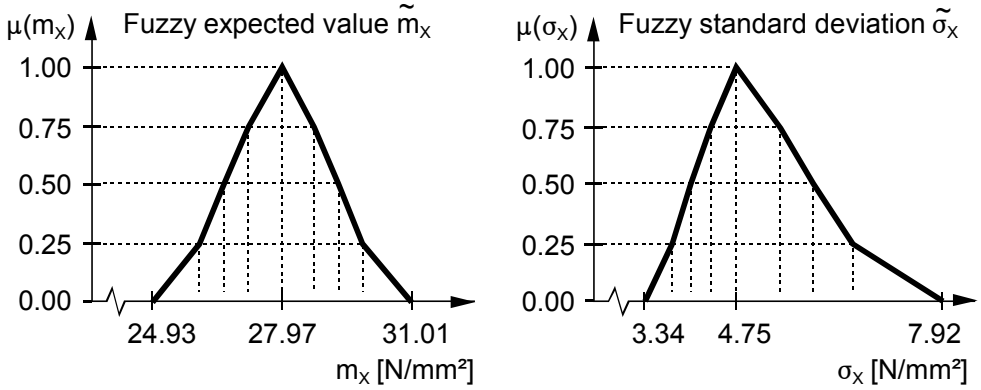


Figure 1. Fuzzy expected value $\tilde{m}_x$ and fuzzy standard deviation $\tilde{\sigma}_x$

3. Imprecise Sample Elements

Imprecision of sample elements may occur, for example, due to imprecise readings of (analog) measuring devices or as a reflection of imprecise individual care of personnel in tests. This imprecision can be expressed in form of fuzzy numbers for the measured values representing the sample elements. It is then possible to construct a fuzzy random quantity $\tilde{X}$ directly from the imprecise data material. The corresponding fuzzy parameters $\tilde{p}_t(\tilde{X})$ for the description of the fuzzy random quantity $\tilde{X}$ can be estimated based on statistical estimations and tests extended to deal with fuzzy arguments. This requires a proper application of fuzzy arithmetic in these algorithms. For a numerical evaluation, the fuzzy analysis based on α-level optimization according to [20] may be utilized. This framework enables an implementation of algorithms of mathematical statistics as the mapping model of a fuzzy analysis. Each fuzzy sample element is then treated as a fuzzy input quantity of the mapping model. The fuzzy result represents the sought parameter $\tilde{p}_t(\tilde{X})$.

As an example, the sample elements from Table 2 are assumed to possess an imprecision of $\pm 2\,\mathrm{N/mm^2}$ due to imprecise readings of the measuring device. This provides information for a modeling of the sample elements as fuzzy triangular numbers denoted by $\tilde{x}_i =< x_{i\,\mu=0l}, x_{i\,\mu=1}, x_{i\,\mu=0r} >$. The values from Table 2 are assessed with $\mu = 1$, from where the linear branches of the membership function decrease down to $\mu = 0$ at the points of the maximum deviation $\pm 2\,\mathrm{N/mm^2}$; see Table 3.

In order to compute the empirical parameters, common statistics (sample functions) are applied with the fuzzy values $\tilde{x}_i$ as arguments. The fuzzy sample mean is then obtained with

$$\tilde{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i\,, \tag{1}$$

in which n is the sample size. The linearity of this mapping model leads to a fuzzy triangular number for the fuzzy sample mean, which is completely specified by the membership levels $\mu = 1$ and $\mu = 0$ as shown

in Fig. 2, $\tilde{\bar{x}} = < 25.97, 27.97, 29.97 > \mathrm{N/mm^2}$. In contrast to this, the mapping model for computing the standard deviation of the sample is nonlinear and even non-monotonic,

$$\tilde{s}_x = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n}\tilde{x}_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\tilde{x}_i\right)^2\right]}. \tag{2}$$

This requires a more sophisticated evaluation technique. In the example, α-level optimization [20] is applied

Table III. Fuzzy sample elements of the cylinder compressive strength f_c of a concrete

Number i of fuzzy realization	Fuzzy compressive strength $\tilde{x}_i = \tilde{f}_{ci}[\mathrm{N/mm^2}]$	Number i of fuzzy realization	Fuzzy compressive strength $\tilde{x}_i = \tilde{f}_{ci}[\mathrm{N/mm^2}]$
1	$< 26.3, 28.3, 30.3 >$	11	$< 24.8, 26.8, 28.8 >$
2	$< 29.5, 31.5, 33.5 >$	12	$< 33.3, 35.3, 37.3 >$
3	$< 33.2, 35.2, 37.2 >$	13	$< 24.3, 26.3, 28.3 >$
4	$< 27.8, 29.8, 31.8 >$	14	$< 21.1, 23.1, 25.1 >$
5	$< 25.6, 27.6, 29.6 >$	15	$< 18.2, 20.2, 22.2 >$
6	$< 28.7, 30.7, 32.7 >$	16	$< 27.2, 29.2, 31.2 >$
7	$< 23.2, 25.2, 27.2 >$	17	$< 23.7, 25.7, 27.7 >$
8	$< 32.6, 34.6, 36.6 >$	18	$< 32.2, 34.2, 36.2 >$
9	$< 26.9, 28.9, 30.9 >$	19	$< 22.8, 24.8, 26.8 >$
10	$< 17.2, 19.2, 21.2 >$	20	$< 20.8, 22.8, 24.8 >$

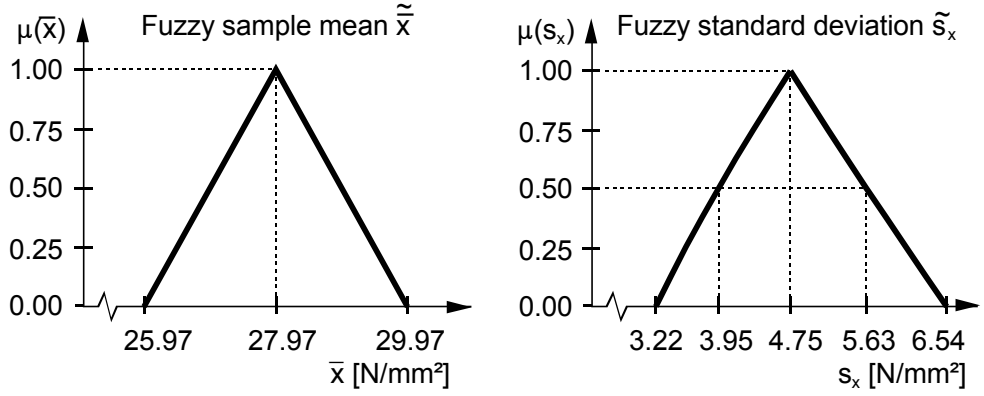


Figure 2. Fuzzy mean $\tilde{\bar{x}}$ and fuzzy standard deviation $\tilde{s}_x$ of the sample from Table 3

to evaluate Eq. (2). The membership function $\mu(s_x)$ is obtained with nonlinear branches; see Fig. 2.

The fuzzy sample elements $\tilde{x}_i$ enter Eq. (1) and Eq.(2), simultaneously. Thus, a relationship exists between the fuzzy sample mean and the fuzzy standard deviation of the sample. This is referred to as interaction

between the fuzzy quantities $\tilde{\bar{x}}$ and $\tilde{s}_x$. This interaction is shown in Fig. 3 for the membership level $\alpha = 0$. Certain combinations of crisp values from $\tilde{\bar{x}}$ and $\tilde{s}_x$ cannot appear. An analytical or numerical determination of this interaction is, however, virtually excluded due to the tremendous computational effort even for a small sample. In the example a numerical approximation solution was determined with the aid of systematic and random-oriented simulations. The effect of the number of fuzzy realizations on the interaction relationship becomes apparent when only the first seven sample elements from Table 3 are considered; see Fig. 4. Not only the position but also the shape of the fuzzy set $\{\tilde{\bar{x}}, \tilde{s}_x\}$ shows a deviation from the illustration in Fig. 3. As a consequence of the same support widths of the fuzzy realizations $\tilde{x}_i$ the minimum and maximum sample means are, in each case, coupled with the same standard deviation of the sample. This property is lost in the general case. As demonstrated for $\tilde{\bar{x}}$ and $\tilde{s}_x$, interaction generally exists between all empirical parameters including the distribution type. The fact that the fuzzy realizations themselves may also be interactive may even lead to non-connected sets for the empirical parameters. Due to the numerical complications in the determination of the interaction, an approximation may be pursued. Or, the interaction may even be neglected; see Fig. 3. Although this means that non-justified parameter combinations are included and thus enter subsequent computations, the "exact" solution is completely contained in this approximation. The negligence of interaction leads to an envelope curve of those parameter combinations, which can actually appear.

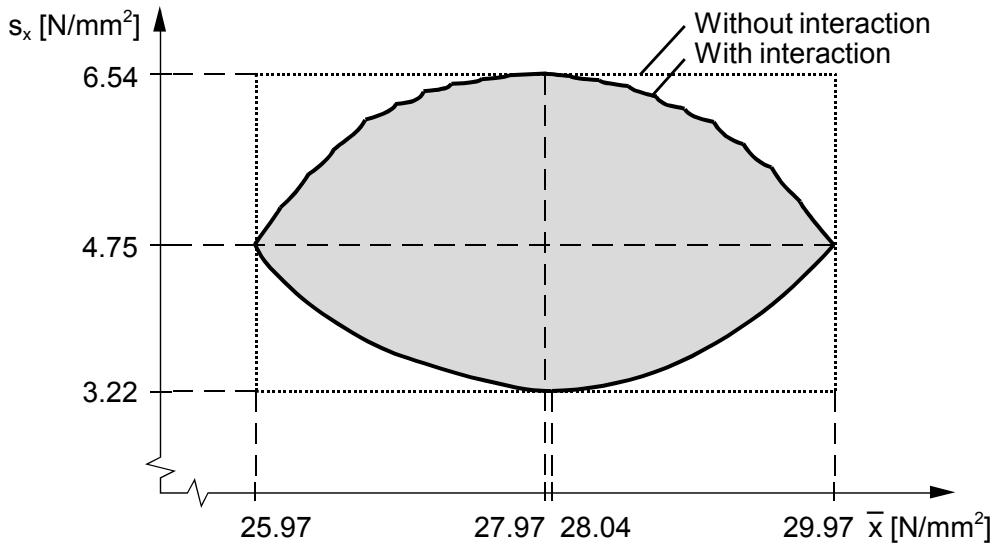


Figure 3. Numerical approximation of the interaction between the fuzzy sample mean $\tilde{\bar{x}}$ and the fuzzy standard deviation $\tilde{s}_x$ for the 20 fuzzy realizations from Table 3

The fuzzy parameters computed from the sample are the basis for the specification of the fuzzy probability distribution function needed for further processing of fuzzy random quantities in engineering computations. In the example, a normal distribution is assumed for the fuzzy random quantity. The functional parameters are then estimated by the fuzzy sample mean $\tilde{\bar{x}}$ as fuzzy expected value $\tilde{m}_x$, and by the fuzzy standard deviation $\tilde{s}_x$ of the sample as fuzzy standard deviation $\tilde{\sigma}_x$ of the fuzzy random quantity. The obtained fuzzy probability density function $\tilde{f}(x)$ and the fuzzy probability distribution function $\tilde{F}(x)$ are

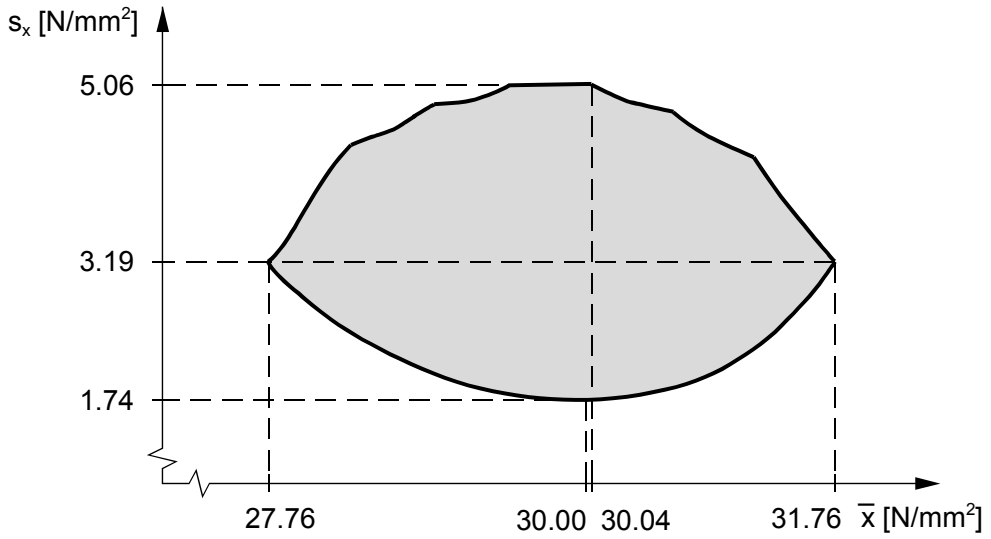


Figure 4. Numerical approximation of the interaction between $\tilde{\bar{x}}$ and $\tilde{s}_x$ for the first seven fuzzy realizations from Table 3

shown in Figs. 5 and 6, respectively. The illustrations show the functions with and without the consideration of the interaction between $\tilde{m}_x$ and $\tilde{\sigma}_x$. Negligence of the interaction between $\tilde{m}_x$ and $\tilde{\sigma}_x$ leads to envelope curves enclosing the exact fuzzy functions $\tilde{f}(x)$ and $\tilde{F}(x)$. The interaction between $\tilde{m}_x$ and $\tilde{\sigma}_x$ excludes the simultaneous occurrence of extrema of the expected value and standard deviation; see Fig. 3. This influences, in particular, the tails of the fuzzy functions $\tilde{f}(x)$ and $\tilde{F}(x)$. The probability mass in the tails is higher if the interaction is neglected. This leads to an overestimation of failure probabilities in a subsequent structural safety assessment. This overestimation is, however, not tremendous and leads to a slightly conservative safety assessment, which is rather welcome.

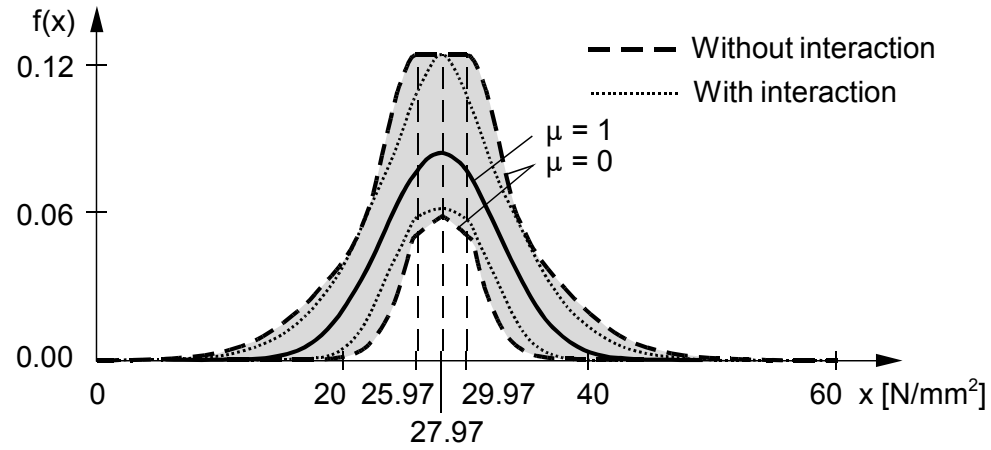


Figure 5. Fuzzy probability density function $\tilde{f}(x)$ with and without consideration of the interaction between $\tilde{m}_x$ and $\tilde{\sigma}_x$

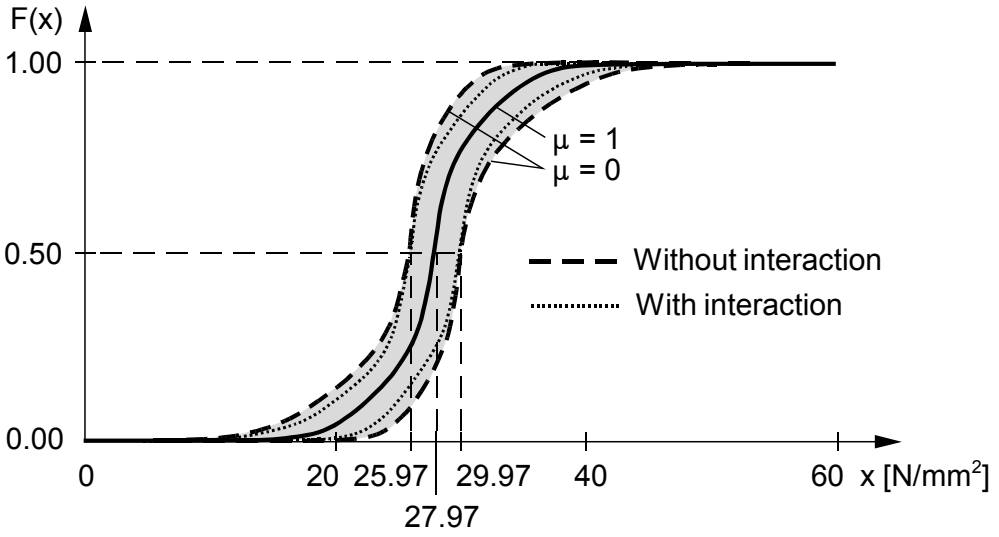


Figure 6. Fuzzy probability distribution function $\tilde{F}(x)$ with and without consideration of the interaction between $\tilde{m}_x$ and $\tilde{\sigma}_x$

4. Inconsistent Environmental Conditions and Expert Knowledge

This situation appears if the sample has been generated under varying environmental conditions. It then defies a traditional statistical evaluation and needs special treatment. The varying environmental conditions may include, for example, involvement of different manufacturers, changes in the type of aggregates / additives from different suppliers, varying hardening conditions (temperature, humidity), and variations in the motivation of the personnel. In those cases, expert knowledge is usually available to separate fuzziness and randomness present in the statistical data material. This separation can be realized by characterizing the environmental conditions with attributes such as a specific supplier for aggregates or a certain team of employees in the production process. Observed realizations with the same attributes are lumped together in a single *group*. These groups are subsets of the population. Each group of realizations with the same attributes is treated as a separate sample. These samples can then be evaluated using statistical methods as they comply with the preconditions in form of constant environmental conditions. The statistical evaluation yields empirical parameter values including a distribution type for each group. For all groups the set $\underline{S}$ of statistical propositions is obtained. Each element of $\underline{S}$ is assigned to a subset of the population. Hence, the set $\underline{S}$ describes the set of real random quantities contained in the observed realizations. The differences between the elements of the set $\underline{S}$ represent imprecision, which may be modeled as fuzziness of the population. The elements contained in $\underline{S}$ and, thus, the associated real random quantities may be assessed with membership values. This results in the fuzzy set $\tilde{\underline{S}}$. The real random quantities together with their membership values form a fuzzy random quantity, which is described by $\tilde{\underline{S}}$.

The fuzzy set $\tilde{\underline{S}}$ can be constructed in parametric or in a non-parametric manner. The parametric construction of $\tilde{\underline{S}}$ involves a distribution assumption from expert knowledge. Then, the membership functions of the empirical distribution parameters may be constructed using histograms. In the non-parametric construction of $\tilde{\underline{S}}$ empirical distribution functions are used, and a direct fuzzification of the probability distribution function curve is pursued.

Parametric Quantification It is presumed that the groups of sample elements with same attributes and their corresponding empirical parameters are known. The parameter values constitute a sample for which a histogram is constructed. The parameter value is plotted along the abscissa, which is subdivided into subsets. In the normal manner the number of sample elements, which is the number of empirical parameter values, per subset is plotted on the ordinate. Then, the histogram can be used as a basis for constructing the membership function of the respective fuzzy parameter.

As an example, let specimens of a concrete be available from different concrete plants. Tests are carried out to measure the cylinder compressive strength f_c. The specimens are labeled, and the concrete plant and work team are registered. Specimens with the same identification (same attributes) are each lumped together in a group. In the example, twelve groups with a different number of specimens (sample size) are identified. By this means, randomness and fuzziness are separated. The statistical evaluation of the measured cylinder compressive strength f_c yields empirical parameters for each group. The sample mean $\overline{x}$ and the standard deviation s_x of the samples are computed; see Table 4.

Table IV. Sample mean $\overline{x}$ and standard deviation s_x of the cylinder compressive strength f_c of the concrete for twelve groups of specimens (twelve samples)

Label of group	Sample size	Sample mean $\overline{x}[\mathrm{N/mm^2}]$	Standard deviation $s_x[\mathrm{N/mm^2}]$
1	54	27.3	5.3
2	48	26.6	4.9
3	42	29.2	4.2
4	38	31.4	3.8
5	44	28.3	5.6
6	48	29.4	3.2
7	55	26.4	5.0
8	47	30.1	4.6
9	64	28.3	5.9
10	53	27.9	3.8
11	75	29.6	6.3
12	52	27.8	4.7

The values listed in Table 4 are used to construct histograms for the sample mean $\overline{x}$ and the standard deviation s_x of the samples; see Fig. 7. The chosen subset widths are $1.0\,\mathrm{N/mm^2}$ for $\overline{x}$ and $0.75\,\mathrm{N/mm^2}$ for s_x. Each of the empirical parameters is modeled using fuzzy triangular numbers. The method of least squares is applied to determine the linear membership functions. The derived fuzzification suggestions are shown in Fig. 7.

Due to the fact that the values $\overline{x}$ and s_x for each group originate from the same sample, interaction exists between the fuzzy quantities $\tilde{\overline{x}}$ and $\tilde{s}_x$. Analog to the analysis of stochastic dependencies between random variables, the interaction relationship may be determined by evaluating the value pairs $(\overline{x}, s_x)$ obtained. These pairs are plotted in a coordinate system, and the interaction relationship is estimated for different membership levels. This procedure is illustrated in Fig. 8 for the membership level $\alpha = 0$. Assuming a normal distribution, the empirical fuzzy parameters $\tilde{\overline{x}}$ and $\tilde{s}_x$ are adopted as the fuzzy distribution parameters

$\tilde{m}_x$ and $\tilde{\sigma}_x$, respectively, of the fuzzy probability distribution. If the assumed distribution type is different for the individual groups, this may be accounted for with a compound distribution and fuzzy parameter for the mixing ratio.

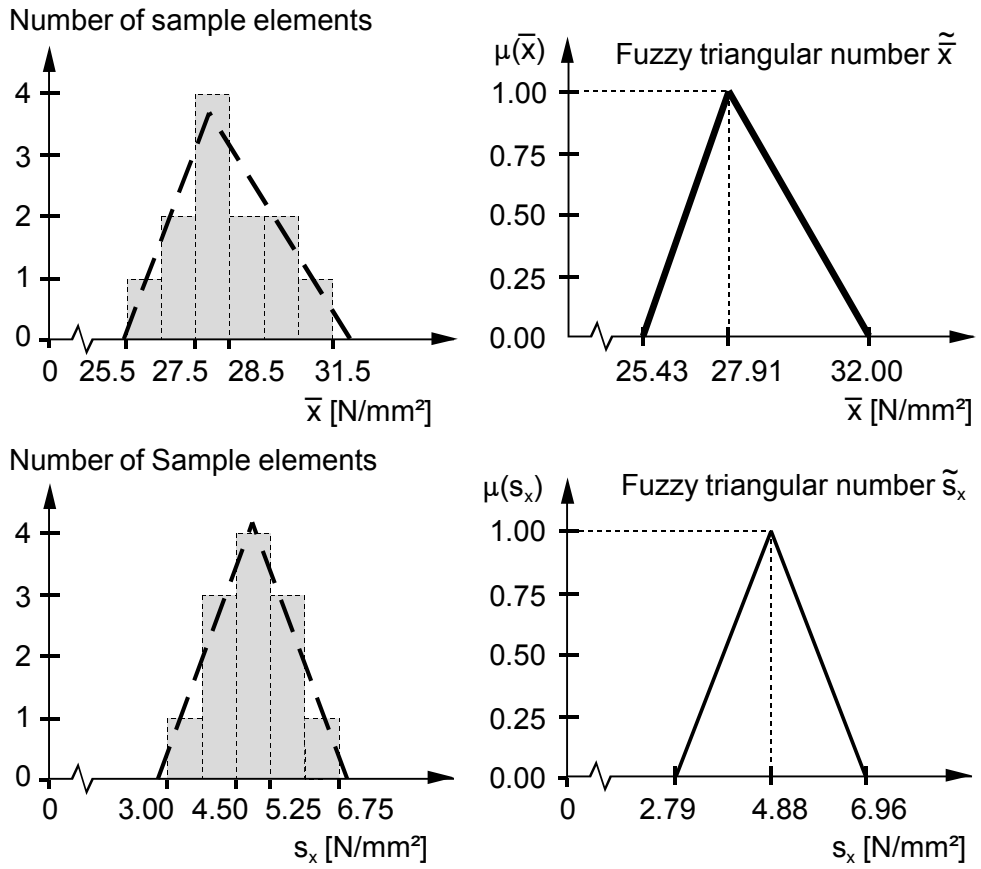


Figure 7. Histograms and fuzzification of the sample mean $\bar{x}$ and the standard deviation s_x assigned to the groups (samples) of the cylinder compressive strength f_c

Non-parametric Quantification The starting point is again the separation of randomness and fuzziness by constructing groups of observed realizations. Then, empirical distribution functions are constructed for the individual groups. The set of empirical distribution functions for all groups is then taken as the basis to determine fuzzy quantities for the functional values of an overall empirical distribution function.

The example from the parametric quantification is reused for demonstration. For each group, a histogram is constructed from the realizations to determine an empirical distribution function. The subset widths and the subset positioning on the abscissa must be the same for all histograms for all groups. The subsets are defined as half-closed intervals $[x_l, x_r)$ on the real number line. The number of observed realizations in the subsets is generally different for the individual groups. The histograms for the first two groups from Table 4 are shown in Fig. 9.

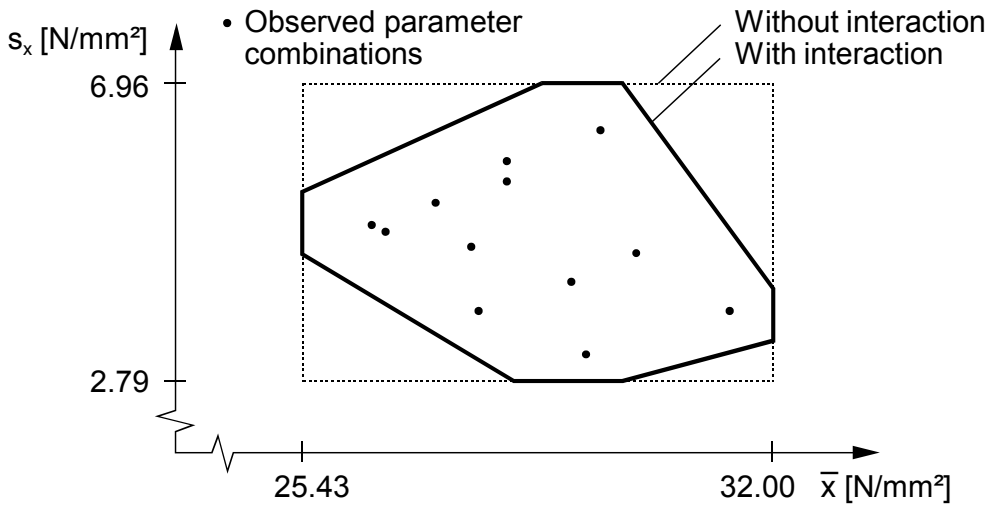


Figure 8. Estimation of the interaction between $\tilde{\bar{x}}$ and $\tilde{s}_x$

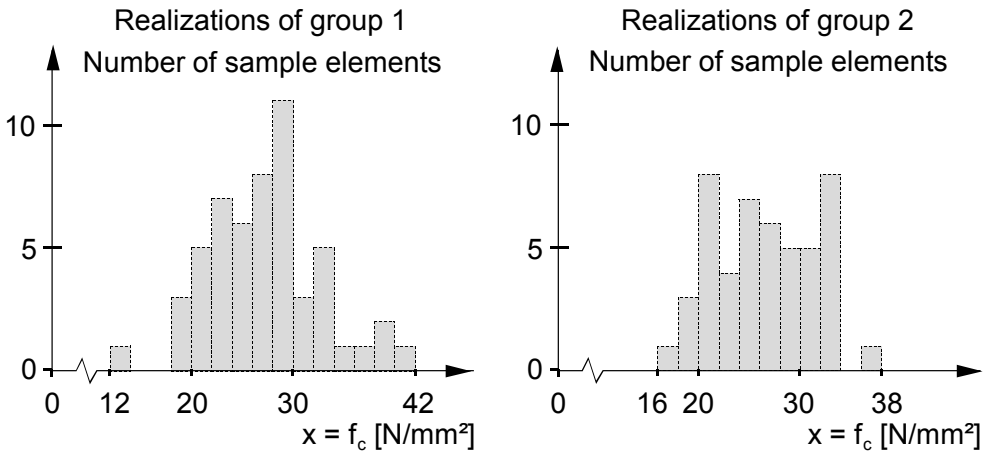


Figure 9. Histograms for the realizations of groups 1 and 2 from Table 4

For each group the empirical probability distribution function

$$F_i^e(x) = \frac{n_{i,k}(x)}{n_i} \tag{3}$$

is developed from the corresponding histogram. In the above, i denotes the group number, n_i is the number of all elements (realizations) in group i, and $n_{i,k}(x)$ is the number of those elements k (in group i), whose values x_k are smaller than x. The values x of the observed realizations are determined by the left-hand subset boundaries (that is, by the x_l of the half-closed intervals $[x_l, x_r)$) in the histograms; these mark discrete positions on the abscissa. The evaluation of all groups yields a bunch of discrete empirical distribution functions. The functional values $F_i^e(x = f_c)$ are listed in Table 4.

Table V. Functional values of the empirical distribution functions $F_i^e(x = f_c)$ for all groups i of specimens

	Group i											
$x = f_c$ $[N/mm^2]$	1	2	3	4	5	6	7	8	9	10	11	12
12	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
14	0.019	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.016	0.000	0.000	0.000
16	0.019	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.016	0.000	0.000	0.000
18	0.019	0.021	0.000	0.000	0.000	0.000	0.036	0.000	0.016	0.000	0.013	0.000
20	0.074	0.083	0.000	0.000	0.023	0.000	0.091	0.000	0.063	0.000	0.067	0.000
22	0.167	0.250	0.000	0.000	0.114	0.000	0.273	0.043	0.172	0.094	0.120	0.096
24	0.296	0.333	0.071	0.026	0.295	0.042	0.327	0.064	0.266	0.170	0.227	0.231
26	0.407	0.479	0.262	0.105	0.409	0.167	0.436	0.191	0.328	0.283	0.320	0.346
28	0.556	0.604	0.476	0.184	0.523	0.354	0.655	0.340	0.422	0.509	0.400	0.577
30	0.759	0.708	0.595	0.395	0.705	0.563	0.764	0.532	0.625	0.717	0.507	0.731
32	0.815	0.813	0.786	0.632	0.750	0.833	0.855	0.702	0.766	0.887	0.653	0.788
34	0.907	0.979	0.810	0.711	0.795	0.917	0.927	0.766	0.844	0.962	0.733	0.904
36	0.926	1.000	0.905	0.816	0.909	0.979	0.964	0.830	0.922	0.981	0.827	0.923
38	0.944	1.000	1.000	1.000	0.932	1.000	1.000	0.979	0.969	0.981	0.933	0.962
40	0.981	1.000	1.000	1.000	0.977	1.000	1.000	0.979	0.969	1.000	0.947	1.000
42	1.000	1.000	1.000	1.000	0.977	1.000	1.000	1.000	0.969	1.000	0.973	1.000
44	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.984	1.000	0.987	1.000
46	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.987	1.000
48	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

For each discrete value x from Table 4 the functional values $F^e(x)$ are taken as a basis to model fuzzy functional values $\tilde{F}^e(x)$ to cover all groups at once. At each (discrete) position x a histogram is constructed using the functional values of the empirical distribution functions. The abscissa is subdivided into suitable subsets in the interval $[0,1)$; and the number of functional values assigned to each subset is plotted on the ordinate. Than, fuzzy numbers are generated from the histograms by simple approximation schemes such as least squares algorithm. In this generation process the properties of the probability measure must be observed. In the present case fuzzy triangular numbers and fuzzy numbers with a polygonal membership function are chosen. The fuzzification process is shown in Fig. 10 for three selected values $x = f_c$. The fuzzification results for all $x = f_c$ are listed in Table 4. The interval bounds of the support as well as the mean value are indicated for each fuzzy probability $\tilde{F}^e(x)$. The obtained fuzzy probabilities $\tilde{F}^e(x)$ for discrete $x = f_c$ are functional values of the sought fuzzy probability distribution function $\tilde{F}(x)$.

This non-parametric representation can finally be replaced by a parametric fuzzy probabilistic model in the form of an envelope. For this purpose, different membership levels α are considered for the determination of fuzzy parameters of the fuzzy probability distribution and for the description of the distribution type. The aim is to determine bounding distribution functions of the fuzzy random variable for each membership level. The entirety of all included probabilistic models then reflects the sought fuzzy probability distribution.

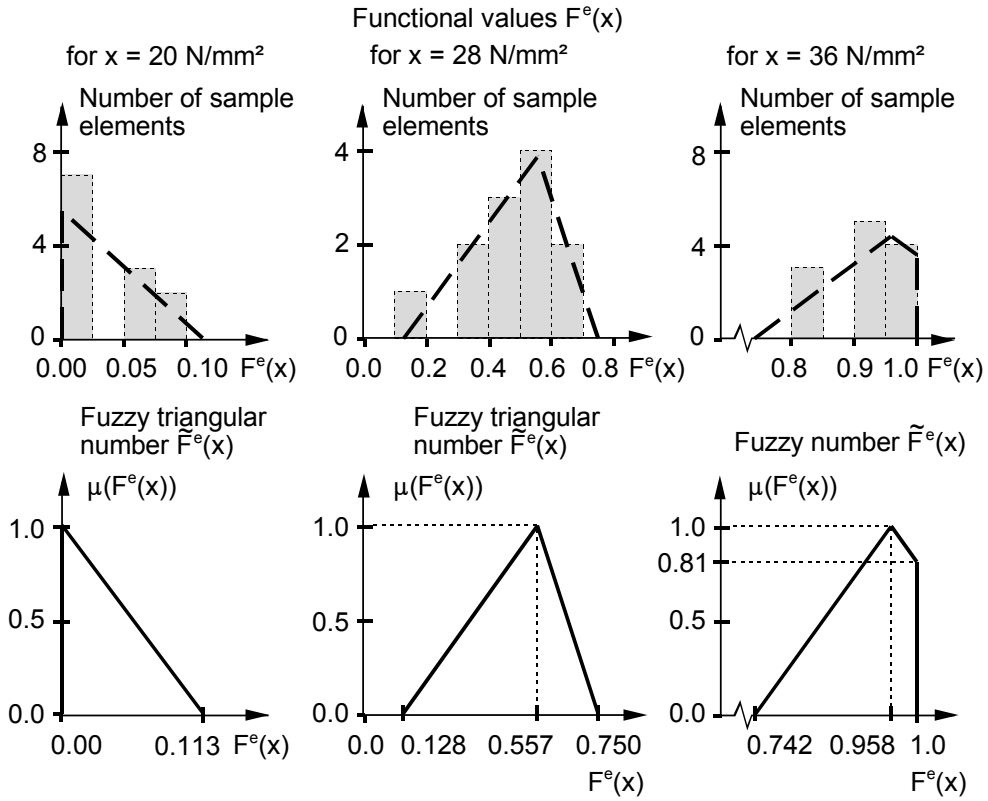


Figure 10. Histograms and membership functions of the functional values of the empirical distribution function $F^e(x)$ for $x = f_c = 20\,N/mm^2$, $x = f_c = 28\,N/mm^2$, and $x = f_c = 36\,N/mm^2$

In this example a compound distribution comprised of a normal distribution (ND) and a logarithmic normal distribution (LND) with a constant ratio of components is adopted. It is assumed that the expected value and standard deviation are the same for both distributions; the minimum value of the component logarithmic normal distribution is specified to be $x_0 = 5\,N/mm^2$. The expected value, standard deviation, and ratio of components are chosen to be free fuzzy parameters of the compound distribution

$$\tilde{F}(x) = \tilde{a} \cdot \tilde{F}^{NV}(x) + (1 - \tilde{a}) \cdot \tilde{F}^{LNV}(x)\,. \tag{4}$$

The subsequent evaluation is restricted to the membership levels $\alpha = 0$ and $\alpha = 1$. The free parameters required for approximating the distribution functions of the originals are determined by the method of least squares. The distribution function $F_1(x)$ for the membership level $\alpha = 1$ is obtained from the values of $F_1^e(x)$. The boundaries of the membership level $\alpha = 0$ are obtained in each case from all values of $F_{0l}^e(x)$ and $F_{0r}^e(x)$, respectively. The following constraints are taken into account:

– All $F_{0l}^e(x) > 0$ lie above the approximation function $F_{0l}(x)$

– All $F_{0r}^e(x) < 1$ lie below the approximation function $F_{0r}(x)$

Table VI. Support bounds $F^e_{0l}(x)$ and $F^e_{0r}(x)$, and mean values $F^e_1(x)$ of the fuzzy probability $\tilde{F}^e(x)$ for all $x = f_c[N/mm^2]$ from Table 4

x	$F^e_{0l}(x)$	$F^e_l(x)$	$F^e_{0r}(x)$	x	$F^e_{0l}(x)$	$F^e_l(x)$	$F^e_{0r}(x)$
12	0.000	0.000	0.000	32	0.603	0.799	0.975
14	0.000	0.000	0.018	34	0.652	0.925	1.000
16	0.000	0.000	0.018	36	0.742	0.958	1.000
18	0.000	0.000	0.035	38	0.913	1.000	1.000
20	0.000	0.000	0.113	40	0.949	1.000	1.000
22	0.000	0.000	0.369	42	0.966	1.000	1.000
24	0.000	0.283	0.417	44	0.983	1.000	1.000
26	0.025	0.358	0.492	46	0.984	1.000	1.000
28	0.128	0.557	0.750	48	1.000	1.000	1.000
30	0.331	0.763	0.825	–	–	–	–

The following values are obtained for the free distribution parameters and the functional parameter *a* of the implemented distribution function:

- Approximation of $F^e_1(x)$: $m_x = 27.66 N/mm^2$, $\sigma_x = 4.34 N/mm^2$, $a = 0.00$
- Approximation of $F^e_{0l}(x)$: $m_x = 34.29 N/mm^2$, $\sigma_x = 4.81 N/mm^2$, $a = 0.00$
- Approximation of $F^e_{0r}(x)$: $m_x = 23.30 N/mm^2$, $\sigma_x = 4.44 N/mm^2$, $a = 1.00$

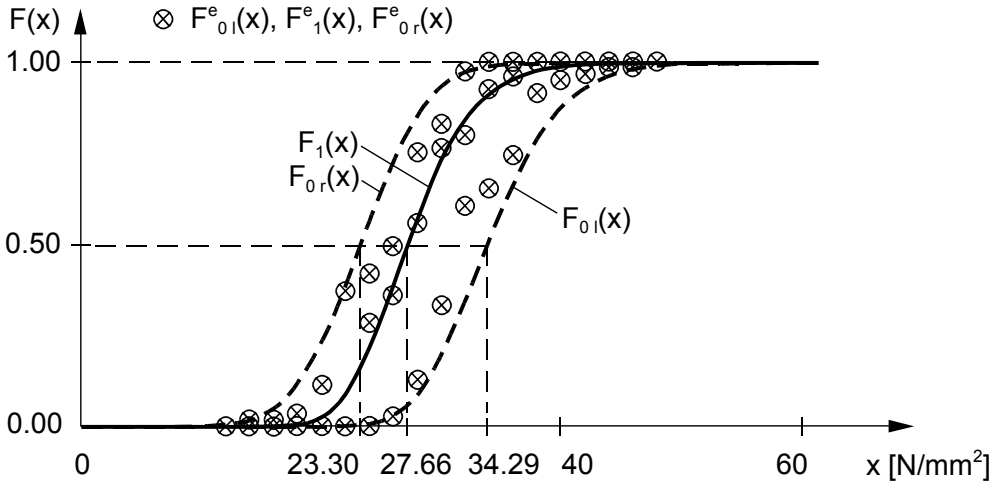


Figure 11. Functional values of the empirical probability distribution functions $F^e_1(x)$, $F^e_{0l}(x)$, and $F^e_{0r}(x)$, as well as the approximation functions $F_1(x)$, $F_{0l}(x)$, and $F_{0r}(x)$

The computed distribution functions $F_1(x)$, $F_{0l}(x)$, and $F_{0r}(x)$ are shown in Fig. 11 together with the adopted functional values of the empirical distribution function from Table 4.

The fuzzy distribution parameters and the fuzzy functional parameter $\tilde{a}$ of the sought fuzzy probability distribution according to Eq. (4) may be expressed as fuzzy triangular numbers (confined to $\alpha = 0$ and $\alpha = 1$):

- $\tilde{m}_x =< 23.30, 27.66, 34.29 > \mathrm{N/mm^2}$,
- $\tilde{\sigma}_x =< 4.34, 4.34, 4.81 > \mathrm{N/mm^2}$, and
- $\tilde{a} =< 0.00, 0.00, 1.00 >$.

The interaction relationship between $\tilde{m}_x$, $\tilde{\sigma}_x$ and $\tilde{a}$ may be determined numerically (Sect. 3), or may be approximately estimated on the basis of the available information. A possible estimation of the interaction is shown in Fig. 12.

In the example, the interaction between $\tilde{m}_x$, $\tilde{\sigma}_x$ and $\tilde{a}$ has only a very slight effect, and may be neglected without a significant effect. The fuzzy probability density functions and the fuzzy probability distribution functions are compared in Figs. 13 and 14, with and without consideration of interaction. The approximation functions $F_{0l}(x)$ and $F_{0r}(x)$ as well as the corresponding probability density functions $f_{0l}(x)$ and $f_{0r}(x)$ are also shown in the figures.

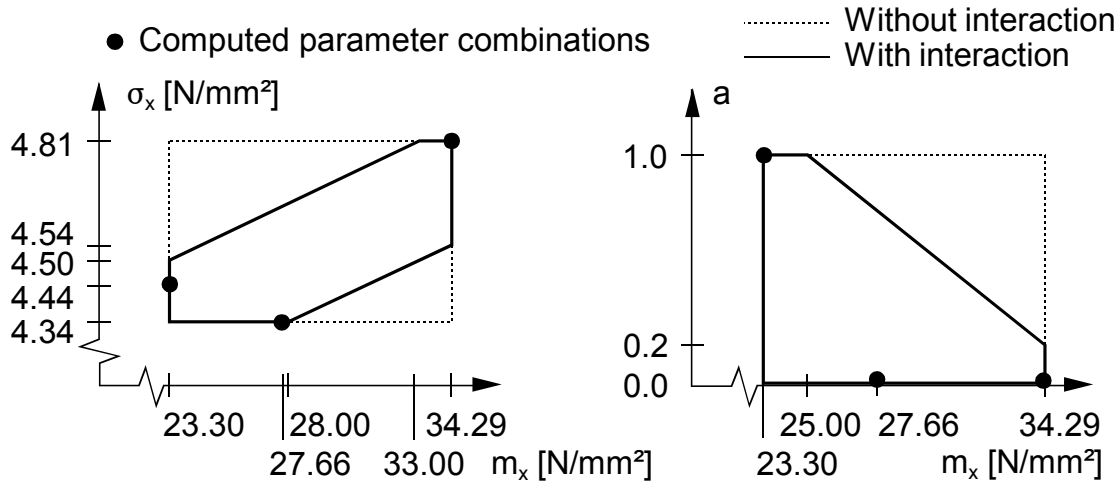


Figure 12. Estimation of the interaction between $\tilde{m}_x$, $\tilde{\sigma}_x$ and $\tilde{a}$

5. Conclusions

Inconsistent data represent a common case of available information in civil engineering practice. These data must be properly evaluated and described numerically to obtain realistic results in a subsequent structural analysis, safety assessment or structural design. The evaluation of inconsistent data is, however, problematic.

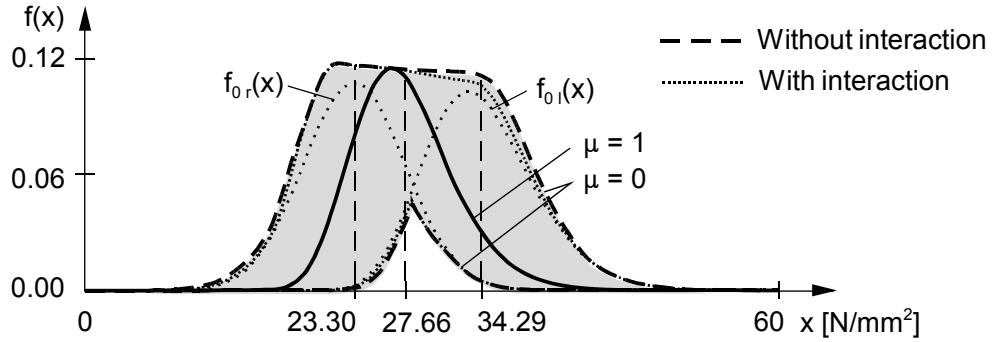


Figure 13. Fuzzy probability density function $\tilde{f}(x)$ with and without consideration of interaction between $\tilde{m}_x$, $\tilde{\sigma}_x$ and $\tilde{a}$; probability density functions $f_{0l}(x)$ and $f_{0r}(x)$ belonging to $F_{0l}(x)$ and $F_{0r}(x)$

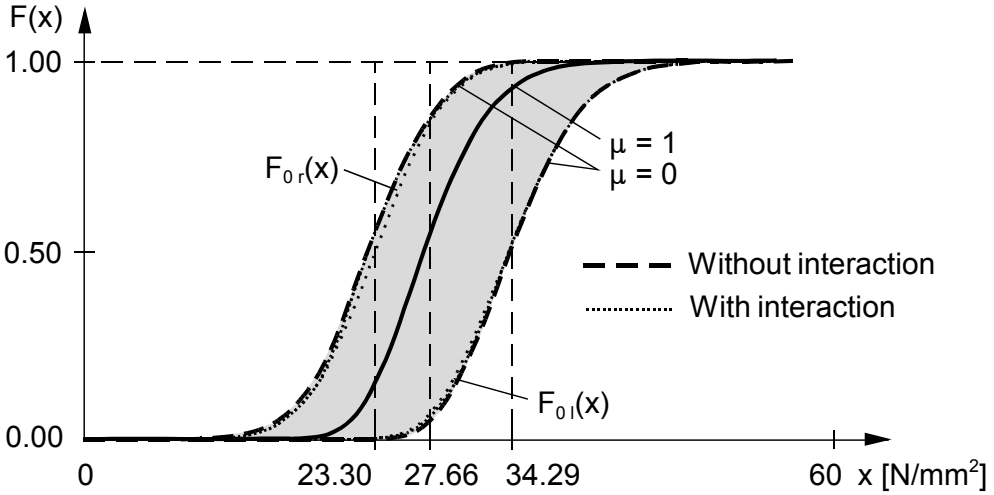


Figure 14. Fuzzy probability distribution function $\tilde{F}(x)$ with and without consideration of interaction between $\tilde{m}_x$, $\tilde{\sigma}_x$ and $\tilde{a}$; probability distribution functions $F_{0l}(x)$ and $F_{0r}(x)$

Stochastic uncertainty and imprecision appear simultaneously and in various configurations. For a proper treatment of this type of information, the model fuzzy randomness is proposed. This enables a separate and simultaneous treatment of statistical uncertainty and imprecision. Due to the variety of possible forms of available information, a general quantification algorithm cannot be formulated. The quantification has to be realized according to the conditions in each particular case. In the paper quantification guidelines for three selected typical cases of inconsistent data in civil engineering were presented by way of examples. Algorithms from traditional statistics have been utilized and combined with fuzzy methods for the inclusion of expert knowledge. The quantification results reflect the stochastic uncertainty and the imprecision of the available information in form of a fuzzy probability. This represents an envelope of all real-valued probabilistic models which meet the available information.

Further developments are focused on the development of a hybrid quantification algorithm for inconsistent data, which includes, simultaneously, more components beyond traditional statistics and fuzzy methods

to extend the spectrum of cases covered and to further improve the quality of the quantification results. This leads, eventually, to a minimization of risks due to modeling errors and associated misinterpretations of structural behavior and safety.

Acknowledgements

The author gratefully acknowledges the financial support by National University of Singapore through the Ministry of Education Academic Research Fund.

References

1. Ayyub, B.: 1998, *Uncertainty Modeling and Analysis in Civil Engineering*. Boston London New York: CRC Press.
2. Bandemer, H. and S. Gottwald: 1995, *Fuzzy Sets, Fuzzy Logic Fuzzy Methods With Applications*. Wiley.
3. Bandemer, H. and W. Näther: 1992, *Fuzzy Data Analysis*. Dordrecht: Kluwer Academic Publishers.
4. Beer, M. and M. Liebscher: 2007, 'Designing Robust Structures – A Nonlinear Simulation Based Approach'. *Special Issue of Computers & Structures*. in press.
5. Ben-Haim, Y.: 2004, 'Uncertainty, probability and information-gaps'. *Reliability Engineering & System Safety* **85**(1-3), 249–266.
6. Ben-Haim, Y. and I. Elishakoff: 1990, *Convex Models of Uncertainty in Applied Mechanics*. Amsterdam: Elsevier.
7. Elishakoff, I.: 1999, *Whys and Hows in Uncertainty Modelling Probability, Fuzziness and Anti-Optimization*. Wien, New York: Springer.
8. England, J., J. Agarwal, and D. Blockley, 'The vulnerability of structures to unforeseen events'. *Special Issue of Computers & Structures*. , submitted.
9. Farkas, L., D. Moens, D. Vandepitte, and W. Desmet, 'Application of fuzzy numerical techniques for product performance analysis in the conceptual and preliminary design stage'. *Special Issue of Computers & Structures*. , submitted.
10. Fellin, W., H. Lessmann, M. Oberguggenberger, and R. Vieider (eds.): 2005, *Analyzing Uncertainty in Civil Engineering*. Berlin Heidelberg New York: Springer.
11. Hall, J. W., E. Rubio, and M. Anderson: 2004, 'Random sets of probability measures in slope hydrology and stability analysis'. *Special Issue of ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik* **84**(10–11), 710–720.
12. Helton, J., J. Johnson, W. Oberkampf, and C. Sallaberry: 2006, 'Sensitivity analysis in conjunction with evidence theory representations of epistemic uncertainty'. *Reliability Engineering & System Safety* **91**, 1414–1434.
13. Helton, J. C. and D. E. Burmaster (eds.): 1996, *Special Issue on the Treatment of Aleatory and Epistemic Uncertainty*, Vol. 54 of *Reliability Engineering & System Safety*.
14. Helton, J. C. and W. L. Oberkampf (eds.): 2004, *Special Issue on Alternative Representations of Epistemic Uncertainty*, Vol. 85 of *Reliability Engineering & System Safety*.
15. Klir, G. J.: 2006, *Uncertainty and information : foundations of generalized information theory*. Hoboken: Wiley-Interscience.
16. Kreinovich, V. and S. A. Ferson: 2004, 'A new Cauchy-based black-box technique for uncertainty in risk analysis'. *Reliability Engineering & System Safety* **85**(1-3), 267–279.
17. Kruse, R. and K. Meyer: 1987, *Statistics with Vague Data*. Dordrecht: Reidel.
18. Möller, B. and M. Beer, 'Engineering Computation Under Uncertainty – Capabilities of Non-Traditional Models'. *Special issue of Computers & Structures*. in press.
19. Möller, B. and M. Beer: 2004, *Fuzzy Randomness – Uncertainty in Civil Engineering and Computational Mechanics*. Berlin: Springer.
20. Möller, B., W. Graf, and M. Beer: 2000, 'Fuzzy structural analysis using alpha-level optimization'. *Computational Mechanics* **26**, 547–565.

21. Muhanna, R. and R. Mullen: 1999, 'Formulation of Fuzzy Finite Element Methods for Solid Mechanics Problems'. *Computer-Aided Civil and Infrastructure Engineering* **14**, 107–117.
22. Muhanna, R. L., R. L. Mullen, and H. Zhang: 2007, 'Interval Finite Element as a Basis for Generalized Models of Uncertainty in Engineering Mechanics'. *Journal of Reliable Computing* **13**(2), 173–194.
23. Neumaier, A.: 2004, 'Clouds, Fuzzy Sets, and Probability Intervals'. *Reliable Computing* **10**(4), 249–272.
24. Pedrycz, W., A. Skowron, and V. Kreinovich (eds.): 2008, *Handbook of Granular Computing*. New York: Wiley.
25. Ross, T. J.: 2004, *Fuzzy Logic with Engineering Applications*. Wiley, 2nd edition.
26. Schenk, C. A. and G. I. Schuëller: 2005, *Uncertainty Assessment of Large Finite Element Systems*. Berlin Heidelberg: Springer.
27. Spanos, P. D. and G. Deodatis (eds.): 2007, *Computational Stochastic Mechanics*. Rotterdam: Millpress.
28. Viertl, R.: 1996, *Statistical Methods for Non-Precise Data*. Boca Raton New York London Tokyo: CRC Press.
29. Viertl, R. and W. Trutschnig: 2006, *Fuzzy histograms and fuzzy probability distributions*.
30. Zimmermann, H.-J.: 1992, *Fuzzy set theory and its applications*. Boston London: Kluwer Academic Publishers.

PROCEEDINGS: Author Index


NSF


Sun
microsystems


SIPTA


GT STRUDL®


Georgia Institute
of Technology
Savannah