EDITOR
Marco De Angelis

CHAIR
Scott Ferson, University of Liverpool

FOUNDERS
Rafi L. Muhanna
Robert L. Mullen

CO-CHAIRS
Marco De Angelis, University of Liverpool
Edoardo Patelli, University of Liverpool

SCIENTIFIC COMMITTEE
Siu-Kui Au, University of Liverpool, United Kingdom
Ekaterina Auer, University of Wismar, Germany
André T. Beck, University of São Paulo, Brazil
James L. Beck, California Institute of Technology, USA
Michael Beer, University of Liverpool, United Kingdom
Eleni Chatzi, ETH Zurich, Switzerland
Jianye Ching, National Taiwan University, Taiwan
Bruce Ellingwood, Colorado State University, USA
Scott Ferson, University of Liverpool, UK
Steffen Freitag, Ruhr University Bochum, Germany
Wolfgang Graf, Technische Universität Dresden, Germany
Achintya Haldar, University of Arizona, USA
Michael Hanss, University of Stuttgart, Germany
Hongwei Huang, Tongji University, China
Héctor A. Jensen, University Santa Maria, Chile
Ioannis A. Kougioumtzoglou, Columbia University, USA
Vladik Kreinovich, University of Texas at El Paso, USA
Rudolf Kruse, Otto-von-Guericke University Magdeburg, DE
Anička Kučerová, Czech Technical University in Prague
Tom Lahmer, Bauhaus-Universität Weimar, Germany
David Lehký, Brno University of Technology, Czech Republic
Mehdi Modares, Illinois Institute of Technology, USA
David Moens, Catholic University of Leuven, Belgium
Giuseppe Muscolino, University of Messina, Italy
Michael Oberguggenberger, University of Innsbruck, Austria
Edoardo Patelli, University of Liverpool, United Kingdom
Kok Kwang Phoon, National University of Singapore, Singapore
Mallela Venkata Rama Rao, Vasavi College of Engineering, India
Siegfried M. Rump, Hamburg University of Technology, Germany
Christian Soize, Marne la Vallée University, France
Lev Utkin, St. Petersburger State Forest Technical University, RU
Marcos Valdebenito, University Santa Maria, Chile
Reinhard Viertl, Vienna University of Technology, Austria
Hao Zhang, The University of Sydney, Australia
Enrico Zio, Polytechnic of Milan, Italy

LOCAL COMMITTEE
Marco De Angelis
Scott Ferson
Héctor Diego Estrada Lugo
Caroline Morais
Jonathan Sadeghi
Noémie Le Carrer
Uchenna Oparaji
Domenico Altieri
Barry Smith
Dominic Calleja
Aloussa Khodadadyan
Bright Uchenna Oparaji
Ander Gray
Andrea Jones
Edoardo Patelli
Roberto Rocchetta
Silvia Tolo
Pat Mercadante
Tania Gricel Benitez López
Hindolo George-Williams

SPONSORS
ESRC - Economic and Social Research Council, https://esrc.ukri.org/
SRA - Society for Risk Analysis, www.sra.org/
ESRA - European Safety and Reliability Association, www.esrahomepage.org/
Agenda

MONDAY 16

Session 1

1 STOCHASTIC METHODS IN DAMAGE DETECTION
   M. Oberguggenberger and M. Schwarz

13 POSSIBILISTIC IDENTIFICATION OF RELIABLE FINITE IMPULSE RESPONSE MODELS
   D. Hose, M. Hanss

21 INFLUENCE OF UNCERTAINTY AND NUMERICAL ERRORS IN THE CONTEXT OF MIMO SYSTEMS
   E. Auer, C. Benavente-Peces, A. Ahrens

31 SURROGATE MODELING FOR SOLVING OPTIMIZATION PROBLEMS WITH POLYMORPHIC UNCERTAIN DATA
   S. Freitag, P. Edler, K. Kremer, G. Meschke

Session 2

41 INTERVAL FIELD FOR SPATIALLY AND TEMPORALLY DEPENDENT UNCERTAINTY-SUPervised MACHINE LEARNING APPROACH
   D. Betancourt, R. L. Muhanna, R. L. Mullen

51 UNCERTAINTY QUANTIFICATION FOR RANDOM FIELDS ESTIMATED FROM EFFECTIVE MODULI OF ELASTICITY

63 POLYMORPHIC UNCERTAINTY MODELING FOR OPTIMIZATION OF TIMBER STRUCTURES
   N. Schietzold, W. Graf, and M. Kaliske

79 PRODUCT RELIABILITY UNDER PROCESS VARIABILITY: AN INTEGRATED PROCESS-PRODUCT RELIABILITY BASED DESIGN OPTIMISATION APPROACH
   M. Faes, B. Van Doninck, M. Imholz, D. Moens

Session 3

87 RISK ANALYSIS OF INFINITE SLOPE FAILURE USING ADVANCED BAYESIAN NETWORKS
   L. He, A. Topa Gomes, M. Broggi, M. Beer

103 Bayesian learning in expert systems using plausible Petri nets
   J. Chiachío-Ruano, M. Chiachío-Ruano, J. Andrews

111 BUILDING ENVELOPE INSPECTION MANAGEMENT USING A HYBRID STATISTICAL RANDOM SAMPLING AND BAYESIAN UPDATING APPROACH
   M. Modares, J. Mohammadi

TUESDAY 17

Session 1

117 DETERMINATION OF TASK TEMPORAL VARIATIONS IN CONSTRUCTION SCHEDULING USING IMPRECISE PROBABILITY
   M. Modares, R. L. Muhanna, and R. L. Mullen

126 WHEN IS PROPAGATION OF INTERVAL AND FUZZY UNCERTAINTY FEASIBLE?
   V. Kreinovich, A. M. Pownuk, O. M. Kosheleva, A. Belina

137 DEEP LEARNING USING INTERVAL NEURAL NETWORKS WITH GUARANTEED GENERALISATION PROPERTIES
   J. Sadeghi, M. De Angelis, E. Patelli

147 ITERATIVE IMPORTANCE SAMPLING FOR ESTIMATING EXPECTATION BOUnDS UNDER PARTIAL PROBABILITY SPECIFICATIONS
   M. Troffaes, T. Fetz, M. Oberguggenberger

Session 2
GLOBAL SENSITIVITY ANALYSIS OF UNCERTAIN LINEAR STRUCTURES SUBJECT TO STATIC LOAD
M. Valdebenito, I. V. Gonzalez, H. A. Jensen

AN EFFICIENT METHOD FOR ESTIMATING CONDITIONAL FAILURE PROBABILITIES
D. Altieri, E. Patelli

WHY TAYLOR MODELS AND MODIFIED TAYLOR MODELS ARE EMPIRICALLY SUCCESSFUL:
A SYMMETRY-BASED EXPLANATION
M. Joldes, C. Lauter, M. Ceberio, O. M. Kosheleva, V. Kreinovich

WHAT IS THE ECONOMICALLY OPTIMAL WAY TO GUARANTEE INTERVAL BOUNDS ON CONTROL?
A. Vaccaro, M. Ceberio, V. Kreinovich

Session 3

WORST-CASE SOLUTION SPACES FOR SYSTEMS DESIGN UNDER UNCERTAINTIES
M. Daub, F. Duddeck

IMPLEMENTATION OF A DIGITAL TWIN OF A NUCLEAR WASTE MANAGEMENT FACILITY
M. Bankhead, K. Fritsch, J. MacDonald-Taylor, G. Ecclestone, W. J. O’Brien

HUMAN INVOLVEMENT – THE UNFORESEEN AND ROBUSTNESS
F. Knoll

WEDNESDAY 18

Session 1

ROBUST OPTIMISATION OF CARGO LOADING AND SHIP SCHEDULING IN TIDAL AREAS
N. Le Carrer, S. Ferson, P. L. Green

Session 2

A REINFORCEMENT LEARNING FRAMEWORK FOR OPTIMISATION OF POWER GRID OPERATIONS AND MAINTENANCE
R. Rocchetta, M. Compare, L. Bellani, E. Zio, E. Patelli

Index of Authors
Stochastic methods in damage detection

M. Oberguggenberger,1* and M. Schwarz1

1 Unit for Engineering Mathematics, University of Innsbruck, Austria
*Corresponding author: michael.oberguggenberger@uibk.ac.at

Abstract
This paper reports on a new method for analyzing wave propagation in two- and three-dimensional random materials. The applicative targets are damage detection, system identification, and calibration of models for randomly perturbed structures. The novelty of the approach is that the dynamic response of the structure is represented by means of Fourier integral operators. Parameter calibration and identification as well as describing the random behavior is achieved through the coefficients in the Fourier integral operators. Damage detection proceeds in three steps: (a) calibration of the nominal material parameters at the undamaged, but possibly randomly perturbed structure; (b) calibration of the stochastic parameters of the random fields perturbing the material properties of the structure; (c) damage such as overall deviation of the nominal parameters or such as cracks are detected by finding the features of the dynamic response outside the confidence region computed from steps (a) and (b). The applicability of the method is demonstrated by means of computer simulations of an elastic body under plane strain.

Keywords: Waves in random media, Fourier integral operators, numerical simulation, parameter calibration, damage detection.

1 Introduction
The paper reports on the results of a research project1 addressing linear wave propagation in random media by means of stochastic Fourier integral operators. The applicative targets are reliability analysis, damage detection, system identification, and calibration of models for randomly perturbed structures in elasticity and strength of materials. The set-up applies to materials with stochastically varying properties.

The dynamic response of an elastic medium is described by the equations of motion, a system of hyperbolic differential equations. Randomness of the medium causes the coefficients of the equations to become random fields. The direct, bottom-up approach would start by modelling the coefficients as random fields, solving the equations, and inferring stochastic properties of the solution. However, the direct approach faces the difficulty that the realizations of the random fields commonly in use in elastostatics do not have the degree of smoothness required to construct solutions in the propagation case. Thus the stochastic characteristics of the solutions are hardly tractable explicitly. Our approach consists in representing the dynamic response of the structure by means of stochastic Fourier integral operators and in shifting the stochastic modelling from the coefficients to stochastic building blocks of the Fourier integral operator.

A Fourier integral operator consists of a complex exponential term containing the phase function (which describes the propagation geometry) and an amplitude. In the linear, constant coefficient case, the response of the system can be expressed explicitly by means of Fourier integral operators based on so-called half wave equations. In the case of a spatially varying medium, we add parametrically described random terms to the phase function and amplitude of the Fourier integral operator, which carry the burden of modelling the random medium and can still be computed numerically. Since neither the stochastic models of the coefficients nor the ones of the phase function and amplitude are known a priori, but have to be determined by data fitting, our top-down approach appears as justified as the bottom-up approach. Parameter calibration is shifted to the terms constituting the Fourier integral operators and comparing the solution produced by the Fourier integral operator with measured data.

The paper starts with setting up the required theory: the equations of motion of three-dimensional linear elasticity, Fourier integral operators (FIOS) and the decomposition of scalar wave equations into two half wave equations, stochastic Fourier integral operators, numerical methods for evaluating the action of a Fourier integral operator, and the description of a two-dimensional test example.

The second part of the paper addresses the central theme of the project, namely stochastic parameter estimation and damage detection. The test runs are performed on data generated by a finite element


8th International Workshop on Reliable Engineering Computing, “Computing with Confidence”
16–18 July 2018, University of Liverpool, Liverpool, UK
model (FEM), whereby the FE-output simulates measured data. The section starts with calibrating the nominal values of the material parameters and continues with calibrating the stochastic parameters of the underlying spatial random perturbations. Note that the Fourier integral operators have a deterministic part, in which the material properties, such as the modulus of elasticity, the Poisson ratio and the density, enter through the wave speed. Comparison of the FE-output with the solution produced by the FIO representation thus admits calibration of the nominal values. The stochastic part of the Fourier integral operator is modeled through a Gaussian random field with a correlation length and a variance to be determined. Assuming a sample of FE-data given in sufficiently many points, the correlation length can be estimated. Concerning the variances, it turned out that the variance of the random field applied in the FE-model does not coincide with the variance needed for the random fields in the Fourier integral operators. This obstacle was overcome by introducing a scaling curve which relates the two variances so that the FIO solution reproduces the stochastic properties of the FEM-solution. This scaling curve was obtained empirically by a large number of parameter variations and Monte Carlo simulations.

Finally, having calibrated the Fourier integral operator to the undamaged, but possibly randomly perturbed structure, this information can be used to produce confidence regions around undamaged solutions. With the help of these confidence regions, statistical tests can be set up that admit deciding about larger deviations in the overall material parameters as well as deciding about the presence of cracks. Using a grid of (virtual) sensors, the crack location can also be determined. This method works if the excitation source, the sensors, and the cracks are sufficiently close to each other. For larger scale distances, the FIOs have to be evaluated over a longer time period, which requires increasing numerical effort. Some of the results have been announced in [1, 2].

2 Theory

This section is devoted to collect the background material required for the methods employed. Starting with the equations of motion of linear elasticity, we turn to Fourier integral operators (FIOs) and the representation of the solution to scalar wave equations, followed by introducing stochastic terms in the FIOs and fast numerical methods for the evaluation of FIOs. The section concludes by presenting a plane strain test example to be used in Section 3.

2.1 Three dimensional linear elasticity

The dynamic displacement \( u(x, t) \) of a three-dimensional homogeneous, isotropic, linearly elastic body is given by the equations of motion

\[
\frac{\lambda + \mu}{\rho} (\nabla \times \nabla \times u(x, t)) + \frac{\mu}{\rho} \Delta u(x, t) - \partial_t u(x, t) = -f(x, t),
\]

where \( f \) is the body force, \( \lambda \) and \( \mu \) are the Lamé constants and \( \rho \) the density. The lateral resp. the transverse wave speeds are \( c_l^2 = \frac{\lambda + 2\mu}{\rho} \) resp. \( c_t^2 = \frac{\mu}{\rho} \). In this case, there exist [3, 4] potentials \( \Phi \) and \( \Psi = (\Psi_1, \Psi_2, \Psi_3)^T \) with

\[
\nabla \Phi(x, t) + \nabla \times \Psi(x, t) = u(x, t)
\]

\[
\nabla \cdot \Psi(x, t) = 0,
\]

where \( \Phi \) and \( \Psi \) satisfy the wave equations

\[
\partial_t \Phi(x, t) - c_l^2 \Delta \Phi(x, t) = -\varphi(x, t)
\]

\[
\partial_t \Psi_j(x, t) - c_t^2 \Delta \Psi_j(x, t) = -\psi_j(x, t), \quad j = 1, 2, 3,
\]

and \( \varphi \) and \( \psi \) have to be determined from

\[
\nabla \varphi(x, t) + \nabla \times \psi(x, t) = -f(x, t)
\]

\[
\nabla \cdot \psi(x, t) = 0.
\]

In this way, the equations of motion (1) are decoupled into four scalar wave equations for the wave potentials. One can compute \( \varphi \) and \( \psi \) by inverting the Laplace equation as follows, using the ansatz

\[
\varphi(x, t) = \partial_x g_1(x, t) + \partial_y g_2(x, t) + \partial_z g_3(x, t)
\]

\[
\psi_1(x, t) = \partial_x g_2(x, t) - \partial_y g_3(x, t)
\]

\[
\psi_2(x, t) = -\partial_z g_1(x, t) + \partial_y g_3(x, t)
\]

\[
\psi_3(x, t) = \partial_y g_1(x, t) - \partial_z g_2(x, t),
\]
for functions $g_j, j = 1,2,3$ to be determined. Then
\[
\nabla \varphi(x, t) + \nabla \times \psi(x, t) = (\Delta g_1(x, t), \Delta g_2(x, t), \Delta g_3(x, t))^\top.
\]
By inverting the Laplace operator one gets $g_j(x, t) = \Delta^{-1} f_j(x, t)$.

### 2.2 Fourier operator solution of the scalar wave equation

The general form of a time-dependent Fourier operator $A$ acting on functions $w(x)$ of a $d$-dimensional space variable $x$ is
\[
A w(x, t) = \frac{1}{(2\pi)^d} \int \int e^{i\Phi(x, y, \xi, t)} a(x, y, \xi, t) w(y) \, dy \, d\xi
\]  
(2)

with a so-called phase function $\Phi(x, y, \xi, t)$ and an amplitude $a(x, y, \xi, t)$. The phase function and the amplitude have to satisfy certain regularity properties so that (2) can be given a meaning as an oscillatory integral, see e.g. [5]. In the cases needed here, the phase function is of the special form $\Phi(x, \xi, t) - \xi \cdot y$ and the amplitude does not depend on $y$. In these cases, the Fourier integral operator can be written in the simpler form
\[
A w(x, t) = \frac{1}{(2\pi)^d} \int \int e^{i\phi(x, \xi, t)} y a(x, \xi, t) w(y) \, dy \, d\xi = \frac{1}{(2\pi)^d} \int \int e^{i\phi(x, \xi, t)} a(x, \xi, t) \hat{w}(\xi) d\xi
\]  
(3)

where $\hat{w}(\xi)$ denotes the Fourier transform.

To represent the solution to the scalar wave equation with constant propagation speed (in any space dimension $d$)
\[
u_{ct}(x, t) - c^2 \Delta u(x, t) = f(x, t)
\]
by means of Fourier integral operators, it is first decomposed in a decoupled system of half wave equations. Setting $v = u_t - ic\sqrt{-\Delta}u$ one obtains
\[
v_{v}(x, t) = \pm c\sqrt{-\Delta} v(x, t) = f(x, t).
\]
Here $c\sqrt{-\Delta}$ is the pseudodifferential operator with symbol $c|\xi|$. The solution of these half wave equations is given as a sum of Fourier integral operators of the form
\[
v(x, t) = \frac{1}{(2\pi)^d} \int e^{i\phi(x, \xi, t)} \left( \hat{v}_0(\xi) + \int_0^t e^{i\pm c|\xi|s} \hat{f}(\xi, s) \right) d\xi
\]
where $\hat{f}(\xi, s)$ denotes the Fourier transform with respect to the spatial variables, and $v_0$ encodes the initial data. This solution is exact. In the general case of non-constant propagation speed one may still construct a so-called parametrix (a solution up to a smooth error) in this way, using Fourier integral operators. However, we shall pursue a more tractable perturbation approach in this paper to deal with non-constant propagation speed.

### 2.3 Stochastic perturbations

According to the principles laid out in the Introduction, we use a top-down approach. Instead of modelling the randomness through the coefficients of the equations of motion, we solve a deterministic model by means of Fourier integral operators and add parametrized randomness into the solution operator. To construct the perturbed Fourier integral operator, we use the following strategy: Let
\[
\frac{1}{(2\pi)^d} \int \int e^{i\Phi(x, \xi, t)} y a(x, \xi) w(y) \, dy \, d\xi
\]
be a term of the solution operator for the wave equation with constant wave speed $c_0$. Adding random perturbations $r_1(x, \xi, t)$, $r_2(x, \xi)$ to the solution operator results in the perturbed, stochastic solution operator
\[
\frac{1}{(2\pi)^d} \int \int e^{i\Phi(x, \xi, t) + r_1(x, \xi, t)} y (a(x, \xi) + r_2(x, \xi)) w(y) \, dy \, d\xi
\]
This representation is very convenient, since one can interpret $r_1$ as the perturbation geometry of the wave propagation, and $r_2$ as the perturbation amplitude of the wave number.

In order to guarantee convergence of the perturbed integral, one has to make sure that $r_1$ and $r_2$ of the right class, i.e., $\Phi + r_1$ must be a non-degenerate phase function and $a + r_2$ must be of Hörmander symbol class $S^m_{\rho, \delta}$ with $0 < \rho \leq 1, 0 \leq \delta < 1$ (see e.g. [5]). Alternatively, one can take $w(y)$ sufficiently smooth and decaying at infinity to arrive at a converging iterated integral.
2.4 Fast Fourier integral operators

The numerical evaluation of the action of a Fourier integral operator is aided by means of the butterfly algorithm developed in [6]. A typical term is subjected to a spatial discretization at fixed time \( t \) as follows:

\[
\int_{\mathbb{R}^d \times \mathbb{R}^d} e^{i \mathbf{\phi} \cdot \mathbf{\xi} \cdot t} \frac{1}{i} y \, \mathbf{\xi} w(y) \, d\mathbf{\xi} \approx \Delta_t \Delta_y \sum_{k \in \mathcal{U}_y} e^{i \mathbf{\phi} \cdot \mathbf{\xi}_k \cdot t} \sum_{j \in \mathcal{U}_x} e^{-i y_j \cdot \mathbf{\xi}_k} w(y_j),
\]

where \( \mathcal{G}_y \) is a spatially discrete and truncated grid in \( \mathbb{R}^d \), e.g. \( \{ \frac{2\pi}{N} (k_1, \ldots, k_d) \mid k_i = 1, \ldots, N \} \) and \( \mathcal{G}_x \) the corresponding grid on the phase side, e.g. \( \{ (j_1, \ldots, j_d) \mid j_i = 1, \ldots, N \} \). Furthermore, \( \Delta_y \) and \( \Delta_\xi \) is the volume of a single grid cell in the corresponding discretization. The second sum can be computed by the fast Fourier transform algorithm. So terms of the form

\[
\sum_{k \in \mathcal{G}_x} e^{i \mathbf{\phi} \cdot \mathbf{\xi}_k \cdot t} g(\mathbf{\xi}_k)
\]

are remaining. A direct evaluation of such a term is very costly and is of complexity \( \mathcal{O}(N^d) \). The butterfly algorithm uses a divide and conquer strategy. Thereby the complexity is reduced to \( \mathcal{O}(N^d \log(N)) \), which is quite tractable for medium-sized domains and space dimensions \( d = 1, 2, 3 \).

2.5 A plane strain test model

In order to keep the computational cost low, we use a plane strain problem for the numerical experiments. This refers to a body with infinite extension in \( z \)-direction and a driving point force of the form \( f(x, t) = [\delta(x) \delta(y) \sin(2\pi t), 0, 0]^T \). By this, \( u \) does not depend on \( z \) and the displacement is only in \( (x, y) \)-direction, i.e., \( u_3 \equiv 0 \). By that, \( g_1(x, y, t) = E(x, y) \sin(2\pi t) \), where \( E(x, y) = -2(2\pi)^{-1} \log(\sqrt{x^2 + y^2}) \) is the fundamental solution of the two dimensional Laplace operator, and \( \Psi = [0, 0, \psi]^T \) with \( \psi = \psi_3 \). Thus the problem is reduced to a two-dimensional one, with displacements \( u_1(x, y, t), u_2(x, y, t) \). Due to the assumed isotropy of the material, it is no restriction of generality to let the driving force act only in \( u_2 \)-direction.

After these simplifications, we may assert that there exist a pressure potential \( \Phi \) and a shear potential \( \Psi \), satisfying the two-dimensional wave equations

\[
\begin{aligned}
\partial_t \Phi(x, y, t) &= c^2_t \Delta \Phi(x, y, t) - \varphi(x, y, t), \\
\partial_t \Psi(x, y, t) &= c^2_s \Delta \Psi(x, y, t) - \psi(x, y, t),
\end{aligned}
\]

with

\[
\begin{bmatrix}
    u_1(x, y, t) \\
    u_2(x, y, t)
\end{bmatrix} = \nabla \Phi(x, y, t) + \begin{bmatrix}
    -\partial_y \Psi(x, y, t) \\
    \partial_x \Psi(x, y, t)
\end{bmatrix}.
\]

The functions \( \varphi \) and \( \psi \) are to be determined by

\[
\begin{bmatrix}
    0 \\
    f(x, y, t)
\end{bmatrix} = \nabla \varphi(x, y, t) + \begin{bmatrix}
    -\partial_y \psi(x, y, t) \\
    \partial_x \psi(x, y, t)
\end{bmatrix}.
\]

The pressure resp. shear wave speed are given by

\[
c_t = \sqrt{\frac{(\nu - 1) E}{(1 + \nu)(-1 + 2 \nu)}}, \quad \text{resp.} \quad c_s = \sqrt{\frac{E}{2(\nu + 1)}},
\]

where \( E \) is the Young modulus and \( \nu \) is the Poisson ratio. Furthermore, the Dirac point mass \( \delta(x) \delta(y) \) is regularized for numerical purposes, the terminal time is set to \( T_{\max} = 2 \), and thus the actually applied force reads

\[
f(x, y, t) = \frac{1}{2\pi \varepsilon^2} \exp \left( -\frac{x^2 + y^2}{2\varepsilon^2} \right) f_0(t), \quad f_0(t) = \sin(2\pi t) \mathbb{I}_{[0, 2]}(t),
\]

where \( \varepsilon \) is a small scaling parameter. Since \( f \) is a product of a function in time and a function in space, one can see from Section 2.1 that \( \varphi \) and \( \psi \) can be represented as

\[
\begin{aligned}
\varphi(x, y, t) &= \varphi_0(x, y) f_0(t), \\
\psi(x, y, t) &= \psi_0(x, y) f_0(t).
\end{aligned}
\]
Applying the spatial Fourier transform to equation (4) leads to the ordinary differential equations

\[ \partial_t \Phi(\xi, \eta, t) = c^2(\xi^2 + \eta^2)\Phi(\xi, \eta, t) - \Phi_0(\xi, \eta)f_0(t) \]
\[ \partial_t \Psi(\xi, \eta, t) = c^2(\xi^2 + \eta^2)\Psi(\xi, \eta, t) - \Psi_0(\xi, \eta)f_0(t). \]

Solving this and applying the inverse Fourier transform we end up with the following solution operators:

\[ \Phi(x, y, t) = \frac{1}{8\pi^2} \int_0^\infty \int_{\mathbb{R}^2} e^{ix\xi + iy\eta}e^{-ic_\ell(t-s)}\sqrt{\xi^2 + \eta^2} \Phi_0(\xi, \eta)f_0(s) d(\xi, \eta) ds \]
\[ + \int_{\mathbb{R}^2} e^{ix\xi + iy\eta}e^{-ic_\ell(t-s)}\sqrt{\xi^2 + \eta^2} \Phi_0(\xi, \eta)f_0(s) d(\xi, \eta) \]
\[ \Psi(x, y, t) = \frac{1}{8\pi^2} \int_0^\infty \int_{\mathbb{R}^2} e^{ix\xi + iy\eta}e^{-ic_\ell(t-s)}\sqrt{\xi^2 + \eta^2} \Phi_0(\xi, \eta)f_0(s) d(\xi, \eta) ds \]
\[ + \int_{\mathbb{R}^2} e^{ix\xi + iy\eta}e^{-ic_\ell(t-s)}\sqrt{\xi^2 + \eta^2} \Phi_0(\xi, \eta)f_0(s) d(\xi, \eta) \]

(7) \hspace{1cm} (8)

3 Application to parameter identification and damage detection

For the purpose of demonstration of the method, we use the plane strain model presented in Subsection 2.5. The basic data are specified as follows: The material is assumed to be aluminum with the nominal material parameters \( E_{\text{mean}} = 70 \) GPa and \( \nu_{\text{mean}} = 0.35 \). The infinite cuboid is excited by a centered dynamic force acting along the \( z \)-axis with frequency \( 2\pi \times 7 \) MHz. The assumed base domain is \( 10 \) cm \( \times \) 10 cm and \( T_{\text{max}} = 7 \) \( \mu \)s. The dimensions of the domain are larger than \( aT_{\text{max}} \), thus we do not have to worry about the boundary conditions during the time interval under consideration.

The idea is to calibrate the Fourier integral operator model at data generated by a finite element model, programmed in Abaqus. The numerical Fourier integral operator always has a periodic boundary, whereas the finite element model is equipped with infinite elements at the boundary. The dynamic response is assumed to be measured at eight sensor locations \( (x_j, y_j), j = 1 \ldots 8 \) at all times \( t \), as shown in Figure 1 and the adjacent table.

3.1 Stochastic parameter estimation

In order to get a realistic setting, we assume that the parameters \( E \) and \( \nu \) are randomly perturbed by background random fields \( R_E, R_\nu \), assumed to be independent, centered and Gaussian with autocovariance functions

\[ \text{cov}(R_E(x_1, y_1), R_E(x_2, y_2)) = \sigma_E^2 e^{-d_{12}/L_E} \]
\[ \text{cov}(R_\nu(x_1, y_1), R_\nu(x_2, y_2)) = \sigma_\nu^2 e^{-d_{12}/L_\nu} \]

where \( d_{12} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \) is the Euclidean distance. Further, \( \sigma_E, \sigma_\nu \) and \( L_E, L_\nu \) denote the respective standard deviations and correlation lengths. We set \( E(x, y) = E_{\text{mean}} + R_E(x, y) \) and \( \nu(x, y) = \nu_{\text{mean}} + R_\nu(x, y) \).

The procedure is as follows: The finite element model is fed with a realization of the described random field; the time-dependent output of the FE-simulation is recorded at the eight sensor locations and serves as measured data. The parameters \( E, \nu \) as well as the stochastic parameters \( \sigma_E, \sigma_\nu \) and \( L_E, L_\nu \) of the background field are estimated by comparing the data with the solution obtained by the Fourier integral operator representation (5), (7), (8). In Subsection 3.2 the procedure is further extended by adding random terms in the Fourier integral operators in order to get estimators that allow for damage detection (i.e., significant deviations of the material properties).

---

Figure 1: The driving force \( f \) and the position of the eight (virtual) sensors.
3.1.1 Estimating the nominal values

The first task is to estimate the nominal values $E_0 = E_{\text{mean}}$ and $\nu_0 = \nu_{\text{mean}}$ based on the data given in the sensor locations. One expects that the estimator for $E_0$ and $\nu_0$ varies more, if the variances of the random fields in the FE-model increase. It may also depend on the correlation lengths. This is about to be examined in the following.

To shorten notation, we let $g_j(t) = u_{\text{FEM}}(x_j, y_j, t)$ be the solution given by the FEM simulation with the given realization of the random field at sensor location $(x_j, y_j)$, recorded over the time interval $[0, T_{\text{max}}]$. Thus $g = [g_{d,j}], d = 1, 2$ is a matrix-valued function. The FIO solution $h_j(t, E_0, \nu_0) = u_{\text{FIO}}(x_j, y_j, t, E_0, \nu_0)$, solves the deterministic, constant coefficient case with parameters $E_0$ and $\nu_0$, and $h = [h_{d,j}], d = 1, 2$.

To examine the connection between $\sigma_{E,\text{opt}}$ on the one hand and $\sigma_E$ and $L_E$ on the other hand. Figure 2

A sample size of $N = 100$ is recommendable. However, for Young’s modulus the analogous estimate turned out to fail:

$\sigma_{E,\text{opt}}^2 \neq \sigma_{E_0,\text{opt}}^2 = \text{var}(E_0, \nu_0)$.

A sample size of $N = 100$ is recommendable. However, for Young’s modulus the analogous estimate turned out to fail:

$\sigma_{E,\text{opt}}^2 \neq \sigma_{E_0,\text{opt}}^2 = \text{var}(E_0, \nu_0)$.

The solution we adopted was to obtain a correction term empirically by Monte Carlo simulation. We examine the connection between $\sigma_{E_0,\text{opt}}$ on the one hand and $\sigma_E$ and $L_E$ on the other hand. Figure 2
shows the dependence of $\sigma_{E_0, \text{opt}}^2$ and $\sigma_E^2$ for a fixed $L_E$. One can clearly observe a linear dependence. The calculation was done with a sample size of $N = 100$ for each $\sigma_E$ from the list

$$\{0.001, 0.005, 0.01, 0.03, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40\} \cdot 70 \text{ GPa}. $$

Figure 3 shows the dependence of $\sigma_{E_0, \text{opt}}^2$ and $L_E$ for a fixed $\sigma_E^2$. This graph was used to fit a function $\tilde{f}(L_E)$. Combining this with the linear dependence of $\sigma_{E_0, \text{opt}}^2$ on $\sigma_E^2$ at fixed $L_E$, we arrive at the functional relation

$$\sigma_{E, \text{opt}} = \sigma_E \tilde{f}(L_E),$$

and so for a given $L_E$ we can estimate

$$\sigma_E = \sigma_{E_0, \text{opt}} / \tilde{f}(L_E).$$

(11)

Thus it remains to estimate the correlation length $L_E$. Since we have different sensor locations, one can also estimate the correlation between the sensors and from there the correlation length. We assume that the parameters in the sensor locations derive from a stationary random field obeying the covariance function

$$\text{cov}(E_{0, \text{opt}, j_1}, E_{0, \text{opt}, j_2}) = C(r) = \sigma_{E_0, \text{opt}}^2 \exp \left(-\frac{r}{L_{E_0, \text{opt}}} \right),$$

(12)

where $r = \sqrt{(x_{j_1} - x_{j_2})^2 + (y_{j_1} - y_{j_2})^2}$. We pair the sensors as shown in Table 1 and estimate

$$C(r_i) \approx \frac{1}{N |\mathcal{P}_i|} \sum_{k=1}^N \sum_{j_1,j_2 \in \mathcal{P}_i} \left(E_{0, \text{opt}, j_1}^{(k)} - \bar{E}_{0, j_1}(L_{E_0, \text{opt}}) \right) \left(E_{0, \text{opt}, j_2}^{(k)} - \bar{E}_{0, j_2}(L_{E_0, \text{opt}}) \right).$$

By fitting the empirical covariances to the autocovariance function, one gets an estimator for $L_{E_0, \text{opt}}$. Figure 4 shows the results of a Monte Carlo simulation with $N = 100$ for each $L_E$ from

$$\{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1, 1.5, 2, 3, 5, 6, 7.5, 9, 10, 20\} \text{ cm},$$

as well as the fitted scaling curve, resulting in a relation

$$L_E = g(L_{E_0, \text{opt}}).$$

(13)

To summarize, in order to estimate $\sigma_E$ and $L_E$, one starts by computing $\sigma_{E_0, \text{opt}}$ from (10). Then the empirical correlation length $L_{E_0, \text{opt}}$ is evaluated by fitting model (12) to the sensor data. Next, $L_E$ is estimated using the empirical relation (13), and finally $\sigma_E$ is obtained from the relation (11).
3.2 Testing for damage

This subsection is dedicated to devising statistical tests, indicating whether the properties of a given material are within certain bounds or not. The strategy is as follows. One generates a sample of signals in the sensor locations by means of stochastic Fourier integral operators. If the measured signal is not too different from the Monte Carlo sample, the material is considered “good”. The generation of the Monte Carlo sample can be done in advance of the testing.

To generate the sample we perturb the deterministic phase functions of the solution FIOs (see (7), (8))

$$
\varphi_w(x, y; \xi, \eta, s, t) = x\xi + y\eta + ic^2_w(t - s)\sqrt{\xi^2 + \eta^2}, \quad w = l, s
$$

by making $c_l$ and $c_s$ space dependent and setting

$$
c_l(x, y) = \sqrt{\frac{(\nu(x, y) - 1)E(x, y)}{(1 + \nu(x, y))(-1 + 2\nu(x, y))}}, \quad \text{resp.} \quad c_s(x, y) = \sqrt{\frac{E(x, y)}{2(\nu(x, y) + 1)}},
$$

that is, we replace the constant parameters $E$ and $\nu$ from formula (6) by random fields

$$
E(x, y) = E_{\text{mean}} + R_E(x, y), \quad \nu(x, y) = \nu_{\text{mean}} + R_\nu(x, y).
$$

Here $R_E$ and $R_\nu$ are Gaussian random fields of the form (9) with parameters corresponding to the undamaged state (the null hypothesis described below).

Since for engineering purposes usually the Young’s modulus is the critical factor we focus on detecting changes in $E_{\text{mean}}$ or $\sigma_E$.

The parameter values for the undamaged state were assumed to be $E_{\text{mean}} = 70$ GPa and $\nu = 0.35$; slight variations were admitted following background random fields with $\sigma_E = 3.5$ GPa, $\sigma_\nu = 0.005$ and $L_{E_{\text{opt}}} = L_{\nu_{\text{opt}}} = 3$ cm. These were the data entered in the FE-model as well as the wave speeds of the FIOs according to (14). In practice, estimated values would be entered in (14) according to Subsection 3.1.
The coefficients of variations were so small that no problems with the square roots in (14) could arise, and further cut-offs were not needed.

The null hypothesis was that the material is in the undamaged state, i.e., it has the parameters described above. The first step was to generate a sample of size $N = 1000$ of system responses, distributed according to the null hypothesis, by means of the FIO-solution operator with wave speeds (14). The null hypotheses is tested against FE-output generated with perturbed parameters or with inclusion of a crack.

In order to test the material one has to choose characteristic features of the signal. This choice is very critical, since a bad set of features can lead to a poor distinction between good and bad material. In our case it turned out that the total spectral energy of the signal, the phase angle and the spectral energy density of selected frequencies are working well enough. In fact, we had eight sensor locations with signals in $x$- and $y$-direction. The four signals in $x$-direction at sensors 2, 4, 5, 7 were close to zero. (Due to the special excitation, no shear waves arrive at sensors 4 and 5, and no pressure waves arrive at sensors 2 and 7.) Thus 12 signals remained for the analysis. Of each signal 5 features were extracted: the total spectral energy density and the phase angle and spectral energy density at the first two nonzero frequencies in the DFT-spectrum (in our case $\omega_1 = 2\pi/7 \approx 0.9$ MHz and $\omega_2 = 4\pi/7 \approx 1.8$ MHz). Thus a total number of 60 features was used for damage detection.

In principle, a significance test works as follows. One first chooses a significance level $\alpha$, say 1% or 5% and computes the $(1-\alpha)$-confidence interval of the feature under consideration from the sample as generated above. The null hypotheses is rejected if the measured value lies outside the confidence interval. In our situation such a test can be tuned in several ways:

(a) Choice of significance level $\alpha$. A small $\alpha$ means that rejection is more affirmative, while acceptance is easier.

(b) Decision about the number of signals leading to rejection or acceptance (out of the 60 available features).

(c) The overall choice of features taken into account.

Further, one could also use the $p$-values and a possible ranking of them in place of the test based on confidence intervals. In addition, a bias due to a possible modeling difference in the sample and the measurement may arise. In our case, the amplitude of the FIO-signal was systematically too small, because reflected waves could arise in the FE-model, but not in the FIO-representation. Notwithstanding these words of caution, we present the results of several scenarios, using a significance level of $\alpha = 1\%$, hence 99%-confidence intervals and a sample of FIO-solutions computed with the parameters of the undamaged state described above.

(1) Scenario 1: The material is having the desired properties (undamaged state).

(2) Scenario 2: The material is not having the desired properties: $E_{\text{mean}}$ is too small (60 MPa).

(3) Scenario 3: The material is not having the desired properties: $E_{\text{mean}}$ is too large (80 MPa).

(4) Scenario 4: The material is having the desired properties, but is suffering a crack. The crack is modeled by a small region in the FE-model having a very small Young’s modulus.

Figure 5 shows realizations of the random field $E(x, y)$ in Scenarios 1–4. As mentioned, the confidence bounds for the null hypotheses were generated by a sample of size 1000 simulated stochastic FIO solutions. Using a 99%-confidence interval in all scenarios, the following results were obtained in repeated test trials:

(1) Scenario 1: The null hypothesis is not rejected. In most of all trials, all of the 60 features accepted the null hypothesis. In some trials, acceptance happened in slightly fewer, but well above 55 features. This would also have been the case, had we used the more strict 95%-interval.

(2) Scenario 2: The null hypothesis is rejected in almost all sensor locations. Since $E_0$ is small, the signal arrives late, and the phase angle is out of the 99%-interval.

(3) Scenario 3: The null hypothesis is rejected in almost all sensor locations. Since $E_0$ is large, the signal arrives early, and the phase angle is out of the 99%-interval.

(4) Scenario 4: Since the crack lies between sensor 3 and the origin, the pulse changes. Thus the sensor in location 3 diagnoses the damage by rejecting the null hypothesis due to a large spectral energy, in total and at both frequencies $\omega_1$ and $\omega_2$, while the others do not. Since there is only one sensor location rejecting the null hypothesis, one can interpret this as the detection of the crack location.
Exemplarily, we show the graphical analysis for Scenarios 1 and 4 in Figures 6 and 7. The upper parts of the figures show the temporal evolution of the FEM-signal (red line) and of the whole Monte Carlo sample (shaded region) in sensor location 3. The histograms show the 99% confidence interval (black lines) and the value of the measured signal (red line), referring to the phase angle and the spectral energy density at frequency $\omega_2 = 1.8$ MHz, and the total spectral energy density (displacement in $x$-direction).

In our trial cases, the method was capable of arriving at the correct diagnosis in all four scenarios. Further trial cases aiming at the detection of changes in $\sigma_E$ were successfully undertaken as well. As a final remark, let us note that Gaussian random fields were used for the sake of demonstration. This has the advantage that the statistical properties of Gaussian random fields are fully controlled by their second moments, both in calibration and simulation. The strategy outlined here may be used with non-Gaussian random fields as well, should the data suggest it.

4 Summary

In this article, a new method for analyzing wave propagation in random media using records of the time dependent response of a dynamically excited structure has been presented. The method is based on the Fourier integral operator representation of the solution to the equations of motion of a linear elastic material. It has been demonstrated by means of data generated by a finite element model that the material parameters can be identified by fitting the solution computed through the Fourier integral operator representation. Random perturbations have been incorporated as additive terms in the phase functions of the Fourier integral operators. In this way, confidence intervals around the nominal response of the unperturbed structure can be calculated. This allows one to detect fatigue or damage in the material through hypothesis tests with the signals of the damaged structure. The intended goals of parameter identification and damage detection have been attained in benchmark examples.

The models can be implemented in real case applications after suitable adaptations. While in the paper, a two-dimensional problem has been addressed, the method is applicable in any space dimension. In the fully three-dimensional case, high computing power is required. A further topic of research could be to reduce the computational cost in the three-dimensional case. The computational cost in the one- and two-dimensional case is low enough to be performed on a standard computing equipment.
Figure 6: Comparison of signal (displacement in x-direction) and Monte Carlo sample in sensor location 3, showing acceptance of the null hypothesis at the 99% confidence level (Scenario 1, undamaged material).

Figure 7: Comparison of signal (displacement in x-direction) and Monte Carlo sample in sensor location 3, showing rejection of the null hypothesis at the 99% confidence level (Scenario 4, crack).

Acknowledgments

We thank our industrial partners INTALES Engineering Solutions, Natters, Austria, for important input and advice on the project. We gratefully acknowledge support by The Austrian Research Promotion Agency (FFG) in the framework of the Bridge Program, Project No. 4602529.

References


Possibilistic identification of reliable finite impulse response models

Dominik Hose* and Michael Hanss

Institute of Engineering and Computational Mechanics, University of Stuttgart, Germany

*Corresponding author: dominik.hose[at]itm.uni-stuttgart.de

Abstract

In this contribution, a novel approach to system identification based on concepts from possibility theory is presented. It can be employed to infer a fuzzy-valued finite-impulse response from input/output data yielding reliable predictions about the behavior of the identified dynamical system. Furthermore, a procedure for the analysis of the system properties in a possibilistic context is provided.

Keywords: Possibilistic Regression Analysis, System Identification, Finite Impulse Response, Fuzzy-Valued Modeling, Robust Control.

1 Introduction

System identification is an important task in a variety of disciplines where a full white-box model of the system under consideration is not easily obtained. One of the most simple approaches for black-box modeling, i.e. when no prior information about the system is available, is to consider the time-discrete finite-impulse response of the system [1]. This is often the first step in any system identification procedure and requires a regression analysis to determine the correlation between the inputs and outputs of the system.

In classical statistics, the aim is to find an estimator for the finite-impulse response, discriminating between what is considered the actual signal and what is assumed to be noise. The standard estimator for the finite-impulse response is the solution of an ordinary least-squares problem [2]. While providing a reasonable fit depending on the number of parameters and the training signal that is being employed, this estimation technique is not concerned with the necessary parameter variations to cover the whole output training signal. The residuals are attributed to noise even though the respective output may still contain system-relevant signals.

To overcome these limitations, the assumed model is considered to be imperfect and not able to account for all of the system dynamics by using just one vector of parameters. Consequently, a (fuzzy) set of parameter vectors is identified that is able to provide reliable bounds for the system output and enable e.g. the design of robust controllers.

To the authors’ knowledge, currently, few approaches to identifying fuzzy-valued parameters of dynamical systems exist. However, fuzzy regression analysis has recently received much attention. The multitude of proposed fuzzy regression algorithms (e.g. [3] or [4]) could be employed to identify fuzzy-valued finite-impulse-response models. Their major drawback is that they typically fail to provide a motivation for the choice of the membership functions of the regression parameters which are usually assumed to be of triangular shape. Other approaches for identifying fuzzy-valued model parameters from data can be found in [5], [6] and [7]. In addition to the aforementioned problems, they deal with highly non-convex optimization problems which can lead to computational infeasibilities. A methodology for identifying interval predictor models has been presented in [8] which yields comparable results – yet it does not provide a fuzzy membership function encoding further information about the involved uncertainties.

The presented approach is based on possibilistic regression analysis [9] and provides a tool to analyze the necessary parameter variations under the assumption that possible residuals stem from unmodeled dynamics – rather than from noise – leading to the conclusion that all the dynamics contained in the data should be accounted for by the fuzzy-valued parameters. Simultaneously, meaningful information is encoded in their membership function.

The resulting fuzzy-valued finite-impulse-response model is then employed to perform a possibilistic stability analysis and for the synthesis of control laws that are robust against the identified uncertainties. An illustrative example showing the advantages compared to the classical least-squares approach is provided.
2 Possibility Theory and Fuzzy Sets

Possibility theory provides a framework for measuring uncertainties similar to probability theory. The possibility measure $\text{Pos} : 2^\Omega \rightarrow [0, 1]$ on the universe of discourse $\Omega$ is based on the three axioms

1. $\text{Pos} (\Omega) = 1$,
2. $\text{Pos} (\emptyset) = 0$,
3. $\text{Pos} (A \cup B) = \max \{\text{Pos} (A) , \text{Pos} (B)\}$ for two disjunct subsets $A, B \subseteq \Omega$.

From these axioms, it follows that the possibility of an event $A$ is defined by the maximum possibility of all elements contained therein, i.e.

$$\text{Pos} (A) = \max_{a \in A} \text{Pos} (\{ a \}) = \max_{a \in A} \pi (a) ,$$

where $\pi$ is called a possibility distribution. If $A$ is uncountably infinite, the max-operator is replaced by the sup-operator. Since this measure is not self-dual, the necessity measure defined by

$$\text{Nec} (A) = 1 - \text{Pos} (\Omega / A) \quad \forall A \subseteq \Omega$$

is introduced additionally. For a detailed discussion of the extensive theoretic background of possibility theory, refer to e.g. [10].

Fuzzy sets were initially introduced by Zadeh [11]. They differ from classical sets in the way that the classical characteristic function, which assumes values of either 1 or 0, is generalized to a membership function $\mu : \Theta \rightarrow [0, 1]$ where $\Theta$ is a universal set (here $\Theta = \mathbb{R}^n$), formalizing the notion of gradual memberships. Each fuzzy set is uniquely defined by its membership function. The space of fuzzy sets on the reals is denoted $\mathcal{F} \mathbb{R}$, and the space of fuzzy vectors on $\mathbb{R}^n$ is denoted $\mathcal{F} \mathbb{R}^n$. A detailed discussion of fuzzy sets and the important subclass of fuzzy numbers can be found in [12]. Additionally, the support of a fuzzy set $\tilde{x} \in \mathcal{F} \mathbb{R}^n$ is the closure of all elements with non-zero membership

$$\text{supp} (\tilde{x}) = \{ x \in \mathbb{R}^n : \mu_{\tilde{x}} (x) > 0 \} .$$

Fuzzy theory is strongly connected to possibility theory [13] as hinted above. More precisely, the membership function $\mu_{\tilde{x}}$ of a possibilistic variable, i.e. a fuzzy variable $\tilde{x} : \Omega \rightarrow \mathbb{R}$, induces a possibility distribution $\pi_{\tilde{x}}$, according to

$$\mu_{\tilde{x}} (x) = \text{Pos} (\{ \omega \in \Omega : \tilde{x} (\omega) = x \}) = \pi_{\tilde{x}} (x) ,$$

in the same way a probability density function of a random variable induces a probability distribution. The value of $\pi_{\tilde{x}} (x)$ is the possibility that $\tilde{x}$ assumes the value $x$, and the resulting possibility measure is

$$\text{Pos}_{\tilde{x}} (X) = \sup_{x \in X} \mu_{\tilde{x}} (x) \quad \forall X \subseteq \mathbb{R} .$$

The subscript denotes a possibility measure induced by a fuzzy variable.

3 Possibility Propagation

Fuzzy arithmetic, i.e. the forward propagation of fuzzy variables or possibility distributions through deterministic mappings, is fundamentally based on the extension principle formulated by Zadeh in [14]. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a mapping and let $\tilde{x} \in \mathcal{F} \mathbb{R}^n$ be a fuzzy vector with membership function $\mu_{\tilde{x}}$. The membership function $\mu_{\tilde{y}}$ of the fuzzy output $\tilde{y} \in \mathcal{F} \mathbb{R}^m$ defined by

$$\tilde{y} = f (\tilde{x})$$

is given by

$$\mu_{\tilde{y}} (y) = \sup_{\tilde{x} : y = f (\tilde{x})} \mu_{\tilde{x}} (x) .$$

If $f$ is not surjective, it is not possible to find an $x$ satisfying $y = f (x)$ for all $y \in \mathbb{R}^m$. In this case, the membership of the respective element $y$ is simply zero. A proof of the extension principle employing possibilistic arguments is given in [15]. Several implementations for efficient computations with fuzzy sets and fuzzy numbers exist and are discussed e.g. in [12] or in [5].
4 Finite-Impulse-Response Models

Any time-discrete LTI (linear time-invariant) system \( G \) is uniquely defined by its impulse-response coefficients \( g[:\cdot] \), allowing to compute the output signal \( y[:\cdot] \) at the \( k \)-th time instant for a given input signal \( u[:\cdot] \) by the infinite discrete-time convolution equation

\[
y[k] = \sum_{n=-\infty}^{\infty} g[n]u[k-n]. \tag{8}
\]

A thorough discussion of this type of system representation can be found e.g. in [1]. Therein, it is stated that "merely knowing the impulse response \( g[:\cdot] \) is sufficient to predict the response of the system \( G \) to an arbitrary input."

Here, causality of the system in question, i.e. \( g[k] = 0 \) for all \( k < 0 \), is assumed. For asymptotically stable systems, it can furthermore be shown that the impulse response coefficients \( g[k] \) tend to zero for large \( k \) and can be considered negligible after some \( M \) time instants. The infinite sum then reduces to the finite convolution

\[
y[k] = \sum_{n=0}^{M} g[n]u[k-n] \tag{9}
\]

which is known as the finite-impulse-response (FIR) model of the system.

5 System Identification

Given a measured input \( u[:\cdot] \) and output \( y[:\cdot] \) at \( N + 1 \) sampling instants, the standard approach to the identification of the FIR model given in Eq. (9) is to include some additive measurement noise \( \nu \). Accordingly, the measured output signal is the superposition of the true output signal being the average of the input signal weighted with the finite-impulse response and the perturbation, i.e.

\[
y[k] = \sum_{n=0}^{M} g[n]u[k-n] + \nu[k]. \tag{10}
\]

Since the input for negative time instants is generally unknown, this equation can usually only be evaluated for \( k \geq M \). Defining the vector of FIR coefficients \( \theta \), the output vector \( y \) and the input matrix \( \Psi \) respectively by

\[
\theta = \begin{pmatrix} g[0] & \ldots & g[M] \end{pmatrix}^T, \quad y = \begin{pmatrix} y[M] & \ldots & y[N] \end{pmatrix}^T, \quad \Psi = \begin{pmatrix} u[0] & \ldots & u[M] \\ \vdots & \ddots & \vdots \\ u[N-M] & \ldots & u[N] \end{pmatrix}, \tag{11}
\]

the noise vector \( \nu \) can be expressed by \( \nu = y - \Psi \theta \). If the noise is white, i.e. it is uncorrelated, has a zero mean and finite variance, the maximum likelihood estimator for \( \theta \) is given by the solution to the ordinary least-squares problem

\[
\min_{\theta} \| y - \Psi \theta \|_2, \tag{12}
\]

whose analytic solution is given by

\[
\theta^{LS} = (\Psi^T \Psi)^{-1} \Psi^T y. \tag{13}
\]

A detailed description of the problem formulation, several solution methods and further implications are presented in [2]. This estimator is able to reproduce a reasonable approximation of the actual system dynamics. Evidently, the quality of the estimated FIR model depends largely on the quality of the data, i.e. the frequencies contained in the test signal, the signal-to-noise ratio, the 'whiteness' of the noise, etc. The choice of good test signals is a science in itself and covered e.g. in [1].

It is a justifiable point of criticism of this method that it is not able to provide robust predictions. If, for instance, the data did not contain noise but were obtained from the identification of a high-order system with complex (perhaps non-linear) dynamics and long settling time, one would still not expect the predictions to exactly match the data. This can partially be attributed to the fact that the assumption of uncorrelated noise is unwarranted in this case since the prediction error is actually resulting from system dynamics that are not accounted for in the model.
6 Possibilistic Regression Analysis

Possibilistic regression analysis [9] provides a tool to overcome the problems mentioned above by including the possibility that every sample is actually an undisturbed result of the system dynamics.

The first step is to find the $K = N - M + 1$ elementary parameters $\theta^i$ that exactly reproduce the output samples

$$y_i = y[i + M - 1], \quad \forall i = 1, \ldots, K$$

(14)

from the correlating inputs and yield the minimum deviation from the remaining data points. The $i$-th row of $\Psi$ is denoted by

$$\psi_i = (u[i - 1] \ldots u[i + M - 1])^T, \quad \forall i = 1, \ldots, K$$

(15)

and provides these inputs. Hence, all admissible parameters that produce a predicted output signal passing through the respective output sample $y_i$ are geometrically located on the line

$$S^i = \{ \theta \in \mathbb{R}^M : \psi_i^T \theta - y_i = 0 \}, \quad \forall i = 1, \ldots, K.$$  

(16)

Minimum deviation from the remaining data is ensured by employing e.g. the least-squares objective function. Of course, it is possible to choose other $p$-norms, that may be more appropriate for a given problem, such as the 1- or the $\infty$-norm. Even the minimization of the distance to the least-squares estimate

$$J(\theta) = \| \theta - \theta^{LS} \|_2$$

(17)

would be acceptable in order to ensure a fuzzy set with a tight support.

Choosing the 2-norm, the problem of finding the elementary parameters can be formulated as solving the $K$ convex optimization problems

$$\min_{\theta^i \in S^i} \| y - \Psi \theta^i \|_2, \quad \forall i = 1, \ldots, K.$$  

(18)

The Lagrange function of this constrained quadratic optimization problem is given by

$$L(\theta^i, \lambda) = (\Psi \theta^i - y)^T (\Psi \theta^i - y) + \lambda_i (\psi_i^T \theta^i - y_i), \quad \forall i = 1, \ldots, K$$

(19)

and the analytic solution yields

$$\begin{pmatrix} \theta^i \\ \lambda^i \end{pmatrix} = \left( 2\Psi^T \Psi \psi_i \psi_i^T 0 \right)^{-1} \begin{pmatrix} 2\Psi^T y \\ y_i \end{pmatrix}, \quad \forall i = 1, \ldots, K.$$  

(20)

It is also convenient to set $\theta^0 = \theta^{LS}$. Evidently, the support of a fuzzy-valued parameter $\tilde{\theta}$ should be an outer approximation of the convex hull of these elementary parameters. The most intuitive – yet weak – approximation is the hypercube

$$C = \{ \theta \in \mathbb{R}^M : \theta_j^{\min} \leq \theta_j \leq \theta_j^{\max}, \forall j = 1, \ldots, M \}$$

(21)

where

$$\theta_j^{\min} = \min_{i=0, \ldots, K} \theta_j^i \quad \text{and} \quad \theta_j^{\max} = \max_{i=0, \ldots, K} \theta_j^i$$

(22)

are the element-wise minima and maxima of the elementary parameters, respectively. Generally, other techniques, leading to alternative approximations of the convex hull, such as bounding hyperspheres or -ellipsoids, are also possible.

Intuitively, the objective values

$$J^i = \| y - \Psi \theta^i \|_2 \quad \forall i = 0, \ldots, K$$

(23)

represent a measure for the plausibility of the corresponding elementary parameters. The larger $J^i$ is, the less plausible, i.e. consistent with the data, the respective elementary parameter is due to a larger deviation measured by the 2-norm. E.g. the least plausible elementary parameter has the objective value

$$J_{\max}^i = \max_{i=0, \ldots, K} J^i.$$  

(24)
This sensitivity information can be encoded in the membership function of the fuzzy FIR coefficients $\tilde{\theta}$ by constructing a minimizing set [16]

$$m : J \mapsto \frac{J - J_{\text{max}}}{J_{\text{LS}} - J_{\text{max}}}$$

(25)

and defining the membership function

$$\mu_{\tilde{\theta}} : \theta \mapsto \max \{0, m(\|y - \Psi\theta\|_2)\}.$$ 

(26)

This particular choice for the membership function yields the least-squares solution as the nominal parameter and every elementary parameter is contained in the support by construction. Most importantly, the membership of the parameters is directly derived from their consistency with the data. Moreover, it ensures that the reference signal $y[\cdot]$ is contained in the fuzzy output domain of the fuzzy-valued finite-impulse-response (fFIR) model. A detailed discussion of the possibilistic regression method is presented in [9].

7 Possibilistic System Analysis

The presented identification procedure allows for robust predictions of the output of the identified system. This follows from the fact that all dynamics that are possibly contained in the data are accounted for by at least one parameter contained in the support of the fFIR coefficients $\tilde{\theta}$. If the $M + 1$ past inputs are known, it is possible to predict $\tilde{y}[k]$ by evaluating

$$\tilde{y}[k] = \sum_{n=0}^{M} \tilde{\theta}_n u[k - n].$$

(27)

This fFIR model can be used to perform a possibilistic system analysis. As an example, consider the task to design a stable feedback control law $u[k] = -Cy[k]$. Stability can be achieved if the largest pole of the closed loop $\lambda_{\text{max}}$ is located inside the unit circle. The fuzzy-valued largest closed-loop pole $\tilde{\lambda}_{\text{max}}$ is computed by applying the extension principle to the solution operator $E : \theta \mapsto \lambda_{\text{max}}$ finding the largest root of the closed-loop characteristic polynomial

$$p(z) = z^M + C \sum_{n=0}^{M} \tilde{\theta}_n z^{M-n},$$

(28)

i.e. by evaluating $\tilde{\lambda}_{\text{max}} = E[\tilde{\theta}]$. The (stability) event

$$S : \|\tilde{\lambda}_{\text{max}}\|_2 < 1$$

(29)

depends on the uncertain fFIR coefficients $\tilde{\theta}$, and thus, it is uncertain itself. Consequently, it has to be evaluated possibilistically, i.e. the degrees of possibility and necessity to which the event $S$ occurs have to be evaluated according to

$$\text{Pos}_{\tilde{\lambda}_{\text{max}}}(S) = \sup_{\lambda : \|\lambda\|_2 < 1} \mu_{\tilde{\lambda}_{\text{max}}} (\lambda).$$

(30)

In general, the following cases can be distinguished:

<table>
<thead>
<tr>
<th>Pos=0</th>
<th>Pos&gt;0</th>
<th>Pos=1</th>
<th>Pos=1</th>
<th>Pos=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nec=0</td>
<td>Nec=0</td>
<td>Nec=1</td>
<td>Nec&gt;0</td>
<td>Nec=1</td>
</tr>
</tbody>
</table>

- strict instability
- possible stability
- nominal stability
- guaranteed stability
- for certain realizations
- for certain realizations
- robust stability
- stability

In particular, if stability is to be guaranteed robustly, then the closed loop has to fulfill

$$\text{Nec}(S) = 1.$$ 

(31)

This translates to the whole support of $\tilde{\lambda}_{\text{max}}$ being inside the unit circle. Evidently, the identified fFIR model admits a smaller feasible set of robust controllers compared to the least-squares FIR model since the latter is contained in the former. More precisely, the fFIR model will not admit any controller that would destabilize any of the elementary dynamics described by the elementary parameters $\theta'$. All the possible dynamics contained in the data are stabilized. For further insight and a detailed explanation of the concept of possibilistic stability analysis, refer to [17].


8 Application

For the sake of clearness, a simple example will be given in this section. Suppose a time-discrete LTI system is given by the difference equation

\[ y[k] - \frac{3}{2} y[k - 1] + \frac{11}{16} y[k - 2] - \frac{3}{32} y[k - 3] = u[k - 1] - u[k - 2] + \frac{2}{9} u[k - 3]. \]

It will be shown that it suffices to identify an fFIR model of order \( M = 4 \) for a reasonable approximation. In order to do so, the system is excited with a random input signal of length \( N = 50 \) with zero mean and unit variance.

The signals predicted by the identified elementary FIR coefficients \( \theta^i \) are shown in Fig. 1. The close fits suggest that the model is chosen with a sufficient number of degrees of freedom permitting a tight approximation of the input/output behaviour.

![Figure 1: Signals predicted by the elementary parameters.](image1)

The resulting fFIR coefficients \( \tilde{\theta} \) are shown in Fig. 2. The true value of the impulse response is contained in the fuzzy coefficients and the identified uncertainty in the fFIR coefficients compensates for the neglected coefficients of the true impulse response.

![Figure 2: (Fuzzy-valued) Finite-impulse-response coefficients.](image2)

For validation purposes, the system is excited with a chirp input \( u[k] = \sin(\frac{1}{20} k^2) \). The identified fFIR model reliably predicts the system output as shown in Fig. 3 since the support of the predicted signal covers the entire reference signal with an adequate amount of uncertainty contained in the prediction supporting the confidence in the identified fFIR model.

Finally, a controller \( u[k] = -Cy[k] \) can be synthesized for the identified model. The fuzzy-valued largest closed-loop pole \( \lambda_{\text{max}} \) depending on the feedback gain \( C \) is shown in Fig. 4. A possibilistic evaluation of the stability event \( S \) is depicted in Fig. 5. Notice, that the procedure provides a reliable estimate for the region of stabilizing controllers. The possibilistic regression algorithm guarantees stability in the region where \( \text{Nec}(S) = 1 \) which is contained in the true region where the system is stable. In contrast, the FIR model identified by the classical least-squares procedure predicts stability where the true system would actually be unstable.
Figure 3: Validation: Predicted response to chirp input.

Figure 4: Largest closed-loop pole depending on the feedback gain $C$.

Figure 5: Evaluation of closed-loop stability for different feedback gains $C$. 


9 Conclusion

Despite the exemplary nature of the proposed procedure, it is generally applicable to a broad range of problems. Any other identification technique based on linear regression, such as auto-regressive modeling, can be performed in the same manner and other properties of the fuzzy-valued dynamical systems can be analyzed by means of possibilistic measures. The use of the proposed framework is recommended when reliable predictions about the completely unknown system behavior are required and the uncertainties are not assumed to stem from stochastic noise.

Ultimately, the presented approach extends the paradigm of building reliable input/output models by finding (fuzzy) set-valued model parameters in order to account for system-inherent uncertainties. Increasing computational power facilitates the evaluation of these models, and thus, their analysis can become a valuable tool for any engineer and scientist aiming at overcoming the limitations inherent to crisp-valued models.

Acknowledgements

The authors gratefully acknowledge the support provided by the German Research Foundation (DFG) in the framework of the research project no. 319924547.

References

Influence of uncertainty and numerical errors in the context of MIMO systems

Ekaterina Auer,1* Andreas Ahrens,1 André Sandmann,1 and César Benavente-Peces2

1 Dep. of Electrical Engineering, U. of Applied Sciences: Technology, Business and Design, Wismar, Germany
2 Dep. of Signal Theory and Communications, Universidad Politécnica de Madrid, Spain
*Corresponding author: ekaterina.auer[at]hs-wismar.de

Abstract

Multiple-input multiple-output (MIMO) systems can be considerably afflicted by uncertainty and numerical errors from various sources at different stages in the modeling and simulation process. In this paper, we perform a verification and validation analysis of the process for frequency selective MIMO systems and establish the respective current verification degree for the subtasks. Our special focus is on the stage of channel estimation, for which we show in detail how the verification degree can be improved using methods with result verification and analyse the uncertainty using Monte-Carlo and interval techniques. However, we also touch upon the stages of channel simulation and power allocation from the same point of view.

Keywords: Verification and validation, Methods with result verification, Simulation, frequency selective MIMO, Channel estimation

1 Introduction

To demonstrate how development in modern engineering works, it is commonly accepted nowadays [1] to consider the so called modeling and simulation cycle [2]. It comprises the three phases analysis, formalization and simulation, which can be reiterated in this order a number of times, the process known as the optimization (calibration) phase. The same cycle can be used to illustrate the verification and validation (V&V) activities which should accompany the development process. Devising and testing methodologies and tools for V&V assessment is a long-standing tradition originally from the area of computational fluid dynamics, supported by the American Society of Mechanical Engineers. The terms verification and validation are defined in the context of modeling and simulation tasks, of software engineering, and of numerical mathematics. Basically, the verification activities are those performed on the way from the formalization to the simulation to make sure that the implementation and its result are correct. Validation is performed en route from the simulation to the new analysis iteration to determine whether the results of the implementation have a good correspondence to reality, whatever ‘good’ and ‘reality’ mean in a given engineering context.

The researches in the V&V area develop compilations of requirements for users to categorize and classify processes simply by employing an assessment procedure or a criterion from a catalog [1, 3, 4]. The known methodologies usually rely on analytically solvable special cases, simplifications and benchmark examples if finding theoretical proofs is too difficult. That is, they do not provide a definitive step-by-step V&V procedure immediately applicable by the engineer. A possible step in this direction was described in [5, 6]. The major difference to the usual V&V methodologies is at the level of verification. Without developing specialized verification benchmarks to test the process, we suggest to perform the verification stage for the appropriately decomposed process itself (or at least, its vital parts), for example, using methods with result verification [7]. Such methods provide a mathematical guarantee that the results obtained on a computer – which, in the usual case, can turn out to be completely inaccurate due to finite precision of the computer arithmetic – are inside a certain region and not outside it. There are several successful examples of such computerized proofs, for example, [8]. As a kind of a by-product, we can propagate bounded uncertainty in parameters through systems using methods with result verification. Moreover, the areas of qualification/analysis and validation/simulation can profit from their use. For example, it can be demonstrated with their help whether the (numerical) formal model is chosen adequately, whether the parameter ranges are determined correctly, or whether the parameters to be neglected are selected appropriately.

In the times when wireless communications are omnipresent, multiple-input multiple-output (MIMO) systems play an important role. MIMO is well established in many wireless and wired communication standards and is used to increase the channel capacity without the need to increase the channel bandwidth.
or the transmit power [9]. The channel capacity is the information theoretic limit of the bit rate, that is, the number of bits per second, that can be transmitted through a physical channel error free. With MIMO, multiple data streams are transmitted on the same frequency band and at the same time. For separating the data streams, spatially separated channels are used, which can be achieved, for example, by placing multiple antennas at the transmitter and receiver side at different locations. Signal processing techniques such as singular value decomposition (SVD) are then used to remove the occurring inter-channel as well as inter-symbol interferences.

MIMO systems can be considerably afflicted by uncertainty and numerical errors from various sources at different stages in the modeling and simulation process. There are several conventional studies dealing with this problem especially at the stage of channel estimation, see for example, [9–12]. Moreover, there are several contributions dealing with formal verification of systems from the area of signal processing. For example, in the report [13], the authors give details about the ongoing work on formalization and property proofs of FAUST (Functional Audio Stream) programs using the tool COQ (coq.inria.fr). Note that formal verification (formal proofs, model checking) is different from result verification mentioned above since its goal is to prove the correctness of algorithms with respect to a certain formal property, usually using logic (or so called formal methods of mathematics). In this sense, formal verification and result verification complement each other and do not compete at the stage of verification. A reference of how to use interval methods [7] for identification of MIMO autoregressive systems with exogenous input is in [14].

All this work notwithstanding, a systematical treatment of MIMO modeling and simulation from the point of view of V&V still seems to be lacking. In this paper, we analyse the task of digital wireless communications using V&V techniques as described in [6]. A short overview of main concepts from the area of V&V is in Section 2. Our special focus in this paper is on the verification stage. In particular, we analyse in detail the influence of uncertainty and numerical errors for the subtask of broadband MIMO channel estimation, which we address using methods with result verification in Section 4. Moreover, we give a general overview of what needs to be done for verifying the subtasks at the power allocation and channel simulation stages in Section 3. A summary of the main results and our future work is in Section 5.

2 V&V: A Short Overview

Standard terminology [1] defines verification as the process of determining that a model implementation accurately represents the developers conceptual description of the model and the solution to the model. Validation is understood as the process of determining the degree to which a model is an accurate representation of the real world from the perspective of its intended uses. Note that this terminology, although widely spread, is not always used correctly or consistently. For example, the term 'channel verification' is actually used to mean 'channel validation' in [11]; the title 'Validated Numerics' in [15] should be 'verified numerics' according to the definitions above.

In [5], a classification of computer models (or simulations, abridged as CM in the following) associated with a given system (process) was proposed for easy identification of the verification degree.

C4: CM uses standard floating point or fixed point arithmetic (the lowest verification degree).

C3: CM is subdivided into appropriate subtasks and uses at least standardized IEEE 754 floating point arithmetic. Additionally, sensitivity analysis is carried out or uncertainty is quantified using traditional methods (e.g., Monte Carlo).

C2: Relevant subtasks are implemented using tools with result verification or delivering reliable error bounds. The convergence of numerical algorithms is proved via analytical solutions, computer-aided proofs, or fixed point theorems.

C1: The whole CM is verified using tools and algorithms with result verification or using real number algorithms, analytical solutions, or computer-aided existence proofs. Uncertainty is quantified and propagated throughout the CM using verified or stochastic (or both) approaches; sensitivity analysis is carried out.

The degree is tagged with a minus if the sensitivity analysis and uncertainty quantification were not carried out. A plus alongside a degree means that the code verification (e.g., using literate programming [16]) is performed additionally. For example, a CM categorized as having degree C5- corresponds to a situation common in small-scale engineering applications, where the model and the simulation use IEEE 754 floating point software and neither the sensitivity analysis nor uncertainty quantification need to be carried out.
Degree C3+ means that, in addition to C3, the code of the CM was verified, that is, it was proven somehow that the program for the model did not contain any implementation errors. Note that it does not necessarily mean that the result of a simulation with such a CM would be correct.

In [17, 18], a questionnaire was developed for determining the accuracy and correctness issues of a process and tested in a biomechanical context. This questionnaire is also useful for MIMO systems (cf. Section 4.1) since it helps to regulate the V&V assessment and to assign/improve the verification degree of the CM. In general, a typical questionnaire should aid in organizing a complex process with respect to initial data sources, their accuracy and its parameter and result ranges. Moreover, it should require precise description of special functions, algorithms (e.g., as a UML activity diagram), and employed mathematical operations. Finally, it should help to find an optimal output data format (or the appropriate means of data exchange).

3 Digital MIMO Communications from the Angle of V&V

In wireless systems, placing multiple antennas at the transmitter and receiver sides improves both the capacity and the integrity of a communications link, a strategy known as the MIMO method [19, 20]. A model for a frequency selective MIMO link consisting of \( n_T \) transmitting and \( n_R \) receiving antennas is

\[
y_k = \sum_{i=0}^{L_h-1} H_i \cdot c_{k-i} + n_k, \quad y_k, n_k \in \mathbb{C}^{n_R}, \quad c_{k-i} \in \mathbb{C}^{n_T}, \quad H_i \in \mathbb{C}^{n_R \times n_T},
\]

where \( k \in \mathbb{Z} \) is a specific instant of time, \( L_h \) is the number of channel taps per I/O path, \( y_k \) is the received data vector, \( c_{k-i} \) is the transmitted signal vector, \( n_k \) is the additive white Gaussian noise vector at the receiver side with a variance \( \sigma^2 \) in both real and imaginary parts, and \( H_i \) is the channel matrix.

To work with this model, solving the subtasks shown in Figure 1 is necessary. After the system configuration is established, the channel matrices \( H_i \) need to be estimated. After that, the frequency selective MIMO link is decomposed into a number of independent weighted frequency flat single-input single-output (SISO) links, for example, using the SVD technique [21]. The weights of such SISO links are usually not equal so that a need for external power allocation might arise. The goal during the power allocation stage is to optimize a certain quality criterion, for example, to minimize the bit error rate (BER). One possibility for that is to use the Lagrange multipliers method. In the following subsections, we describe the models for each of the subtasks in more detail.

![Figure 1: Subprocesses during the modeling and simulation of a (MIMO) digital channel](image)

### 3.1 Channel Estimation Stage

The first stage in the process is the channel estimation stage. One possibility for identifying the values of the channel matrices \( H_i \) is based on the least squares minimization. For frequency selective MIMO channels, a so-called pilot sequence of a certain length \( L_s \) is chosen for each MIMO input first. Usually, a sequence fulfilling a certain optimality condition should be preferred, but a random sequence of (complex) values can also be chosen. Denote by \( S \) the matrix containing pilot sequence vectors which are transmitted at the inputs \( \mu = 1 \ldots n_T \). Further, denote by \( r_\nu \) the corresponding received signal for each output \( \nu = 1 \ldots n_R \). The following model describes the interconnection between the received and transmitted signals:

\[
r_\nu = (S_1 \ S_2 \ \ldots \ S_{n_T}) \cdot \begin{pmatrix} h_{\nu 1} \\ \vdots \\ h_{\nu n_T} \end{pmatrix} + n_\nu \text{ or } r_\nu = Sh_\nu + n_\nu,
\]

where \( n_\nu \) is the white Gaussian noise vector and \( h_{\nu \mu} \in \mathbb{C}^{L_s} \). Here, \( h_{\nu \mu} \) denotes the \( L_s \) symbol-spaced channel taps between the MIMO input \( \mu \) and the output \( \nu \). This system has \((L_s - L_h + 1)\) equations.
Uncertainty and numerical errors in the context of MIMO systems

and \( n_T L_h \) unknowns, where \( L_h \) can also be understood as the number of matrices \( H_i \) to be estimated. That is, to be able to get a practically meaningful solution, the condition \( L_s - L_h + 1 \geq n_T L_h \) should be fulfilled. That means that the minimum length of the pilot sequence is \( L_s = n_T L_h + L_h - 1 \). In the linear system (2), the response \( r_ν \) and the information \( S \) can be considered as known. Moreover, \( n_ν \) has a known density with the zero mean and variance \( σ \). The channel matrices \( H_i \) consisting of \( h_ν \) need to be estimated, the exact relationship between \( h_ν \) and \( H_i \) being

\[
H_i = \begin{pmatrix}
h_{11}[i] & \cdots & h_{1nT}[i] \\
\vdots & \ddots & \vdots \\
h_{n1}[i] & \cdots & h_{nRnT}[i]
\end{pmatrix},
\]

where the notation \( h_ν[i] \) means the \( i \)-th component of the vector \( h_ν \). Since the system can be overdetermined, the optimal solution in the sense of the least squares minimization can be found by the following procedure (\( S^ν \) is the conjugate transposed of \( S \)):

\[
\begin{align*}
\begin{bmatrix} r_ν + n_ν \end{bmatrix} &= \begin{bmatrix} S^ν \end{bmatrix} \begin{bmatrix} r_ν + n_ν \end{bmatrix} = S^ν S^ν \\
&= (S^ν S)^{-1} S^ν r_ν + (S^ν S)^{-1} S^ν n_ν \\
&= \begin{bmatrix} h_ν \end{bmatrix} = (S^ν S)^{-1} S^ν (r_ν + n_ν).
\end{align*}
\]

Numerically more feasible is solving the system of linear matrix equations \( S^ν (R + N) = S^ν S H \), where \( R = (r_ν) \) for all \( ν \) and \( N \) is the matrix of noise. In the presence of the Gaussian noise, it is advantageous to choose pilot sequences with higher lengths for achieving more accurate approximations for \( H \). On the other hand, higher pilot lengths reduce the capacity for transmitting real, useful information [12].

Pilot sequences can be chosen such that \( S^ν.S = αI \). Here, \( I \) is the identity matrix, \( α = (L_s - L_h + 1)U_s^2 \) and \( U_s \) is the half level transmit amplitude. This optimality condition simplifies formula (3) even further:

\[
\begin{bmatrix} h_ν \end{bmatrix} = \frac{1}{(L_s - L_h + 1)U_s^2} \cdot S^ν \begin{bmatrix} r_ν + n_ν \end{bmatrix}.
\]

More complex models can be considered, but for the purposes of the first study, we analyse this simple estimation procedure from the point of view of V&V. Since this analysis is in the focus of this paper, we give the details separately in Section 4. There, we show how to improve the overall verification degree to C2 for this part of the process.

### 3.2 Channel Simulation Stage

For ease of presentation, we consider w.l.o.g. \( L_h = 1 \) in this section and in Section 3.3. The resulting frequency flat model can be readily extended to frequency selective channel conditions as shown in [21]. To obtain a number of independent, weighted SISO links from a given MIMO channel \( H = H_0 \), a well-established SVD technique can be used. It is assumed that the matrix \( H \) can be decomposed as \( H = U \cdot Σ \cdot V^H \). Here, \( U \) and \( V \) are unitary matrices and \( Σ \) is the diagonal matrix with real elements. The matrix \( Σ \) contains the positive square roots of the eigenvalues of \( H^H H \) in descending order on the main diagonal. They are called singular values and denoted by \( √ξ_r \) in the following. If we consider a pre-processed data vector \( x := V \cdot c \) and post-process the corresponding receive signal \( z := Hx + n \) by multiplying it by \( U^H \), then

\[
u := U^H z = U^H (U^H V^H) V^H c + U^H n = Σ c + w.
\]

In this way, the MIMO system is (theoretically) transformed into \( L = \min \{n_T, n_R\} \) independent, non-interfering layers \( u_i \) having unequal weights \( √ξ_r \):

\[
u = √ξ_r c_i + u_i \quad \text{for} \quad i = 1 \ldots L.
\]

Other methods for achieving the same goal of separating the SISO links can be employed, for example, the geometric mean decomposition (GMD). There are a number of difficulties during this stage due to numerics or uncertainty. For example, \( U^H \cdot U \) or \( V^H \cdot V \) are not exactly identity matrices due to numerical errors, creating interference between the links. Moreover, the matrix \( H \) is strictly speaking uncertain, making the weights also uncertain. It is therefore a challenging and important task to quantify their influence at this stage. The verification degree is C3 at the moment.
The improvement of the verification degree is challenging since it is not easy to perform this stage using methods with result verification, especially if SVD is employed. There is a certain amount of research on bounding real singular values of (interval) matrices, for example, [22–26]. However, an open access interval implementation of a singular value decomposition does not seem to exist. Note that the module VERTHINSVD [27] (uivtx.cs.cas.cz/~rohn/matlab/) is not available online at the moment. Moreover, the interval equivalents of matrices $U$ and $V$ are not especially meaningful, as explained in [28].

There is a possibility to (approximately) decompose a MIMO into several SISO links based, for example, on the GMD [29]. Moreover, the SVD decomposition can be performed using the CORDIC algorithm [30]. There is a certain amount of research on bounding real singular values of (interval) matrices, for example, [22–26]. However, an open access interval implementation of a singular value decomposition does not seem to exist. Note that the module VERTHINSVD [27] (uivtx.cs.cas.cz/~rohn/matlab/) is not available online at the moment. Moreover, the interval equivalents of matrices $U$ and $V$ are not especially meaningful, as explained in [28].

In our current paper [31], we analysed the power allocation for frequency flat MIMO channels using a combined analytical-verified approach. If the Lagrange multipliers method is used to minimize the BER as the quality criterion, then the following cost function needs to be minimized:

$$J(p_1 \ldots p_L, \lambda) = \frac{2}{\sum \log_2 M_l} \sum_{l=1}^{L} \left(1 - \frac{1}{\sqrt{M_l}}\right) \cdot \text{erfc} \left( \frac{1}{2\sigma} \sqrt{\frac{3p_l \cdot \xi_l \cdot P_s}{L(M_l - 1)}} \right)$$

(7)

Here, $p_1 \ldots p_L$ are the power allocation parameters with which we modify the weights $\sqrt{\xi_l}$ from Eq. (6) in order to improve the overall BER. Moreover, $M_l$ is the number of symbols for the layer $l \in \{1, \ldots L\}$, $\lambda$ is the Lagrange multiplier assigned to the constraint $\sum_{l=1}^{L} p_l = L$ and $P_s$ is the overall transmit power. This problem can be solved using interval methods with the help of the general-purpose global optimizer, for example, that implemented in the C-XSC Toolbox [32]. As a result, a verified upper bound on the lower bound for the BER can be computed. This leads to the overall verification degree of C2- for this part of the process. However, we could show that much better results (tighter interval bounds on $p_1 \ldots p_L$) could be obtained by computing the derivatives analytically and solving the resulting system of nonlinear equations numerically using the corresponding verified solver provided, again, by the C-XSC Toolbox. The verification degree remains the same.

The verified upper bound on the BER (not on the lower bound of the BER) can be computed if the interference between the SISO layers of a MIMO link (cf. Section 3.2) is taken into account in the proper way. This is a part of our future work.

4 Estimating the Channel Matrix

In this section, we analyse the stage of the channel estimation in detail. First, the questionnaire leading to the assessment of the verification degree is reproduced. Next, the conventional procedure for channel estimation based on the least squares method is described. Finally, the results for the verified version of this subprocess are given.

4.1 Questionnaire for Channel Estimation

In this subsection, we reproduce the answers to the verification questionnaire for the stage of channel estimation. In practice, the algorithms can be implemented on an Application Specific Integrated Circuit (ASIC) or a Digital Signal Processor (DSP) which usually store data and carry out computations using a fractional/fixed point number format. We call it practice mode in the following. Alternatively, our goal can be testing the developed algorithms, in which case a PC simulation is sufficient (simulation mode).

The questions concern four distinct areas inside the subprocess: input data, models, algorithms and output data. They help to establish the current verification degree as C4- for the practice mode and C3 for the simulation mode. We improve the degree to C2- by using methods with result verification as described in Section 4.3.

4.1.1 Description of input data

The following initial information is necessary for this stage of the process: $n_T$, $n_R$, $L_h$ (integer numbers), the pilot sequences $S$ stored in a look-up table, and the received data $R$ from an analog to digital
The source of the initial received data can be an oscilloscope (for both the practice and the simulation mode) or a computer simulation; everything else is predefined.

Description form: The data is stored in a text file or in MATLAB script files. The pilot sequences’ bits have the boolean data type. Symbols obtained by combining the corresponding bits can become integer or real or even complex depending on the application. The oscilloscope data is in Hierarchical Data Format (HDF5), the basic storage data type being fractional/fixed point (practice mode) or double IEEE 754 type (simulation mode).

Pre-selection: There is no data selection in this subprocess.

Accuracy: The pilot sequences are known exactly. The received data is provided in a discretised form by the ADC. Moreover, the receiver is disturbed by the Gaussian noise.

4.1.2 Description of models

The formulas are given in Section 3.1. The typical values of parameters are $n_T > 1$, $n_R > 1$ (2, 3 or 4). The length of a pilot sequence $L_s$ depends on the application. At the moment, there are no transmission standards using frequency selective channel estimation. In our example simulation in this section, we use the lengths 5, 33, 321. The sequences themselves can be devised analytically to fulfill the optimality condition $S^H S = a I$ [12] or can be generated randomly.

4.1.3 Description of algorithms

Type: The conventional algorithm described in Section 4.2 is of the numeric type for the solution of linear systems and of the stochastic type for the Monte Carlo simulation taking care of the noise.

Parallelisation: The algorithm is parallelisable. For example, the channel coefficients can be computed in parallel for different MIMO outputs.

Architecture: ASIC or DSP in the practice mode. The current lab implementation uses a standard PC (simulation mode).

Operations: (Matrix) addition, multiplication. Additionally for arbitrary pilot sequences: transpose, complex conjugation.

Sub-algorithms: Least squares solution for overdetermined linear systems (for arbitrary pilot sequences).

There are no known nonnumerical factors the algorithm is especially sensitive to. The UML description of the algorithm is omitted for the sake of brevity.

4.1.4 Description of output

Data type: The output data (the estimated coefficients of $H_i$) are in the IEEE double format for the simulation mode, fixed point format for the practice mode; channel coefficients are unitless.

Accuracy: See the the mean squared error in Eqs. (8)-(9).

High accuracy: The estimations of the coefficients are the more accurate the higher the length of the pilot sequence $L_s$. However, this reduces the amount of transmitted useful information.

Failures: Aside from the possibility of the usual failures connected to the floating/fixed point numerics, the algorithm is robust.

Information exchange: The data is stored in a text file using the decimal number representation with 15 digits after the decimal point (simulation mode). In the practice mode, they stay in the memory in the usual fixed point format. There is no immediate need to plot the output data. The estimated coefficients are used afterwards for channel simulation.
4.2 Conventional Channel Estimation
In this section, the results achieved by the least squares broadband MIMO channel estimation algorithm are compared for different pilot sequence lengths $L_s$ by means of a Monte Carlo simulation and a conventional floating-point implementation. For validation purposes, we choose a two tap frequency selective $(2 \times 2)$ MIMO channel, i.e., $L_h = 2$ and $n_R = n_T = 2$, with the known matrices $H_0$ and $H_1$ defined by

$$H_0 = \begin{pmatrix} 4 & 12 \\ 5 & 25 \\ 2 & 16 \\ 5 & 25 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.48 \\ 0.4 & 0.64 \end{pmatrix}, \quad H_1 = \begin{pmatrix} 2 & 6 \\ 5 & 25 \\ 1 & 8 \\ 5 & 25 \end{pmatrix} = \begin{pmatrix} 0.4 & 0.24 \\ 0.2 & 0.32 \end{pmatrix}.$$  

The resulting empirical probability density functions (PDFs) of the estimated channel taps are shown in Fig. 2 for three different pilot sequence lengths and $\sigma$ corresponding to a fixed signal-to-noise ratio of $30$ dB. The results show that a higher pilot sequence length reduces the variance of the estimated

![Figure 2: Empirical probability density functions of the estimated MIMO channel taps comparing different pilot sequence lengths — dash-dotted line: $L_s = 5$, solid line: $L_s = 33$, dashed line: $L_s = 321$](image)

amplitudes. Specifically, the mean squared error (MSE) between the estimated channel taps in vector $\hat{h}_\nu$ and the original taps in vector $h_\nu$ can be taken as a quality indicator as defined by

$$\text{MSE} = \frac{1}{n_T n_R L_h} \cdot E \left\{ \sum_{\nu=1}^{m_n} (\hat{h}_\nu - h_\nu) \right\}^H \cdot (\hat{h}_\nu - h_\nu) \right\} ,$$

with $E\{ \cdot \}$ denoting the expectation function. Under additive white Gaussian noise disturbance and the use of optimal pilot sequences the MSE can be expressed analytically [12] by

$$\text{MSE}_{LS} = \frac{\sigma^2}{P_s} \cdot \frac{n_T}{L_s - (L_h - 1)} .$$

This shows that choosing a longer pilot sequence reduces the estimation error, which confirms the Monte Carlo simulation results.

4.3 Computing Verified Bounds on the Channel Matrix
We implemented the whole channel estimation process using interval arithmetic and, in particular, the linear matrix equations solver provided by the C-XSC Toolbox. Instead of the Gaussian noise, we used the corresponding bounding intervals of $[-\sigma, \sigma]$, $[-2\sigma, 2\sigma]$ and $[-3\sigma, 3\sigma]$. Using such intervals means that we suppose that the density inside the bounds corresponds to the uniform distribution. The real numbers that we give are rounded to the third digit after the decimal point, if appropriate. The intervals are rounded outwards to the same number of digits. The length of the pilot sequence for the results shown below is $L_s = 5$. The channel matrices are the same as in the previous subsection, with the estimated values using Eq. (3) for the uncertainty of $\pm \sigma$ ($\pm 0.0316$) equal to

$$H_0 = \begin{pmatrix} 0.755, 0.845 & 0.435, 0.525 \\ 0.355, 0.445 & 0.595, 0.685 \end{pmatrix}, \quad H_1 = \begin{pmatrix} 0.355, 0.445 & 0.195, 0.285 \\ 0.155, 0.245 & 0.275, 0.365 \end{pmatrix},$$

$$\text{mid}(H_0) = \begin{pmatrix} 0.8 & 0.48 \\ 0.4 & 0.64 \end{pmatrix}, \quad \text{mid}(H_1) = \begin{pmatrix} 0.4 & 0.24 \\ 0.2 & 0.32 \end{pmatrix},$$

REC 2018 – E. Auer, A. Ahrens, A. Sandmann, C. Benavente-Peces
Uncertainty and numerical errors in the context of MIMO systems

where the notation mid(·) means the matrix of the midpoints. The maximal width of the interval components of matrices $H_0$, $H_1$ is equal to 0.089443. Note that the midpoint matrices correspond to the real chosen channel. The estimated $H_0$, $H_1$ for the uncertainty of ±σ under the consideration of the optimality condition of the pilot sequences are

$$H_0 = \begin{pmatrix} 0.755, 0.845 & 0.435, 0.525 \\ 0.355, 0.445 & 0.595, 0.685 \end{pmatrix}, \quad H_1 = \begin{pmatrix} 0.355, 0.445 & 0.195, 0.285 \\ 0.155, 0.245 & 0.275, 0.365 \end{pmatrix}$$

with the same midpoint matrices. That is, the estimated matrices are also numerically the same for both computational variants from Eqs. (3),(4). $H_0$, $H_1$ for the uncertainty of ±2σ (±0.063) are

$$H_0 = \begin{pmatrix} 0.710, 0.890 & 0.390, 0.570 \\ 0.310, 0.490 & 0.310, 0.490 \end{pmatrix}, \quad H_1 = \begin{pmatrix} 0.310, 0.490 & 0.150, 0.330 \\ 0.110, 0.290 & 0.230, 0.410 \end{pmatrix},$$

$$\text{mid}(H_0) = \begin{pmatrix} 0.8 \ 0.4 \\ 0.4 \ 0.64 \end{pmatrix}, \quad \text{mid}(H_1) = \begin{pmatrix} 0.4 \ 0.24 \\ 0.2 \ 0.32 \end{pmatrix},$$

with the maximal width of 0.178885. Finally, $H_0$, $H_1$ for the uncertainty of ±3σ (±0.095) containing 99.7% of the possible values for the Gaussian noise are

$$H_0 = \begin{pmatrix} 0.665, 0.935 & 0.345, 0.615 \\ 0.265, 0.535 & 0.505, 0.775 \end{pmatrix}, \quad H_1 = \begin{pmatrix} 0.265, 0.535 & 0.105, 0.375 \\ 0.065, 0.335 & 0.185, 0.455 \end{pmatrix},$$

again with the same midpoint matrices. The maximal width is 0.268328.

The estimated $H_0$, $H_1$ for $L_s = 33$ and $L_s = 321$ are almost exactly the same. That is, the length of the pilot sequence does not play any role for the quality of the obtained intervals and can be chosen to be minimal ($L_s = 5$). The result is not surprising in this setting since the artificially generated data for the received vectors in the matrix $R$ are obtained not by adding a specific value of the noise to the coefficients of $H_1$ (as in Section 4.2) but by considering an interval enclosing relevant values which is always the same for a given $\sigma$. That is, the optimal solution of the overdetermined system in Eq. (2) in the sense of the least squares method is the same independently of the pilot sequence length. However, we expect similar behaviour in the real life settings using actual measurements for the responses $R$.

In Figure 3, the relationship between the PDFs obtained with the Monte Carlo simulation and the interval enclosure containing 99.7% of the possible noise values is exemplified for the coefficient in the first row and the first column of the channel matrix $H_0$. The conservativeness of the interval enclosure in the figure is explained by the width of the input uncertainty which is of order $2 \cdot 10^{-1}$ (±3σ). The result has the width of the same order of magnitude which is quite good for interval methods considering that the task is linear.

![Figure 3: Channel coefficient $H_{01}(1,1)$: Empirical PDFs for pilot sequence lengths $L_s = 5$ (dash-dotted line), $L_s = 33$ (solid), $L_s = 321$ (dashed) and the interval enclosure (blue line) for ±3σ](image)

5 Conclusions

In this paper, we analysed the task of digital MIMO communications using V&V techniques. We described how to improve the verification degree of the three possible stages of this process with a special focus on
the channel estimation for frequency selective MIMO systems. We could show that verified techniques could help to reduce the pilot sequences’ lengths at the same time taking care of numerical errors and enclosing the additive uncertainty. As future work, the proposed interval estimation procedure should be tested for real life systems with unknown channels in the simulation mode. Our long-term goal is the implementation of interval procedures for the practice mode (i.e., on ASIC chips).

References


Surrogate modelling for solving optimization problems with polymorphic uncertain data

Steffen Freitag,* Philipp Edler, Katharina Kremer, and Günther Meschke
Institute for Structural Mechanics, Ruhr University Bochum, Germany
*Corresponding author: steffen.freitag[at]sd.rub.de

Abstract
The solution of optimization problems with polymorphic uncertain data requires to combine stochastic and non-stochastic approaches. In this paper, a concept is presented, which allows one to consider uncertain a priori parameters and uncertain design parameters quantified by stochastic numbers and intervals. To solve optimization problems in structural mechanics by means of iterative optimization algorithms (e.g. particle swarm optimization), often multiple runs of nonlinear finite element models with varying a priori and design parameters have to be performed. For each design to be optimized, an interval analysis in combination with Monte Carlo simulations is necessary. This can only be realized by substituting the nonlinear finite element model by numerically efficient surrogate models. In this paper, a strategy for neural network based surrogate modelling is presented. Instead of just replacing the deterministic finite element simulation, it is focused on surrogate models to replace the stochastic simulation. The approach is verified by an analytical solution and applied to optimize the concrete cover of a reinforced concrete bridge structure taking the variability of material parameters and construction imprecision into account.

Keywords: Surrogate Model, Artificial Neural Network, Optimization, Polymorphic Uncertainty

1 Introduction
Uncertain parameters can be quantified by stochastic distributions within the structural design process. In [1], an overview on reliability-based design optimization approaches is presented. In general, a surrogate optimization problem has to be defined, because the solution of an optimization problem requires deterministic measures. Such measures are mean values, variances or quantile values of the original objective function, which allows one to consider the robustness by optimizing the mean value and minimizing the variability, see e.g. [2], and [3].

In addition to stochastic models, uncertain parameters can be described by intervals or fuzzy numbers. Surrogate objectives can be formulated by means of worst case scenarios, e.g. minimizing the upper bound of an interval objective function, in case of interval parameters or by means of defuzzified measures (e.g. fuzzy mean, fuzzy centroid) in case of fuzzy parameters. Stochastic and non-stochastic models can be combined to polymorphic uncertainty models within structural optimization approaches, see e.g. [4] and [5].

In this paper, optimization approaches are presented, which allow to combine stochastic and interval design and a priori parameters. A particle swarm optimization algorithm [6] is applied to solve surrogate optimization problems, e.g. to minimize the worst case mean value of an objective function. This requires to perform Monte Carlo simulations and optimization-based interval analyses of deterministic structural simulations for each optimization run. In order to reduce the computational effort of the deterministic simulation and the Monte Carlo simulation, artificial neural network (ANN) surrogate models are applied. Whereas ANN surrogate models have already been developed to replace deterministic simulations within stochastic analyses, see e.g. [7–10], or to replace interval simulations within interval stochastic analyses, see e.g. [11], an approach for replacing the stochastic simulation with ANN surrogate models is introduced in this paper.

After a verification test, the presented surrogate modelling strategy is applied to minimize the crack widths of a reinforced concrete bridge structure. A nonlinear finite element model is utilized to compute the load bearing capacity and the crack patterns of the structure. Based on the finite element simulation results, an ANN surrogate model is trained to approximate the objective function. Within the optimization, the Young’s modulus of concrete is modelled as a stochastic a priori parameter and the position of the reinforcement layers are defined as interval design parameters, with a fixed radius and midpoints to be optimized. Based on Monte Carlo simulation results of the ANN surrogate model, another ANN is trained to replace the Monte Carlo simulation within the optimization runs.
2 Optimization with polymorphic uncertain data

Optimization with polymorphic uncertain data means, that stochastic and non-stochastic uncertain parameters are combined to solve an optimization problem. It is classified into uncertain a priori parameters, which cannot be optimized, and design parameters, which are to be optimized. Here, optimization approaches combining stochastic numbers and intervals are presented, which results in the following possible parameter representations:

- uncertain a priori parameters
  - stochastic numbers $A$
  - intervals $a$
- design parameters
  - deterministic numbers $d$
  - stochastic numbers $D$
  - intervals $d$

Whereas a stochastic number $X$ is defined by its probability density function $f(x)$, an interval $x = [l_x, u_x]$ is quantified by its lower and upper interval bounds $l_x$ and $u_x$, respectively, or by its midpoint

$$m_x = \frac{1}{2} \cdot (l_x + u_x),$$  \hspace{1cm} (1)

and radius

$$r_x = \frac{1}{2} \cdot (u_x - l_x).$$  \hspace{1cm} (2)

To solve optimization problems with uncertain a priori or design parameters, surrogate objectives have to be defined, because the minimization or maximization of an objective function $Z$ requires a deterministic representation. In case of stochastic uncertainty, typical surrogate objectives to be optimized are the mean value $\mu(Z)$, the standard deviation $\sigma(Z)$ or quantiles of $Z$. The surrogate objectives can also be combined resulting in a multiple objective optimization, e.g. minimizing the mean value and the standard deviation of an objective function to obtain a robust optimal design. In case of interval uncertainty, the midpoint $m_z$, the radius $r_z$, the lower bound $l_z$ or the upper bound $u_z$ of the objective function can be defined as surrogate objectives, e.g. resulting in a worst case optimization, if the upper bound of the interval objective function is minimized.

If both, intervals and stochastic numbers are considered within an optimization problem, the surrogate objectives for stochastic uncertainty and interval uncertainty can be combined, e.g. worst case mean value $\min\{\max\{\mu(Z)\}\}$, which means that the upper bound of the mean value of the objective function is minimized. This is visualized in Figure 1, for the combination of an interval design parameter $d$ (with fixed radius $r_d$ and midpoint $m_d$ to be optimized) and stochastic a priori parameters. It can be seen, that for each deterministic design $d$, the value of the objective function $Z$ is a stochastic distribution with a mean value $\mu(Z(d))$ and for an interval design $d$, an interval of mean values is obtained. The optimal design is marked by the optimal interval midpoint $m_d$.

Optimization problems with stochastic and interval parameters can be solved by two possible approaches, see Figure 2. In both approaches, four computational loops are required and in the first loop,
an optimization algorithm, e.g. a particle swarm optimization algorithm [6], is applied to solve the optimization problem with a surrogate objective, where also constraints can be considered. The fourth loop contains the deterministic structural simulation. The interval dominated approach requires an interval analysis (e.g. by using an optimization-based approach according to [12]) with stochastic realizations in the second loop and a stochastic analysis (e.g. by using Monte Carlo simulation) in the third loop, see Figure 2a. Within the stochastics dominated approach, a stochastic analysis with interval samples (e.g. Interval Monte Carlo simulation [13]) is performed in the second loop and an interval analysis (e.g. by using an optimization-based approach according to [12]) in the third loop, see Figure 2b.

Figure 2: Computational schemes for optimization tasks with interval and stochastic parameters.

If numerical simulation techniques are applied for the second and third loops, such as the Monte Carlo simulation for the stochastic analysis and an optimization-based interval analysis, a high number of deterministic structural simulations (samples) are required. In case of detailed numerical models, e.g. based on the finite element method, numerically efficient surrogate models can help to solve optimization problems with polymorphic uncertain parameters.

3 Neural network based surrogate modelling

Several artificial neural network (ANN) approaches have been developed to approximate time consuming FE simulations in structural mechanics, see e.g. [14]. In [15] and [16], different network architectures are presented. Here, it is focused on feedforward neural networks. They are used to map the inputs (realizations of the uncertain a priori and design parameters) onto the outputs (corresponding value of the objective function). A feedforward neural network consists of an input layer, a number of hidden layers and an output layer. The neurons of each layer have synaptic connections to the neurons in the previous and following layers. Whereas the number of input and output neurons are given by the approximation problem (i.e. number of design and uncertain a priori parameters and number of objective functions), the number of hidden layers and hidden neurons have to be defined according to the complexity of the objective functions to be approximated.

Starting from the input layer to the output layer, the signals of a neuron in a feedforward ANN are computed by

\[
x_i^{(m)} = \varphi_i^{(m)} \left( \nu_i^{(m)} \left( \sum_{h=1}^{H} x_{ih}^{(m-1)} \cdot w_{ih}^{(m)} + b_i^{(m)} \right) \right),
\]

where \( \varphi_i^{(m)}(.) \) is the activation function of neuron \( i \) in layer \( (m) \). The argument \( \nu_i^{(m)} \) of the activation function contains the sum of all output signals \( x_{ih}^{(m-1)} \) of the previous layer \( (m-1) \) multiplied by the corresponding synaptic weights \( w_{ih}^{(m)} \) and adding a bias value \( b_i^{(m)} \). Different activation functions \( \varphi_i^{(m)}(.) \) can be used, e.g. linear function, logistic function (sigmoid function), hyperbolic tangent function or area hyperbolic sine function.

The weights and bias values of the ANN are determined within the network training. For the ANN approximation of an objective function, a sufficient number of supporting points are defined (e.g. by regular grids, random sampling or Latin hypercube sampling) to create patterns of input and output data. The whole data set is divided into training, testing and verification data to guarantee a good approximation performance and to avoid overfitting. Backpropagation algorithms, see e.g. [15], are the most commonly used approaches to train an ANN by minimizing the error between the network outputs and the desired responses at the supporting points.

In general, ANN surrogate models are used to approximate the deterministic simulation model. However, in case of simulations with polymorphic uncertain parameters, also the interval analysis or the stochastic analysis can be replaced by surrogate models. To solve optimization problems, it would be beneficial, if all a priori uncertain parameters are part of the neural network approximation, because they are fixed and cannot be changed during the optimization, and just the design parameters are defined as...
ANN inputs. In Figure 2, four possible surrogate modelling strategies are shown for optimization tasks with interval and stochastic parameters:

- surrogate model for deterministic simulation
- surrogate model for interval analysis
- surrogate model for stochastic analysis
- surrogate model for interval stochastic analysis

To replace the interval analysis, see Figure 2b, either the surrogate objective (e.g. midpoint, lower or upper bound of the objective function) is defined as ANN output and a ANN is trained with deterministic data or the whole interval analysis is replaced by an ANN with interval signal processing, see e.g. [17]. The second approach has been applied in [11] to replace the interval analyses within an Interval Monte Carlo simulation.

In this paper, a concept to replace the stochastic analysis according to Figure 2a is presented. The ANN is trained to approximate the surrogate objective (e.g. the mean value) to be minimized or maximized. This allows one to avoid time consuming Monte Carlo simulations during the optimization. In case of FE simulations, it may be necessary to work with two levels of surrogates. At the first level, the deterministic FE simulation is replaced by an ANN, which is then used to train another ANN replacing the stochastic simulation at the second level. This means that at each supporting point of the second level ANN, a Monte Carlo simulation is performed with the first level ANN to get the desired responses (e.g. mean values).

Both concepts, i.e. to replace the interval analysis and to replace the stochastic analysis, can be combined within interval stochastic surrogate models. In [18], a similar concept has been introduced to map fuzzy bunch parameters with ANNs for fuzzy stochastic analyses.

4 Examples

4.1 Verification with analytical solution

The proposed neural network surrogate modelling approach for the stochastic simulation is verified by an analytical solution of an optimization problem

$$\min \{ \max \{ \mu(Z(d_1, d_2, D_3)) \} \}, \text{ with } Z(d_1, d_2, D_3) = d_1^2 + d_2^2 + D_3 + \sigma_1^2 + A_2 + 1,$$  \quad (4)

considering three design parameters (deterministic number $d_1$, interval $d_2$ and stochastic number $D_3$, see Table 1) and two additional uncertain a priori parameters (interval $\sigma_1$ and stochastic number $A_2$, see Table 2).

Table 1: Design parameters of the optimization problem according to (4).

<table>
<thead>
<tr>
<th>parameter</th>
<th>type</th>
<th>search space</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
<td>deterministic</td>
<td>$[-5, 5]$</td>
</tr>
<tr>
<td>$d_2$</td>
<td>interval, midpoint $m \cdot d_2$ to be optimized, fixed radius $d_2 = 1$</td>
<td>$[-9, 9]$</td>
</tr>
<tr>
<td>$D_3$</td>
<td>Gaussian, mean value $\mu(D_3)$ to be optimized, fixed standard deviation $\sigma(D_3) = 0.5$</td>
<td>$[-4, 4]$</td>
</tr>
</tbody>
</table>

Table 2: Uncertain a priori parameters of the optimization problem according to (4).

<table>
<thead>
<tr>
<th>parameter</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_1$</td>
<td>interval, midpoint $m \cdot a_1 = 1.2$, radius $a_1 = 1.8$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>Gaussian, mean value $\mu(A_2) = 0$, standard deviation $\sigma(A_2) = 0.75$</td>
</tr>
</tbody>
</table>

The analytical solution of the optimization problem (4) with polymorphic uncertain parameters is $\mu(Z(d_1, d_2, D_3)) = 7$ with the optimal design parameters $d_1 = 0$, $m \cdot d_2 = 0$, $\mu(D_3) = -4$.

In [19], the optimization problem (4) has been solved by a particle swarm optimization algorithm [6] for the interval dominated and the stochastics dominated approaches according to Figure 2. It has been investigated, that the interval dominated approach was about 10 times faster than the stochastics dominated approach and that for both approaches the relative deviation $Z_{rel} = \frac{Z_{num} - Z_{an}}{Z_{num}}$ of the numerically computed optimum $Z_{num}$ with respect to the analytically computed optimum $Z_{an}$ converges to...
zero with an increasing number of samples of the Monte Carlo simulation. But for the interval dominated approach, the optimum $Z_{\text{num}} > Z_{\text{an}}$, whereas for the stochastics dominated approach $Z_{\text{num}} < Z_{\text{an}}$, i.e. the interval dominated approach is more reliable than the stochastics dominated approach, see [19].

In order to verify the presented ANN surrogate modelling approach, a feedforward neural network with one hidden layer (4-5-1 architecture) is trained to approximate the objective function $\mu (Z(d_1, d_2, D_3))$, see Figure 3. This means that the Monte Carlo simulation loop within the optimization is replaced by the ANN. Inputs of the neural network are realizations $d_1, m_2$ and $\mu (D_3)$ of the three design parameters and realizations $a_1$ of the interval a priori parameter $A_1$, see Figure 3. Here, 1296 supporting points (a regular grid $6 \times 6 \times 6 \times 6$ of the four dimensional input space) are used to train, test and verify the ANN with 60%, 20% and 20% of the data, respectively. For each of the 1296 supporting points, a Monte Carlo simulation with varying stochastic design parameter $D_3$ and varying stochastic a priori parameter $A_2$ is performed to obtain the corresponding target output value. In order to investigate the sensitivity of the sample size within the Monte Carlo simulation, four ANNs are created with 1,000, 3,000, 5,000 and 10,000 samples, respectively. The prediction performance of all ANNs is of high quality with a coefficient of determination close to 1.0.

Figure 3: Feedforward neural network for the approximation of the objective function $\mu (Z)$ according to the optimization problem (4).

In Table 3, the relative deviation $Z_{rel}$ of the numerically computed optimum $Z_{\text{num}}$ with respect to the analytically computed optimum $Z_{\text{an}}$ is presented for the four ANNs created with different sample sizes. It can be seen, that the optimization problem can be solved with the ANN surrogate model with similar accuracy compared to the original solution without ANN. The ANN solution could further be improved, if more supporting points are used to generate the ANN.

Table 3: Relative deviation $Z_{rel}$ of the numerically computed optimum $Z_{\text{num}}$ with respect to the analytically computed optimum $Z_{\text{an}}$; comparison of the interval dominated approach without and with ANN surrogate model.

<table>
<thead>
<tr>
<th>number of samples</th>
<th>$Z_{\text{rel}}$ without ANN $\cdot 10^{-3}$</th>
<th>$Z_{\text{rel}}$ with ANN $\cdot 10^{-3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>7.86</td>
<td>3.49</td>
</tr>
<tr>
<td>3,000</td>
<td>3.00</td>
<td>3.03</td>
</tr>
<tr>
<td>5,000</td>
<td>2.69</td>
<td>−2.37</td>
</tr>
<tr>
<td>10,000</td>
<td>0.80</td>
<td>−2.43</td>
</tr>
</tbody>
</table>

4.2 Optimization of the reinforcement layout of a reinforced concrete bridge structure
The proposed ANN surrogate modelling approach is applied to optimize the reinforcement layout of a reinforced concrete bridge structure. The structural system and the cross section of the two-span bridge is shown in Figure 4. To investigate the cracking behaviour under long-term loading, the bridge is subjected to its self weight $g = 67.25$ kN/m and a constant traffic loading $q = 22.2$ kN/m² over the whole bridge deck.

The load bearing capacity and the crack patterns of the structure are analysed by a finite element model. In this model, the reinforcement bars are considered by a smeared formulation taking the bond slip mechanism into account. The concrete material model is based on [20] and for the steel reinforcement, an elasto-plastic model is adopted. Due to the double symmetric system and loading conditions only $\frac{1}{4}$ of the structure has to be computed.

In Figure 5, the distribution of the internal crack variable $\alpha_R$ is shown. In a post-processing, the crack width $w_i$ for each finite element is calculated using the internal crack variable $\alpha_R$, see [21]. This
allows one to evaluate the exposed lateral surface of reinforcement $M$ [mm$^2$] based on the crack widths at the reinforcement layers, see [19], which is defined as a durability measure in this work.

In order to minimize the exposed lateral surface of reinforcement, an optimization task has to be solved. Within this optimization, the Young’s modulus $E_c$ of concrete is considered as a stochastic a priori parameter (Gaussian distribution with mean value $\mu = 33,300$ N/mm$^2$ and standard deviation $\sigma = 400$ N/mm$^2$). The design parameters $\bar{h}_{\text{bottom}}$ and $\bar{h}_{\text{top}}$ (position of the reinforcement rebars with respect to the upper and lower edge of the cross section, respectively) are defined as intervals with fixed radii and midpoints to be optimized. To consider both, stochastic and interval uncertainty within the optimization, the minimization of the worst-case mean value of the exposed lateral surface of reinforcement is defined as objective

$$
\min \{ \max \{ \mu(M(\bar{h}_{\text{bottom}}, \bar{h}_{\text{top}})) \} \} .
$$

As a constraint of the optimization problem, the accepted failure probability with respect to the load bearing capacity of the bridge structure is defined as $P_{f,\text{tol}} \leq 10^{-4}$.

In [19], optimization runs for different reinforcement layouts (number and diameter of rebars) of the described bridge structure have already been performed for deterministic, interval and the described polymorphic uncertain conditions, i.e. stochastic a priori parameter and interval design parameters. The influence of the radius of the interval design parameters has been investigated in [22]. Here, it is focused on the proposed multilevel ANN surrogate modelling approach.

Based on the results of 90 FE simulations with varying realizations of the interval design parameters $\bar{h}_{\text{bottom}}$ and $\bar{h}_{\text{top}}$ and the stochastic a priori parameter $E_c$, two neural networks are trained to predict the exposed lateral surface of the reinforcement $M$ (objective function) and the load bearing capacity $lbc$ (to evaluate the failure probability constraint), respectively. After training and testing, the first neural network with a 3-10-5-1 architecture is applied to compute the mean values $\mu(M)$ of the exposed lateral surface of the reinforcement in the design space, see Figure 6a. The sensitivity of the sample size to the objective function has been investigated. Here, 100,000 samples of the Gaussian distributed Young’s modulus $E_c$ are used to approximate the objective function.

To solve the optimization problem, a particle swarm optimization algorithm is applied. Here, a swarm with six particles is used and the search for the global optimum is repeated three times. In each search
step, for each particle an optimization-based interval analysis is required and a Monte Carlo simulation has to be performed for each run of this additional internal optimization. In order to replace the Monte Carlo simulation within the optimization, another neural network with 2-10-4-1 architecture is trained to approximate the dependency of realizations of the interval design parameters $h_{\text{bottom}}$ and $h_{\text{top}}$ (inputs) to the objective function $\mu(M)$ (output). Based on the first neural network, 6,561 supporting points (regular $81 \times 81$ grid) are used to train (60%), test (20%) and verify (20%) this additional ANN with 100,000 samples of the stochastic a priori parameter $E_c$. The resulting ANN approximation of the objective function in the design space is shown in Figure 6b, which shows a very good agreement with the reference solution to be approximated, see Figure 6a. This ANN is then applied to compute the optimal design $\min \{ \max \{ \mu(M(h_{\text{bottom}},h_{\text{top}})) \} \}$ just by solving an optimization-based interval analysis for each optimization run. This reduces the computation time approximately by a factor of three to four.

The results of both approaches, i.e. the ANN approximation of the deterministic simulation and the additional ANN approximation of the stochastic simulation, are presented in Table 4 and Table 5, respectively. The additional ANN approximation of the stochastic simulation leads to almost the same results compared to the pure ANN approximation of the deterministic simulation. Here, the constraints have only been checked at the end of the optimization, because the evaluation of the failure probability in each iteration step of the optimization is too time consuming. It should be noted that for both approaches the constraint $P_{f,tol} \leq 10^{-4}$ (accepted failure probability with respect to the load bearing capacity of the bridge structure) is slightly exceeded for $r_{h_{\text{bottom}}/h_{\text{top}}} = 10\, \text{mm}$.

Table 4: Results of the optimization with ANN approximation of the deterministic simulation for varying interval radii $r_{h_{\text{bottom}}/h_{\text{top}}}$ of the design parameters; value of the objective function $\max \{ \mu(M) \}$, midpoints of the best designs, computation time for three optimization runs and failure probability $P_f$.

<table>
<thead>
<tr>
<th>$r_{h_{\text{bottom}}/h_{\text{top}}}$ [mm]</th>
<th>$\max { \mu(M) }$ [mm$^2$]</th>
<th>$m_{h_{\text{bottom}}}$ [mm]</th>
<th>$m_{h_{\text{top}}}$ [mm]</th>
<th>time [s]</th>
<th>$P_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7.88E-05</td>
<td>55.0000</td>
<td>142.8696</td>
<td>20.5</td>
<td>2.29E-05</td>
</tr>
<tr>
<td>5</td>
<td>1.59E-04</td>
<td>60.0000</td>
<td>142.9992</td>
<td>4273.5</td>
<td>7.10E-05</td>
</tr>
<tr>
<td>10</td>
<td>4.86E-04</td>
<td>65.0000</td>
<td>140.2368</td>
<td>3708</td>
<td>1.39E-04</td>
</tr>
</tbody>
</table>

Table 5: Results of the optimization with additional ANN approximation of the stochastic simulation for varying interval radii $r_{h_{\text{bottom}}/h_{\text{top}}}$ of the design parameters; value of the objective function $\max \{ \mu(M) \}$, midpoints of the best designs, computation time for three optimization runs and failure probability $P_f$.

<table>
<thead>
<tr>
<th>$r_{h_{\text{bottom}}/h_{\text{top}}}$ [mm]</th>
<th>$\max { \mu(M) }$ [mm$^2$]</th>
<th>$m_{h_{\text{bottom}}}$ [mm]</th>
<th>$m_{h_{\text{top}}}$ [mm]</th>
<th>time [s]</th>
<th>$P_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7.68E-05</td>
<td>55.0000</td>
<td>138.1468</td>
<td>7.5</td>
<td>1.27E-05</td>
</tr>
<tr>
<td>5</td>
<td>1.47E-04</td>
<td>60.0000</td>
<td>142.6607</td>
<td>1076.5</td>
<td>7.27E-05</td>
</tr>
<tr>
<td>10</td>
<td>4.82E-04</td>
<td>65.0000</td>
<td>140.4340</td>
<td>1175.5</td>
<td>1.39E-04</td>
</tr>
</tbody>
</table>
5 Conclusions

In this paper, a neural network based surrogate modelling strategy has been presented to solve optimization problems in structural mechanics considering stochastic and interval parameters. A first artificial neural network is trained to approximate the deterministic finite element analysis. Based on this neural network, a second neural network is created to replace the stochastic analysis (Monte Carlo simulation) within the optimization runs. This surrogate modelling strategy has been verified by an analytical solution, and it has been applied to optimize the cracking behaviour of a reinforced concrete bridge structure. The results show good approximation and prediction capabilities and a reduction of the computation time by a factor of three to four.

In future works, the approximation quality of the surrogate models can further be improved by generating additional training and test points close to the optimal designs and close to the limit state functions adaptively during the optimization. It would also be beneficial, if the time consuming constraints check can be incorporated into the surrogate model for the stochastic analysis. Moreover, the whole interval stochastic simulation could be replaced by another surrogate model, where only the design parameters are inputs and the influence of all a priori uncertain parameters are captured by the neural network parameters.

Acknowledgments

This research is supported by the German Research Foundation within Subproject 6 of the Priority Programme SPP 1886 'Polymorphic uncertainty modelling for the numerical design of structures', which is gratefully acknowledged by the authors.

References

Interval field for spatially and temporally dependent uncertainty–machine learning approach

David Betancourt1*,2, Rafi L. Muhanna2, and Robert L. Mullen3

1School of Computational Science and Engineering, Georgia Institute of Technology, U.S.A.
2School of Civil and Environmental Engineering, Georgia Institute of Technology, U.S.A.
3School of Civil and Environmental Engineering, University of South Carolina, U.S.A.
*Corresponding author: david.betancourt[at]gatech.edu

Abstract

A major shortcoming to interval uncertainty approaches in computational mechanics is the lack of an interval field to represent uncertainty. This paper introduces the supervised interval field (SIF), a model to quantify spatially and temporally dependent uncertainty using machine learning. We introduce deep learning model architectures that are used to develop the SIF for domains of any dimension using deep recurrent neural networks. We demonstrate how to unify the SIF with the Interval Finite Element Method (IFEM) and show an experiment using soil layer data.

Keywords: Interval field, Spatially dependent uncertainty, Machine learning, Finite elements, Computational mechanics, Materials

1 Introduction

The analysis and design of complex engineering systems is exposed to a great number of uncertain parameters in the system inputs. Insufficient modeling of these parameters can lead to inaccurate and/or disastrous expectations of the system’s performance [1–3]. In the context of computational mechanics, random fields [4, 5] have been widely used within a probabilistic framework in order to model the spatially and temporally dependent uncertainty in the system. Despite its widespread adoption, random field theory is inadequate for cases where there is significant epistemic uncertainty and/or not enough data to select a distribution [6–8]. Under such cases, non-probabilistic approaches such as interval models have been employed to model both aleatory and epistemic uncertainty.

One of the main drawbacks within the non-probabilistic interval framework is, unlike random fields, not being able to model spatially- or temporally-varying uncertainty in the system. Given this limitation, the spatially varying uncertainty is modeled by assigning the same interval variable over the entire domain. Therefore, the Interval Finite Element Method (IFEM) solution can be overestimated because it does not account for parameter dependency in the field [9]. As a result, interval field models have been recently proposed [10–12] in order to model spatially-varying dependency within the interval model framework. Nonetheless, the interval field methods proposed are limited in their applicability and scalability.

Consequently, this paper introduces the supervised interval field, a data-centric model to quantify spatially and temporally dependent uncertainty using machine learning integrated with the IFEM. In computational mechanics, the prime applications are for spatially-varying material uncertain properties and uncertain time-varying loading. In order to account for the uncertain properties, machine learning models are presented that are able to represent the spatial and temporal dependencies in the data. Once the predictions of the machine learning model are obtained, the supervised interval field is formed, discretized, and integrated in one framework with the IFEM. The proposed method is capable of extending to domains of any Euclidean dimension and it’s independent of the finite element mesh discretization.

Numerical experiments are conducted using soil layer data in order to model spatially-varying material uncertainty with the proposed interval field. We demonstrate that our method more accurately predicts the actual measured data (ground truth) of the soil material properties. In addition, the presented method can be used in multiple engineering applications, scales, and data types.

Our contribution is as follows: (1) our method reduces the field parameter dependency conservatism by using the proposed interval field (SIF); (2) extension of our method to any domain dimensionality; (3) introduction of machine learning methods that work well with computational mechanics applications; (4) provide an interval enclosure for machine learning model predictions; and (5) bridge the gap between engineering and machine learning.
2 Related Work

The first to incorporate the concept of spatially varying uncertainty within IFEM was Moens et al [13–15] through different methods. In [14] an interval field concept was developed which allowed to use dependency between elements, which is otherwise not possible using interval parameters. To that end, two methods were introduced: The Local Interval Field Decomposition (LIFD), and the Inverse Distance Weighting Interpolation (IDW). For LIFD, the dependency parameters are determined from a gradient method developed by Imholz et al, which defines the dependency between parameters as the maximum gradient that can occur between them. In this definition, perfect dependency corresponds to a zero gradient and perfect independency corresponds to the maximum gradient between parameters. The dependency parameters are discretely defined at each element of the mesh—thus the dimensionality of the interval field equals the numbers of elements in the mesh. The discrete values for the dependency parameters are determined by a smooth second-order polynomial function presented in [14]. For the IDW method, the base functions needed for the interval field are determined and spatially selected based on prior engineering knowledge.

The pitfalls for the Moens et methods are as follows: The LIFD is computationally expensive as the dimensionality of the uncertainty equals the number of elements in the mesh. Additionally, the theoretical arguments for the determination of the dependence are not fully justified. The IDW is computationally cheaper, and could produce good results where enough a-priori domain engineering knowledge is available. As a result, expert hand-engineering would be required to produce good results thus rendering it non-scalable. Also, both methods were only presented for 1D cases.

Wu et al [12] developed a hybrid method for static FE analysis (X-UISS method) that combines random fields and intervals fields. It has an algorithm that easily extends to 2D and 3D domains. For the interval field implementation, the X-UISS method defines the upper-bound $UB$ and lower-bound $LB$ functions from the extrema of a set of measurements at a spatial coordinate $x$. Then, the $UB$ and $LB$ of measured points are linearly interpolated between the $x$ points where data is unavailable. Then, the conventional IFEM [16] procedure is implemented using the linearly interpolated values for the domain. The pitfall of this method is that it is computationally expensive by having to solve both Stochastic Finite Element Method (SFEM) and IFEM methods but more importantly because it uses linear interpolation between data points, which is an oversimplification of the spatial uncertainty variability in a domain.

More recently, Sofi et al [10, 11] developed a 1D interval field approach for IFEM which uses the extra unitary interval (EUI) in order to reduce the overestimation of the IFEM due to mutual dependency. In contrast to the Hybrid method by Wu et al. and the LIFD method by Moens et al., the dimension of uncertainty does not depend of the FE mesh size of the model. However, their underlying interval field method is based on a closed-form solution of a determinate beam or an approximate solution for statically indeterminate beams. A general EUI method for 2D and 3D domains under general boundary conditions has not been developed to the authors’ knowledge.

In the area of machine learning, work that seeks to include data dependencies in the prediction includes speech and image recognition, among others that make predictions on sequential data. We borrow some of these ideas for the present work and present our own contributions in order to quantify the uncertainty. To our knowledge, our work presents the first application of machine learning to model interval fields for spatial and temporal dependency.

3 Background

A brief introduction to the different theories and techniques used in this paper follows.

**Supervised Learning (SL)** Supervised learning [17, 18] seeks to learn a predictive model $f : X \rightarrow Y$ in some n-dimensional Euclidian space $\mathbb{R}^n$, given a set of training examples. In order to achieve this goal, the SL model is trained using the X inputs (features) with known Y outputs (target values). This composes the training set $T = (X, Y)$ where $X \in \mathbb{R}^n$ is the feature space and $Y \in \mathbb{R}^d$ is the output space. More specifically, the set of feature vectors $X$ is used as input to the SL algorithm, which produces at each instance an output $f(x_i)$. During training the algorithm reduces the difference between the true target $y_i$ and its prediction $\hat{f}(x_i)$ by minimizing a loss function $L(\Theta)$, such as squared loss, for parameters $\Theta$. Following the training phase, the testing phase uses feature instances not present during training and it predicts $\hat{f}(x_i)$ using the learned model without access to its targets $y_i$. After successfully testing the algorithm, the ultimate goal of SL is to generalize to predictions of unseen data where target values are not available.
Deep Recurrent Neural Networks  Recurrent neural networks (RNN) are extensions to neural networks used for sequential data [19]. For a sequence of inputs \((x_1, \cdots, x_N)\), a recurrent neural network (RNN) computes a sequence of outputs \((y_1, \cdots, y_N)\) using a recurrence equation at every time step

\[ h_t = f_W(h_{t-1}, x_t) \]  

In Eq. 1, \( h_t \in \mathbb{R}^n \) is the new hidden state of the RNN, \( f_W \) is a function with weight parameter matrix \( W \), \( h_{t-1} \) is the previous hidden state, and \( x_t \) is the input vector at time \( t \). RNNs can be represented as a deep multilayer model and are generally optimized using gradient-based methods. Vanilla RNNs have the pitfall of exploding gradients due to long-term dependencies in the data when processing long sequences [19]. Long short-term memory [20] (LSTM) units are used to build the deep RNN network (or deep LSTM network) to bypass this problem.

Long Short-Term Memory Units  Long short-term memory (LSTM) [20] units are designed to process and predict sequential data while addressing the numerical shortcomings of the exploding gradients in vanilla RNNs [21]. In order to do so, LSTMs maintain a memory cell state \( c_t \) along with the hidden state \( h_t \) at every time \( t \). At each time \( t \), the LSTM can choose to remember or forget information by using the following gating mechanisms, and shown on Fig 1. When combined in a deep architecture, deep LSTM networks are a formidable machine learning predictive model [19, 21].

Random Fields  Random fields [4] are generalizations of random processes where the random variable \( x \in \mathbb{R} \) and \( \theta \) is the outcome of the random phenomenon \( H(x, \theta) \). Random fields are used to represent the spatial variation of a system property by a random variable defined over the region on which the variation occurs.

Stochastic Finite Elements  In the context of computational mechanics, random field theory has been adopted with the Finite Element Method (FEM) to form the Stochastic FEM (SFEM) [5] in order to find the response of a stochastic system. In particular for most SFEM problems, the Random Field is Gaussian and discretized through a Kauhunen-Loève expansion, while the response is found with a polynomial chaos expansion. Together this procedure is known as the spectral approach to SFEM [5].

Interval Finite Elements  The Interval Finite Element Method (IFEM) is a numerical analysis method that describes the uncertainty by using interval valued parameters. IFEM is used in situations where a probabilistic characterization of the system is not possible. The IFEM finds guaranteed upper and lower bound enclosures of the response [16].

4 Methods

4.1 Supervised Interval Field

We now present the first data-centric approach to generate interval fields using supervised machine learning (SL) models. Up to this point, spatially- and temporally-varying uncertainty in computational mechanics is modeled using the probabilistic framework of the SFEM and, to a lesser extent, with non-probabilistic frameworks using IFEM. Both approaches require direct prior knowledge about the uncertain properties as a global value for the domain in question. To do so, SFEM uses a random field to distribute the variability of the uncertain property (with prior mean \( \mu \) and standard deviation \( \sigma \)), while IFEM uses interval bounds obtained from the measurement devices the obtained the data. The recently proposed interval field techniques [10–12, 14] also require some direct information about the uncertain properties for the domain and the spatial variability is determined in a number of ways, some of which are probabilistic in nature.

On the other hand, once trained, the SIF only requires data in the form of features in the subspace of the domain where we must find the uncertain properties (i.e., the target values)—given that the model has been trained on enough data. In this fashion, we are able to predict the uncertain properties without making assumptions on its probability distribution, such as would be the case with a Random field.

In order to develop the SIF, the trained models presented in Sections 4.1.1 are used to obtain a prediction at each spatial coordinate \( x \) in the domain. Interval uncertainty bounds can be obtained with expert knowledge or directly from the measuring devices [22]. In this work, interval bounds are set at 10% for the experiments.
4.1.1 Deep LSTM Network

Deep LSTMs are the driving force behind recent advancements in time-dependent applications such as speech recognition [23], time series prediction [24], and image captioning [25]. Using deep LSTMs, a model can learn to extract complex abstractions as sequential data representations. Additionally, using deep LSTMs can allow for direct inference on images, which can be very helpful for structural health monitoring and other applications where image data are routinely obtained.

In order to obtain the hidden state vector representation at time $t$, each of the gates in the LSTM units performs the following computations:

\[
\begin{align*}
    i_t &= \text{sig}(W_{hi}h_{t-1} + W_{ix}x_t) \\
    f_t &= \text{sig}(W_{hf}h_{t-1} + W_{fx}x_t) \\
    c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{hc}h_{t-1} + W_{cx}x_t) \\
    o_t &= \text{sig}(W_{ho}h_{t-1} + W_{ox}x_t + W_{co}c_t) \\
    h_t &= o_t \odot \tanh(c_t)
\end{align*}
\]  

(2)

where the $i_t$, $f_t$, $o_t$ are the input, forget, and output gates, respectively. The $\text{sig}(\cdot)$ and $\tanh(\cdot)$ are the logistic sigmoid activation function and hyperbolic tangent function, respectively. The weight matrices are represented by the recurrent weight matrices $W_h$ and the input-to-hidden weight matrices $W_x$. Fig. 1 shows a diagram of an LSTM unit.

To train the model, we use the Adam optimizer [26] and add dropout regularization [27, 28] to prevent overfitting.

4.1.2 XGBoost

This section presents Extreme Gradient Boosting (XGBoost), a state-of-the-art ensemble algorithm that is broadly easier to implement than deep LSTMs while being a robust predictor. Nonetheless, unlike deep LSTMs, XGBoost as well as many SL algorithms do not explicitly handle spatial or temporal dependencies. Hence, we present it to use as comparison with the deep LSTM model and because it can be valuable for some engineering SIF applications.

Ensemble algorithms work by combining different models in order to reduce the overall variance and bias of the prediction. The most popular ensemble methods are bagging [29] and boosting [30]. In Bagging, different models or different algorithms are combined in a way that reduces the variance of the prediction, while in Boosting, different weak models are combined to reduce the overall bias and variance of the prediction. For regression problems, generally an average of the learners is taken as the final prediction for each instance.

XGBoost is a tree ensemble algorithm that uses CART regression trees [31] as its base learner. XGBoost [32] was introduced in 2014 and has become a dominant algorithm for structured data that has won numerous data science competitions such as the HiggsML challenge [33], and ACM RecSys Challenge [34]. XGBoost is an updated variant of Gradient Tree Boosting [35] and provides higher speed, more regularization options, and greater scalability.

In order to train the model, the objective function to be minimized is a regularized loss composed of the training loss $L(\Theta)$ and a regularization term $\Omega(\Theta)$ to prevent overfitting. XGBoost optimizes the objective function additively. For the data instance $i$ in iteration $t$, the prediction $\hat{f}^{(t)}(x_i)$ of the model is the recursive sum of the individual predictions of the $k$ trees in the ensemble and function of parameters $\{\Theta\} = f^{(t)}$, as follows:

\[
\hat{f}^{(t)}(x_i) = \sum_{k=1}^{t} \hat{f}_k(x_i) = \hat{f}^{(t-1)}(x_i) + f^{(t)}(x_i)
\]  

(3)

The objective function to be minimized at every iteration $t$ is

\[
L^{(t)} = \sum_{i=1}^{N} L(y_i, f^{(t)}(x_i)) + \Omega(f^{(t)})
\]  

(4)
This optimization problem is solved using first and second-order gradient statistics as shown on [32]. In our experiments we show that the Deep LSTM network is superior to XGBoost. Nonetheless, engineers might find that XGBoost performs well for their particular application and gain time savings from its implementation and training time.

4.1.3 Model Training and Architectures

Algorithm 1 is used for the SIF.

We show results for XGBoost and deep LSTM network on the experiment in Section 5. Figure 2 shows the different learning models $f$ used for the SIF.

**Algorithm 1 SIF Algorithm**

1. Obtain training set $T$
2. Obtain testing set $D$
3. Choose learning model $f$
4. Choose loss function $L$
5. Obtain interval uncertainty bounds $U$

\[\text{for } i = 1 : M \text{ instances do} \]
6. Train $f$ to minimize a training loss $L(\Theta)$
7. Obtain prediction $\hat{f}(x_i)$

\[\text{end for}\]

Figure 2: Model Architectures
Left: Simple XGBoost model
Right: Deep LSTM model stacked with feed-forward layer

4.1.4 A Unified Framework: SIF Discretization for IFEM

The SIF predictions are independent of the IFEM mesh. Thus a post-processing discretization of the SIF to match the elements in the mesh of the IFEM is required. The procedure is as follows: the spatially- and/or temporally-varying uncertain properties of the domain obtained with the SIF are midpoint input values to the IFEM in order to obtain the interval responses. In particular, the discretization of the SIF for the purpose of IFEM analysis is performed by taking the expected value of the predictions of the SIF corresponding to the size $l$ of each finite element in the mesh, which gives the interval midpoint value $I_{\text{mid}}^n$ at element $n$, as follows:

\[I_{\text{mid}}^n = \mathbb{E}[\hat{f}(x_i)], \quad \text{where } \hat{f}(x_i) \text{ is the model prediction for sample } i.\]  

(5)

After the discretization, a conventional IFEM is performed with the postulated uncertainty bounds.

5 Experiments

The experiments performed consist of using the SIF method to predict the material properties of soil-column layers at various locations in a 3D domain to produce the interval field, which later is discretized
as material input to the IFEM. The soil layer properties were obtained at the University of Massachusetts-Amherst by Prof. Dr. Paul W. Mayne of the Georgia Institute of Technology. They were obtained via cone penetration tests (CPTs) \cite{36}, which are performed with two separate measurement devices inserted from the surface into each soil-column obtaining soil properties with respect to depth for approximately 15 m. In particular, Test 1 measures three different soil properties obtained at about every 0.05 m: the cone resistance \(q_c\), sleeve friction \(f_s\), and porewater pressure \(u_2\) \cite{36, 37}. In parallel, Test 2 measures the shear wave velocity \(V_s\) in the soil column at every 1 m \cite{36, 38}.

For the learning models, we use a subset of the data and denominate it as the training set \(T\), with the feature space \(X\) consisting of the measurements of Test 1 \(\{q_c, f_s, u_2\}\), and the target space \(Y\) consisting of the measurements of Test 2 \(\{V_s\}\). Once the model is trained, we predict the target values for shear wave velocity \(V_s\) for a subset of unseen data (testing set \(D\)) by querying the trained model with only the features \(X\) of subset \(D\). Thus, by using the SIF method, we can produce the interval field based on feature data alone without making any assumptions about the probability distribution of the target values. A hypothetical scenario for this application would be one where a subset of the data in Test 2 are corrupted or missing, or that only performing the Test 1 on a portion of the domain results in significant cost and time savings. Further, for other applications, such as aerospace batch manufacturing and structural health monitoring, the SIF approach is superior to the random fields assumptions of the SFEM, since SIF does not require assumptions on the probability distributions of the unknown properties.

### 5.1 Datasets

The dataset consist of text files corresponding to the CPT results for 15 different ‘soil-column’ locations in the same site. Fig. 3 shows examples for features and target values at two distinct locations.

[Figure 3: Dataset Samples](#)

- Top and bottom: examples of data features (left) and corresponding targets (right).

### 5.2 Model Evaluation

Randomized Search \cite{39} was used for the hyperparameter optimization of the models. In order to account for the effect of the uncertainty bounds, the inputs \(X\) for the testing set \(D\) were queried under four different cases using the trained deep LSTMs, as follows: (1) Raw inputs \(X\) without any postulated uncertainty added (i.e. midpoint input); (2) Lower bound inputs \(X\) with -5% uncertainty bounds added;
(3) Upper bound inputs $X$ with +5% uncertainty bounds added; (4) Raw inputs $X$ with noise multiplier $\epsilon \sim N(0,0.05)$. Table 1 shows that the deep LSTM outperforms XGBoost in these experiments. Fig 4 shows the predictions for the testing column using the different LSTM model inputs. Given that the uncertainty bounds are small (10% total) the predictions are similar. In the SIF-IFEM example show in Sec. 5.3, the raw input $X$ is used to obtained the material predictions.

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2$ Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM (X)</td>
<td>0.915</td>
</tr>
<tr>
<td>LSTM (\overline{X})</td>
<td>0.909</td>
</tr>
<tr>
<td>LSTM (X)</td>
<td>0.875</td>
</tr>
<tr>
<td>LSTM (X+noise)</td>
<td>0.896</td>
</tr>
<tr>
<td>XGBoost*</td>
<td>0.228</td>
</tr>
</tbody>
</table>

*XGBoost is only tested with raw features.

$R^2$ is the coefficient of determination for the model prediction.

5.3 IFEM Example using SIF

The trained model for the SIF was used to predict the shear wave velocities in the testing set $\mathcal{D}$ using raw inputs $X$. A comparison is done with the SFEM, where the parametrization consists of the mean $\mu$ and standard deviation $\sigma$ of a Gaussian prior distribution. To obtain the parameters of the prior for the SFEM, we use the shear wave velocities stored in the training set $\mathcal{T}$ and compute its mean and standard deviation ($\mu = 117 \text{ MPa}; \sigma = 46 \text{ MPa}$). In order to proceed to the FE analysis, the shear wave velocities $\{V_{s,i}\}$ are converted to moduli of elasticity $E_i$ for every sample $i$, as follows [36]:

$$E = 2G_{max}(1 + \nu)$$

$$G_{max} = V_s^2 \rho_t$$

where $\rho_t = \gamma_t/g$

$$\gamma_t = 26 - \frac{14}{1 + (0.5 \log(f_s + 1))^2}$$

$$f_s = \text{CPT measurement (kPa)}.$$  

$$\nu = 0.15$$  

Then, without loss of generality, a cantilever structure meshed with 10 beam equal-length elements is employed to show the difference between the SIF-IFEM and the SFEM, as shown on Figure 2. The length of the beam is $L = 10 \text{ m}$, the length of each element $n$ is $l = 1 \text{ m}$, the distributed load is $q = 1 \text{ kN}$, and the mid-point interval material elasticity for each element $E_n$ is determined using Eq. 5, and shown on Table 2 (obtained from raw input predictions). 10% uncertainties are assigned to all the elements $E_n$. 

Figure 4: Predictions on the testing set for the different LSTM models used.
The comparison shows that the SFEM has a 22.7% deviation from the deterministic displacement solution at the tip of the cantilever, while the SIF-IFEM closely encloses the solution with a postulated 10% level of uncertainty. Moreover, using the SIF discretization procedure for IFEM (SIF-IFEM) of Sec. 4.1.4, smoothens the mesh element material properties and brings it to close agreement with ground truth values for the moduli of elasticity $E_n$, as shown on Fig. 4. The large deviation of the SFEM from deterministic displacement values based on ground truth $E_n$ could even be larger if the soil layers had a more drastic stiffness change (e.g., from soft soil to hard rock).

Table 2: Material Properties used for IFEM at each beam element

<table>
<thead>
<tr>
<th>Element Number</th>
<th>$E$ (Ground Truth) (MPa)</th>
<th>$E$ (Prediction) (MPa)</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>139.4</td>
<td>142.7</td>
<td>2.4</td>
</tr>
<tr>
<td>2</td>
<td>134.2</td>
<td>132.4</td>
<td>1.3</td>
</tr>
<tr>
<td>3</td>
<td>101.3</td>
<td>97.8</td>
<td>3.5</td>
</tr>
<tr>
<td>4</td>
<td>74.1</td>
<td>72.9</td>
<td>1.6</td>
</tr>
<tr>
<td>5</td>
<td>78.1</td>
<td>71.4</td>
<td>8.6</td>
</tr>
<tr>
<td>6</td>
<td>76.3</td>
<td>72.5</td>
<td>5.0</td>
</tr>
<tr>
<td>7</td>
<td>68.3</td>
<td>73.9</td>
<td>8.2</td>
</tr>
<tr>
<td>8</td>
<td>75.5</td>
<td>73.9</td>
<td>2.1</td>
</tr>
<tr>
<td>9</td>
<td>77.5</td>
<td>72.7</td>
<td>6.2</td>
</tr>
<tr>
<td>10</td>
<td>73.6</td>
<td>70.1</td>
<td>4.8</td>
</tr>
</tbody>
</table>

Figure 5: Deflection Comparison of SIF-IFEM and SSEM with Cantilever Beam

IFEM with 10% uncertainties
For SFEM: $\mu = 117MPa$ and $\sigma = 46MPa$

6 Conclusions and Discussion

Engineering systems must contend with great deal of uncertainty. Whenever the spatial- or temporally-varying uncertainty is epistemic and cannot be characterized by a random field, the non-probabilistic interval analysis framework presents an alternative. However, the corresponding concept of an interval field was not well-developed yet in one unified framework that could scale to any domain dimension. The introduction of the SIF and the unified framework of the SIF-IFEM, both presented in this paper, offer a robust method for an interval field. The power of the SIF lies in not having to make a-priori assumptions on the probability distribution of the uncertainties, while the IFEM is computationally more efficient in the finite element solution than the SFEM. Moreover, the SIF-IFEM framework allows to bound the machine learning model predictions provided by the SIF by any given level of uncertainty bounds.

REC 2018 – D. Betancourt, R.L. Muhanna, R.L. Mullen
Acknowledgments
We would like to thank Dr. Paul W. Mayne for graciously providing the soil layer data.

References


Uncertainty quantification for random fields estimated from effective moduli of elasticity

Jack Pierce-Brown,1* Luis C. Neves,1 and Donald L. Brown2

1 Faculty of Engineering, University of Nottingham, United Kingdom
2 School of Mathematical Sciences, University of Nottingham, United Kingdom
*Corresponding author: jack.piercebrown[at]nottingham.ac.uk

Abstract

The stochastic finite element method is a useful tool to calculate the response of systems subject to uncertain parameters and has been applied extensively to analyse structures composed of randomly heterogeneous materials. The methodology to estimate the parameters of the random field underlying a stochastic finite element model often utilises the midpoint approximation wherein material properties that are measured over a sample volume are treated as point observations of the random field at the centroid of the sample volume. This paper investigates the error induced by this approximation for the case of effective moduli of elasticity resulting from tensile loading as well as 3 and 4-point bending. A computer experiment has been performed consisting of the generation of synthetic stiffness profiles from a lognormal stochastic process, the calculation of effective properties as weighted harmonic averages and the estimation of random field parameters through the method of moments. The uncertainty in the parameter estimates is quantified and a recommendation is made as to which bending test is superior for obtaining random field parameter estimates with reference to the statistics of the base process and the tensile loading condition.

Keywords: Uncertainty Quantification, Effective Elastic Modulus, Midpoint Approximation, Random Field Theory, Stochastic Finite Element Method

1 Introduction

The stochastic finite element method is a useful tool to calculate the response of systems subject to uncertain parameters [1] and has been applied extensively to analyse structures composed of randomly heterogeneous materials such as timber [2], concrete [3] and soil [4]. The mathematical model for spatial variation of material properties that underlies stochastic finite element modelling is random field theory [5]. This paper considers the variation of modulus of elasticity $E$ with spatial parameter $v$ modelled by a random field $E(v)$ parametrised by a location parameter $\mu$, scale parameter $\sigma$ and a correlation parameter $\theta$ known as the correlation length [6]. The accuracy of the response distribution obtained by the stochastic finite element method depends on the uncertainty associated with the estimation of the parameters of the base random field. Conventional statistical methodologies to estimate the parameters of the random field model require point observations of the base random field $E(v)$ however experimentally determined material properties are in fact effective material properties $E^\ast$ constituting an average over the sample volume. The effective property $E^\ast$ of a heterogeneous material is the property of the homogeneous equivalent that produces identical behaviour with respect to some measurable quantity such as displacement in the case effective modulus of elasticity, flow in the case of effective hydraulic permeability and voltage in the case of effective conductivity [7].

The problem of the incompatibility between the inferential requirement for point observations of the underlying random field $E(v)$ and the reality that material properties can only be measured over a non-zero volume is often resolved through the use of the midpoint approximation [8]. This approximation treats the effective property $E^\ast$ that is obtained over the sample volume as a point observation of the base process $E(v)$ at the centroid $v_c$ of the sample volume. Figure 1 illustrates the error $\epsilon = E(v_c) - E^\ast$ that results from the midpoint approximation and it is argued that this approximation is responsible for significant uncertainty in the estimated parameters of the random field $E(v)$. This paper presents the results of a computer experiment that has been conducted to quantify parameter uncertainty by comparing the distribution of the parameter estimates obtained from the effective properties under the midpoint approximation to the distribution of statistics of the random field. The total uncertainty is composed of two sources [9] namely aleatoric uncertainty which is the statistical uncertainty due to limited sample size and epistemic uncertainty which is the systematic uncertainty due to estimator bias primarily resulting from the utilisation of the midpoint approximation.
Uncertainty quantification for random fields estimated from effective moduli of elasticity

The motivation for selecting modulus of elasticity $E$ as the material property considered in this paper is due to the importance of this material property to the grading of timber which seeks to assign timber elements into strength classes for structural use. However the methodology and results contained within this paper are also applicable to other effective properties which are averages over the underlying random field. In particular this paper considers the effective elastic modulus $E^*$ of a simply supported beam obtained by three different tests namely 3-point $P_3$ and 4-point $P_4$ bending resulting in midspan deflection $U_B$ as well as tension loading $P_T$ resulting in total displacement $U_T$ as shown in Figure 2.

![Figure 2: 1D model of a simply supported beam of length $L$ under 3-point loading $P_3$, 4-point loading $P_4$ resulting in midspan deflection $U_B$ as well as tensile loading $P_T$ resulting in total deflection $U_T$.](image)

This paper presents the formulation of the effective modulus of elasticity $E^*$ of a Euler-Bernoulli beam as a weighted harmonic mean. This formulation offers a compelling alternative to the finite element method for obtaining the displacement of a heterogeneous beam at a single spatial location. The derivation for the case of the 4-point effective property $E^*_4$ is shown in the Appendix otherwise for the case of the 3-point effective property $E^*_3$ see Bechtel [10] or the case of tension $E^*_T$ see Fenton and Griffiths [11]. Under Euler-Bernoulli beam theory [12] the effective modulus of elasticity for all loading conditions can be formulated as:

$$E^* = \left( \int_0^L \omega(v) \cdot E(v)^{-1} \, dv \right)^{-1}$$  \hspace{1cm} (1)

where $(.)^{-1}$ denotes the pointwise reciprocal such that $E(v)^{-1} = \{v \in D | \frac{1}{E(v)}\}$ where $D = [0, L]$ is the domain of the integral. The weighting function $\omega(v)$ satisfies the following property:

$$\int_0^L \omega(v) \, dv = 1$$  \hspace{1cm} (2)

where the units of the weighting function $\omega(v)$ are given by the reciprocal of the units of the domain $D$ so that Equation 2 is dimensionless. The weighting function is dependent on the loading and boundary conditions of the beam and consequently a separate weighting functions for the case of tension $\omega_T(v)$, 3-point $\omega_3(v)$ and 4-point bending $\omega_4(v)$ are presented. It should be noted that this paper does not suggest an equivalence between the profiles of tensile and bending modulus of elasticity but rather the intention of this work is to assess the influence of the midpoint approximation for different loading scenarios. In the case of tension every point in the domain has a equal contribution to the tensile effective modulus of elasticity $E^*_T$ since the stress distribution is spatially constant and consequently $E^*_T$.
is given by an unweighted harmonic mean [11] and consequently every point in the domain $D$ has an equal weighting. The weighting function $\omega_T(v)$ of the tensile effective elastic modulus $E_T^*$ determined from total displacement $U_T$ is given by:

$$
\omega_T(v) = \frac{1}{L} \quad \text{if } 0 \leq v \leq L 
$$

(3)

The weighting function $\omega_3(v)$ for the 3-point effective elastic modulus $E_3^*$ determined from midspan deflection $U_B$ is given by [10]:

$$
\omega_3(v) = \begin{cases} 
\frac{12v^2}{L^3} & \text{if } 0 \leq v \leq \frac{L}{2} \\
\frac{12(L - v)^2}{L^3} & \text{if } \frac{L}{2} \leq v \leq L 
\end{cases} 
$$

(4)

The weighting function $\omega_4(v)$ for the 4-point effective elastic modulus $E_4^*$ determined from midspan deflection $U_B$ is given by (derivation in Appendix):

$$
\omega_4(v) = \begin{cases} 
\frac{324v^2}{23L^3} & \text{if } 0 \leq v \leq \frac{L}{3} \\
\frac{108v}{23L^2} & \text{if } \frac{L}{3} \leq v \leq \frac{L}{2} \\
\frac{108(L - v)}{23L^2} & \text{if } \frac{L}{2} \leq v \leq \frac{2L}{3} \\
\frac{324(L - v)^2}{23L^3} & \text{if } \frac{2L}{3} \leq v \leq L 
\end{cases} 
$$

(5)

A comparison of the three weighting functions is shown in Figure 3 for a beam of length $L = 1$ m. It is observed that the 3-point weighting function $\omega_3(v)$ assigns the highest weighting to the centroid of the sample volume $v_c = 0.5$ m followed by the four-point weighting function $\omega_4(v)$ whilst the tensile weighting function $\omega_T(v)$ assigns the lowest centroidal weighting. The 4-point weighting function $\omega_4(v)$ is more uniform over the domain than the 3-point weighting function $\omega_3(v)$ and this is attributed to the constant central bending moment induced by four-point bending (see Equation 27). The derivative of the tension weighting function $\omega'_T(v)$ is continuous over the domain of integration whilst the derivatives of the bending weighting functions, $\omega'_3(v)$ and $\omega'_4(v)$, are discontinuous at the points where deflection is observed and load is applied. It is observed that as the weighting function $\omega(v)$ approaches the Dirac-delta function $\delta(v - v_c)$ [13] then the effective property $E^*$ becomes a point sample of the random field at the centroid $v_c$ and consequently it is expected that the parameter estimates obtained from $E_3^*$ will be the most accurate whilst the parameter estimates obtained from $E_T^*$ will be the least accurate.

![Figure 3: Weighting functions for tensile $\omega_T(v)$, 3-point $\omega_T(v)$ and 4-point $\omega_4(v)$ loading conditions ($L = 1$ m).](image-url)
2 Computer experiment
A number of researchers have conducted physical experiments into the variation of elastic modulus of timber for the case of tensile loading [14] as well as 3-point [15] and 4-point [16] bending. A computer experiment has been designed to replicate these experiments and quantify the uncertainty associated with estimating the parameters of the random field model directly from the measured effective properties. The experimental procedure involves making \( N = 7 \) observations of the effective elastic modulus \( E^* \) over an effective length of \( L \approx 0.43 \text{ m} \) on a beam of total length \( L_0 = 3 \text{ m} \) for a total of \( M = 20 \) beams as shown in Figure 4. The effective elastic modulus \( E^* \) is no longer a constant but rather a piecewise continuous function notated by \( E^*(v) \). The Monte Carlo method is used to replicate the experimental procedure a total of \( R = 1 \times 10^6 \) times in order to analyse the uncertainty associated with the parameter estimates. In particular the aleatoric and epistemic uncertainty is quantified for the estimate \( \hat{\mu} \) of the location parameter \( \mu \), the estimate \( \hat{\sigma} \) of the scale parameter \( \sigma \) and the estimate \( \hat{\theta} \) of the correlation parameter \( \theta \).

The experimental procedure requires the generation of synthetic stiffness profiles designed which model spatial variation in stiffness along the beam. Since negative values for elastic modulus are not physically permissible the marginal distribution of \( E(v) \) was selected as lognormal since this distribution is strictly non-negative. Thus the stiffness profile is a lognormal stochastic process \( E(v) \) with location parameter \( \mu = 10 \text{ GPa} \), scale parameter \( \sigma = 3 \text{ GPa} \) and the exponential correlation function [17]:

\[
\rho(\tau) = \exp \left( -\frac{2|\tau|}{\theta} \right)
\]

where \( \tau = |v_1 - v_2| \) is the absolute separation between points in the domain \( D = [0, L_0] \) and the value of the correlation length \( \theta \) is taken to be \( 0.6 \text{ m} \). The exponential correlation function is selected since it produces realisations which are not smooth which is realistic of material properties [18]. A single realisation of the base random field \( E(x) \) is shown in Figure 5 along the 3-point effective elastic moduli \( E^*_3(v) \) determined according to the experimental procedure illustrated in Figure 4.

![Figure 4: Setup of the computer experiment where effective properties \( E^*_{n,m} \) are observed over \( N = 7 \) segments of length \( L \approx 0.43 \text{ m} \) for a total of \( M = 20 \) beams of length \( L_0 = 3 \text{ m} \).](image)

![Figure 5: One realisation of the base process \( E(v) \) and its piecewise constant profile of 3-point effective elastic modulus \( E^*_3 \).](image)
3 Simulation
The approach adopted to simulate the lognormal process \( E(v) \) is to first simulate the underlying Gaussian process \( E_G(v) \) and then obtain \( E(v) \) through:

\[
E(v) = \exp(E_G(v))
\]  

(7)

Inconveniently the operation of exponentiation alters the statistical moments of \( E_G(v) \) and consequently in order to obtain a lognormal process \( E(v) \) with specified statistics \( \mu, \sigma^2, \rho(\tau) \) the approach is to backconfigure the properties \( \mu_G, \sigma_G^2, \rho_G(\tau) \) of the Gaussian process \( E_G(v) \) using the following relationships [6]:

\[
\mu_G = \ln \left( \mu \left( 1 + \frac{\sigma^2}{\mu^2} \right)^{-\frac{\mu}{2}} \right)
\]  

(8)

\[
\sigma_G^2 = \ln \left( 1 + \frac{\sigma^2}{\mu^2} \right)
\]  

(9)

\[
\rho_G(\tau) = \frac{\ln \left( \rho(\tau)(\exp(\sigma^2) - 1) + 1 \right)}{\sigma^2}
\]  

(10)

The numerical simulation of the Gaussian process \( E_G(v) \) requires that it be discretised and a common method to achieve this also happens be the midpoint approximation. It should be recognised that the midpoint approximation is not only widely used for estimation but also for the simulation of random fields and it is the ubiquitousness of this approximation that motivates this research into the uncertainty associated with it. Under the midpoint approximation \( E_G(v) \) is approximated in each element \( \Omega_k \) with a piecewise constant function \( \tilde{E}_G(v) \) given by the value of the field at the centroid \( v_c \) of the element:

\[
\tilde{E}_G(v) = E_G(v_c), \ v \in \Omega_k
\]  

(11)

The discretised field \( \tilde{E}_G(v) \) is then given by a random vector \( E_G = \{E_G(v_1^k), \ldots, E_G(v_K)\} \) where \( K = 1050 \) is the number of elements. Since \( E_G(v) \) is a Gaussian random field then \( E_G \) is multivariate normal random vector with mean \( \mu_G \), standard deviation \( \sigma_G \) and correlation matrix \( \Sigma_G \). The method adopted to simulate the Gaussian random vector \( E_G \) is covariance matrix decomposition [19] which is based on the following decomposition of \( E_G \):

\[
E_G = \mu_G + \sigma_G \psi \xi
\]  

(12)

where \( \mu_G \) and \( \sigma_G \) are the mean and standard deviation of the underlying normal random field \( E_G(v) \), \( \xi \) is a column vector of \( K \) uncorrelated mean-zero, unit variance normal random variables and \( \psi \) is a \( K \) by \( K \) matrix that satisfies the following decomposition of the correlation matrix \( \Sigma_G \):

\[
\psi \psi^T = \Sigma_G
\]  

(13)

where \( \psi \) is obtained as a lower-triangular matrix through Cholesky decomposition [20]. The discretised approximation to \( E(v) \) is then obtained through \( E = \exp(E_G) \). Note that since \( E_G(v) \) was discretised according to the midpoint approximation method then \( E(v) \) is also similarly discretised. In each segment the random field is discretised into \( K^* = 150 \) elements and the effective property of the \( n^{th} \) segment of the \( m^{th} \) beam \( E_{n,m}^* \) is obtained by discretising the integral in Equation 1 for instance the effective property of the first segment of the \( m^{th} \) beam is given by:

\[
E_{1,m}^* = \frac{K^*}{L} \sum_{k=1}^{K^*} \frac{\omega(v_k)}{E_m(v_k)}
\]  

(14)

4 Estimation
Estimates of the location \( \mu \) and scale \( \sigma \) parameter of the random field model can be obtained from the observed effective properties \( E_{n,m}^* \) using the following method of moments estimators:

\[
\hat{\mu} = \frac{1}{NM} \sum_{m=1}^{N} \sum_{n=1}^{M} E_{n,m}^*
\]  

(15)
Uncertainty quantification for random fields estimated from effective moduli of elasticity

\[ \hat{\sigma}^2 = \frac{1}{NM-1} \sum_{m=1}^{M} \sum_{n=1}^{N} (E_{n,m}^* - \hat{\mu})^2 \]

(16)

where \( N \) is the number of segments over which \( E^* \) is observed and \( M \) is the number of beams. These estimators are simply the sample mean and variance and are optimal for the case that the samples are statistically independent however since the effective properties are correlated along the length of the beam this is another source of epistemic bias that is quantified in the Section 5.

A common approach to estimate the correlation parameter \( \theta \) is to compute the sample correlation function also known as the correlogram \[21\] and then use regression to estimate the value of \( \theta \) that minimises the misfit between the sample correlation function \( \hat{\rho}(\tau) \) and theoretical model \( \rho(\tau) \) \[22\]. However this procedure is computationally expensive for the number of simulations \( R = 1 \times 10^6 \) required by the computer experiment and consequently a computationally simpler approach is pursued here. Instead of considering the full autocorrelation structure only the lag-1 correlation between adjacent effective properties is considered. An estimator of the lag-1 correlation coefficient \( \rho_1 \) is given by:

\[ \hat{\rho}_1 = \frac{1}{M(N-1)\hat{\sigma}} \sum_{m=1}^{M} \sum_{n=1}^{N-1} (E_{n,m}^* - \hat{\mu}) (E_{n,m}^* - \hat{\mu}) \]

(17)

An estimate of the correlation length of the exponential correlation function is obtained from the lag-1 correlation coefficient \( \hat{\rho}_1 \) by changing the subject of Equation 6 to \( \theta \) and substituting the length over which the effective property was observed \( \tau = L \):

\[ \hat{\theta} = -\frac{2L}{\ln (\hat{\rho}_1)} \]

(18)

where \( L \approx 0.43 \) m. In the case of high aleatoric uncertainty due to limited sample size a negative value may be obtained for the lag-1 correlation coefficient \( \hat{\rho}_1 \) in which case the correlation length estimate \( \hat{\theta} \) would be complex which is not physically meaningful. The approach utilised to resolve this issue is to set the \( \hat{\theta} \) equal to zero in the case that \( \hat{\rho}_1 \) is estimated as having a negative value.

5 Results

This section presents the results of the computer experiment into the distribution of the parameter estimates \( \{\hat{\mu}, \hat{\sigma}, \hat{\theta}\} \) of the random field. The statistics of these estimated parameters are referred to as meta-statistics in order to distinguish them from the statistics of the random field. Two non-normalised meta-statistics are considered for each parameter estimate namely the sample mean Mean[,] and standard deviation Std[,] of the parameter estimates from the true value of the parameters \( \mu = 10 \text{ GPa}, \sigma = 3 \text{ GPa}, \theta = 0.6 \text{ m} \) is a measure of the uncertainty due to epistemic error. A normalised meta-statistic, percentage bias Bias[,] is introduced to facilitate comparison of the epistemic uncertainty between different parameter estimates:

\[ \text{Bias} \{\hat{\mu}\} = 100\% \frac{R}{R} \sum_{r=1}^{R} \left( \frac{\hat{\mu}_r - \mu}{\mu} \right) \]

(19)

where \( R \) is the number of simulation, \( \hat{\mu}_r \) is the parameter estimate associated with the \( r^{th} \) simulation and \( \mu \) is the true value of the parameter. The total bias estimated by the meta-statistic Bias[,] consists of bias due to both the midpoint approximation and the bias associated with the estimators presented in Section 4. The meta-statistics of point estimates from the base process \( E(v) \) are also presented since the bias in these estimates is only due to estimator bias and this enables inferences on what proportion of the bias is due to estimator bias in comparison to the midpoint approximation for the estimates obtained from effective properties. The meta-statistic Std[,] quantifies the aleatoric uncertainty in each of the parameter estimates of the random field. This uncertainty is due to a number of factors including the non-ergodicity of each realisation of \( E(v) \), the limited number of beams \( M \) and sections \( N \) over which the effective properties \( E^* \) are observed. The normalised meta-statistic maximum absolute percentage error MAPE[,] \[23\] is utilised to evaluate total uncertainty resulting from both aleatoric and epistemic uncertainty:

\[ \text{MAPE} \{\hat{\mu}\} = 100\% \frac{R}{R} \sum_{r=1}^{R} \left| \frac{\hat{\mu}_r - \mu}{\mu} \right| \]

(20)
5.1 Distribution of sample mean

Figure 6 presents the distribution of the parameter estimate \( \hat{\mu} \) for each of the tests whilst Table 1 details the four meta-statistics obtained for each parameter estimate. The bias in \( \mu \) for the base process \( E(v) \) has been found to be negligible despite the sample mean not being an unbiased estimator for correlated samples. All of the effective properties \( \{ E_T, E_3, E_4 \} \) have decreased expected value compared to the base process \( E(v) \) with expected value \( \mu = 10 \) GPa. This is due to the effective properties being given by a harmonic mean which tends towards the least elements of the set and consequently the effective properties are disproportionately affected by regions of low stiffness. This effect is weaker for bending behaviour because weak-zones near the supports do not significantly affect expected value however in the case of tensile loading the effect of weak-zones on the effective property \( E^* \) is independent of spatial position due to the uniform tensile weighting function \( \omega_T(v) \). The deviation in the expected value of the location parameter \( \hat{\mu} \) is slightly greater for the case of 4-point bending since the weighting function \( \omega_4(v) \) is closer to uniform than the 3-point weighting function \( \omega_3(v) \).

A similar variability, as quantified by the meta-statistic Std[], has been observed for all the tests suggesting that the aleatoric uncertainty in the location parameter estimate \( \hat{\mu} \) is not significantly does not vary significantly between tests. In particular the aleatoric uncertainty is similar for estimates of \( \mu \) from the base process \( E(v) \) and estimates of \( \mu \) from its effective property \( E^* \). This is due to the fact that the variability of the arithmetic mean of a set of random variables is not significantly different than the variability of the harmonic mean of the same set. The standard deviation is slightly greater for 3-point bending and this is explained by considering that as the weighting function \( \omega(v) \) approaches the Dirac delta function \( \delta(v - v_c) \) the effective property \( E^* \) is in effect averaging over an increasingly small statistical region and consequently variability in the estimate increases. MAPE[] is similar for all tests suggesting that the epistemic uncertainty due to the midpoint approximation is small compared to the aleatoric uncertainty associated with \( M = 20 \) realisations of the beam stiffness profile \( E(v) \) for the experimental setup detailed in Figure 4.

![Figure 6: Distribution of location parameter estimate \( \hat{\mu} \) for tension \( E_T^*(v) \), 3-point \( E_3^*(v) \) and 4-point \( E_4^*(v) \) bending tests with base distribution \( E(v) \) shown for comparison.](image.png)

<table>
<thead>
<tr>
<th>Observed process</th>
<th>Mean[( \hat{\mu} )] (GPa)</th>
<th>Std[( \hat{\mu} )] (MPa)</th>
<th>Bias[( \hat{\mu} )] (%)</th>
<th>MAPE[( \hat{\mu} )] (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base ( E(v) )</td>
<td>10.00</td>
<td>284</td>
<td>0.00</td>
<td>2.27</td>
</tr>
<tr>
<td>Tension ( E_T^*(v) )</td>
<td>9.71</td>
<td>276</td>
<td>-2.85</td>
<td>3.30</td>
</tr>
<tr>
<td>3-point ( E_3^*(v) )</td>
<td>9.83</td>
<td>287</td>
<td>-1.75</td>
<td>2.72</td>
</tr>
<tr>
<td>4-point ( E_4^*(v) )</td>
<td>9.81</td>
<td>285</td>
<td>-1.87</td>
<td>2.76</td>
</tr>
</tbody>
</table>
5.2 Distribution of sample standard deviation

Figure 7 presents the distribution of the scale parameter estimate $\hat{\sigma}$ for each of the tests whilst Table 2 details the four meta-statistics obtained for each parameter estimate. Figure 7 shows that the estimates of the scale parameter $\hat{\sigma}$ obtained from the effective properties $\{E_3^*, E_4^*, E_T^*\}$ underestimate the true variability of the base process $E(v)$. This underestimation of the true variability of the base process $E(v)$ is unconservative since under random field theory the beam would be modelled as being inappropriately homogeneous. The trends are similar to those observed for $\hat{\mu}$ and the reduction in the expected value of $\hat{\sigma}$ is once again greatest for tensile loading $E_T^*$ and at a minimum for 3-point bending $E_3^*$. Table 2 shows that the estimator of standard deviation is negatively biased for the base process $E$ and $E^*$ is attributable to estimator bias and not the epistemic bias due to the midpoint approximation.

There is not a significant variation between the meta-statistic $\text{Std}[\hat{\sigma}]$ for each effective property and consequently it is concluded that estimating the location and scale parameters of the base process from its effective properties does not significantly increase the aleatoric uncertainty associated with estimates of either the location $\mu$ or scale parameter $\sigma$. The MAPE of the meta-statistics suggests that for $\hat{\sigma}$ the epistemic uncertainty is significantly greater than the aleatoric uncertainty whilst for $\hat{\mu}$ they are of similar magnitude. Both of the parameter estimates $\hat{\mu}$ and $\hat{\sigma}$ experience a reduction in expected value when estimated from the effective properties however it should be noted that whilst underestimation of $\mu$ is conservative the underestimation of $\sigma$ is not and consequently there is a necessity to develop appropriate methodologies to correct for the epistemic error in the scale parameter estimate $\hat{\sigma}$ so that the stochastic finite element model appropriately quantifies the risk associated with heterogeneous structures.

![Figure 7: Distribution of scale parameter estimate $\hat{\sigma}$ for tension $E_T^*(v)$, 3-point $E_3^*(v)$ and 4-point $E_4^*(v)$ bending tests with base distribution $E(v)$ shown for comparison.](image)

Table 2: Meta-statistics of the scale parameter estimate $\hat{\sigma}$.

<table>
<thead>
<tr>
<th>Observed process</th>
<th>Mean[$\hat{\sigma}$] (GPa)</th>
<th>Std[$\hat{\sigma}$] (MPa)</th>
<th>Bias[$\hat{\sigma}$] (%)</th>
<th>MAPE[$\hat{\sigma}$] (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>name</strong></td>
<td><strong>symbol</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>$E(v)$</td>
<td>2.98</td>
<td>197</td>
<td>0.7</td>
</tr>
<tr>
<td>Tension</td>
<td>$E_T^*(v)$</td>
<td>2.34</td>
<td>198</td>
<td>-22.0</td>
</tr>
<tr>
<td>3-point</td>
<td>$E_3^*(v)$</td>
<td>2.60</td>
<td>215</td>
<td>-13.4</td>
</tr>
<tr>
<td>4-point</td>
<td>$E_4^*(v)$</td>
<td>2.57</td>
<td>213</td>
<td>-14.3</td>
</tr>
</tbody>
</table>

5.3 Distribution of correlation length

Figure 8 presents the distribution of the parameter estimate $\hat{\theta}$ for each of the tests whilst Table 3 details the four meta-statistics of $\theta$ for each test. Figure 8 shows that all of the parameter estimates $\hat{\theta}$ obtained from the effective properties $\{E_3^*, E_4^*, E_T^*\}$ exhibit large positive biases in comparison to the negative biases observed for both $\hat{\mu}$ and $\hat{\sigma}$. This is due to the effective properties acting to average and therefore...
smooth the base process $E(v)$ thus greatly increasing measured correlation. The Bias$[\hat{\theta}]$ in $\hat{\theta}$ for the base process $E(v)$ in Table 3 indicates that the lag-1 correlation length estimator given by Equation 18 has a negative bias of approximately 1%. The small magnitude of this underestimation indicates the lag-1 estimator is a good estimator of the correlation length $\theta$ provided the number of samples is sufficiently large. However this negative bias implies that the true increase in correlation caused by estimating from the effective properties is greater than that suggested by the Bias$[\hat{\theta}]$ meta-statistic in Table 3 by approximately a single percentage point.

The uncertainty associated with estimating the correlation parameter $\theta$ from effective properties is significantly higher than estimating $\theta$ directly from the base process $E(v)$. This is unsurprising since the estimator of lag-1 correlation given by Equation 18 depends on both $\hat{\mu}$ and $\hat{\sigma}$ and consequently the uncertainty associated with both of these parameter estimates is propagated into the estimate $\hat{\theta}$ of the correlation length. Estimating from the effective properties $E^*$ did not lead to a significant increase in aleatoric uncertainty for $\hat{\mu}$ and $\hat{\sigma}$ however the aleatoric uncertainty is significantly increased for $\hat{\theta}$ and this is due to the significant distortion of the correlation information caused by the weighted harmonic mean in comparison to the distortion observed for either location or scale information. The MAPE$[\hat{\theta}]$ meta-statistic shows that both the aleatoric and epistemic uncertainties have a similar contribution to total uncertainty however it is not possible to make a general conclusion as to whether the overestimation of the correlation length is conservative or unconservative.

Figure 8: Distribution of correlation parameter estimate $\hat{\theta}$ for tension $E^*_T(v)$, 3-point $E^*_3(v)$ and 4-point $E^*_4(v)$ bending tests with base distribution $E(v)$ shown for comparison

<table>
<thead>
<tr>
<th>Observed process</th>
<th>Mean$[\hat{\theta}]$</th>
<th>Std$[\hat{\theta}]$</th>
<th>Bias$[\hat{\theta}]$</th>
<th>MAPE$[\hat{\theta}]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>symbol</td>
<td>(m)</td>
<td>(mm)</td>
<td>(%)</td>
</tr>
<tr>
<td>Base</td>
<td>$E(v)$</td>
<td>0.59</td>
<td>61</td>
<td>-1.1</td>
</tr>
<tr>
<td>Tension</td>
<td>$E^*_T(v)$</td>
<td>1.00</td>
<td>219</td>
<td>+67.5</td>
</tr>
<tr>
<td>3-point</td>
<td>$E^*_3(v)$</td>
<td>0.74</td>
<td>181</td>
<td>+23.7</td>
</tr>
<tr>
<td>4-point</td>
<td>$E^*_4(v)$</td>
<td>0.76</td>
<td>183</td>
<td>+26.5</td>
</tr>
</tbody>
</table>

6 Conclusion

The computer experiment has shown that estimating random field parameters from the effective properties $\{E^*_T, E^*_3, E^*_4\}$ using the midpoint approximation results in a minor negative epistemic bias in the location parameter estimate $\hat{\mu}$, a moderate negative epistemic bias in the scale parameter estimate $\hat{\sigma}$ and a large positive epistemic bias in the correlation parameter estimate $\hat{\theta}$. The aleatoric uncertainty associated with estimates of $\mu$ and $\sigma$ obtained from effective properties is similar to base aleatoric uncertainty whilst the aleatoric uncertainty associated with the estimate of $\theta$ from the effective properties is significantly greater than base aleatoric uncertainty. The computer experiment is only run for one set of random field
parameters ($\mu = 12$ GPa, $\sigma = 3$ GPa, $\theta = 0.6$ m) however the trends are expected to hold for different parameter values with the magnitude of the trend increasing as either the scale parameter $\sigma$ increases of the correlation ratio $\theta/L$ decreases since this increases the heterogeneity of the beam.

The trends in the meta-statistics of the parameter estimates are consistent for all three tests (tension $E_T^*$, 3-point $E_3^*$ and 4-point $E_4^*$ bending) however the magnitude of the trend is always greatest for the tension test and smallest for three-point bending. It is concluded that the midpoint approximation is least appropriate for tension due to the uniformity of its weighting function $\omega_T(v)$ and most appropriate for 3-point bending due to the central concentration of its weighting function $\omega_3(v)$. The error induced by the midpoint approximation is lower for the 3-point bending test than the 4-point bending test however this increased accuracy has been found to be insignificant compared to the magnitude of the epistemic error induced by midpoint approximation. It is the recommendation of this paper that further research should focus on the development of a methodology to correct the epistemic biases in the parameter estimates obtained from effective properties. There has been some research conducted on this topic however it has been under the assumption that the effective properties of a beam is an unweighted arithmetic mean [24] which neither accounts for the disproportionate contribution of zones of low stiffness to effective modulus of elasticity or the spatial weighting associated with bending tests.

Acknowledgments

The authors would like to acknowledge the support of both the Resilience Engineering Research Group and the Faculty of Engineering, University of Nottingham in conducting this research.

Appendix

This appendix contains the derivation of the weighting function $\omega_4(v)$ for the effective property $E_4^*$ associated with midspan deflection $U_B$ of a simply supported beam under 4-point loading $P_4$ as shown in Figure 2. The variables $u$ and $v$ describe spatial variation in the internal bending moment $M$, elastic modulus $E$ and spatial weighting $\omega$ whilst the variable $x$ describes variation in displacement $U$. Under Euler-Bernoulli beam theory the relationship between beam deflection $U(x)$ and modulus of elasticity $E(v)$ is given by:

$$\frac{d^2 U(x)}{dx^2} = -\frac{M(v)}{E(v)I} \tag{21}$$

where $I$ is the area moment of inertia which is assumed to be constant. Under the standard theory of double integration beam deflection $U(x)$ is given by:

$$U(x) = \frac{1}{l} \int_0^x \int_0^u \frac{M(v)}{E(v)} \, dv \, du + Cx + D \tag{22}$$

where $C$ and $D$ are the first and second constants of integration respectively. Applying boundary conditions for a simply supported beam, namely that $U(x = 0) = 0$ and $U(x = L) = 0$, gives:

$$U(x) = \frac{1}{l} \int_0^x \int_0^u \frac{M(v)}{E(v)} \, dv \, du - \frac{x}{I} \int_0^L \int_0^u \frac{M(v)}{E(v)} \, dv \, du \tag{23}$$

Through an interchange in the order of integration the double integral can be reduced to a single integral:

$$U(x) = \frac{1}{l} \int_0^x \frac{M(v)}{E(v)} \int_v^x \, du \, dv - \frac{x}{l} \int_0^L \frac{M(v)}{E(v)} \int_v^L \, du \, dv \tag{24}$$

$$= \frac{1}{l} \int_0^x \frac{M(v)(x - v)}{E(v)} \, dv - \frac{x}{l} \int_0^L \frac{M(v)(L - v)}{E(v)} \, dv \tag{25}$$

An expression for midspan deflection $U_B$ is obtained by substituting $x = L/2$ into Equation 24:

$$U_B = \frac{1}{l} \int_0^{L/2} \frac{M(v)(L/2 - v)}{E(v)} \, dv - \frac{1}{2l} \int_0^L \frac{M(v)(L - v)}{E(v)} \, dv \tag{26}$$
The internal bending moment $M_4(v)$ for four-point loading as shown in Figure 2 is given by:

$$M_4(v) = \begin{cases} 
-Pv/2 & \text{if } 0 \leq v \leq \frac{L}{3} \\
-PL/6 & \text{if } \frac{L}{3} \leq v \leq \frac{2L}{3} \\
-P(L-v)/2 & \text{if } \frac{2L}{3} \leq v \leq L 
\end{cases}$$  \hspace{1cm} (27)

Substituting the equation for expression for bending moment $M_4(v)$ into Equation 26 yields:

$$U_B = \frac{P}{T} \left[ -\int_0^{L/3} \frac{2v(L/2-v)}{4E(v)} \, dv - \int_{L/3}^{L/2} \frac{2L(L/2-v)}{12E(v)} \, dv \\
+ \int_{L/3}^{L/3} \frac{v(L-v)}{4E(v)} \, dv + \int_{L/3}^{2L/3} \frac{L(L-v)}{12E(v)} \, dv + \int_{2L/3}^{L} \frac{(L-v)^2}{4E(v)} \, dv \right]$$  \hspace{1cm} (28)

Combining terms over similar domains of integration and simplifying gives:

$$U_B = \frac{P}{T} \left[ \int_0^{L/3} \frac{v^2}{4E(v)} \, dv + \int_{L/3}^{L/2} \frac{Lv}{12E(v)} \, dv \\
+ \int_{L/3}^{2L/3} \frac{L(L-v)}{12E(v)} \, dv + \int_{2L/3}^{L} \frac{(L-v)^2}{4E(v)} \, dv \right]$$  \hspace{1cm} (29)

Under the assumption that the beam is homogeneous with constant modulus of elasticity $E_4^*$ the midspan displacement is given by:

$$U_B = \frac{23L^3P}{648IE_4^*} \hspace{1cm} (30)$$

An equation for the effective property $E_4^*$ is obtained in terms of midspan displacement $U_B$ by changing the subject of Equation 30:

$$E_4^* = \frac{23L^3P}{648IU_B} \hspace{1cm} (31)$$

Substituting the heterogeneous result given by Equation 29 into the homogeneous result given by Equation 31 yields:

$$E_4^* = \frac{23L^3}{648} \left( \int_0^{L/3} \frac{v^2}{4E(v)} \, dv + \int_{L/3}^{L/2} \frac{Lv}{12E(v)} \, dv + \int_{L/3}^{2L/3} \frac{L(L-v)}{12E(v)} \, dv + \int_{2L/3}^{L} \frac{(L-v)^2}{4E(v)} \, dv \right) \hspace{1cm} (32)$$

The effective property $E_4^*$ for four-point bending is then expressed as a weighted harmonic mean:

$$E_4^* = \left( \int_0^L \omega_4(v) \cdot E(v)^{-1} \, dv \right)^{-1} \hspace{1cm} (33)$$

where the weighting function $\omega_4(v)$ is given by:

$$\omega_4(v) = \begin{cases} 
\frac{324v^2}{23L^3} & \text{if } 0 \leq v \leq \frac{L}{3} \\
108v & \text{if } \frac{L}{3} \leq v \leq \frac{L}{2} \\
108(L-v) & \text{if } \frac{L}{2} \leq v \leq \frac{2L}{3} \\
324(L-v)^2/23L^3 & \text{if } \frac{2L}{3} \leq v \leq L 
\end{cases}$$  \hspace{1cm} (34)
References


Polymorphic uncertainty modeling for optimization of timber structures

Friedemann Niklas Schietzold, Wolfgang Graf, and Michael Kaliske*

Institute for Structural Analysis, Technische Universität Dresden, Germany
*Corresponding author: Michael.Kaliske[at]tu-dresden.de

Abstract

Uncertainty characteristics of wood are mainly originated in natural variation. Out of this, traditional approaches of random variables can be expanded to polymorphic uncertainty models. Therefore, e.g. fuzzy probability based randomness is used in such a way, that random variables are extended by fuzzy variables in parameterization of the distribution functions, see [1] and [2]. In this manner, coupling of both aleatoric and epistemic uncertainty is involved in uncertainty analysis.

The FEM is applied as a basic solution of particular load situations and the focused timber structures. A local orthotropic formulation is used and the properties are evaluated of each integration point with respect to a specified location of the tree trunk axis.

In this contribution, an approach to polymorphic uncertainty modeling for timber structures is introduced. According to [3], models representing the spatial variation and interdependencies of material parameters are required for a realistic representation in numerical simulation. For this purpose, on the one hand interactions between fuzzy variables, on the other hand correlations among random variables are considered. Random fields are utilized to capture spatially varying material properties in context with the discretization of FE. Approaches to both, spatially and structurally depending autocorrelations along with crosscorrelations based on [4], are presented.

The preliminary steps aim at an optimization in design of timber structures, provided that polymorphic uncertain design as well as a priori parameters are considered. The tools for uncertainty analysis and the basic FEM solution are prepared as a basis for an automated optimization processing, whereas they are preferably parallelized, incorporating methods for reducing the numerical effort. Results of the uncertainty analysis of a timber structure are shown exemplary.

Keywords: Polymorphic Uncertainty, Fuzzy Probability based Randomness, Random Fields, Structural Depending Correlation

1 Introduction

A fundamental aspect of natural materials, such as in wooden structures, is the underlying uncertainty. Natural variability, uncertainties in experimental data and lack of knowledge may not be neglected in numerical simulations of those in engineering tasks. The material parameters of timber components naturally scatter due to growth, environmental conditions and spatial variation. Therefore, models incorporating the different characteristics of uncertainty are needed in the analysis of timber structures. This is on the one hand the natural variation, characterized by aleatoric uncertainty. On the other hand, lack of knowledge and experimental discrepancies, which are characterized by epistemic uncertainty. Additionally, spatial variations are relevant for most engineering tasks, especially for the uncertainty modeling of material properties defined on the macroscale such as strength and stiffness. In the context of imprecise probability, those uncertainty characteristics are combined and applied to a spatial polymorphic uncertainty modeling of timber structures.

The design of timber structures can be supported by optimization strategies. In traditional approaches, deterministic design variables and a priori variables are considered and, therefore, deterministic optimization results are obtained. The incorporation of spatial polymorphic uncertainty models leads to an optimization on top of an extensive uncertainty analysis. Approaches to this framework are points of interest of recent researches. As a preliminary step to the extension of optimization, in this contribution a comprehensive spatial polymorphic uncertainty modeling for timber structures is defined and uncertainty analysis approaches are presented.

This contribution is divided into four main parts. Firstly, an overview on the mathematical fundamentals is given regarding definitions of several basic uncertainty models. Those approaches are based on fuzzy and random variables. The latter one is subsequently extended by the definition of the spatial random field. The uncertainty model fundamentals are concluding in the definition of the fuzzy prob-
ability based random variable as a polymorphic uncertainty model, according to combinations of basic uncertainty characterizations.

These definitions are secondly followed by the description of different approaches of taking interdependencies into account in aleatoric and epistemic uncertainties. After a short introduction to the structure of the uncertainty analysis, therefore, fuzzy interactions, random field autocorrelations and random field crosscorrelations are defined.

The definitions of uncertainty models are thirdly applied to timber structures, where special characteristics of wood, regarding structure and uncertainty, are analyzed. The approaches are applied in an extensive example of an uncertainty analysis of a rafter tie beam capentry joint.

Finally, conclusions of the presented contribution are outlined. The links for expansion to an optimization of timber structures with uncertain design variables and a priori variables are shown as well as possible objectives of optimization.

2 Polymorphic Uncertainty Model

2.1 Basic Uncertainty Models

2.1.1 Fuzzy Variables

The fuzzy variable $A^f$ in the sense of [5] is given by association of a membership function $\mu_{A^f}(x)$ to values $x$ with

$$
\mu : \mathbb{R} \to [0; 1] : x \mapsto \mu(x), \\
\forall x \in \mathbb{R} : 0 \leq \mu_{A^f}(x) \leq 1, \\
\exists x_l; x_r \in \mathbb{R} : \forall x \in [x_l; x_r] : \mu_{A^f}(x) = 1.
$$

(1) (2) (3)

The convexity regarding [6] is an optional restriction to the membership function, where its monotonic decrease is required in $x$ for both sides of the peak values $\mu_{A^f}(x) = 1$. This facilitates the addition of further weighting information to an interval by associating a convex membership function $\mu_{A^f}(x)$.

Fundamental for numerical analysis with fuzzy variables is the representation by $\alpha$-levels, e.g. described by [3]. For each $\alpha$-level with $\alpha \in (0; 1]$, cuts of the membership function as intervals $I_{A^f, \alpha}$ can be derived, where each

$$
I_{A^f, \alpha} = [x_{\alpha,l}; x_{\alpha,r}], \\
\forall x \in I_{A^f, \alpha} : \mu_{A^f}(x) \geq \alpha.
$$

(4) (5)

With the set of all cuts $I_{A^f} = \{I_{\alpha} \mid \alpha \in (0; 1]\}$, a discretization of fuzzy variables is treated. If the set of $n$ discrete $\alpha$-level cuts, with $n \geq 2$, is sorted in $i = 1, 2, \ldots, n - 1$, it holds that

$$
0 \leq \mu_{A^f}(x_{\alpha,i}) \leq \mu_{A^f}(x_{\alpha,i+1}) \leq 1, \\
\text{where } \mu_{A^f}(x_{\alpha,1}) = 0 \text{ and } \mu_{A^f}(x_{\alpha,n}) = 1.
$$

(6) (7)

![Figure 1: Membership functions $\mu(x)$ of convex fuzzy variables, left with an $\alpha$-level pointed out, right with linguistic weighting information in three categories (from [3]).](image)

The definition of left and right interval bounds of all regarded $\alpha$-levels is sufficient for the discretized modeling of fuzzy variables. For a fuzzy variable $A^f$ assuming $n$ discrete $\alpha$-levels,

$$
A^f = \{x_{\alpha_1,l}; x_{\alpha_1,r}; \ldots; x_{\alpha_n,l}; x_{\alpha_n,r} \} : x_{\alpha_1,l} \leq x_{\alpha_n,r}
$$

(8)

leads to a short notation. Two special cases arise from this. If $n = 2$,

$$
B^f_{n=2} = \{x_{\alpha_1,l}; x_{\alpha_2,l}; x_{\alpha_2,r}; x_{\alpha_1,r}\}
$$

(9)
describes a fuzzy trapezoidal variable. Involving the condition \( x_{\alpha_{2}} = x_{\alpha_{2},l} = x_{\alpha_{2},r} \) on \( \alpha \)-level \( \alpha_{2} = 1 \), it concludes in a fuzzy triangular variable and

\[
C_{n=2}^{\alpha} = \langle x_{\alpha_{1};l}; x_{\alpha_{2}}; x_{\alpha_{1};r} \rangle
\]

is the short notation with three parameters. The interval \( S = [x_{\alpha_{1};l}; x_{\alpha_{1};r}] \) is known as the support of the fuzzy variable (see [1]).

The nature of uncertainty described by fuzzy variables as a possibilistic measure is called epistemic uncertainty. A fuzzy variable with an \( \alpha \)-level is illustrated in Figure 1 as well as one representation with specific linguistic weighting information associated with the membership function of possibility.

The numerical treatment of fuzzy variables, based on the discretization of \( \alpha \)-levels, can be performed via an optimization strategy for every interval in accordance with the observed \( \alpha \)-level-cuts, see [7]. In case of at least weak monotonic behavior of the fundamental solution of system response over the observed interval, the extreme values will be found at the corners. While this monotonic behavior is not given in general, an appropriate optimization method is required.

### 2.1.2 Random Variables

Random uncertainty models (probabilistic models) used in this context are based on the common probability space \((\Omega, \Sigma, P)\). A random variable \(X^r\) is a mapping

\[
X^r : \Omega \rightarrow \mathbb{R} : \omega \mapsto X^r(\omega)
\]

and

\[
P : \Sigma \rightarrow [0; 1]
\]

is the probability, or rather the probabilistic measure for the occurrence of an event \(x \in \Sigma\). Therefore,

\[
F(x) = P(X^r \leq x)
\]

is the cumulative distribution function (cdf) in \(x\). The probability density function (pdf)

\[
f(x) = \frac{dF(x)}{dx}
\]

is obtained by differentiation of the cdf.

Each random variable \(X^r\) is formed by the associated functions \(F_{X^r}\) and \(f_{X^r}\), which are specified by parameters \(\lambda_{X^r}\). Those parameters, e.g. moments like mean value \(E(X^r)\) and standard deviation \(S(X^r)\), are deterministic for common random variables and yield

\[
F_{X^r}(x) = F(x; \lambda_{X^r}) \quad \text{and} \quad f_{X^r}(x) = f(x; \lambda_{X^r}).
\]

Random variables are characterized by probabilistic uncertainty and, therefore, classified as aleatoric uncertainty models. With the exception of non-parametric approaches, the choice of an underlying distribution function is obligatory for modeling a random variable.

The numerical simulation of random variables is treated by generating common independent random numbers and transforming them according to [3], while applying a priori defined distributions.

### 2.2 Random Fields

In [8] and [9], a random process is defined as a set of random variables

\[
\{H^r(\theta), \ \theta \in \Theta\},
\]

where \(H^r(\theta)\) is a function that can be interpreted as random variable at every point \(\theta\) in the space of observation \(\Theta\). The space of observation may be associated with a timeline, where \(\Theta \subseteq T\) or with an Euklidian space, where \(\theta \in \Theta \subseteq \mathbb{R}^3\). For the latter one, the uncertainty model is also known as random field, where \(\theta\) represents a coordinate vector of the observed point in space. Both are significantly important in the field of engineering tasks in use of the stochastic finite element method (SFEM).

Based on the common probability space \((\Omega, \Sigma, P)\), a random field is

\[
H^r = H^r(\theta, \omega), \ \omega \in \Omega
\]
following the definition of [3]. The random field is called univariate, if for every point \( \theta \), the field \( H^i(\theta, \omega) = H^i \) is a random variable. Otherwise, if \( H^i(\theta, \omega) = H^i \) is a vector of random variables, the field is multivariate and the associated probability functions of the underlying random variables are called marginal distributions. Different marginals \( F_{H^i}(h_i, \theta) = P(H^i \leq h_i) \) are possible for every random variable. The random field is stationary, if all of them are independent of the point of observation \( \theta \).

The trend function (see [8]), also known as mean function (see [3]) of a random field is \( E(H_{\theta, \omega}) \{. \) Based on this, \( \text{Kov}(\theta_A, \theta_B) = E(H^i(\theta_A, \omega), H^i(\theta_B, \omega)) \) defines the auto-covariance function for two points \( A \) and \( B \) in the space of observation.

### 2.3 Fuzzy Probability Based Random Variables

The combination of both epistemic and aleatoric uncertainty models leads to polymorphic uncertainty. Fuzzy probability based random variables are polymorphic uncertainty models and are classified as imprecise probability.

Assume \( X^i \) to be a random variable referring to Section 2.1.2. If at least one of the distribution parameters \( \lambda_X^i \) is a fuzzy variable (see Section 2.1.1), the parameters become a set of fuzzy parameters \( \lambda_{X^i}^f \), defining a fuzzy-valued probability distribution function. The uncertain variable turns into a fuzzy probability based random \( (fp-r) \) variable \( X^{fr} \). Referring to [3], an \( fp-r \) variable is based on a fuzzy probability space \( (\Omega, \Sigma, P^f) \) and \( F_{X^fr}(x) = P^f(X^{fr} \leq x) \) is the related fuzzy-valued \( cdf \) and fuzzy valued probability, respectively. It holds for every \( \alpha \in (0; 1] \) of the underlying fuzzy variable that they become intervals on the \( \alpha \)-level-cuts. Hence, a fuzzy probability based random variable is a mapping from the set of elementary events in the fuzzy probability space to the real numbers \( X^{fr} : \Omega \rightarrow \mathbb{R} \).

Given that the random variable \( X^i \) is an observation of an univariate random field or part of a multivariate one, a fuzzy probability based random field is modeled, which performs as a polymorphic uncertainty function.

### 3 Interdependencies in Uncertainty Models

Approaches to interdependencies are well known for random variables and random fields, see Section 3.2. In addition to this, a simple method for the inclusion of interdependencies between fuzzy variables is shown in Section 3.1. The desired relationships are taken into account in numerical analysis in the appropriate order, which is illustrated in Figure 3.
F.N. Schietzold and W. Graf and M. Kaliske

stochastic uncertainty models is used to deal with the whole bandwidth of probabilities, occurring especially for natural materials like wood. With the epistemic characteristic of fuzzy variables, an expansion to this is done. This enables the treatment of lack of knowledge and permits the involvement of expert knowledge, e.g. by the modeling of a membership function. All those steps lead to a more and more extensive polymorphic uncertainty model. This results with mixed aleatoric and epistemic characteristics in the uncertainty model as well as in the observed problem solution, which in this case is a response of an uncertainty analysis incorporating on a structural analysis of the structural problem of interest.

Figure 3: Structure chart for the uncertainty analysis featuring interdependencies

Keeping this extensive model in mind, the second aim of interdependencies turns out as the reduction of uncertainty. If no interdependencies are taken into account, every specific combination of variable realizations is admissible and equals in possibility or probability for random or fuzzy variables, respectively. This means in effect an independent occurrence. Therefore, interdependencies can generally be interpreted as restrictions to the possibility or probability of those possible combinations in the total input space of variables.

3.1 Interaction of Fuzzy Variables

Interactions of fuzzy variables are described in [7] and [10]. According to this, let $A^f$ and $B^f$ be a fuzzy variable, respectively. The sampling domain $X$ is spanned by the Cartesian products

$$X = \{ A^f_\alpha \times B^f_\alpha \mid \alpha \in (0, 1) \}. \quad (23)$$

An illustration of such a fuzzy sampling domain is given in Figure 4a), where no interactions are applied and, therefore, all combinations of the two fuzzy variables are feasible. A simple approach of applying interactions is shown in Figure 4b). Therein, a spatial plane is constraining the sampling domain, excluding specific sampling combinations.

Figure 4: Spatial sampling spaces $X$ of two fuzzy variables

While the numerical analysis is treated by means of an optimization of discrete $\alpha$-levels, the interaction with planes in the three-dimensional sampling domain can be discretized to multiple two-dimensional constraint functions on every $\alpha$-level. Both the “Trial and Error” and the closest point projection can be used during the sampling to include the constraints into $\alpha$-level optimization.

A definition for up to four linear constraint functions via eight support points is shown in Figure 5 as well as an illustration of the top view associated with Figure 4b). By this, it becomes clear that the constraint function $I_{\alpha_1}$ defined on the supporting $\alpha$-level $\alpha_1 = 0$ as well has to be valid for all overlying $\alpha$-levels. Besides, arbitrary nonlinear constraint functions are conceivable.

8th International Workshop on Reliable Engineering Computing
16–18 July 2018, University of Liverpool, Liverpool, UK
Depending on the \(\alpha\)-level optimization strategy, the constraining of the fuzzy sample domain provides a beneficial reduction of the numerical effort. The domain of optimization is directly connected to this. It may also occur that multiple constraining leads to an extreme decrease of the optimization domain. In this case, it has to be ensured, that the optimization strategy is not failing, but rather still performs a feasible sampling.

### 3.2 Auto- and Crosscorrelation of Random Fields

Autocorrelation functions are used to describe interdependencies within random fields. An autocorrelation function \(R_{Hrf}(\theta_A; \theta_B)\) evaluates the autocorrelation coefficient between every two points \(\theta_A\) and \(\theta_B\) in the observation space of a random field \(H_{rf}\). It is defined as the mapping

\[
R : \Theta \times \Theta \to [-1; 1],
\]

\[
: \theta_A; \theta_B \mapsto R(\theta_A; \theta_B).
\]

In \([3, 4]\) and \([11]\), besides others the following common autocorrelation functions are mentioned

- **exponential**
  \[
  R_{\text{exp}}(\theta_A; \theta_B) = \exp \left( -\frac{d_\Theta(\theta_A; \theta_B)}{l_\Theta} \right),
  \]

- **squared exponential**
  \[
  R_{\text{exp}^2}(\theta_A; \theta_B) = \exp \left( -\frac{(d_\Theta(\theta_A; \theta_B))^2}{l_\Theta^2} \right),
  \]

- **cardinal sine**
  \[
  R_{\text{sin}}(\theta_A; \theta_B) = \sin \left( \frac{d_\Theta(\theta_A; \theta_B)}{l_\Theta} \right),
  \]

In those autocorrelation functions, \(d_\Theta\) is denoting the distance measure on the observation space, which may be anisotropic or isotropic with regard to the dimensions of \(\Theta\). Following \([3]\), a problem dependent distance measure is possible, although the isotropic Euklidian norm

\[
d_{\text{Euklid}}(\theta_A, \theta_B) = \sqrt{\sum_i^n (\theta_{B,i} - \theta_{A,i})^2}
\]

is suitable in most cases of \(n\)-dimensional engineering tasks, e.g. used in SFEM. An anisotropic distance measure is introduced and applied in Section 4.2.

In numerical simulation of autocorrelation structures, all autocorrelation coefficients between all discretized points of observation are arranged to the autocorrelation matrix \(F\). In \([3]\), the autocovariance matrix \(K_{Kov}\) of a field \(H_{rf}\) is associated with this while

\[
K_{Kov;H_{rf}} = \sigma_{H_{rf}}^2 \cdot F_{H_{rf}}.
\]

It holds that

\[
\forall \theta_A; \theta_B \in \Theta : R(\theta_A; \theta_B) = R(\theta_B; \theta_A),
\]

because directional indepencency is required and, out of this, the autocorrelation as well as the covariance matrix remain symmetric.
Interdependencies between different random fields are called crosscorrelations. They are defined as simple crosscorrelation coefficients $c(H_{rf}; J_{rf})$ between two random fields $H_{rf}$ and $J_{rf}$. Similar to the autocorrelation coefficients, it holds $c \in [-1; 1]$. The crosscorrelations are taken into account in numerical simulation of multivariate random fields by merging the autocorrelation matrices of the underlaying random fields to form blocks on the main diagonal of a full correlation matrix. Then, the crosscorrelation coefficients are added to the off-diagonal parts. While a factorization of the full correlation matrix is needed to numerically generate realizations of the auto- and crosscorrelated random fields, this leads to very high numerical effort. According to [3], [4], [11] and [12], the Karhunen-Loève-expansion for discretized random fields is used. This is mainly depending on an eigenvalue/eigenvector decomposition of the full correlation matrix. In [4], a simplified approach with a satisfying reduction of the numerical problem is presented. Therein, the numerical problem of eigenvalue/eigenvector decomposition is decomposed in two parts. Instead of decomposing the full correlation matrix, two significant smaller decompositions of both the autocovariance and the crosscorrelation structure are carried out. By ordering and combining the possibly truncated series of eigenvalues and eigenvectors in appropriate block matrices, the numerical effort is substantially reduced. This method is used in Section 4.

4 Application for Uncertainty Modeling of Timber Structures

4.1 Structural Analysis of Timber

![Figure 6: Timber characteristics](image)

Timber is a natural building material. It is used natural raw, in contrast to materials such as concrete, steel or plastic. It is mainly characterized due to its macro-structure. According to [13], timber is a porous composite material of parallel single fibers with wood rays as matrix and both lignin and hemicellulose are binders. The tree trunk axis according to the parallel aligned fibers is called longitudinal direction $l$. The mechanical properties of wood are characterized by three main directions and a coordinate system in combination with the radial direction $r$ and the tangential direction $t$ forms the basis for the orthotropic mechanical behavior of wood. The three directions of local orthotropy are illustrated in Figure 6a).

According to Figure 6b), illustrating the shape of the annual rings (growth rings), the coordinate system equals a cylindrical polar coordinate system. By this, it becomes clear that the coordinates of orthotropy have to be evaluated separately at every observation point with regard to the tree trunk axis. This is called local orthotropy, see [13], [14] and [15].

The orientation and location of the tree trunk axis does not have to be parallel, neither centric, in relation to the timber component. If the tree trunk axis is located e.g. similar to Figure 6c), capable mechanical models and element formulations are needed in FEM.

An extensive approach in structural analysis of timber structures is described in [16]. The shown element formulation and constitutive laws allow the specific definition of the location of the tree trunk axis for every component of the FEM model. The local orthotropic material properties are evaluated at every GAUSS integration point with respect to the tree trunk axis location. This model facilitates hygro- and thermomechanical modeling and calculation of timber structures. Furthermore, nonlinear elastic and plastic behavior of large deformations is taken into account.
4.2 Uncertainty Model for Wood

Industrial materials require more or less complex production steps. The desired material structures are predefined and thus are well known. Because of the natural growth of trees, uncertainties exist regarding the actual structure and spatial varying material properties of the timber components, see [17].

Regarding [14], the material properties of wood can be described by random variables, taking into account their aleatoric characteristics. While modeling of material properties requires experimental data, it is shown e.g. in [1], [2] and [18], that stochastic uncertainty models itself do not provide sophisticated fitting.

Anyway, the prerequisites for pure stochastic modeling, which amongst others are massive amount of experimental data, identical testing samples, identical environmental conditions and, therefore, the ability of experimental reproduction, are in fact not given at any rate, especially for timber. This is also related to the fact, that nondestructive tests of material properties in multiple directions are not trivial and non standard.

A sophisticating uncertainty modeling can be achieved by including those lacks of knowledge. According to [1], [2] and [18], a fuzzy probability based random model is suggested. Therein, based on a comprehensive experimental database [19], a polymorphic uncertainty modeling is performed for material properties of spruce wood.

The geometries of timber components are as well affected by uncertainties. Especially for optimization of timber structures, it is meaningful to model essential design variables as uncertain design variables, e.g. if the cuttings of timber beams are not known to be absolutely precise. Natural defects in the geometry, like knots or a priori cracks, can be described uncertain as well, see e.g. [1] and [17].

If fuzzy probability based random fields, according to Sections 2.2 and 2.3, are utilized for polymorphic uncertainty modeling, a distance measure has to be chosen. Due to the anisotropic nature of timber, an anisotropic distance measure is assumed in the following. The structure of wood can be characterized by a cylindric polar coordinate system, see Section 4.1. The modified distance measure is evaluated in polar coordinates according to Figure 6c). In addition, specific weighting factors are involved for the three directions, which are

- \( m_z \) for the longitudinal direction, parallel to the tree trunk axis,
- \( m_\phi \) for the tangential direction, along the same annual ring, orthogonal to the tree trunk axis and
- \( m_r \) for the radial direction, from one annual ring to another, as well orthogonal to the tree trunk axis.

\[
d_{\text{mod}}(a, b) = \sqrt{[ (b_z - a_z) \cdot m_z]^2 + \left[ (b_\phi - a_\phi) \cdot \frac{a_r + b_r}{2} \cdot m_\phi \right]^2 + \left[ (b_z - a_z) \cdot m_z \right]^2}.
\]

It is assumed, that higher correlation exists in longitudinal direction, according to the membership to the same fiber, and in tangential direction, according to the membership to the same year growth conditions. On the contrary, the correlation in radial direction is assumed to be lower.

This modified distance measure, evaluated for two points of observation \( a \) and \( b \), is defined as

![Figure 7: Autocorrelation functions \( \kappa_a \) discretized on 4 integration points respectively for each element sized 10 x 10 on rectangular grids [0; 60] x [0; 100]](image)
The weighting factors \( m_z \), \( m_\phi \) and \( m_r \) artificially enlarge or shrink the distances in the associated directions. A high weighting factor results in lower correlation relationship for the associated direction.

Resulting autocorrelation functions, featuring the exponential kernel (see Section 3.2), are shown exemplary in Figure 7a) and Figure 7c) reduced to 2D grids and, therefore, evaluated for weighting for annual rings and fibers, respectively. The illustrations are according to the SFEM discretized at specific integration points. Other discretization of random fields is possible, regarding [4], e.g. the midpoint discretization.

Sample realizations of uncorrelated and autocorrelated random fields in midpoint discretization, featuring the Euclidean norm and the modified distance measure, are shown in Figure 8.

![Figures 8a-d: Realizations of standard normal distributed random fields on rectangular grids [0; 60] \times [0; 100] with \( z \) showing the realization of the random function on every pixel.](image)

With this modified distance measure, a spatially depending autocorrelation for polymorphic uncertainty models can be obtained. The quantification of the weighting factors is as difficult as the quantification of correlation lengths or autocorrelation kernels. Nondestructive test methods have to be used, to determine those properties for different directions and in a meaningful spatial scope. In [20] and [21], experimental approaches using the X-ray computed tomography method are presented. Especially the latter one focuses on the determination of density, which is an important property. Interdependencies between density and other material properties like stiffness and strength exist. An extensive knowledge about the spatial distribution of density would lead to reliable assumptions regarding the autocorrelation structure.

### 4.3 Example

A rafter tie beam carpentry joint, consisting of two timber beam parts, is modeled by an FEM simulation. The geometry, load and bearing conditions as well as the predefined tree trunk axes are illustrated in Figure 9. The geometry is assumed to be deterministic, see Table 1, and is chosen according to [22].

<table>
<thead>
<tr>
<th>length</th>
<th>( a )</th>
<th>( b )</th>
<th>( c )</th>
<th>( d )</th>
<th>( e )</th>
<th>( h )</th>
<th>( h_d )</th>
<th>( k )</th>
<th>( m )</th>
<th>( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>[mm]</td>
<td>550</td>
<td>100</td>
<td>200</td>
<td>250</td>
<td>40</td>
<td>120</td>
<td>120</td>
<td>75</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

The objective is the observation of the maximal rafter load \( S_d \) under uncertain material properties. The material model from [16], see Section 4.1, is used for the FEM calculation, which is the basic solution for...
the uncertainty analysis, regarding Section 2.3. The calculation is performed with the software FEAP and hygro- and thermomechanical processes are neglected, or rather a quasi static calculation is performed. Load stepping is applied and the failure condition is set to be the first load step that starts evolving the history variables of plastic strains, see [16].

![Diagram of rafter and tie beam carpentry joint](image)

Figure 9: Rafter and tie beam carpentry joint for the FEM model

A mesh with 1260 nodes on 696 elements is chosen. The elements for the timber parts are spatial eight-node with eight GAUSS integration points in each. The two separate material blocks of rafter and tie are connected by interface elements in the joint planes. The boundary conditions are chosen in such a way, that the beams are considered to be end parts of further continuing ones. The total load $S_d$ is equivalently distributed on the element nodes at the upper part of the rafter, where the inner nodes have doubled load with regard to the outer ones.

The material properties are described by parameters from the proposed polymorphic uncertainty models [1], [2]. The uncertainty model is expanded to random fields for two material properties of every beam. Due to the very high numerical effort, interval parameters are used instead of fuzzy parametrization and some of the material properties are depicted fully stochastically to avoid some samples in the fuzzy analysis. The parametrization is given in Table 2:

<table>
<thead>
<tr>
<th>material parameters for rafter and tie</th>
<th>probability distribution</th>
<th>distribution parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{cl, rafter}$, $f_{cl, tie}$ [N/mm^2]</td>
<td>log. normal</td>
<td>$E_u = 3.7737$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_u = 0.0491$</td>
</tr>
<tr>
<td>$f_{cq, rafter}$, $f_{cq, tie}$ [N/mm^2]</td>
<td>GUMBEL</td>
<td>$a = 2.9687$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$b = 5.9334$</td>
</tr>
<tr>
<td>$f_{t1, rafter}$, $f_{t1, tie}$ [N/mm^2]</td>
<td>normal</td>
<td>$E = 121.64$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S = 18.2$</td>
</tr>
<tr>
<td>$f_{cq, rafter}$, $f_{cq, tie}$ [N/mm^2]</td>
<td>WEIBULL</td>
<td>$\theta = 2.7771$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$k = 9.7439$</td>
</tr>
<tr>
<td>$f_{v, rafter}$, $f_{v, tie}$ [N/mm^2]</td>
<td>log. normal</td>
<td>$E_u = [1.7209; 1.7634]$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_u = [0.1095; 0.1406]$</td>
</tr>
</tbody>
</table>

From the sensitivity analysis in [18], it is known that the shear strengths $f_{c, rafter}$, $f_{c, tie}$ and the radial/tangential compression strengths $f_{cq, rafter}$, $f_{cq, tie}$ are highly influencing the results of an analytical solution for the rafter tie beam joint. Therefore, those two strengths are identified as random fields for each beam, respectively. The correlation lengths $l_{\theta, rafter}$, $l_{\theta, tie}$ for the autocorrelation of the fields are assumed to be interval variables. They are listed in Table 3, together with the LAMÉ parameters, which are assumed with common values for spruce wood.

A crosscorrelation is assumed between the two fields with the crosscorrelation coefficient

$$c(f_{v, rafter}, f_{cq, tie}) = 0.5$$ (33)
and, for the autocorrelation, the modified distance measure, according to Section 4.2, is used with a squared exponential autocovariance kernel (see Section 3.2) and the three weighting factors are assumed to be

\[ m_r = 20.0, \quad m_\phi = 1.0, \quad m_z = 1.0. \]

Due to the requested non-Gaussian marginals, a correction factor for all correlation coefficients is advised by [23], called the Nataf transformation. Due to the relative similarity of the corrected correlation coefficients to the original ones for most marginals, see [4], the Nataf transformation is neglected in this first approach and the correction factor is assumed to be \( \kappa \approx 1 \) in all cases.

Table 3: Lamé parameters and interval parametrization of correlation lengths for spruce wood

<table>
<thead>
<tr>
<th>description</th>
<th>parameter for rafter and tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Lamé-parameter</td>
<td></td>
</tr>
<tr>
<td>( \lambda_q )</td>
<td>43.58 ([N/mm^2])</td>
</tr>
<tr>
<td>( \lambda_l )</td>
<td>10,000 ([N/mm^2])</td>
</tr>
<tr>
<td>2. Lamé-parameter</td>
<td></td>
</tr>
<tr>
<td>( G_q )</td>
<td>51.16 ([N/mm^2])</td>
</tr>
<tr>
<td>( G_l )</td>
<td>650 ([N/mm^2])</td>
</tr>
<tr>
<td>correlation length</td>
<td>( l_{\Theta,\text{rafter}}; l_{\Theta,\text{tie}} = [50; 200] ) ([mm])</td>
</tr>
</tbody>
</table>

Fuzzy interactions according to Section 3.1 are considered exemplary between the correlation lengths of the fields. The strength parameters share the same correlation length parameter for the two fields. An interdependency is assumed, that combinations of very high correlation lengths in the one and very low correlation lengths in the other random field are excluded. The interaction between \( l_{\Theta,\text{rafter}} \) and \( l_{\Theta,\text{tie}} \) is defined by two linear functions. With regard to the definition in Figure 5a), the setup constraints are illustrated in Figure 10.

From preliminary calculations, it is known, that the maximal load, leading to plastic deformations, will almost never exceed 20 kN. While the discretization of the result is directly depending on the discretization of load steps, a specific load function is applied, which is shown in Figure 11a).

For the expansion of the random fields, eigenvalue decompositions of the autocovariance structure are required for every change in the correlation length, which means for every fuzzy sample. The eigenvalue decomposition is performed with LAPACK routines for all or several eigenvalues. The plot in Figure 11b) shows results of preliminary computations of eigenvalue decompositions for the tie random field (4800 observation points) on the maximum and minimum level of correlation length. By this, a linear function for the amount of calculated eigenvalues is chosen for final calculation depending on the correlation length. The quality of field approximation is calculated according to [4].

To achieve a quality of \( q > 0.95 \) in every case of random fields for the tie, 240 eigenvalues are computed for \( l_{\Theta} = 200 \) mm, 1440 eigenvalues are computed for \( l_{\Theta} = 50 \) mm and linearly interpolated in between. For the random fields in rafter (567 observation points), all eigenvalues are computed for all correlation lengths in less than 2 seconds and, therefore, are considered totally.

The uncertainty analysis is parallelized on the level of the stochastic analysis. Therefore, up to 16 basic solutions (FEM calculations) are carried out simultaneously. The stochastic analysis is performed
with a Monte-Carlo-simulation using 560 samples in each. Niederreiter quasi-random numbers are used for random variables and pseudo-random numbers for the random fields involving new time dependent seeds for every realization.

The fuzzy analysis is featuring a modified evolutionary algorithm for the \( \alpha \)-level optimization. For this, three chains each with three children are considered in up to six evolution steps (generations). Since only interval parameters are considered, this reduces to two \( \alpha \)-level optimizations in each fuzzy analysis, one for minimum and one for maximum response. The optimization is only performed for the median quantile, and the other quantiles are assumed to behave similarly. All over all, this results in up to 180 fuzzy samples in every fuzzy analysis and sums up to totally 140 400 samples of basic solution for the complete uncertainty analysis.

In Figure 12, the calculated material parameters are visualized for one sample realization with high correlation lengths in the random fields. The annual ring and fiber characteristics are apparent therein.

Sample results in failure load step are illustrated in Figure 13 with failure from plastic deformation in \( xz \)-shear on the downside of the tie, just between the joist hanger and the joint area.

The final result of the uncertainty analysis is shown in Figure 14. Due to the very low amount of samples in the stochastic analysis, only quantiles in the interval \( \hat{F} \in [0.125; 0.875] \) are plotted. The optimization results in interval valued quantiles of an empirical probability distribution function. Vertical cuts in this plot at a given fractile can also be interpreted as interval valued probabilities of failure.

In Figure 14, the epistemic and the aleatoric characteristics are visible. The first one results from the interval valued parameters, reduced by considered interactions. The latter one follows from the random variables and random fields, reduced by the considered auto- and crosscorrelations.
Figure 12: Sample random field realizations with Gauss-point values projected to element nodes

Figure 13: Sample basic FEM solution in failure load step

Figure 14: Interval valued quantiles of the resulting empirical probability distribution function
5 Conclusions and Outlook

In simulation of natural materials, such as timber, multiple uncertainty models can be used standalone or combined. Uncertainty models involving both, epistemic and aleatoric uncertainties are presented. Multiple origins cause uncertainties in wooden materials. On the one hand, the growth of trees is inducing natural variations for timber components. On the other hand, facts as the irreproducibility of experiments and missing nondestructive testing methods lead to additional lack of knowledge about the material properties of wood. Furthermore, uncertain geometries, due to nature or manufacturing can be considered. For those reasons, polymorphic uncertainty models are introduced, treating aleatoric as well as epistemic uncertainties.

The spatial variation of material properties is incorporated into the uncertainty model by using spatial functions. The polymorphic uncertainty model of fuzzy probability based random fields is used as approach. The characterization is based on a fuzzy valued parametrization for the underlying probability functions and properties of the random fields.

The aim of including additional information by incorporating interdependencies and, thereby, reducing the uncertainty is shown. Interdependencies within fuzzy variables are called fuzzy interactions and interdependencies within stochastic uncertainty models are provided by correlation, which in detail leads to auto- and crosscorrelations.

Exclusion of devious fluctuations in spatial relationships of low distance is a further consequence of consideration of autocorrelations. Common autocorrelation functions as well as underlying distance measures are presented. A modified distance measure in polar coordinates with weighting factors for the three directions of local orthotopy is introduced. Fitted to the characteristic macrostructure of wood, the distance measure allows a structurally depending autocovariance. Interdependencies between several random fields are considered as crosscorrelations, which are handled as an extension of the autocorrelation matrices of single random fields to a multivariate full correlation matrix. Methods for reducing the numerical effort in conjunction with the eigenvalue decomposition of the full correlation matrix for expansion of multivariate random fields are discussed, mainly referring to [4].

A method for fuzzy interactions as simple constraints in the fuzzy sampling domain is presented according to [10]. This allows interdependencies between fuzzy variables in common frameworks for fuzzy analysis with α-level optimization approaches, according to [24].

To show the capabilities of this extensive spatially depending polymorphic uncertainty model, an uncertainty analysis for an exemplary timber structure is carried out. The uncertainty model is based on data from [1] and the basic solution is an FEM calculation with significant numerical effort. A framework of parallelized analysis, performing multiple basic solutions simultaneously, is modified and used for this. Results for a rafter tie beam joint equipped with a partly reduced uncertainty model are presented. The structure of uncertainty analysis and intermediate steps are depicted.

Figure 15: Structure chart for the expansion of the uncertainty analysis with a multi-objective optimization loop

In summary, a comprehensive polymorphic uncertainty model for timber structures is introduced and an uncertainty analysis of it is presented. The aim of the related research project is the optimization of timber structures described by the shown uncertainty models. Due to the fully automated uncertainty analysis, the preliminary steps for an extension to optimization of timber structures are given. Figure 15 shows the structure chart, already mentioned in Figure 2b) and Figure 3. It is expanded with two elements, which are object of further research. This is on the one hand uncertainty reduction in the sense of representative uncertainty reducing measures. They are needed to provide results of each uncertainty analysis, that are comparable in fitness, for the additional optimization loop on top on the other hand. This results in optimization with uncertain design variables and a priori parameters and is developed.
according to the approaches to optimization with uncertainties of [3] and [25]. The presented uncertainty modeling and structure of analysis can be applied to arbitrary timber structures and several objectives of optimization can be examined.

Acknowledgments
The demonstrated results and research activities arise from the subproject “Multi-objective optimization of wooden structures with polymorphic uncertain parameters” GR 1504/10 and KA 1163/36 as a part of the priority program SPP 1886. The funding by the German Research Foundation (DFG) is gratefully acknowledged.
References
Product reliability optimization under plate sheet forming process variability

M. Faes1* B. Van Doninck,2 M. Imholz,1 and D. Moens1
1KU Leuven, Department of Mechanical Engineering, Belgium
2Flanders Make, Belgium
*Corresponding author: matthias.faes[at]kuleuven.be

Abstract
This paper presents an approach to perform product reliability optimization taking uncertainty that stems from the production process explicitly into account. An integrated workflow combining production process simulation with a structural model of the resulting component is presented. Then, based on probabilistic and interval models of an uncertain plate thickness, which were constructed using direct measurements, a design optimization is performed. Specifically, reliability based design optimization and interval based design optimization are applied and critically compared. It is shown that both methods deliver similar safety margins on the resulting component, although that the probabilistic method is more conservative. Furthermore, it is shown that in this specific case, interval based design optimization is an order of magnitude more efficient.

Keywords: RBDO, Integrated process-product analysis, interval analysis

1 Introduction
In order to ensure that a structure is both reliable and not over-conservatively designed, the non-determinism that is attributed to different parameters of the numerical design model can readily be taken into account by applying non-deterministic numerical methods and uncertainty quantification (UQ). Two complementary philosophies for UQ exist to date: the probabilistic [1] and the possibilistic approach [2]. Following a probabilistic approach, the non-determinism in the model parameters is modeled as a (joint-) probability density function (PDF), which is propagated through the numerical model in order to infer the likelihood of obtaining a certain response value. From this quantified uncertainty in the model responses, for example the robustness and reliability of the structure can be computed, and hence used as objective function in an optimization problem. Possibilistic approaches on the other hand, such as interval methods, only consider crisp parameter values to bound the range of possible parameter values. These bounds are propagated through a numerical model to infer the worst-case structural responses of the design. These methods typically require less data to provide an analyst with an objective estimation of the uncertainty in the model responses. Also, generalized hybrid methods have been recently presented [3], which aim at combining the strengths of both frameworks. The truthfulness of the non-deterministic models can be enhanced following inverse uncertainty quantification approaches where the non-deterministic quantities are calibrated based on measurement data [4–7].

Very often, the root cause of model non-determinism lies in inherent variability during the production process. For instance, literature shows that a large part of the variability in the mechanical performance of parts produced via Additive Manufacturing can be traced back directly to the production process [8, 9]. Similarly, also in more conventional production techniques such as sheet metal forming, a non-negligible part of mechanical performance variability can be traced back to the production process. As such, when optimizing the design and reliability of functional end-use components, the effect this process variability should be accounted for. Commonly this is performed implicitly by modeling the resulting non-determinism at the component level. In contrast hereto, this paper presents a coupled process-product simulation approach for the reliable design of sheet metal formed products that takes process variability explicitly into account. Hereto, starting from initial process parameters, the dimensional accuracy of a sheet metal formed product is predicted using a data-driven process model. Starting from this geometry, a functional performance evaluation of the product is executed using a finite element model of its dynamic behavior. Then, since data to objectively quantify model non-determinism are in practice usually very scarce, the impact of non-deterministic model choices on the optimized product is studied. The paper is structured as follows. Section 2 recalls some aspects of probabilistic and interval design optimization. Section 3 presents the case study and discusses the application of both concepts to the closed-loop process-product model. Finally, the conclusions of this paper are presented in section 4
2 Non-deterministic product analysis

To perform a virtual structural design optimization on a component level, the set of governing partial differential equations that describe the structural performance is usually approximated by means of a numerical model \( \mathcal{M}(\mathbf{x}) \). This model is parametrized by a parameter vector \( \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^k \) with \( \mathcal{X} \) the set of physically admissible parameters and \( k \in \mathbb{N} \). In case \( \mathcal{M}(\mathbf{x}) \) is constructed following a finite element approach, the model domain \( \Omega \) is discretized by means of a set of linear or quadratic finite elements, yielding \( d \) structural degrees of freedom (DOF).

The model \( \mathcal{M}(\mathbf{x}) \) provides a vector of model responses \( \mathbf{y} \in \mathcal{Y} \subset \mathbb{R}^l \), with \( \mathcal{Y} \) the set of admissible model responses and \( l \in \mathbb{N} \), through a set of function operators \( m_i, i = 1, \ldots, l \), which are defined as:

\[
\mathcal{M}(\mathbf{x}) : y_i = m_i(\mathbf{x}) \quad i = 1, \ldots, l
\]

with \( m_i : \mathbb{R}^k \mapsto \mathbb{R} \). In case \( \mathbf{y} \) consists of nodal responses, the number of responses usually matches the number of degrees of freedom (i.e., \( l = d \)). In the context of incorporating non-determinism into \( \mathcal{M}(\mathbf{x}) \), generally two strategies are commonly applied:

**Probabilistic approach:** Variable model parameters \( \mathbf{x} \) are in a probabilistic context modeled as a random variable \( \mathbf{x}(\theta) \). Each such a random variable \( \mathbf{x}(\theta) \) provides a mapping \( \mathbf{x}(\theta) : (\Theta, \sigma, \mathbb{P}) \mapsto \mathbb{R} \) with \( \theta \in \Theta \) a coordinate in sample space \( \Theta \) and \( \sigma \) the sigma-algebra.

**Interval approach:** The uncertainty on the model parameters \( \theta \) are represented as intervals \( \mathbf{x}(\tau) \). Interval parameters are either represented using the bounds of the interval \( \mathbf{x}(\tau) = [\mathbf{x}\_\text{min}, \mathbf{x}\_\text{max}] \) or the centre point \( \bar{x} = \frac{\mathbf{x}\_\text{max} + \mathbf{x}\_\text{min}}{2} \) and the interval radius \( r_x = \frac{\mathbf{x}\_\text{max} - \mathbf{x}\_\text{min}}{2} \cdot \mathbb{R} \). The set of real interval-valued scalars is denoted \( \mathbb{IR} \).

In the context of design optimization, \( \mathbf{x} \) is usually split into a set of design variables \( \mathbf{x}\_d \in \mathbb{R}^{k_d} \) and non-design variables \( \mathbf{x}\_\text{nd} \in \mathbb{R}^{k_{\text{nd}}} \). This selection is usually made pragmatically, considering which quantities lie in the control of the designer (e.g., plate thickness) and which have to be taken as is. Note that non-determinism enters the model through both sets of parameters, and as such should be accounted for during the design optimization. Similarly, \( \mathbf{y} \) is split into a set of objective variables \( \mathbf{y}\_o \in \mathbb{R}^{l_o} \) and performance variables \( \mathbf{y}\_p \in \mathbb{R}^{l_p} \). Examples of the former include for instance the total mass or cost of the component, whereas examples of the latter include stresses or deflections which should be limited to ensure structural performance.

### 2.1 Reliability based design optimization

In a probabilistic context, the optimization of the design under parameter non-determinism is usually performed using the framework of reliability based design optimization (RBDO). The superscript \( \pi \) is in the following discussion omitted for notational clarity, since all variables are considered to be modeled as a random variable. In the context of RBDO, following optimization problem is considered:

\[
\begin{align*}
\text{minimize} & \quad \mathbf{y}(\mathbf{x}\_d, \mathbf{x}\_\text{nd}) \\
\text{subject to} & \quad P_f(\mathbf{y}(\mathbf{x}\_d, \mathbf{x}\_\text{nd})) < c
\end{align*}
\]

with \( b_i \) deterministic bounds on certain performance variables and \( P_f \) the probability of failure with upper bound \( c \), and which is defined as:

\[
P_f = \int_{\mathbb{R}^{l_o}} 1_{\mathcal{X}}(\mathbf{y}) f_{\mathbf{y}}(\mathbf{y}) d\mathbf{y}
\]

Herein, \( f_{\mathbf{y}}(\mathbf{y}) \) is the probability density function of \( \mathbf{y} \) and \( 1_{\mathcal{X}}(\mathbf{y}) \) is the indicator function, which is defined as:

\[
1_{\mathcal{X}} = \begin{cases} 
0 & \text{if } \mathbf{y} \in \{ \mathbf{y} \mid \mathbf{y} = m(\mathbf{x}), \mathbf{x} \in \mathcal{X}, g(\mathbf{y}) > 0 \} \\
1 & \text{if } \mathbf{y} \in \{ \mathbf{y} \mid \mathbf{y} = m(\mathbf{x}), \mathbf{x} \in \mathcal{X}, g(\mathbf{y}) \leq 0 \}
\end{cases}
\]

with \( g : \mathbb{R}^{l_o} \mapsto \mathbb{R} \) the performance function of the design. In practice, the multidimensional integral in eq. (3) is intractable to solve either analytically or using numerical integration [10]. Therefore, research in the last decades presented various approximative methods, ranging from Monte Carlo integration and highly efficient moment-based approaches [11] to advanced techniques such as Subset simulation [12]. Of key importance herein is the efficiency of the determination of the probability of failure, since this quantity has to be determined in each iteration of the algorithm that solves eq. (2).
2.2 Interval design optimization

In an interval context, a similar optimization problem can be defined as in the probabilistic case. However, since no relative likelihood of a parameter value between the interval boundaries is inferred, the concept of failure probability is not applicable. Therefore, the optimization problem formulated in eq. (2) is reformulated as:

\[
\begin{align*}
\text{minimize} & \quad y^d(x^d, x^{nd}) \\
\text{subject to} & \quad y^d(x^d, x^{nd}) > b \\
& \quad \overline{y}^d(x^d, x^{nd}) < c
\end{align*}
\]

(5)

with \(b\) and \(c\) possible upper and lower bounds on the performance variables. Note that herein the superscript \(\cdot I\) is omitted for the sake of notational clarity since all considered parameters are interval-valued. This problem needs to be solved in a nested double loop optimization approach:

**Inner loop:** find the solution of the interval finite element model, where the bounds on \(y\) are found as:

\[
\begin{align*}
\underline{y}_i &= \inf_{x^d} m_i(x^d, x^{nd}) & i = 1, \ldots, l \\
\overline{y}_i &= \sup_{x^d} m_i(x^d, x^{nd}) & i = 1, \ldots, l
\end{align*}
\]

(6)

In the case when all \(m_i\) are strictly monotonic, this optimization problem can be replaced by evaluating the vertices of the hyper-cubic set \(x\). As such, a deterministic amount of \(2^k\) deterministic model evaluations are needed. Otherwise, dedicated optimization techniques such as discussed in [2] should be applied.

**Outer loop:** solve the optimization problem formulated in eq. (5). Since this problem is usually sufficiently smooth, often gradient based optimizers, such as for example based on the BFGS representation of the Hessian, can be applied. This is obviously not the case for heavily bifurcated or complicated sets of equations.

3 Case Study

3.1 Case introduction

In this study, a deep drawn cup (Erichsen cup) is chosen as the use case because it is a generic sheet formed product, with a simple geometry, that has industrial relevance. A picture of a cup can be seen in figure 1. For the deep drawn cup, typical parameters to take into account in the design process are the initial sheet thickness, the material type and process parameters such as blank holder force and drawing depth, as well as the constitutive material properties in the elastic and plastic region of the constitutive material model. Even with this simple, one-step forming process, the influence of variability on these parameters on the final product performance variability is significant. Without loss of generality, the study focuses on the dynamic behavior of the cup for illustrative reasons.

![Figure 1: A deep drawn cup (Erichsen cup)](image)

3.2 Integrated process-product simulation

Before the methods described in section 2 can be applied on this use case, the process and product simulations need to be coupled. Therefore we use Noesis Optimus [13], a process integration and design optimization software platform. The process simulation is a data driven model which estimates the material thinning due to the forming process. The data driven model is trained using the random forest
modeling approach on the available data: material and process parameters as inputs and the respective thickness distribution measurements as output for a set of 100 cups. The model outputs averaged thinning information for five zones defined within the product. The five zones are shown in figure 2-b. The product simulation is a finite element analysis (FEA) for the dynamic behavior of the cup, which outputs the mass and the eigenfrequencies of the cup.

Connecting the process model with the product simulation requires a number of steps. First, a nominal product shape is selected as a reference, i.e. the designed shape (CAD) or a measured product shape (figure 2-a). Second, the nominal shape is divided into zones based on a set of criteria, i.e. radius and height (figure 2-b). Next, the average thinning in each zone, as predicted by the data driven process model based on the material and process parameters, is applied to each zone. This results in a product with a discrete thickness distribution (figure 2-c). In the last step, morphing techniques can be used to take into account the geometrical process results (figure 2-d). This can range from scaling the cup to adjust for the drawing depth to morphing the flange to account for wrinkling and orthotropic material effects, depending on the specific use case. For the use case of the deep drawn cup, the wrinkling effect on the flange is captured by fitting a model through the flange outer boundaries of the 100 cups as derived from the FEM simulation models. Based on the process parameters and the fitted model, the cup flange of the axisymmetric model is adapted to better represent the actual product.

Figure 2: Connecting process model with the product model: (a) Nominal shape; (b) Divided into thickness zones; (c) Mapping of the process results; (d) Morphing of the geometry

Figure 3 shows the implementation of the integrated process-product simulation in Noesis Optimus. Based on the selected input parameters, the data driven model calculates the averaged thinning, which is then applied to the divided model with five zones. The number of zones is determined by gradually increasing the number of zones in the simulation models, starting from one, until sufficient accuracy was reached compared to the reference simulation models with detailed thickness information. No significant increase of accuracy was noticed when using more than 5 zones. Once the thinning is added to the product and the geometry is adjusted, a FE model is constructed and the modal analysis is started. When finished, the mass and the eigenfrequencies are automatically extracted from the result files. The advantage of having an integrated process and product simulation is that it can be used in an optimization to find the required process parameters (material type, initial plate thickness and machining settings) to produce the final product which will have the desired/optimal behavior. With the current state of practice, an optimization for a certain product attribute will result in an optimized geometry that has a uniform thickness. This poses two problems to be solved by the designer:

1. It is not guaranteed that this optimized geometry can actually be produced, and if so, what process parameters are needed to produce it.
2. A formed product with a uniform thickness is impossible to produce and therefore the designer has to take into account a large safety factor on the thickness.

3.3 Process parameter optimization

This section illustrates the application of RBDO and IBDO on the sheet metal formed cup. The goal of the optimization study is to minimize the total mass of the cup. As additional design constraint, the first eigenfrequency should be greater than 3500 Hz. The input for the design study is the initial plate thickness before sheet metal forming, which is not deterministically known. The selection for the initial plate thickness as optimization parameter is based on the computation of the first order Sobol indices of the integrated workflow. The design optimization under this uncertain parameter is conducted following a probabilistic (RBDO) and interval approach (IBDO) and the results are critically compared.
The uncertainty models of the initial plate thickness are based on thickness measurements on 12 replica of the sheet metal formed parts.

### 3.3.1 Reliability based design optimization

The optimization problem in a probabilistic context is given in eq. (2). In this specific case, the first eigenfrequency of the design should lie above 3500 Hz with a probability of 99.7% (i.e., $\beta > 3$). For computational reasons, and due to the relative simplicity (i.e., linear) of the failure domain, the probability of failure (i.e. $P(f_1 < 3500)$) is computed via the First Order Reliability Method (FORM) approximation. The input parameters of the optimization problem are listed in table 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Distribution</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{\text{init}}$</td>
<td>Initial plate thickness</td>
<td>Normal $N(0.7734, 0.012)$ mm</td>
<td></td>
</tr>
<tr>
<td>$F$</td>
<td>Clamp holding force</td>
<td>deterministic</td>
<td>45 N</td>
</tr>
<tr>
<td>$d$</td>
<td>Deep draw depth</td>
<td>deterministic</td>
<td>31.4 mm</td>
</tr>
</tbody>
</table>

Eq. (2) is solved using a sequential quadratic programming approach with a BFGS representation of the Hessian. The reliable optimum was obtained after three iterations of the optimization solver, where the FORM approximation of $P_f$ is computed during each iteration. As such, a total of 78 deterministic model evaluations are needed. Due to the non-determinism in the initial plate thickness, the designed thickness of the plate was increased from 0.773 mm to 0.926 mm to ensure product reliability. The variance on the model responses corresponding to the reliable optimum is obtained by means of Monte Carlo simulation with 1000 samples on the optimized structure, and serve as additional validation of the product.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Mean Value</th>
<th>Var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$</td>
<td>Mass</td>
<td>65.8 g</td>
<td>0.824 g</td>
</tr>
<tr>
<td>$t$</td>
<td>Plate thickness</td>
<td>0.927 mm</td>
<td>0.011 mm</td>
</tr>
<tr>
<td>$f_1$</td>
<td>Eigenfrequency 1</td>
<td>3613.0 Hz</td>
<td>37.6 Hz</td>
</tr>
</tbody>
</table>

### 3.3.2 Interval design optimization

The optimization problem in an interval context is given in eq. (5). In this specific case, the lower bound of the first eigenfrequency of the design should lie above 3500 Hz. Since the rate of change of eigenvalues of a linear system is constant with respect to the input parameters of that system, the model is considered to be perfectly monotonous. Therefore, the vertex (or Transformation) method [14] guarantees the exact bounds of the interval valued problem. Therefore, the inner optimization loop is for this specific problem replaced by 2 deterministic solves corresponding to the bounds of the input intervals. The interval on the initial plate thickness is set as $t_{\text{init}} = [0.754; 0.794]$ mm. The midpoint of this interval $\hat{t}$ is used as input for the optimization problem, and it is furthermore assumed that the range of the interval uncertainty $r_t$
remains constant throughout the design optimization, since it stems directly from the non-determinism in the initial plate thickness. A total of six deterministic model evaluations were needed to obtain the interval optimum. The results of the IBDO are listed in table 3.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Lower bound</th>
<th>Upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$</td>
<td>Mass</td>
<td>63.47 g</td>
<td>66.35 g</td>
</tr>
<tr>
<td>$t$</td>
<td>Plate thickness</td>
<td>0.893 mm</td>
<td>0.934 mm</td>
</tr>
<tr>
<td>$f_1$</td>
<td>Eigenfrequency 1</td>
<td>3500 Hz</td>
<td>3633 Hz</td>
</tr>
</tbody>
</table>

3.3.3 Comparison

Considering the results of both approaches presented, it is clear that the computational cost of performing a RBDO is considerably higher as compared to an IBDO for this case. Note that when non-monotonic models are considered in an IBDO context, or more accurate methods for the computation of the failure probability are applied in an RBDO context (such as e.g., SubSet for highly non-linear limit state functions), this conclusion may no longer hold, and the computational advantage might be shifted more or less towards one of the two approaches.

From these tables, it is also clear that although both approaches are based on philosophically different foundations, the obtained results in terms of the safety margin on the first eigenfrequency are comparable. Indeed, when comparing the interval boundaries obtained by IBDO with the $3 \cdot \sigma$ bounds of the RBDO result, it is clear that they are approximately similar. This is a direct result from the fact that the constraints in both optimization problems are modeling the same physical constraint (i.e., maintaining a crisp safety margin on the first eigenfrequency), which is as such taken into account by the optimization procedure.

However, as is visualized in figure 4, the estimate on the first eigenfrequency of the RBDO approach is a lot more conservative as compared to the fitted interval. Consequently, also the optimized initial plate thickness and the mass of the structure, are higher as compared to the interval approach. This is not a result of the optimization method in se, but is rather due to the assumptions that were made when estimating the non-determinism in the initial plate thickness. Namely, Gaussianity in the probabilistic case, and that the uncertainty is captured by the 12 considered replica in the interval case. As such, both results should also be interpreted with care.

![Figure 4: RBDO and IBDO optimum of the first eigenfrequency (left) and the corresponding mass (right)](image-url)
4 Conclusions
This paper presents an integrated work-flow coupling a production process model with a structural mechanical analysis. This approach is applied to the reliable design of a steel sheet metal formed cup by optimizing the production process. Hereto, both interval and probabilistic approaches are applied. Starting from the same data set, a non-deterministic model is constructed for both approaches, and the initial plate thickness is optimized following both a reliability based design optimization (RBDO) and interval based design optimization (IBDO) approach. The results of both approaches are compared. The conclusions of this paper can be summarized as follows:

- it is possible to infer structural reliability of sheet metal formed products, based on a coupled work-flow containing both production process and structural mechanical analysis.
- RBDO and IBDO give similar results as concerns the obtained safety margin on the structural performance, since the employed objective functions represent similar physical bounds on the structural problem.
- for this specific case, the interval approach significantly outperforms the probabilistic method in terms of computational efficiency.
- the optimized structure relies heavily on the quantification of the non-determinism in the uncertain model quantities. In case insufficient data are available for a probabilistic analysis, it is shown that the results become slightly over-conservative with respect to the interval approach.
- care should be taken in the interpretation of the results, as both techniques come with assumptions. In the probabilistic case, the validity of the Gaussianity-assumption can be validated by testing more distributions. The interval results can be made more robust by for example inflating the identified bounds on the initial plate thickness using Chebyshev’s inequality theorem.

Acknowledgements
The authors want to acknowledge Flanders Make and the Flanders agency for Innovation & Entrepreneurship (VLAIO) for their support in the framework of the ICON project ‘ViDesPro’ (Virtual Design Platform for Sheet Material Products’). The authors would furthermore like to acknowledge the financial support of the Flemish Research Foundation (FWO) in the framework of the project ‘HiDIF: High Dimensional Interval Fields’ (project number G0C2218N).

References


Risk analysis of infinite slope failure using advanced Bayesian networks

Longxue He,1 António Topa Gomes,2 Matteo Broggi,1 and Michael Beer1,3,4*

1 Institute for Risk and Reliability, Leibniz Universität Hannover, Germany
2 CONSTRUCT, Faculty of Engineering, University of Porto, Portugal
3 Institute for Risk and Uncertainty, University of Liverpool, UK
4 CERSM, Tongji University, China
* Corresponding author: beer@irz.uni-hannover.de

Abstract

Slope stability is affected by multiple factors. To assess the risk of slope failure, it is critical to identify the key influence factors and accurately predict the probability of slope failure. For that purpose, Bayesian networks (BNs) are an effective tool. This paper presents the application of advanced BNs to evaluate the geotechnical risk of slopes. The first part of the paper presents the theoretical background regarding BNs. It includes the simple BNs, and explains further developments. Particularly, enhanced Bayesian networks, able to cope with continuous input parameters, and Credal networks, specially used for incomplete input information, are presented and discussed. The second part of the paper deals with two geotechnical examples implemented to demonstrate the feasibility and predictive effectiveness of BNs. The developments regarding the use of BNs in geotechnical engineering are presented. The ability of BNs to deal with the slope stability risk is discussed as well. The paper also evaluates the influence of several geotechnical parameters. Besides, it discusses how the different types of BNs contribute to assessing the reliability and risk of real slopes, and how new information could be introduced in the analysis.

Keywords: Risk analysis, Infinite slope, enhanced Bayesian networks, Credal networks

1 Introduction

The occurrence of slope failure has the catastrophic potential for causing landslides such as debris flows and rock falls, which threatens the human life and property, attracting public attention and requiring research in geotechnical engineering [1, 2]. Various attempts have been carried out in literature to analyze the stability of slopes and estimate the probability of slope failure. In these approaches, infinite slope model is widely utilized to study shallow landslides under the varying position of the groundwater table, random variations of soil parameters and so on [3–5].

The probabilistic method has played an important role in the estimation of the probability of failure of slopes since the very early studies [6–10] due to the unavoidable uncertainties existing in vague environmental condition, varying soil properties as well as insufficient information affecting the slope failure.

Numerous numerical methods have been applied in slope stability analysis. For instance, slope stability problems associated with structural reliability methods (SRMs) have been conducted by means of first-order reliability methods (FORMs) [8] and simulation approaches, such as Monte Carlo Simulation [11], Importance Sampling [12] and Subset sampling [13], etc. These studies demonstrated the feasibility of structural reliability analysis for calculating the probability of slope failure in geotechnical engineering. However, the interactive influences of the multiple parameters in the analysis of a slope failure cannot be identified and further singled out the critical factors with these approaches.

Artificial neural networks also have been adopted to predict slope stability of slopes with the geometric or geological data of the influential factors [14, 15]. This approach is not good at quantifying the uncertainty and characterizing the impact of individual risk factors on the slope stability using information updating.

As is known, the identification of the most important factors leading to failure and the estimation of the probability of slope failure can prevent potential geological disasters. A robust model presented in recent researches, Bayesian networks (BNs) have been increasingly employed in slope stability analysis [16–20]. BNs, as the causal probabilistic models, have been developed and successfully applied to natural hazards, safety, and reliability engineering for over two decades since their first introduction by Pearl [21]. Compared to the aforementioned numerical tools, BNs carries huge advantages over other available methods to calculate the probability of slope failure. In particular, they show the following advantages.
• Simple graphical visualization. The failure of a slope can be affected by geo-environmental parameters, weather condition, natural hazards (e.g. earthquakes and storm) as well as human activities. The combination of these factors can negatively impact on slope stability. BNs can not only integrate these elements into a rigorous framework but provide a visual cause-effect relation among events in a graphical model. In particular, BNs help decision makers and even non-expert without a strong background in geotechnical engineering to gain a good understanding of the failure mechanisms. For a detailed overview on how to construct a graphical framework for risk assessment of rock-fall hazard with a BN model is given by [22].

• Uncertainty quantification. BNs are developed successfully to capture the uncertainties affecting the problem and benefit from the capability of the forward and backward propagation of probabilities according to the axioms of Bayesian probability theory [23].

• Information update from new observation. Updating of the event probabilities in BNs can be efficiently performed in near-real-time by mean of Bayesian updating to respect the information carried by the new observation. Thanks to this, the BNs model can provide the decision makers with up-to-date information on the slope failure mechanisms as soon as new evidence is presented.

In previous studies, traditional BNs (i.e., only discrete probability values and binary event are considered) have been applied to analyse slope stability. However, the slope stability problem is clearly influenced by both discrete events and continuous variables, thus it is impractical to obtain discrete probabilities of all the factors affecting a slope. Thanks to the development of enhanced Bayesian networks (eBNs), proposed by Straub & Der Kiureghian [24], it is possible to deal also with a causal relationship to continuous quantities in a BN framework.

An additional limitation of BNs is that the employed probabilistic models are precise, hindering the application in geotechnical engineering where the available information is often scarce. Thus, Credal networks (CNs), an extension of BNs to take imprecise probabilities into account, have been introduced in the geotechnical problem to assess debris flow hazard [25]. Nonetheless, no detailed investigations are attempted to apply the advanced model into the analysis of slope stability.

In this paper, the study aims at presenting the graphical models of the slope to estimate the failure probability, in which both of eBNs and CNs implemented are based on the infinite slope, and then new observations are inserted to update the model in order to identify the effect of factors on slope stability.

2 Methodology

2.1 Bayesian Networks

Bayesian networks, also known as Bayesian belief networks or causal networks, originate from artificial intelligence and statistics [21, 26]. They were developed as a powerful modeling tool for decision support and quantification of uncertainties, especially for low probability events. They have been applied to risk analysis in many studies since 2001 [27].

In a nutshell, a Bayesian network (see Figure 1) is a directed acyclic graph, in which a set of variables are represented by nodes. The relation between each node is represented in terms of parent-child and linked by an arrow, denoting the conditional dependencies between these variables. Conditional Probability Tables (CPTs) are attached to each node and consider all the possible states of a variable. Then, the probabilities of the nodes are determined by factorization of the joint probabilities by means of Bayes’ theorem. The joint probability is the function of all the random variables in BNs. For any BN, it can be given mathematically by a product of the CPTs entries,

$$P(X_i) = \prod_{i} P(X_i|pa(X_i))$$

(1)

where $X_i = X_1, \ldots, X_n$ denote the nodes of the BN, $pa(X_i)$ are the set of parents of $X_i$, and $P(X_i|pa(X_i))$ represent the entries of the CPTs. The effective methods for general inference in BNs can be achieved in [28], and it is also applicable for probability updating. For instance, in the case where evidence is assigned to the observed nodes $X_j = e$, this information will propagate through the prior probabilities to the posterior probabilities as follows,

$$P(X_i|e) = \frac{P(X_i,e)}{P(e)} = \frac{\prod_{i} P(X_i|pa(X_i), e)}{\sum_{X_i \backslash X_j} P(X_i,e)}$$

(2)
note that the joint distribution $P(X_i, e)$, obtained by using equation (1), associates with the evidence value $e$, and compute $P(e)$ from $P(X_i, e)$ by marginalizing out all the variables except the node $X_j$. If a node with no children has no associated evidence, it is called “barren node” meaning that the conditional probability is useless for the calculation of the marginal probabilities of non-barren nodes [29].

In general, as for the ability of belief propagation in the network, marginal posterior probabilities of the query nodes can be achieved through both top to bottom and inverse reasoning by means of the inference algorithms [21, 30], including exact algorithms and approximate algorithms. In comparison to approximate algorithms, exact algorithms, which are suitable for computing discrete BNs, are guaranteed to gain correct answers and hence it is a more robust computational method. In case of continuous variables in a BN, however, given the difficulty of defining the prior probability distributions as the discrete form, unavoidably impeding the application of BNs to the practical field.

BNs consisting of discrete and continuous variables are referred to as hybrid BNs. With consideration of exact algorithms, there are three special approaches to extending discrete BNs to continuous BNs or hybrid BNs. The first is to restrict continuous nodes to Gaussian random variables while allowing them to link only towards their non-discrete children. The second method is to define the continuous nodes as a mixture of truncated exponential distributions (MTEs), which is a generalization allowing to approximate any distribution function, but still requires further scrutiny [31]. The final methodology is enhanced Bayesian networks, implemented by joining BNs with SRMs, and was successfully applied in risk and reliability analysis [32, 33]. An introduction to this method is given in details in the following section.

2.2 Enhanced Bayesian Networks

In a system reliability problem, the outcome domain of an event, determined by a set of continuous random variables with known distributions, can be divided into failure and safe region by the relevant limit state functions. The failure probability of an event is the integral of the probability density function in the failure domain. This is a very complex integral and in general intractable analytically. Its value can be approximated by means of first or second order reliability methods or simulation approaches. Enhanced BNs approach is to combine SRMs with BNs. Specifically, in a BN, the continuous nodes must have at least an offspring, which is a discrete node defined as a domain in the outcome space of these continuous nodes. That is, the continuous nodes should meet the requirement of well-established SRMs, and it is the key condition for using eBNs approach.

Then these continuous nodes lose the causal dependencies with their deterministic nodes by means of FORM and SORM or simulation approaches. Finally, hence, all the continuous nodes can be removed from eBNs according to node elimination algorithm [24], and thus hybrid BNs are reduced to discrete BNs.

An example of computation of the total probability of an eBN and the process of node elimination is described by equation (3) to equation (5) for the simple case represented by Figure 2. From of equation (1), the joint probability of all the nodes for the eBN can be written as

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_4|X_2, X_3)f(X_3)f(X_2|X_1)$$

in which $P(X_4)$ and $P(X_4|X_2, X_3)$ represent conditional probabilities of discrete nodes $X_4$ while $f(X_3)$ and $f(X_2|X_1)$ are the probability density functions of continuous nodes $X_2$ and $X_3$, respectively.
The joint probability of the discrete nodes can be obtained by marginalization calculation of equation (3) as:

\[
P(X_1, X_4) = \int \int_{X_2, X_3} P(X_1 | X_2) P(X_1) P(X_4 | X_2, X_3) f(X_3) f(X_2 | X_1) dX_2 dX_3
\]

in the case the domain of node \(X_4\) can be determined by the outcome space of its parent nodes, then equation (4) can be rearranged as:

\[
P(X_1, X_4) = P(X_1) \int \int_{\Omega_{X_4}(X_2, X_3)} f(X_3) f(X_2 | X_1) dX_2 dX_3
\]

where \(\Omega_{X_4}(X_2, X_3)\) represents variable \(X_4\) as a domain in the outcome space of variables \(X_2 \) and \(X_3\).

The form of equation (5) are in line with the definition of structural reliability problems, and hence can be estimated by means of SRMs.

### 2.3 Evidence on Continuous nodes

As already stated, BNs show a powerful capability in updating probabilistic propagation through given observations. As previously discussed, evidence is inserted to replace certain priori probability on observed nodes, and the probabilities of the other nodes are updated using exact algorithms in discrete BNs. In a similar way, in eBNs, it is necessary to discretize continuous nodes with evidence at first, and then the corresponding discrete nodes are kept in place of the continuous nodes in the reduced BNs.

A plethora of discretization methods for continuous nodes in the BNs has been being investigated in [31, 34–37]. Currently, there are no formalized approaches for discretization of continuous random variables. Thus, in terms of the study problem here, a credible discretization approach for eBNs [24] is used in this study.

In terms of the study problem in this paper, only discretization of continuous nodes without any parent is introduced here. The previously introduced example is here reintroduced to explain how to discretize continuous nodes in eBNs. As shown in Figure 3, node \(X_3\) is substituted with two nodes, a discrete variable \(X_3\) _discrete_ and a continuous variable \(X_3\) _continuous_.

\(X_3\) _discrete_ has \(i\) states that are defined by the outcome space of \(X_3\) with conditional cumulative distribution function \(F_{X_3}[x_3]\), and the number of its states is identical to corresponding intervals of the divided domain of \(X_3\). Each sub-domain of \(X_3\) can be represent by \([x_{3i}, \bar{x}_{3i}]\), where \(x_{3i}\) and \(\bar{x}_{3i}\) denote the lower and upper bounds of the interval, respectively. Then the probability mass function of \(X_3\) _discrete_ given the state \(i\) can be achieved as,

\[
P(X_3 \mid x_3) = F_{X_3}[\bar{x}_{3i}] - F_{X_3}[x_{3i}]
\]

On the other hand, \(X_3\) _continuous_, as the child of \(X_3\) _discrete_, inherits all the descendants and outcome space of \(X_3\). The continuous variable \(X_3\) _continuous_ is eliminated from the model after it becomes a barren node by used of SRMs, and the discretized node \(X_3\) _discrete_ is retained to facilitate new observations updating the model.

In the same way, for inserting the evidence on \(X_3\), the process of discretization is to split the domain of \(X_3\) given the evidence into the sub-domains, each of which is obtained with a discrete probability value. In this study, the number of the sub-domains on the observed continuous node is defined by five with the same interval.
2.4 Credal Networks

In the case imprecise probabilities are introduced to Bayesian networks, they are referred as Credal networks (CNs) since the node corresponding to an imprecise event is associated with a credal set instead of a CPT or PDF. Credal sets are defined as closed convex sets associated with a set of probability distribution functions, which are used to represent imprecise probabilities in the graphical models. Fagiuoli and Zaffalon [38] used convex sets to compute posterior probabilities in a discrete BN with exact algorithms, and first referred to this kind of model as Credal networks. A detailed introduction of credal networks can be found in the literature [39, 40].

The inference in CNs is more complex than BNs, which is studied by some researchers [41–45]. Thanks to the development of inference algorithms in CNs, some exact and approximate inference algorithms can be used for the reasoning of CNs although imprecise probabilities propagation in CNs is still under study. In this paper, an integration of CNs and SRMs [46] is adopted to analyse the slope problems.

2.4.1 Inference computation in CNs

Continuous variables and interval variables in CNs are computed in terms of eBNs approach. The computation condition and elimination procedure (Figure 4) is the same with eBNs. After removing the continuous and imprecise nodes, CNs only contain two types of conditional probability in discrete nodes: point probabilities and bounded probabilities. Afterwards, variable elimination algorithm, as a classical exact inference algorithm, is applied here to estimate probability propagation in CNs.

As an example, a simple Credal network (CN) is shown in Figure 4. It consists of three types of nodes: discrete node $X_1$, continuous node with a known distribution $X_2$, and an imprecise node $X_3$. The deterministic node $X_4$ is dependent of all the other three nodes. Besides, both of discrete nodes $X_1$ and $X_4$ belong to binary variables, and then the joint probability for identifying upper and lower bounds of nodes in the CN can be expressed as,

$$P(X_1, X_4) = P(X_1)P(X_4|X_1)$$  \hspace{1cm} (7)

in which $\overline{X}_4$ denotes the upper and lower bounds in node $X_4$ with two states $x_{41}$ and $x_{42}$. Then, according to variable elimination, exact bounds of marginal probability with upper bound in the state $x_{11}$ of node $X_1$ can be obtained as,

$$P(x_{11})_{exact} = \max \left( \sum_{X_4} P(X_1)P(\overline{X}_4|X_1) \right) = \max \left[ \frac{P(x_{11})P(x_{41}|x_{11}) + P(x_{11})P(x_{42}|x_{11})}{P(x_{11})P(x_{41}|x_{11}) + P(x_{11})P(x_{42}|x_{11})} \right]$$  \hspace{1cm} (8)

The lower bound of the marginal probability can be obtained in a similar way with the minimum operator. Although traditional exact inference algorithms are efficient to compute the exact bounds, the exact inference is highly inefficient and leads to a combinatorial explosion in the case of complex networks, since it requires the evaluation of every possible bound combination for every node.

A novel algorithm has been introduced to avoid this combinatorial explosion encountered by exact inference [47]. The outcome from this approach can get the inner bounds, which can be equal to the exact bounds if no nodes with probability interval are observed. For a query node, briefly, instead of computing the true bound with identifying all of the combinations of the bounds in input, the key step is to compare the conditional probabilities of the query variable given the related nodes in CNs. Therefore,
it is obvious that the result by use of this kind of inner approximation is exact if no evidence involved in the bounded nodes.

In a word, this approach makes the computation low-cost, and it is effective to obtain real-time results with respect to the imprecise nodes in the model.

2.4.2 Simulation methods for the model with continuous or imprecise nodes

Slope stability analysis is commonly based on the evaluation of the factor of safety and involves a large number of random variables [48], and the probability of the slope failure can be obtained by means of SRMs with various simulation methods. Direct Monte Carlo approach is a robust and feasible method to compute the probability of failure. It has been previously applied to slope stability [8] and it is a classical simulation tool suited for the reduction of eBNs. Nevertheless, it requires a very high number of samples in the case of small failure and high dimensions. This is especially the case in the analysis of slope failures, where failure probabilities are typically in the order of $10^{-4}$ or smaller.

Advanced line sampling [49] is a recently developed advanced Monte Carlo methods, based on line sampling [50], and employs an adaptive algorithm to adapt the important direction to the shape of limitation state surface. Most importantly, it allows for sets of probability distributions to be included in the estimation of imprecise failure probabilities, which are bounded with upper and lower probabilities. Because of these advantages, advanced line sampling is adopted for the evaluation of CNs in this paper.

3 Risk assessment of slope stability with advanced BNs

The slope stability analysis considers the driving forces and resisting forces in the slope, comparing them and calculating the Factor of Safety. When the driving forces are larger than the resistant forces, the slope fails. The model introduced in the following section is to construct a cause-effect graph to estimate the probability of slope failure and identify the key landslide-induced factors.

3.1 Factors of safety for an infinite slope

Factors of safety are frequently computed to identify whether a slope is safe, which can be obtained by the ratio of resisting and driving stresses along a potential slip surface. This calculation, however, is not based on a unique equation, since there are a variety of methods [51–54] that can be selected to obtain the factor of safety according to the different conditions. These conditions also depend on the type of failure surface and its extension.

For an infinite slope, as the Figure 5 shown, the equation for the factor of safety in terms of effective
stress analysis is given by

\[ FOS = \frac{c + (\gamma_d Z_d + \gamma_{sat} Z_{sat} - \gamma_w Z_{sat}) \cdot \cos \beta \cdot \tan \phi}{(\gamma_d Z_d + \gamma_{sat} Z_{sat}) \cdot \sin \beta} \]

(9)

here, the drained parameters of cohesion \((c)\) and friction angle \((\phi)\) are the main properties of soil. \(Z_d\) and \(Z_{sat}\) are the thickness of unsaturated and saturated soil layer, respectively, and the sum of them is the total thickness of soil \((Z)\). \(\beta\) is the slope inclination and \(\gamma_w\) is the unit weight of water, \(9.81 \text{kN/m}^3\). For the layer above and below water table, soil unit weight should be split into two parts: dry unit soil weight \((\gamma_d)\) and saturated unit soil weight \((\gamma_{sat})\). This analysis has been completed using in infinite slope stability model [55]. Moreover, \(FOS \leq 1\) means the slope failure, whilst the \(FOS\) larger than 1 indicates the slope is safe. All the calculations are performed in effective stresses but, for the sake of simplicity, the effective parameters, cohesion and friction angle, are simply denominated as \(c\) and \(\phi\), as there is no risk to misunderstand effective and total strength resistances.

3.2 The stability analysis of an infinite slope with BNs

The slope BN (see Figure 6) considers the stability of an infinite slope with its relevant variables and factors of safety. Therefore, it is reasonable to take factors of safety as a target event to infer the crucial events affecting the consequence, and further ascertain the causal relation amongst the events.

![Figure 6: The hybrid BN model of an infinite slope.](image)

The model includes six factors and one failure event. The main strength parameters of the soil, Cohesion and Friction Angle are the resisting forces preventing the occurrence of a failure. Meanwhile, the geometrical parameters of the slope are the slope inclination and slope’s height, being also two important factors for slope stability. The angle of a slope defines how much driving force is distributed in the parallel direction along the slope surface. Small angles mean small pulling force on the downslope movement while large angle provides the large pulling force. In this model, \(Z\) and \(\beta\) are constant, and the height of slope can be obtained by \(Z/\cos \beta\), so they are not considered into this BN.

According to effective stress principle [56], pore water pressure is defined by the unit weight of soil and the corresponding soil thickness. In such conditions, it was also considered the influence of the water table in the slope stability. In light of this, the nodes Unsaturated Unit Weight, Saturated Unit Weight and Saturated Thickness are selected in the slope model.

The position of groundwater table is an unstable variable defining the slope safety. The node Saturated Thickness represents the depth of saturated soil, which is the level of the water table. This random variable is governed by the drainage condition. To be specific, the water table is away when drainage takes place. If not, the depth of saturated soil will assume random values ranging between 0 and \(Z\). In general, the event of Drainage affects the node Saturated Thickness.

Finally, it is evidence that slope sliding along different slip surfaces and accordingly the failure surface may be of any shape [19, 57]. Undoubtedly, an existing form of slope failure can be defined as a result event in BNs. So, in this study, a translational failure event proposed is represented by a bottom node, named Slope Failure in Figure 6, whose probability can be computed by a limit state function \(G(X)\),

\[ G(X) = FOS - 1 \]

(10)

in which the node is a discrete variable with two states: \(G(X) > 0\), the node denotes the probability of a stable slope, otherwise, it is the failure probability of the slope.
3.3 Slope failure analysis with Credal networks

In the analysis of an infinite slope with Bayesian networks approach, each node can be defined with precise probability distribution functions. With the limitation of information for the geotechnical problems, however, it is impossible to define all the nodes as point probabilities. Imprecise probabilities are employed to cope with imprecise problems in engineering analysis by means of transforming imprecise information into the probabilistic form [58]. This approach is also interesting to solve geotechnical problems. A discussion and comparison about how to define the input information using intervals or probabilistic model can be found in [59]. Interval model is the most common way to describe imprecision in geotechnical engineering, such as uncertain soil parameters with non-probabilistic information can be set as the form of the interval to make uncertainty analysis of a slope stability [60]. In this paper, hence, this approach is used to define the scarce input information. A Credal network proposed in this section is to estimate the probability of slope failure with limited information subject to drainage influence and initially demonstrate the suitability of CNs for the slope stability analysis.

As is aforementioned, the shear resistance cohesion and friction angle are two key factors of influencing the slope stability. If there is scarce information provided for an infinite slope, for example, the parameters \( c \) and \( \phi \) change with geological/geotechnical conditions, so without any experiment test, it cannot be known in advance the exact properties of the slope’s material. In this case, they should be two imprecise inputs for the CN.

The model implemented presents nine nodes, including discrete variables, continuous variables, interval variables and parametric p-boxes. These corresponding nodes are represented by rectangular, circle, ellipse and trapezoid, respectively (see Figure 7). In addition, the events and construction of the model remain similar to the previous BN model of the slope. Differently, the nodes Cohesion, Friction Angle, Saturated Unit Weight and Unsaturated Unit Weight associated with imprecise information, and the prior interval-value definitions of these imprecise nodes are achieved from expert judgement.

If the further information is available, such as the information of parametric p-boxes is linked with the nodes Cohesion and Friction Angle. To be specific, the two soil parameters can be described by the known probability distribution, whose parameters are interval variables. Then in this credal network, the imprecise information is presented by a combination of the nodes Vcohesion and Cohesion, Vfriction and Friction Angle. Comparing to the previous BN, the nodes Cohesion and Friction Angle in the CN model are substituted by the respective parametric p-boxes.

Figure 7: The CN model of an infinite slope.

Slope Failure is the node of interest in the CN, whose failure state of the node can predict the occurrence of a shallow landslide. The probability of slope failure is inferred by marginal probability calculation in the reduced CN. Furthermore, an analysis can be conducted to demonstrate the effect of the node Drainage on the slope stability. A simple example is implemented in the next section in order to illustrate the feasibility of this method.

4 Example 1: soil slope

The translational slip shown in Figure 5 is adopted to illustrate example 1. The total thickness of the slope is 4 m at the inclination angle \( \beta = 30^\circ \). The BN model in Figure 6 is adopted in this example.

In the view of quantification, the key parameters of the soil slope in the BN are all set as random variables with known probability distributions according to their uncertain properties. Amongst these, the thickness of unsaturated soil \( (Z_d) \) changes complementarily with the thickness of the saturated slope,
and thus can be expressed by \((4 - Z_{\text{sat}})\). The detailed definitions of variables involved in the BN can be obtained from Table 1. The CPT of querying node of slope failure is evaluated by means of reliability analysis, which is used to update the model with inserting evidence as well.

In light of the above-mentioned method, the CPT of the querying node can be computed by equation (11) to (12), which is shown as follows,

\[
P(SF) = P(D) \int \int_{\Omega_{SF}(\gamma_d, \gamma_{\text{sat}}, Z_{\text{sat}}, c, \phi)} f(\gamma_d) f(\phi) f(c) f(Z_{\text{sat}} | D) d\gamma_d d\phi dc d\gamma_{\text{sat}} dZ_{\text{sat}}
\]

(11)

here, the domain of the slope failure \(\Omega_{SF}(\gamma_d, \gamma_{\text{sat}}, Z_{\text{sat}}, c, \phi)\) can be described by the limited state function according to equation (10), as:

\[
P(SF) = \begin{cases} 
P_f, & FOS - 1 < 0 \\ 
P_s, & FOS - 1 \geq 0 
\end{cases}
\]

(12)

in which \(P_f\) denotes the failure probability of the slope while the safe probability is \(P_s\).

Table 1: Input parameters of the unsaturated soil.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Variable type</th>
<th>CPD*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohesion [kPa]</td>
<td>Continuous</td>
<td>logN(22,10)</td>
</tr>
<tr>
<td>Friction angle [°]</td>
<td>Continuous</td>
<td>N(35,3)</td>
</tr>
<tr>
<td>Unsaturated unit weight [kN/m³]</td>
<td>Continuous</td>
<td>N(17, 0.4)</td>
</tr>
<tr>
<td>Saturated unit weight [kN/m³]</td>
<td>Continuous</td>
<td>N(19, 0.5)</td>
</tr>
<tr>
<td>Saturated thickness of soil [m]</td>
<td>Continuous</td>
<td>U(0, 4) or 0</td>
</tr>
<tr>
<td>Drainage ((D))</td>
<td>Discrete</td>
<td>[0.5, 0.5]</td>
</tr>
<tr>
<td>Slope failure ((SF))</td>
<td>Discrete</td>
<td>([P_f, P_s])</td>
</tr>
</tbody>
</table>

*\(N, \text{logN}, \text{represents normal and lognormal distribution with mean and standard deviation, respectively. U represent uniform distribution with lower and upper bound.}\)

In order to characterize the relationship between slope stability and other influence factors, the easy way is to change these nodes to obtain a new result of the slope stability. Then the new observations with narrowing the distribution range of random variables are inserted into the continuous nodes. On the basis of the expert knowledge, the new bounds of the nodes Cohesion, Friction Angle, Unsaturated Unit Weight, and Saturated Unit Weight, are \([0, 100]\), \([25, 45]\), \([16, 19]\), \([18, 21]\), respectively. Besides, a narrower bound of each observed node also is computed to make a further observation. Monte Carlo simulation is applied to each network updating.

4.1 Results from example 1

Figure 8 shows the reduced BN, where all the continuous variables are removed from the original BN by means of eBNs approach, and only two discrete nodes left: Drainage and Slope Failure. Then the reasoning in the eBN can be inferred with this traditional BN, and the results are exhibited in Table 2. The prior marginal probability of node Slope Failure is 2.74%. Compared with no evidence, the occurrence of failure of the slope in case of drainage is rare, which is much lower than the state of no drainage, whose result with 5.13%. That means that if drainage takes place it can stabilize the slope. Furthermore, it can be reasonably achieved that drainage is of importance to this slope. Then from this information support, the decision maker immediately knows the risk of slope failure can be avoided if he spends money in draining the slope.
In geotechnical problems, it is common that the soil characterization is performed in different phases and, therefore, new observations can be obtained in an advanced step of the study. These new results serve to identify the influence of soil parameters on the slope stability. The adoption of the discretized approach allows considering these new results as evidence, updating the probabilities in the model. From the results in Table 3, the failure probability of the slope varies from 2.55% to 2.67%, which is close to the original result, but the new information indicates the reduced effect on the failure result.

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Cohesion</th>
<th>Friction Angle</th>
<th>Unsaturated Unit Weight</th>
<th>Saturated Unit Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence</td>
<td>0 ≤ c ≤ 100</td>
<td>25 ≤ φ ≤ 45</td>
<td>16 ≤ γd ≤ 19</td>
<td>18 ≤ γsat ≤ 21</td>
</tr>
<tr>
<td>P(FS)</td>
<td>0.0255</td>
<td>0.0267</td>
<td>0.0259</td>
<td>0.0256</td>
</tr>
</tbody>
</table>

The reason for such a small variation in the failure probability results from the large range given for the observations. Therefore, the outcome will be much more evidence if the observations are narrower, what could be a result of additional geotechnical tests, for example, narrowing the considered intervals as Table 4 shows. P(FS) is 1.81% and 2.37% respectively with the corresponding limited range of the nodes Cohesion [10, 90] and Friction Angle [30, 40]. These results reveal that both of them greatly affect the reduction of the risk of the slope, while Unsaturated Unit Weight and Saturated Unit Weight have a smaller effect on the results of P(FS) in comparison with the results of Table 3. Given a slight change in each observed nodes, the value of P(FS) shows more apparently variation in Cohesion and Friction Angle than nodes Unsaturated Unit Weight and Saturated Unit Weight. In this study, therefore, Cohesion and Friction Angle are more important for this slope stability.

5 Example 2: slope in residual soil from Porto

In regions where igneous rock, like granite or gneiss are present, the weathering of the rock produces soils designated as residual soils. These materials are very common in mountainous countries as the case of Portugal, Spain, Brazil, China, Hong Kong, Singapore, and Africa.

In the northern part of Portugal, there has been an extensive geotechnical characterization of granite residual soils [61–63]. Table 5 presents the common strength parameters found in the residual soil from Granite, in the Porto region. The typical values for strength parameters of this type of soil from Porto, such as cohesion and friction angle, are known as the interval-value with lower and upper bounds. Unsaturated and saturated unit soil weight are both defined based on expert knowledge. Besides, for a typical design, a slope in residual soils is typically designed with a fixed inclination of 3H to 2V, and the total thickness of this slope is assumed as 4 m in this study. The failure surface is considered parallel to the surface of the slope, as shown in Figure 9.

Three different situations of information available in Cohesion, Friction Angle, Unsaturated Unit Weight, and Saturated Unit Weight are described in Table 5. With the model, a change in any of the four factors would cause the varying failure probability of a slope. For precise information provided, the eBNs approach is used here as a reference, and interval analysis is adopted to cope with the limited information. If further information about the variables can be achieved, such as input distribution with a bound on its mean, then the parametric p-boxes might be introduced in the imprecise nodes Cohesion and Friction Angle. Thus it is possible to observe the change of the results in comparison with only interval nodes in the model.
L. He, A. Topa Gomes, M. Broggi, M. Beer

Figure 9: A residual soil slope.

Table 5: Input parameters of the residual soil.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Precise information</th>
<th>Imprecise information</th>
<th>P-boxes*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohesion [kPa]</td>
<td>logN(20,4)</td>
<td>0</td>
<td>logN(μ_c,4), μ_c ∈ [16, 22]</td>
</tr>
<tr>
<td>Friction angle [°]</td>
<td>N(37,1.85)</td>
<td>25</td>
<td>N(μ_f,1.85), μ_f ∈ [36, 38.5]</td>
</tr>
<tr>
<td>Unsaturated unit weight [kN/m³]</td>
<td>N(18.5,0.51)</td>
<td>17</td>
<td>[17, 20]</td>
</tr>
<tr>
<td>Saturated unit weight [kN/m³]</td>
<td>N(20,0.6)</td>
<td>18</td>
<td>[18, 22]</td>
</tr>
</tbody>
</table>

*μ indicates the mean of the distributions. The notes of Table 1 also apply here.

Considering the Credal network of the residual soil slope, the CPT of the discrete node Slope Failure is obtained on the basis of the factor of safety, same as equation (11) and (12), where the calculation of the failure probability is made with adaptive line sampling approach [49].

5.1 Results from example II

From the results (Table 6), it can be seen that an exact probability of slope failure can be obtained with the precise input for the conservative model. The 2.47% failure probability indicates a reasonable risk for this existing slope with precisely specific parameters. However, in the case of poor information, the input uncertainty affects the precision of output so that the results are denoted with the probability bounds. When the input nodes Cohesion, Friction Angle, Unsaturated Unit Weight, and Saturated Unit Weight only can be defined as interval variables with the limited information, the probability bound of slope failure is between 0 and 1. The result is too wide to provide the useful information of slope stability. In other words, each combination of the different values of the factors can produce any possibility of the slope states, failure or safe. Hence, the feasible way is to reduce the uncertainty input in order to increase the precision of the output, what can be done by producing additional geotechnical information or by approaching the reliability problem with different methods. For example, a practical common geotechnical solution would result from performing additional boreholes in the slope and laboratory test what would allow to more precise geotechnical parameters.

Table 6: Slope failure probability.

<table>
<thead>
<tr>
<th>Different information</th>
<th>Precise</th>
<th>Imprecision</th>
<th>Parametric p-box</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(FS)</td>
<td>0.0274</td>
<td>[0, 1]</td>
<td>[0, 0.0711]</td>
</tr>
</tbody>
</table>

Comparing to the first two input information, the further observation is to add the probability boxes in the imprecise nodes Cohesion and Friction Angle. As it is shown in Table 6, the probability bound of
failure slope became dramatically tighter after introducing P-boxes. The range of the failure result is from
0 to 7.11%, and the upper bound of the failure slope reveals a steep decrease. Besides, the precise result
with 2.47% is included in this range. It illustrates actual risk can be estimated with the consideration of
the reasonable application of parametric p-boxes in the CN model.

Table 7: Failure probabilities with two states of Drainage.

<table>
<thead>
<tr>
<th>Different information</th>
<th>Precise</th>
<th>Imprecision</th>
<th>Parametric p-box</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P(\text{FS}</td>
<td>D = \text{true}))</td>
<td>0</td>
<td>[0, 1]</td>
</tr>
<tr>
<td>(P(\text{FS}</td>
<td>D = \text{false}))</td>
<td>0.0514</td>
<td>[0, 1]</td>
</tr>
</tbody>
</table>

In Table 7, the risk of slope failure under drained conditions shows a greatly reduced tendency, and
even the risk can decrease to 0 in contrast to the state of no drainage. That is because if water is away,
the percolation forces disappear and the resistant forces also increase, as a result of the increase in the
normal force and, therefore, the friction component of the strength also increases.

The result with the interval [0, 1] based on the very poor information cannot give further information
for decision makers, but the probability bound of Slope Failure with the evidence Drainage makes sense
by ways of p-boxes. Specifically, if drainage is not implemented, the failure result of the residual soil
slope with [0, 15.31%] is much wider than the one with drainage.

6 Conclusions

This study presents the advanced Bayesian networks methods to estimate the risk of failure of the slope
subjected to drainage state. For the purpose of identifying the factors that affect the slope failure the
most, new observations are made in some continuous nodes to update the model. The proposed methods
proved useful and the results provide significant information for the decision makers.

Enhanced Bayesian networks and Credal networks are applied relying on the input information availability. Enhanced Bayesian networks consist of two types of nodes, continuous and discrete nodes, where an integration of Bayesian networks and structural reliability analysis is applied to make the inference in this precise model, while the Credal network, especially for the scenario that there is no enough abundant information to get the precise CPDs for each of nodes. Additionally, discrete variables, random variables, interval variables and p-boxes are presented in the model. After the adoption of adaptive line sampling, the reduced model is computed by exact inference algorithms. The bounds of result provide a rough estimation of the actual risk, and the permission of application of p-boxes in the model contributes to the reduction of the uncertainty in output. Moreover, a discretization process is applied when new evidence enters the continuous nodes. These capabilities ensure the wide flexibility of the model in analyzing the risk of the slope.

The two examples implemented demonstrate that the models have interesting possibilities for assessing
the risk of the slope. The exact failure probability of an infinite slope in the first example indicates a low
risk, and according to the analysis of updating the information in the specific nodes, the conclusion can
be made that the risk of the slope can be significantly improved with drainage. Although interval-value is
a suitable way for representing the non-probabilistic information, the interval results of the residual soil
slope may fail to acquire the usable range of real value. In this case, p-boxes involved obviously narrow
the bound of failure probability. All in all, both of enhanced Bayesian networks and Credal networks are
effective and feasible means to estimate the failure of a slope.

Acknowledgments

This work is supported by Chinese Scholarship Council.

References

[1] H. Rahardjo, S. N. Alfrendo, C. L. Eng, and Y. S. Ng, Effects of groundwater table position and
soil properties on stability of slope during rainfall, Geotechnical and Geoenvironmental Engineering
136, 1555 (2010).
of causal factors for the Majiagou landslide using modern data mining methods, the International
Consortium on Landslides 14, 311 (2016).
Resources Research 44, 404 (2008).


Modelling adaptive systems using plausible Petri nets

J. Chiachío,1* M. Chiachío,1,2 D. Prescott,1 and J. Andrews1

1Resilience Engineering Research Group, University of Nottingham, UK
2Department of Structural Mechanics and Hydraulic Engineering, University of Granada, Spain
*Corresponding author: juan.chiachioruano[at]nottingham.ac.uk

Abstract

One of the main challenges when analyzing and modelling complex systems using Petri nets is to deal with uncertain information, and moreover, to be able to use such uncertainty to dynamically adapt the modelled system to uncertain (changing) contextual conditions. Such self-adaptation relies on some form of learning capability of the Petri net, which can be hardly implemented using the existing Petri net formalisms. This paper shows how uncertainty management and self-adaptation can be achieved naturally using Plausible Petri Nets, a new Petri net paradigm recently developed by the authors [Information Sciences, 453 (2018) pp. 323-345]. The methodology is exemplified using a case study about railway track asset management, where several track maintenance and inspection activities are modelled jointly with a stochastic track geometry degradation process using a Plausible Petri net. The resulting expert system is shown to be able to autonomously adapt to contextual changes coming from noisy condition monitoring data. This adaptation is carried out taking advantage of a Bayesian updating mechanism which is inherently implemented in the execution semantics of the Plausible Petri net.

Keywords: Uncertain information, Bayesian learning, Plausible Petri nets, Infrastructure Asset Management

1 Introduction

A Petri net (PN) is a mathematical and graphical modelling tool first introduced by Carl Petri in 1962 [1] for the analysis of the dynamic behaviour of sequential asynchronous automatons. Since then, the PN methodology has been greatly expanded and combined with other modelling techniques like fuzzy sets, neural networks, etc., for the modelling of complex processes and intelligent systems [2]. The main concepts relative to the theory of PNs are summarised in [3], while a tutorial for practical engineering applications can be found in [4]. A known limitation of the original PN formalism is the difficulty to deal with uncertain information during runtime, and moreover, its inability to use such information to dynamically accommodate the modelled system to new environmental and contextual conditions [5, 6]. In the literature, a number of PN variants have been introduced in order to enhance the original PN formalism with improved rules for uncertainty quantification and self-adaptation [7, 8]. However, none of the existing PN formalisms provides direct means to efficiently deal with uncertainty while considering the hybrid systems, consisting of a combination of discrete and continuous processes whose evolution may be uncertain [9].

This paper overviews a recent PN formalism proposed by the authors in [9, 10], known as Plausible Petri nets (PPNs), whereby discrete events (e.g., go/no-go decisions, intervention activities, etc.) can be modelled together with continuous processes whose evolution may be uncertain (e.g. a component deterioration process). In particular, this paper highlights through a case study how uncertain information and self-adaptation can be achieved naturally using PPNs, since they have an inherent Bayesian updating mechanism implemented within their execution rules. As an illustrative example, a railway track asset management problem is idealised using a PPN model, whereby a number of operational rules (e.g., maintenance operations and inspection activities) are considered together with a continuous-variable stochastic process for track geometry degradation [11]. Through simulation, the operational rules are shown to be autonomously and adaptively triggered based on the degree of belief [12] about the state of degradation of the track, which is sequentially updated as long as contextual changes in the form of (noisy) condition monitoring data become available.

The remainder of the paper is organised as follows. Section 2 presents the fundamentals of the classical Petri nets. In Section 3, an overview of the PPN paradigm is provided. Section 4 is devoted to present the case study about railway track asset management modelling using PPNs. Finally, Section 5 provides concluding remarks.

8th International Workshop on Reliable Engineering Computing, “Computing with Confidence”
16–18 July 2018, University of Liverpool, Liverpool, UK
Modelling adaptive systems using plausible Petri nets

2 Basics of Petri nets modelling

Petri nets (PN) [1] are typically regarded as a powerful modelling tool for complex systems, especially when system-level operational nonlinearities (e.g., resource availability, concurrency and synchronisation of components, etc.) need to be considered in the analysis. PNs are based on a graphical and mathematical language with well-established execution semantics, which greatly increases their suitability for modelling complex distributed systems and systems of systems [2, 13].

From a graphical point of view, PNs are composed by two types of nodes, namely, places (represented by circles) and transitions (represented by boxes). The nodes are connected by arcs either from places to transitions or vice versa. In most practical applications, the places represent particular discrete states of the system being modelled (e.g. the health state of a component), while the transitions represent processes that enable the system to move from one place to another. The places contain tokens that travel through the net to other places depending on the firing of the transitions [3]. The presence of a token in a particular place can be interpreted as holding the truth of the condition or information represented by that place (e.g., "component failed"), and the distribution of tokens over the PN at a specific time is referred to as the marking of the net, which is expressed as a vector indicative of the state of the PN. A particular transition $t$ is fired only if all places leading to that transition have at least one token. Those places define the pre-set of transition $t$, denoted by $\bullet t$. After firing the transition, one token is added to each of its output places, thus defining the post-set of the transition, referred to as $t \circ$. Arcs are labeled with their corresponding weights, which are non-negative integer values indicating the amount of parallel arcs (1 by default). In Figure 1, an illustration of a sample PN comprised of three places ($p_1, p_2, p_3$) and one transition ($t_1$) is depicted.

![Figure 1: Example of Petri net comprised of three places ($p_1, p_2, p_3$) and one transition ($t_1$).](image)

From a mathematical perspective, a PN can be defined as a tuple $\mathfrak{t} = \langle P, T, E, W, M_0 \rangle$, where $P \in \mathbb{N}^n$ is the set of places, $T \in \mathbb{N}^t$ is the set of transitions, $E \subseteq (P \times T) \cup (T \times P)$ is the set of arcs connecting places to transitions and vice versa, $W$ is a set of non-negative numerical values (1 by default) acting as weights applied to each arc within $E$, and $M_0$ is a vector containing the initial distribution of tokens over the set of places (initial marking).

At a particular time $k \in \mathbb{N}$, the dynamics of the overall PN can be described using the state transition equation, which is mathematically expressed as [3]:

$$M_{k+1} = M_k + A^T u_k$$  \hspace{1cm} (1)

where $M_k$ is the marking of the PN at time $k$; $u_k = (u_{1,k}, u_{2,k}, \ldots, u_{n,t,k})^T$ is the firing vector at time $k$, with $u_{i,k} = 1$ if transition $t_i$ is fired, and $u_{i,k} = 0$ otherwise; and $A$ is the $n_t \times n_p$ incidence matrix, whose $(i,j)$-th element is obtained as $a_{ij} = a_{ij}^+ - a_{ij}^-$, $i = 1, \ldots, n_t$, $j = 1, \ldots, n_p$, with $a_{ij}^+$ being the weight of the arc from transition $t_i$ to output place $p_j$, whereas $a_{ij}^-$ is the weight of the arc from input place $p_j$ to transition $t_i$. Therefore, using Equation (1), the evolution of the marking and thus the dynamics of the overall system can be simulated time to time.

In PNs, any transition $t_i$ needs to be enabled as a condition to be fired, which occurs when each input place of $t_i$ is marked with at least $a_{ij}^-$ tokens. Mathematically:

$$M(j) \geq a_{ij}^- \forall p_j \in \bullet t_i$$  \hspace{1cm} (2)

where $M(j) \in \mathbb{N}$ is the marking for place $p_j$. Note that in practical engineering applications, transitions are typically assigned with time delays which are useful for task scheduling modelling and performance evaluation in dynamical systems [3]. In such cases, a transition is fired once its time delay has passed. The resulting PNs are known as Timed Petri nets if the delays are deterministic, and Stochastic Petri nets if the delays are uncertain, represented by probability distributions.
3 Dealing with uncertainty using Plausible Petri nets

PNs, as originally conceived by Carl A. Petri [1], are well-suited to represent the dynamics of complex systems and processes since they provide a graphical support for system idealization, but also because they rely on rigorous mathematical principles which enable their computational implementation and simulation. However, a typical criticism of PNs is that they are not adequate to deal with uncertain information (e.g., uncertain system states, uncertain processes, etc.) [8, 14], nor do they consider the hybrid nature of real-world dynamical systems, consisting of a combination of discrete and continuous processes whose evolution may be uncertain [9].

Plausible Petri nets (PPNs) are a variant of PNs recently developed by the authors in [9, 10], where the uncertainty is rigorously accounted for through states of information, which provide a mapping that assigns to each possible numerical value of the state variable its relative plausibility. Two types of processes can be jointly simulated using PPNs; namely, discrete processes, where the tokens are objects in the sense of integer moving units, as in classical PNs; and continuous processes using numerical variables (e.g., a degradation process), where the tokens are probability density functions (referred to as states of information) which are transferred through the net based on particular execution rules. Thus, in PPNs, the overall net is partitioned into two disjoint subnets, namely the numerical subnet and the symbolic subnet, which evolve interactively under the same execution rules. These are denoted using the superscripts \((N)\) and \((S)\), respectively. Specific details about the PPN paradigm can be found in [9], but here, the key differences between PNs and PPNs are provided for the sake of clarity and better readability.

3.1 Execution rules of PPNs

In PPNs, the referred states of information about a system state variable \(x_k \in X\) are denoted by the probability density functions (PDFs) \(f^P(x_k)\) and \(f^S(x_k)\) for numerical places and transitions, respectively. Thus, the marking \(M_k\) of the overall PPN at a particular time \(k\) consists in a combined vector \(M_k = (M^N_k, M^S_k)\), where \(M^N_k\) is the marking for the numerical subnet (consisting of a column vector of normalised PDFs) and \(M^S_k\) is the marking of the symbolic subnet (consisting of a column vector of discrete values). While the marking evolution of the symbolic subnet is given by the state transition equation in (1), the marking evolution of the numerical subnet relies on two basics operations for information flow dynamics: the conjunction and disjunction of states of information \([12, 15]\). Using these operations, the logic operators \(\text{AND} (\wedge)\) and \(\text{OR} (\vee)\) are invoked to enable the information exchange across the numerical subnet by respectively combining and aggregating states of information [9]. An illustrative explanation of the conjunction and disjunction of states of information operations is given in Figure 2. From this standpoint, the dynamics of PPNs can be formulated by adopting of the following rules [9]:

1. An input arc from place \(p_j^{(N)}\) to transition \(t_i\) conveys a state of information given by \(a_{ij}^-(f^P \wedge f^S)(x_k)\), which remains in \(p_j^{(N)}\) after transition \(t_i\) has fired;

2. Transition \(t_i\) produces to an output arc a state of information given by \(a_{ij}^+(f^S \wedge f^P)(x_k)\), where \(f^S(t_i)(x_k)\) is the PDF resulting from the disjunction of the states of information of the pre-set of \(t_i\), given by \(f^S(t_i)(x_k) = \frac{\beta}{\sum} (f^{p_1} + f^{p_2} + \cdots + f^{p_m})(x_k)\), where \(\beta\) is a normalising constant, and \(p_1, \ldots, p_m \in \bullet t_i \subset D^{(N)}\);

3. After firing the numerical transition \(t_i\), the state of information resulting in place \(p_j^{(N)} \in t_i^*\) is the disjunction of the state of information \(f^P(x_k)\) (the previous state of information) and \(a_{ij}^+(f^S \wedge f^P)(x_k)\), the information produced after firing transition \(t_i\). Mathematically:

\[
\left(f^{P}(x_{k+1}) = \left(f^{P}(x_k) \vee a_{ij}^+(f^S \wedge f^P)(x_k)\right)\right)(x_k)
\]

Note that the execution rules given above for PPNs are analytically intractable except for very simple cases, since the conjunction of states of information requires the evaluation of normalising constants involving an intractable integral. For these cases, a numerical method using particles [17] is proposed in [9] to compute the conjunction of states of information while circumventing the evaluation of the normalising constants. Note also in the execution rules that PPNs have an inherent Bayesian learning mechanism implemented in their execution semantics. In particular, if one observes rule 1 and just assumes that \(f^P(x_k)\) acts as likelihood function for a set of data \(y_k \in D\) (i.e. \(f^P(x_k) = p(y_k|x_k)\)), and that \(p_j^{(N)}\) represents a prior PDF of \(x_k\), the conjunction of both states of information leads to a posterior PDF of the state variable, assuming that the \(X\)-space is a linear space [15]. This interesting property of PPNs is further exploited within the context of an engineering case study.
4 Case study

In this section, an engineering case study about railway track asset management is provided to illustrate how uncertain information and condition monitoring data can be integrated using PPNs to support autonomous and adaptive decisions about infrastructure degradation and maintenance. To this end, a PPN model is developed to idealise an expert system for railway track asset management, incorporating a physics-based model for track geometry degradation, condition monitoring data, along with inspection activities. The PPN is depicted in Figure 3. For this case study, the deterioration of the track is assumed to occur due to traffic loadings (expressed in load cycles), and also, it is assumed that the track geometry degradation can be periodically measured using train-borne sensors. In this sense, every time a new measurement is available, the PPN incorporates that data to update the underlying track degradation model so that further inspections and maintenance activities are autonomously triggered based on up-to-date model predictions, instead of using simply the raw data as a base for decision making.

Observe in Figure 3 that the PPN is comprised of one numerical place, \( p_1^{(N)} \) (containing the track degradation model), five symbolic places, \( p_2^{(S)} \) to \( p_6^{(S)} \) (representing the activation of inspection activities and other management variables), three mixed transitions, \( t_1 \) to \( t_3 \), and two symbolic transitions, \( t_4 \), \( t_5 \). The formulation and implementation details of the adopted track degradation model can be found in \([11]\), so they are not repeated here. The condition monitoring data used for this case study consist in a set of measurements about track settlement \( Y = \{y_i, y_j, \ldots, y_k\} \) taken from the literature \([18]\), which are sequentially introduced to the system at loading cycles \( \{i, j, \ldots, k\} \subset \mathbb{N} \). In Figure 3, a token in place \( p_1^{(S)} \) represents a data point arrived to the system. The dataset is reproduced in Table 1. The measurements are assumed to come with a 5% white-noise type error, therefore \( y_k \sim \mathcal{N}(x_k, \sigma_{w_k}) \), with \( \sigma_{w_k} = 0.05 \|y_k\| \). This PDF represents the likelihood of the measurements, and comprises a state of information within transition \( t_1 \), given by \( f_1^{(1)}(x_k) = \mathcal{N}(x_k, \sigma_{w_k}) \).

By evaluating the proposed PPN-based expert system, changes in the numerical and discrete states of the system are obtained based on a number of automated actions which are activated through firing transitions \( t_1 \) to \( t_5 \). A summary of the functions defining the transitions is provided in Table 2. The execution rules given in Section 3.1 are applied to obtain the overall system evolution described through the marking \( M_k \), \( k > 0 \).

In particular, observe that each time a new measurement is available, transition \( t_1 \) is enabled, which leads to the conjunction of the states of information of \( p_1^{(N)} \) and \( t_1 \). As explained in Section 3.1, this conjunction conveys a Bayesian updating of the degradation variable \( x_k \) using monitoring data, as per Table 1. The results for the updated degradation variable along with its 5% – 95% confidence interval are depicted in Figure 4a (see the leftmost panel). In panel 4b, the temporal evolution of the

Table 1: Railway track settlement (strain) data used for calculations taken from [18].

<table>
<thead>
<tr>
<th>Loading cycles ((\times 10^3))</th>
<th>0</th>
<th>0.625</th>
<th>1.25</th>
<th>2.5</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>50</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plastic strain</td>
<td>0</td>
<td>0.0017</td>
<td>0.0045</td>
<td>0.0058</td>
<td>0.0075</td>
<td>0.0087</td>
<td>0.0104</td>
<td>0.0111</td>
<td>0.012</td>
<td>0.01275</td>
</tr>
</tbody>
</table>

Figure 2: Conjunction (left panel) and disjunction (right panel) of two arbitrary states of information, \( f_a(x) \) and \( f_b(x) \). The term \( \mu(x) \) is the homogeneous density function \([15, 16]\), representing a state of complete ignorance about the stochastic variable \( x \in \mathcal{X} \).
uncertainty (differential entropy, DE) in the estimation of $x_k$ within place $p_1^{(N)}$ is shown, with indication of the reference level when inspections are needed. The observed drops in the sequence of DE values in Figure 4b correspond to the uncertainty reduction due to Bayesian learning of the track degradation model when new measurements become available.

Table 2: Description of the transitions adopted in PPN depicted in Figure 3. PI: Periodic inspections, OI: opportunistic inspections, LC: Line Closure.

<table>
<thead>
<tr>
<th>ID</th>
<th>Type</th>
<th>Rule</th>
<th>State of information</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>Mixed</td>
<td>$H(x_k) \geq -4.8$</td>
<td>$f_k \sim p(y_k</td>
<td>x_k)$ (Likelihood)</td>
</tr>
<tr>
<td>$t_2$</td>
<td>Mixed</td>
<td>$E_{f_1}(x_k) \geq 0.014$ [m]</td>
<td>$f_k \sim I_{C_2}(x_k)$</td>
<td>Activates OI</td>
</tr>
<tr>
<td>$t_3$</td>
<td>Mixed</td>
<td>$\tau_1 \sim N(1.1)$ (delay)</td>
<td>$f_k \sim I_{C_3}(x_k)$</td>
<td>Switches to LC</td>
</tr>
<tr>
<td>$t_4$</td>
<td>Symbolic</td>
<td>$\tau_2 \sim N(24,1)$ (delay)</td>
<td>–</td>
<td>Activates OI</td>
</tr>
<tr>
<td>$t_5$</td>
<td>Symbolic</td>
<td>$\tau_3 \sim N(24,1)$ (delay)</td>
<td>–</td>
<td>Concludes inspections</td>
</tr>
</tbody>
</table>

Observe from these results that there is a period required by the PPN model to learn from the data, which corresponds to the loading cycles in the interval $(0, 5 \cdot 10^3]$. After this learning period, not only does the precision of the prediction of $x_k$ clearly improve with time (predicted values of $x_k$ closer to data $y_k$), but also the uncertainty of the prediction gradually tends to diminish, which is a numerical evidence of the Bayesian learning taking place in $p_1^{(N)}$, and therefore, an evidence of the self-adaptiveness of the PPN from monitoring data. Correspondingly, the activated inspections (represented in $p_2^{(S)}$) are adaptively triggered based on the updated degradation variable, following the rules given in Table 2.

5 Conclusions

This paper provided an overview of the modelling capabilities of the recently proposed Plausible Petri nets (PPNs), with special emphasis on their ability to deal with uncertain information and self-adaptation in complex systems modelling. In particular, an engineering case study about railway track asset management was provided to illustrate the real-world problems that can be modelled using PPNs. The results revealed how uncertain information and condition monitoring data can be integrated using PPNs to support autonomous and adaptive decisions about infrastructure inspection and maintenance. This example reveals that smart infrastructure asset management can be achieved using novel computational tools like PPNs, which has the potential to not only reduce the expenditure of a country in infrastructure asset management, but also to dramatically change the way the infrastructures are designed and managed.
Modelling adaptive systems using plausible Petri nets

Figure 4: Left: Plot of mean values and probability bands of $f^p_k$ for $k = 0 \rightarrow 75,000$. Right: History plot of the differential entropy of $f^p_k$. The dashed-horizontal line represents the threshold value (-4.8) given to activate transition $t_2$.

Acknowledgments
The authors would like to thank the EPSRC and RSSB who jointly and equally support grant EP/M023028/1 "Whole-life Cost Assessment of Novel Material Railway Drainage Systems", and also the Lloyd’s Register Foundation which partially provides support to this work. RSSB is a rail industry body. Through research, analysis, and insight, RSSB supports its members and stakeholders to deliver a safer, more efficient and sustainable rail system. The Lloyd’s Register Foundation is a charitable foundation in the U.K. helping to protect life and property by supporting engineering-related education, public engagement, and the application of research.

References


Determination of task temporal variations in construction scheduling using imprecise probability

Mehdi Modares,1* Rafi L. Muhanna,2 and Robert L. Mullen3

1Department of Civil, Architectural and Environmental Engineering, Illinois Institute of Technology, USA
2School of Civil and Environmental Engineering, Georgia Institute of Technology, USA
3Department of Civil and Environmental Engineering, University of South Carolina, USA

*Corresponding author: mmodares@iit.edu

Abstract

In Critical Path Method (CPM) and Program Evaluation and Review Technique (PERT) for construction scheduling and monitoring, it is crucial to determine the allowable temporal variation for each task such that the project’s total duration is not significantly exceeded. Using the task durations, the overall project completions as well as the allowable float in starting each task are calculated such that the project’s overall completion date is not changed. However, as each event possesses uncertainties, deterministic schemes are not capable of quantification of those uncertainties. Moreover, the traditional probabilistic approaches require a significant level of available data to correctly obtain the probability distribution function for each task. As the available data is limited, these approaches may result in errors when estimating those variations.

In this work, a new formulation of CPM and PERT based on the concepts of imprecise probability is introduced. P-box formulation is employed to represent the imprecise probability; see Williamson and Downs (1990) [1], Ferson et al. (2003) [2]. This method calculates bounds on duration for each task through definition of an imprecise probability structured objective function leading to calculation of bounds on the total project duration. An example problem illustrating the application of this method is presented and compared to the use of calculations using more conventional probabilistic and interval approaches.

Keywords: Imprecise Probability, Construction Scheduling, Critical Path Methods

1 Introduction

The Critical Path Method (CPM) is an approach worldwide utilized for planning and monitoring of both industrial and construction projects. Using conventional CPM, the overall duration of a project is determined. Moreover, the project’s critical tasks or activities are identified. The critical tasks are activities whose temporal variations (float times) may increase and affect the overall project duration.

However, the analysis performed in conventional CPM is deterministic and does not consider the uncertainties in the activities’ durations. The early methods that consider the presence of uncertainty in project scheduling are mostly probabilistic methods which require a significant level of accuracy in the selection of the Probability Density Function (PDF). For instance, in Program Evaluation and Review Technique (PERT) method, the project durations are described in terms of three values, the minimum, expected and maximum duration (Moder et al. [3]). This method has an underlying probabilistic assumption of a Beta distribution for each activity which may exhibit complexities regarding the notion of skewness. Cottrell [4] assumed a normal distribution for each activity. This assumption leads to only two time parameters for each activity, the most likely time and the pessimistic time. However, this reduction in parameters results in an assumed symmetric distribution for each activity that may not be supported by measured data.

In recent years, iterative probabilistic (e.g. Monte Carlo simulations) and possibilistic (e.g. fuzzy) methods with general probability distributions given for the duration of each activity have been explored (Lu and AbouRizk [5]). In a Monte Carlo simulation approach for project scheduling, the probability distribution of project duration and the overall risk associated in meeting a scheduled project completion is determined. Using this approach, the duration of each activity is realized from a random number generated from the probability distribution prescribed for the activity and a deterministic CPM analysis completed. The procedure in repeated a significant number of times and the probability distribution for the project completion is computed. In most Monte Carlo simulations, the random variables associated with each activity are assumed to be independent. This may not be a realistic assumption because of the inherent dependency among project activities. For example, when production rates on an activity are adversely impacted by a weather event, the activity durations may be positively correlated. On the other
hand, when there is a space constraint associated with two work teams in the same area, the activity 
durations may be negatively correlated.

In general, determination of the proper probability distribution for each activity requires large amounts 
of multi-activity data to construct the overall multidimensional probability distribution. An alternative 
to having sufficient data to construct the “correct” distribution is to estimate a range of distributions 
that represent possible activity duration. One of such methods for handling uncertainty in a system 
with no assumptions of a well-defined PDF is imprecise probability theory. Imprecise probability theory 
addresses the extraction of useful information about a process in the case that the probability distribution 
function (PDF) of system inputs or parameters may not be known. Contrasting to fuzzy set theory, as 
more information about the PDF is acquired, imprecise probability methods can sharpen its description 
of parameters; in the limit, of known PDF, imprecise probability reduces to conventional probability.

In this paper, the notion of imprecise probability theory is used to describe the duration of activities 
in a project. In particular, the activity duration is represented as a probability-box (pbox). A pbox is a 
set of cumulative distribution functions (CDF), the set of distributions is given by an upper and lower 
bounding CDFs are represented. Following that, the bounds on the total project duration are calculated.

1.1 Overview
This work is a symbiosis of two historically independent fields, project scheduling (using stochastic analy-
sis), and the imprecise probability theory (based on P-box). Prior to introducing the developed method, 
descriptions of both fields are presented. Following that, the developed methodology for project scheduling 
using imprecise probability is introduced. The methodology is then illustrated using numerical examples.

2 Previous Applications of Stochastic Methods to CPM and PERT
To accurately perform project scheduling and planning, especially for a large-scale project, the time 
duration of each activity must be estimated with sufficient reliability. In order to achieve this reliability, 
numerous probabilistic methods have been utilized to represent and quantify the temporal uncertainty 
of each activity. One of those methods is to use Subjective Beta Distribution Fitting based on the 
available data for each activity (AbouRizk et al. [6]). Using this method, a beta distribution is fitted to 
an activity thru statistical approaches. In case the activities’ end-points are known, the two distribution 
shape parameters are obtained using moment matching or maximum likelihood approaches; whereas, if 
the activities’ end-points are unknown, the method of least squares can be used as the aforementioned 
methods may fail (AbouRizk et al. [7]).

In case there are not sufficient data available, subjective approaches (e.g. expert opinion) can be used. 
Using these approaches, the two shape parameters of the beta distribution can be calculated thru five 
subjective estimation methods. These five methods are based on using: 1) mean and variance, 2) mean 
and mode, 3) mode and variance, 4) mode and an arbitrary percentile, and 5) two arbitrary percentiles. 
The probability parameter (mean and variance) obtained from this method leads to accurate estimation of 
the PDF shape parameters that can be used in simplified CPM/PERT simulations (AbouRizk et al. [6]).

Another probabilistic approach for representing and quantifying uncertainty in fitting the beta distribu-
tion has been thru using “Three Time Estimates” that determine the activity duration used (McLaugh-
lin and Pickhardt [8]). Those estimates include: 1) the “most optimistic” time, i.e., minimum time needed 
to perform an activity, 2) the “most pessimistic” time, i.e., maximum time needed to perform an activity, 
and 3) the “most likely” time i.e., normal time needed to perform an activity (occurrence estimation is 
based on the results from sufficient repetition).

3 Imprecise Probability Structures Based on P-boxes
3.1 Continuously Bounded P-boxes
Consider $F(x)$ as the Cumulative Distribution Function (CDF) for the random variable $X$. When the 
distribution parameters are uncertain, for every $x$, an interval $[\underline{F}(x), \overline{F}(x)]$ generally can be found to 
bound the possible values of $F(x)$, i.e., $\underline{F}(x) \leq F(x) \leq \overline{F}(x)$. Such a pair of two CDFs $\underline{F}(x)$ and $\overline{F}(x)$ 
construct a so-called probability box or probability bounds Ferson et al. [2].

Figure 1 shows the probability box for a normal distribution with an interval mean of $[2.0, 3.0]$ and a 
standard deviation of 0.5. In this simple example, it is easy to verify that $\underline{F}(x)$ is the CDF of the normal 
variable with a mean of 3 and $\overline{F}(x)$ is the one with a mean of 2. Probability box represents a general 
framework to represent imprecisely specified distributions. It can represent not only distributions with 
unknown parameters, but also distributions with unknown type or even unknown dependencies.
3.2 Discrete Bounded P-boxes

An alternative to continuously bounded P-boxes, is the use of discrete, interval based P-box structures as a data type and a library of extensions of standard arithmetic operators on such structures. A discrete P-box structure consists of a collection of interval values, each of which has an associated probability.

In a uniform discretization, the associated probability for each interval is the same. For example, consider a P-box defined by bounding lognormal distribution with interval mean of [2.47, 11.08] and interval standard deviation of [2.76, 12.38]. In Figure 2, the continuous P-box and the interval discretization of the P-box into 10 intervals is shown for these bounding lognormal distributions. While the continuous distributions are not bounded to the right, the discrete P-box is truncated when the cumulative distribution is greater than a specified tolerance. In this example, a tolerance of 0.01 is used (Figures 2 and 3).

While other researchers have used non-conservative mid-point discretization as shown in Figure 3, this paper uses a discretization that encloses the original P-box (Figure 2). The arithmetic of discrete P-Box structures is discussed in the work of Ferson et al. [2], Williamson and Downs [1], who gives detailed description of algorithms for arithmetic operators with either the assumption of independence between variables or the consideration of any-dependency between variables. Copulas can be employed to provide for other dependency (bounds) on random variables. (Ferson and Hajagos. [9]).
4 Numerical Example - Simple Network for Building Construction
4.1 Problem Definition - Simple Network for Building Construction
This example calculation is based on a model from Ahuja et al. [10] (Chapter 15) which was used initially to illustrate a PERT analysis (Table 1). Figure 4 depicts the activity information schematically.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Description</th>
<th>Start Node</th>
<th>End Node</th>
<th>Min Duration</th>
<th>Expected Duration</th>
<th>Max Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Order/prefab metal building</td>
<td>0</td>
<td>2</td>
<td>20</td>
<td>22</td>
<td>25</td>
</tr>
<tr>
<td>B</td>
<td>Clear site Weather</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>C</td>
<td>Underground and foundation</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>D</td>
<td>Erect prefab building</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>E</td>
<td>Finish interior</td>
<td>3</td>
<td>4</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>

Figure 4: Schematic depiction for the activity information

4.2 Analytical Results
Table 2 enumerates the results obtained by conventional CPM and PERT methods and probability approaches through confident limited (as reported by Lu et al. [5]), as well as those using imprecise probability structures including the one with no assumption on probability distribution. For probabilistic and Imprecise Probability approaches, the expected duration for each activity is considered to have an independent normal distribution with the mean set to the PERT time and a standard deviation is: (Max time-Min time)/6.
Table 2: Expected overall duration of the project using different methods

<table>
<thead>
<tr>
<th>Analytical Method</th>
<th>Overall Expected Project Duration (Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPM</td>
<td>42</td>
</tr>
<tr>
<td>PERT</td>
<td>43.5</td>
</tr>
<tr>
<td>Activity Time=(Min Time+4*Expected Time+Max Time)/ 6</td>
<td></td>
</tr>
<tr>
<td>Probability Approach</td>
<td>[43.20, 44.10]</td>
</tr>
<tr>
<td>(The 95% confident limited)</td>
<td></td>
</tr>
<tr>
<td>Imprecise Probability</td>
<td>[43.02, 44.21]</td>
</tr>
<tr>
<td>(Independent-30 point enclosure discretization)</td>
<td></td>
</tr>
<tr>
<td>Imprecise Probability</td>
<td>[43.37, 43.83]</td>
</tr>
<tr>
<td>(Independent-100 point enclosure discretization)</td>
<td></td>
</tr>
<tr>
<td>Imprecise Probability</td>
<td>[41.57, 47.49]</td>
</tr>
<tr>
<td>(With any Dependencies)</td>
<td></td>
</tr>
<tr>
<td>Interval Analysis</td>
<td>[37, 62]</td>
</tr>
<tr>
<td>(No assumption on distribution of activity duration,</td>
<td></td>
</tr>
<tr>
<td>“worst-case interval”)</td>
<td></td>
</tr>
</tbody>
</table>

4.3 Observations
The results in Table 2 show that using an imprecise probability structure for event duration yields a series of more objective bounds for the overall duration mean than conventional CPM, PERT, and probabilistic approaches. Moreover, higher levels of discretization narrow the bounds more precisely. Furthermore, consideration of any dependencies yields wider results. Direct implementation of interval time durations with no assumption on the distribution yields the widest results compared to imprecise probability approach.

5 Summary and Conclusions
In this work, a new formulation for construction scheduling based on the concept of imprecise probability is introduced. This method can obtain the bounds for the mean of overall project duration with consideration of uncertainties in the values and distribution for duration of each activity. The results obtained from this method have a higher level of confidence and robustness due to objective evaluation of variations in the parameter distributions. This objectivity makes it attractive to introduce imprecise probability concepts in the field of construction scheduling and management.

References
[8] F. S. McLaughlin and R. C. Pickhardt, Quantitative techniques for management decisions (Houghton
Mifflin, 1979).


Building envelope inspection management using a hybrid statistical random sampling and Bayesian updating approach

M. Modares1* and J. Mohammadi1

1Department of Civil, Architectural and Environmental Engineering, Illinois Institute of Technology, USA
*Corresponding author: mmodares@iit.edu

Abstract

Condition assessment and inspection planning are crucial for preserving existing buildings, especially those with historical values. Building envelopes are especially vulnerable to damage; and their inspection is an important part of the overall building condition assessment. However, often the existence and extent of damage to envelopes may not be visible in a visual inspection. This will lead to uncertainties in condition assessment and building envelope inspection process and management.

In this work, a new hybrid method for building envelope inspection is introduced that employs statistical random sampling and Bayesian updating to provide a better process for planning for the inspection process. The method is specifically for façades made up of masonry stones, bricks or terra-cotta blocks. For these façade elements, common modes of damage include (1) deterioration of their supporting links to the structure; (2) mortar washout; and (3) construction errors or substandard installation practices. Since the type and extent of damage are not often visible using visual inspections, it is necessary to remove blocks for further evaluation of the hidden areas. The process of opening façade for inspection is costly and involves uncertainty; since an inspector is not sure how many openings will need to be selected and at what locations. A statistical sampling method may be used to better manage the façade inspection process by suggesting the number and locations for the openings. In a prior publication, the authors used this method in a limited application for terra-cotta blocks only. In this paper, the method is extended to other types of façades and also in selecting the locations where there is a higher probability of hidden damage.

In this method: (1) a probability distribution for the proportion of damaged blocks is assumed; (2) a limited number of blocks is then randomly selected for removal to determine their damage condition (3) based on the results of this initial investigation, a Bayesian approach is used for updating information on the proportion of damage blocks and modifying the damage probability distribution, (4) a damage index is then introduced with consideration for the probability of proportion of damaged blocks and the probability that damage areas are within specific areas on the façade; and (5) this index is then adjusted based on the spatial probability of damage in different areas on the façade. The index is used as a decision-making tool in planning for façade inspection and condition assessment. The applicability of this approach in building envelope health condition assessment and planning for future inspections is also presented.

Keywords: Inspection Planning, Sampling, Bayesian Updating, Building Envelope

1 Introduction

Structural condition assessment for building envelopes is considered an important part of the structures overall health monitoring. This is especially important for buildings with historical values that require preservations. Depending on their types, building envelopes perform a variety of functions each requiring a different kind of design expertise (see Fig. 1). The types of damage experienced by building envelopes are unique; and as such, the condition assessment of envelopes requires special attention and procedures that are suitable to address their damage potentials. In US, various municipalities have specific rules on how the inspection must be conducted and prescribe a minimum percentage of area coverage that would suffice an adequate level of inspection. As reported in Mohammadi and Modares [1], these requirements are broad; and most often only provide guidance in general with some additional information regarding the level and extent of examination that is needed. Depending on the rules governed by a municipality, the inspection may only be cursory (using afar examination) or rigorous that may involve close examination with or without additional testing by a licensed structural engineer or architect. Since the requirements on inspection are varied across municipalities, there are not very many well-published studies that have covered the subject including the coverage on (1) the long-term effectiveness and shortcomings of the visual inspection on condition assessment; and (2) suggested procedures for optimizing the process especially when the visual inspection must accompany additional tests or localized detailed examination of target

8th International Workshop on Reliable Engineering Computing, “Computing with Confidence”
16–18 July 2018, University of Liverpool, Liverpool, UK
areas on a façade. Among limited available literature on the subject, specific details on the evolution of building envelope inspections in the City of Chicago are covered by Kurzydlo and Campisi [2] and Chin and Gerberding [3]. It is noted that Chicago is considered the first municipality in the United States mandating inspection requirements for buildings to safeguard the public from hazard of building envelope damage and deterioration. Thomasen and Searls [4] comment on the final estimation of a façade health condition and indicate a combination of several activities such as observations, test results, and analyses of the system and the materials as the basic components of the process.

In case of façades made up of masonry stone, bricks and terra-cotta blocks, in addition to visual inspections, depending on the overall health condition of the system, the detailed investigation will require removing blocks and examining hidden areas. This is because there are damage scenarios that are hidden and not easily visible during visual inspections. Buildings with historical values in most part possess envelopes made up of those façade elements. As such, some may involve hidden damage areas that are simply because of old construction practices. However, the process of opening a façade for inspection is costly and involves uncertainty. This is because there are many causes for damage; and damage occurs at random at different locations on the façade. An inspector does not often have the exact knowledge of how many openings will need to be selected, and at what locations, to have a reasonable coverage of damage areas and to make certain a good idea on the extent of damage to the façade has been obtained. Terra-cotta blocks are especially critical in façade condition assessment. This is because: (1) they have architectural values used in façades mainly in early twentieth century; and (2) they are not being widely used in today’s practice, and thus any repairs must utilize old terra-cottas already in place or custom-made new ones through a few manufacturers that are still in operation. Mohammadi and Modares [1] provide an overview of terra-cotta façade elements and their compositions and installation practice.

An efficient inspection process requires an adequate number of openings that can be made to provide a reasonable estimate of the overall condition of the façade with minimum cost. Specific to terra-cotta blocks and the mode of failure associated with corrosion in metal anchors that are used to attach blocks to the structural system, Mohammadi and Modares [1] demonstrated that a random sampling method along with a Bayesian updating approach may be employed for decision making in regard to the selection of openings for inspections. The specific parameter used for updating is the proportion of damaged blocks within a façade. The process arrives at probability levels for this parameter, which is then updated as new inspection results are obtained. They also introduced a damage index based on these probability levels, which can be used as a decision-making tool in determining the state of the health of the façade and planning for further evaluation of the condition of blocks.

Using the model introduced by Mohammadi and Modares [1] as a basis, this paper extends the model to (1) other types of façades such as masonry stone and bricks, (2) applications where specific areas within the façade will need to be prioritized for inspection because of a higher probability of hidden damage; and (3) modify the damage index and adjust it for areas with higher probability of hidden damage within a façade.

2 Overview

Along with development of the general method of building envelope inspection, this paper provides an overview of façade hidden damage types and their causes, the application of the aforementioned statistical sampling methods and model formulations, and illustrative examples to demonstrate the applicability of the model.
3 Decision Making and Planning for Inspection

In current practice, when detailed investigation of the condition of façade blocks is required, the location and number of openings on the façade is decided by the engineer or architect of the record primarily based on (1) his/her experience; (2) urgency for inspection and target timetable; and (3) budget. There are no specific rules on what percentage of façade blocks must be removed to have a sufficient sample size and to establish confidence on the overall condition of the façade. Unless there are some indications for damage-prone blocks, the location of the openings is selected at random, because of the uncertainty involved in the damage locations. There is even more uncertainty in regard to how many openings must be selected. In the simplest form, the planning includes using a prescribed percentage value for the number of openings. However, it is not uncommon to notice that none of the openings selected indicate any damage. And in this situation, it is not clear on how to proceed with the final conclusion on the condition of the façade. Is it fair to indicate that the entire façade has no damage? Obviously one cannot state this with much confidence, especially if the façade is part of an aging structure using older construction methods and materials. At the same time, one cannot conclude that there is still damage within the system, since the samples collected do not support this notion. These uncertainties may end up with erroneous decision-making, which will result in the loss of effort and resources.

4 Previous Work

To better manage the process of selecting the number of openings, Mohammadi and Modares [1] suggest using a statistical sampling method and the Bayesian updating approach. This model is exclusively used for damage due to corrosion of metal anchors in terra-cotta blocks. They use the proportion of the number of damaged blocks \( p \) as a random variable, which also represents the parameter in the binomial distribution model (Khisty, Mohammadi and Amedkudzi [5]). This can be used to estimate the probability of damage among blocks. In principle, the main trust of the method is to use the outcome of the investigation on a limited number of blocks as the "prior probability" values in the Bayesian process to update the results and arrive at "posterior probability" values for the variable \( p \). The initial probability values for \( p \) can be obtained based on experts opinion, is outlined below.

1. Start with an initial probability distribution for the parameter \( p \) (prior probability).
2. Select a limited number of openings at random for inspection.
3. Use the Bayesian method and obtain posterior probabilities for \( p \).

The authors further introduce the idea of using a "damage index" as a means to utilize the results of the analysis for decision-making and planning for the follow-up steps in the process.

**General formulation of Bayesian updating applied to façade inspection**

The formulation of the Bayesian approach in information updating is governed by the following equation (see for example Ang and Tang [6]).

\[
P''(p = p_i) = \frac{P(\varepsilon | p = p_i)P'(p = p_i)}{\sum_i P(\varepsilon | p = p_i)P'(p = p_i)}
\]

(1)

where:

- \((p = p_i)\) is the event that the proportion of damaged blocks is a value equal to \( p_i \).
- \(P'\) = existing information on the proportion of damaged blocks (prior probability). The parameter \(P'\) represents the probability that the proportion of damaged blocks is equal to \( p_i \) (i.e., the probability that \( p = p_i \)) based on the available information.
- \(\varepsilon\) = new information on the number of damaged blocks discovered when detailed investigations using openings in the façade are made.
- \(P''\) = updated information on the proportion of damaged blocks (posterior probability). It is noted that \(P''\) represents the probability that the proportion of damaged blocks is equal to \( p_i \) (i.e., the probability that \( p = p_i \)) based on new information just acquired. The prior probability values \((P')\) can be prescribed using the inspector’s notion and past experience. In case no prior information on \(P'\) is known, an arbitrary distribution model may be used. This work suggests using a uniform or triangular distribution. The new data (\(\varepsilon\)) is described by the number of damaged blocks identified during detailed investigations.
The conditional probability $P(\varepsilon \mid p = p_i)$ describes the outcome to be $\varepsilon$, if the proportion of damaged metal anchors is $p_i$. With a binomial distribution function, the term $P(\varepsilon \mid p = p_i)$ can be written as:

$$P(\varepsilon \mid p = p_i) = \frac{k!}{q!(k-q)!} (p_i)^q (1-p_i)^{k-q} \tag{2}$$

in which $q$ is the number of damaged blocks in a total of $k$ openings.

It is noted that using the updated probabilities described by Eqs. (1) and (2), the estimate of the mean for the proportion of damaged blocks ($\hat{p}''$), based on the posterior probabilities, can be obtained as:

$$\hat{p}'' = \sum_{i=1}^{n} p_i P''(p = p_i) \tag{3}$$

this value is used as a damage index for decision making on the health condition of the façade as described below:

1. For cases where $\hat{p}'' < 0.05$, the façade is considered to be in good condition. This means that on the average, less than 5% of blocks are damaged.

2. For cases where $0.05 < \hat{p}'' < 0.20$, it is recommended that detailed investigations by implementing more openings be considered; and based on the outcome, $\hat{p}''$ be recomputed for decision making.

3. Finally in cases where $\hat{p}'' > 0.20$, still additional openings are recommended. However, it is indicated that if the new outcome still indicates a large value for the damage index ($\hat{p}''$), major repair may be necessary.

It is further noted that $\hat{p}''$ can also be used to determine how many openings are needed for planning purposes, since the expected number of damaged blocks is equal to the total number of blocks on the façade multiplied by $\hat{p}''$.

5 Methodology

In this paper, the model developed by Mohammadi and Modares [1] is further expanded and applied to other types of façades made up of masonry stone and bricks. Although the mathematics of the updating model is still the same as that used for terra-cotta blocks, there are some differences that need to be considered in the final outcome of the model and the decision making process. These differences primarily stem from the type of damage that would lead to additional information in regard to the location and areas where conceivably there is a higher probability that damaged blocks are found. And as such, the estimate obtained for the proportion of damaged blocks will not be the same on the entire façade and will be subject to a spatial probability distribution. It is emphasized the model introduced by the authors in their prior publication assumes the distribution and estimates for $p$ do no vary from one area on the façade to another.

Masonry stone – For masonry stone blocks, the damage scenarios are very similar to the ones mentioned for terra-cotta blocks. The most critical mode of damage is the corrosion of the metal anchors and deterioration of the spot where the anchor is attached to the stone. The latter causes a loose connection, which is often not visible in a visual inspection unless the damage is widespread on the façade. The method presented for planning and inspection process and the determination of the proportion of damaged blocks are expected to be very similar to the case of terra-cotta blocks.

Bricks – Specific to brick masonry, because of its composition, the way blocks are laid and the type of anchors they used, there may be changes in the appearance or deformations that may indicate a higher probability of damage in some areas on the façade. For example, a persistent prevalence of moisture may appear as efflorescence deposits on brick. Ineffective anchors (because of deterioration and damage) may result in a portion of façade bulging out (see Fig. 2). Although a bulge may not necessarily be resulted from damaged anchors, it can still be considered a “hot spot” on the façade that needs more attention. This indicates that perhaps bricks in the bulged area have a higher probability of anchor damage compared with those located elsewhere on the wall.
6 Significance of Spatial Distribution of Probability of Damage on the Façade

As indicated earlier, specific to brick façades, there may be changes in the appearance of the blocks that may indicate certain areas have a higher probability of damaged bricks than other areas. In a preliminary investigation, the inspection report may note these areas and provide a spatial distribution as a judgmental call on the relative values of the probability of damage. For example, the appearance of a bulge (as shown in Fig. 2) may indicate that the bulge area has a higher probability of possessing damaged blocks. Of course establishing a relative probability for damage for the bulge area depends on the inspector’s notion and intuitive judgment. For example, as shown in Fig. 3 a middle area on the façade (marked as $A_0$) has certain signs of damage; and as such in the opinion of the inspector there is a higher probability that the area contains more damaged blocks.

The inspector may provide his/her own notion (or seek expert opinion) on a relative value for the probability of damaged blocks in areas $A_1$ and $A_2$ compared to that in area $A_0$. Denoting the mean values for the proportion of damaged blocks as $\hat{p}_0$, $\hat{p}_1$ and $\hat{p}_2$, respectively, for the three areas $A_0$, $A_1$ and $A_2$, we can write the following equations.

$$A_0 \cdot \hat{p}_0 + A_1 \cdot \hat{p}_1 + A_2 \cdot \hat{p}_2 = A \cdot \hat{p}''$$

(4)
\[
\hat{p}_1 = a_1 \cdot \hat{p}_0
\]  \hspace{1cm} (5)

\[
\hat{p}_2 = a_2 \cdot \hat{p}_0
\]  \hspace{1cm} (6)

In which \(\hat{p}''\) is the mean value for the proportion of damaged blocks obtained for the entire façade (the value from Eq (3)) and \(a_1\) and \(a_2\) are smaller than 1 and represent, respectively, the relative probability of damaged blocks in the areas \(A_1\) and \(A_2\) compared to that in area \(A_0\). Solving Eqs. (4)-(6), the value for the mean value for the proportion of damaged blocks (\(\hat{p}_0\)) in area \(A_0\) is now adjusted and is estimated as:

\[
\hat{p}_0 = \frac{\hat{p}'' \cdot A}{A_0 + a_1 A_1 + a_2 A_2}
\]  \hspace{1cm} (7)

In deriving Eq. (7), only three areas are considered with one having a higher probability for damaged blocks compared with the other two. In a more general form, assume some \((m+1)\) areas \(A_0, A_1, A_2, \ldots, A_m\) are identified in which the probability of damaged blocks in the area \(A_0\) is the highest among all areas. Furthermore, it is determined that the ratio of the probability of damaged blocks in an areas such as \(A_i\) (where \(i = 1, 2, \ldots, m\)) to that in area \(A_0\) is \(a_i\). Then the proportion of damaged blocks in in the area \(A_0\) can be written in terms of \(\hat{p}''\) in the following form:

\[
\hat{p}_0 = \frac{\hat{p}'' \cdot A}{A_0 + \sum_{i=1}^{m} a_i A_i}
\]  \hspace{1cm} (8)

and

\[
\hat{p}_i = a_i \cdot \hat{p}_0 \text{ (for } i = 1, 2, \ldots, m) \]  \hspace{1cm} (9)

It is further noted that in incorporating the spatial probability distribution for the probability of damage, the areas \(A_i\) (in the special case of Fig. 3 \(i = 0, 1, 2\) and 2) were treated as deterministic values. In a more advanced formulation, these can be treated each as a random variable such that the significance of uncertainties in them can be considered in the formulation.

7 Numerical Illustrations

In a façade inspection, and as initial step, a uniform probability function (prior probabilities) for the proportion of damaged blocks (\(p\)) on the façade is used. With 11 possible values for \(p\) (i.e., \(p_i = 0., 0.1, 0.2, 0.3, \ldots, 1.0\)):

\[
P'(p = p_i) = \frac{1}{11} = 0.091
\]  \hspace{1cm} (10)

for \((i = 0, 1, 2, 3, \ldots, 10)\)

The inspector decides to use 10 openings at random distributed over the entire façade (which has a total area = \(A\)) and removes blocks for investigating whether there are any hidden damage. During these inspections, \(q\) blocks (out of a total of 10) indicate they had hidden damage. If \(q = 0\) (i.e., no damaged blocks were found), using Eqs. (1) and (2), the values for the posterior probabilities, \(P''(p = p_i)\), are computed as summarized in Table 1. Furthermore, using Eq. (3), these values result in a posterior mean value for the proportion of damaged blocks (i.e., \(\hat{p}''\)) = 0.045.

Now assume the inspector observes special conditions on about 50% of the façade, in the middle portion, that indicate there may be a higher probability of damaged blocks. Based on previous experience, the inspector decides that this area has 5 times more possibility to possess damaged blocks. Referring to Fig. 3 thus:

\[
A_0 = 0.5A, \text{ and } A_1 = A_2 = 0.25A
\]
Table 1: Posterior probabilities when new outcome reveals there are no damaged blocks in 10 openings

<table>
<thead>
<tr>
<th>$p_i$</th>
<th>$P''(p = p_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.670</td>
</tr>
<tr>
<td>0.1</td>
<td>0.234</td>
</tr>
<tr>
<td>0.2</td>
<td>0.072</td>
</tr>
<tr>
<td>0.3</td>
<td>0.019</td>
</tr>
<tr>
<td>0.4</td>
<td>0.002</td>
</tr>
<tr>
<td>0.5</td>
<td>0.001</td>
</tr>
<tr>
<td>$\geq 0.60$</td>
<td>0</td>
</tr>
</tbody>
</table>

Furthermore, $a_1 = a_2 = 0.20$

Using Eq. (7),

$$\hat{p}_0 = 1.67 \hat{p}'' = 1.67 \times 0.045 = 0.075$$

and

$$\hat{p}_1 = \hat{p}_2 = 0.33 \hat{p}'' = 0.33 \times 0.045 = 0.015.$$  

These values indicate that even with opening resulting in no damaged blocks, the middle area has an estimated value for the proportion of damaged blocks that is more than 0.05 and thus needs to be investigated further with more openings.

To further explore other possibilities for the number of damaged blocks in a first batch of 10 openings, Table 2 is developed which shows $\hat{p}''$, $\hat{p}_0$, $\hat{p}_1$ and $\hat{p}_2$ for cases of $q = 0, 1, 2, 3, 4$ and $5$. This means upon conducting 10 openings, 1, 2, ... damaged blocks are found. As noticed from results in Table 2, if one or more damaged blocks in 10 openings are found, the estimated proportion of damaged blocks in the middle area $\hat{p}_0 > 0.2$, which indicates decision in favor of repair.

Table 2: Estimated proportion of damaged blocks in different areas of façade of Fig. 3

<table>
<thead>
<tr>
<th>$q$</th>
<th>$\hat{p}''$ (for entire façade)</th>
<th>$\hat{p}_0$ (for area $A_0$)</th>
<th>$\hat{p}_1$ (for area $A_1$)</th>
<th>$\hat{p}_2$ (for area $A_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.045</td>
<td>0.075</td>
<td>0.015</td>
<td>0.015</td>
</tr>
<tr>
<td>1</td>
<td>0.182</td>
<td>0.300</td>
<td>0.060</td>
<td>0.060</td>
</tr>
<tr>
<td>2</td>
<td>0.251</td>
<td>0.419</td>
<td>0.082</td>
<td>0.082</td>
</tr>
<tr>
<td>3</td>
<td>0.333</td>
<td>0.556</td>
<td>0.110</td>
<td>0.110</td>
</tr>
<tr>
<td>4</td>
<td>0.417</td>
<td>0.696</td>
<td>0.138</td>
<td>0.138</td>
</tr>
<tr>
<td>5</td>
<td>0.500</td>
<td>0.835</td>
<td>0.167</td>
<td>0.167</td>
</tr>
</tbody>
</table>

To further demonstrate the significance of updated results, consider using the case of $q = 0$ again. Assume there are some 3,600 blocks on the façade, of which about 1,800 are in the middle area $A_0$ and 900 each at areas $A_1$ and $A_2$ (see Fig. 3). If no updating is used, and with the prior probability values suggested by Eq. (10), the estimated mean value for the proportion of damaged blocks ($\hat{p}'$) is 0.5. Using this value, the estimated number of damaged blocks is $0.5 \times 3,600 = 1,800$, which indicates that an approximately the same number of openings would be needed for further investigations. Obviously, this requires a substantial effort involving considerable costs and resources. Upon updating, after the first round of inspections (with $q = 0$ damaged blocks in a total of 10), the estimated mean for damaged block is reduced to $\hat{p}'' = 0.045$. This in turn reduces the number of required openings to $0.045 \times 3,600 = 162$, which is substantially less than the case without the updating. If there is no notion on target areas for higher probability of damage, these 162 openings will need to be selected randomly on the entire façade. However, if there is a notion that the area $A_0$ has about 5 times higher probability of possessing damaged blocks, compared with those in areas $A_1$ and $A_2$, the mean value for the proportion of damaged blocks in area $A_0$ is adjusted to 0.075 (see Table 2). This requires some $0.075 \times 1,800 = 135$ openings only for area $A_0$ with the rest of openings ($162 - 135 = 27$) considered for the other two areas. By using the adjusted value for the proportion of damaged blocks for area $A_0$, the specific area is targeted that needs more attention by distributing the effort and resources more efficiently.
8 Conclusions

The following presents the main conclusions of the studies presented in this paper.

1. The Bayesian updating introduced in a prior publication by the authors for planning and management of terra-cotta façade inspection can also be extended to façades made up of masonry stone and bricks.

2. Specific to the brick façade application, certain condition of bricks may indicate a higher probability for damage. This notion can be considered for developing a spatial distribution for damage in different areas on the façade.

3. The information on spatial distribution for the probability of damage can be used in adjusting the estimate for the proportion of damaged blocks. The results may be helpful in planning a more refined schedule for further inspections and determination of areas that need to receive more attention when detailed investigations through openings on the façade will be necessary.

References


When is propagation of interval and fuzzy uncertainty feasible?

Vladik Kreinovich,1* Andrew M. Pownuk,2 Olga Kosheleva,3 and Aleksandra Belina4

1 Department of Computer Science, University of Texas at El Paso, USA
2 Department of Mathematical Sciences, University of Texas at El Paso, USA
3 Department of Teacher Education, University of Texas at El Paso, USA
4 Department of Building Structures, Silesian University of Technology, Poland
* Corresponding author: vladik[at]utep.edu

Abstract

In many engineering problems, to estimate the desired quantity, we process measurement results and expert estimates. Uncertainty in inputs leads to the uncertainty in the result of data processing. In this paper, we show how the existing feasible methods for propagating the corresponding interval and fuzzy uncertainty can be extended to new cases of potential practical importance.

Keywords: Interval uncertainty, fuzzy uncertainty, uncertainty propagation, feasible algorithms

1 Introduction

Need for data processing. In many practical situations, we are interested in a quantity \( y \) which is difficult (or even impossible) to measure directly.

For example, when we design a structure – a building, an airplane, etc. – we would like to know the maximal force that can be applied without breaking this structure. For a building, directly measuring this force by trying to topple the building would be too expensive.

Another example is prediction: we want to predict the mechanical properties of an airplane several years from now, without having to wait these several years.

In all such cases, what helps is that we usually know the relation \( y = f(x_1, \ldots, x_n) \) between the desired quantity \( y \) and some easier-to-measure and/or easier-to-estimate quantities \( x_1, \ldots, x_n \). This dependence is sometimes given in terms of an explicit expression, but more often, it is given as an algorithm for computing \( y \) based on \( x_1 \) – e.g., for mechanical or thermal properties, as the algorithm for solving the corresponding partial differential equations.

Then, after measuring and/or estimating \( x_i \), we apply the algorithm \( f \) to the measurement/estimation results \( \tilde{x}_i \), and get an estimate \( \tilde{y} = f(\tilde{x}_1, \ldots, \tilde{x}_n) \) for \( y \). Such an application of the algorithm \( f \) is an important case of data processing.

Need for propagating uncertainty. Neither measurement nor expert estimates are absolutely precise. As a result, the estimates \( \tilde{x}_i \) are somewhat different from the actual (unknown) values \( x_i \). So, even if the algorithm \( f \) is precise, the result \( \tilde{y} = f(\tilde{x}_1, \ldots, \tilde{x}_n) \) of applying this algorithm to the estimates \( \tilde{x}_i \) is, in general, different from the value \( y = f(x_1, \ldots, x_n) \) that we would have gotten if we knew the exact values of the input quantities.

It is therefore desirable to estimate the corresponding “propagated” uncertainty \( \Delta y \overset{\text{def}}{=} \tilde{y} - y \); see, e.g., [21]. How big can this approximation error be?

This is very important in many practical situations. For example, if we are designing a chemical plant, and, based on our computations, we conclude that the predicted level of undesired chemicals in the air \( \tilde{y} \) will not exceed the required threshold \( y_0 \), this does not necessarily mean that we can start building this plant: it all depends on how accurate this estimate is. If \( \tilde{y} \) is slightly smaller than \( y_0 \), but the approximation error \( \Delta y \) can be large, this means that there is a real possibility that the plant will not be functioning safely.

Similarly, if, based on the measurement results, we predict that the building will withstand the earthquake of given magnitude, this does not necessarily mean that people in this building will be safe: it all depend on how accurate were our predictions.

In all such situation, it is important to estimate the approximation error \( \Delta \), i.e., to propagate the uncertainty of measuring/estimating \( x_i \) through the data processing algorithm \( f \).

Need for interval uncertainty. The desired approximation error \( \Delta y \) comes from the measurement/estimation errors \( \Delta x_i \overset{\text{def}}{=} \tilde{x}_i - x_i \). Thus, to estimate \( \Delta y \), we need to have information about the error \( \Delta x_i \).
Ideal case is when for each $i$, we know which values $\Delta x_i$ are possible and what is the probability of getting each possible values - i.e., when we know the probability distribution for each $\Delta x_i$. In principle, we can get this probability distribution if we calibrate the corresponding measuring instrument - i.e., compare its results with the results of a much more accurate ("standard") measuring instrument. However, calibration is a very expensive procedure. As a result, in many engineering applications, we do not know these probabilities. At best, we know the upper bound $\Delta_i$ on the corresponding measurement errors: $|\Delta x_i| \leq \Delta_i$. In this case, based on the measurement result $\tilde{x}_i$, all we can conclude about the actual (unknown) value $x_i$ is that this value is somewhere on the interval $[\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]$. This situation is therefore known as a situation of interval uncertainty; see, e.g., [11, 16, 17].

Different values $x_i$ from the corresponding intervals lead, in general, to different values $y$. It is therefore desirable to find the range of all possible values $y$, i.e., the interval

$$\lbrack y_\text{min}, y_\text{max} \rbrack = \{ f(x_1, \ldots, x_n) : x_i \in [\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i] \text{ for all } i \}.$$  

The problem of computing this range is known as the main problem of interval computations.

It is known that, in general, this problem is NP-hard; see, e.g., [14]. This means that, unless $P=NP$ (which most computer scientists believe to be impossible), it is not possible to have a feasible algorithm that exactly computes the range $\lbrack y_\text{min}, y_\text{max} \rbrack$ for all possible functions $f(x_1, \ldots, x_n)$. It is therefore desirable to find cases when feasible algorithms are possible.

**Need for fuzzy uncertainty.** Often - e.g., for expert estimates - we do not even have a guaranteed upper bound. Instead, we have several bounds with different levels of certainty. As a result, instead of explicitly saying which values of $\Delta x_i$ are possible and which are not - as in the case of interval uncertainty - we have, in effect, for each possible value $\Delta x_i$, a degree to which, according to the experts, this value is possible. In mathematical terms, this means that we have a function $\mu$ that maps every possible value of $\Delta x_i$ into a degree $\mu(\Delta x_i)$ from the interval $[0, 1]$, so that:

- 1 corresponds to full confidence, and
- 0 corresponds to complete absence of confidence.

Such a function is known as a fuzzy set, and the corresponding uncertainty is known as fuzzy uncertainty; see, e.g., [1, 12, 19, 24].

**What is known and what we do in this paper.** Due to the fact that the problem of propagating interval and fuzzy uncertainty is of great practical importance, there exist many techniques for such a propagation. In this paper, we list some of the known techniques - and show how these known techniques can be extended to new cases of potential practical importance.

**2 Linearized Case: Interval Uncertainty**

**Possibility of linearization.** In many practical situations, the measurement errors $\Delta x_i = \tilde{x}_i - x_i$ are small. In such situations, we can expand the expression

$$\Delta y = \tilde{y} - y = f(\tilde{x}_1, \ldots, \tilde{x}_n) - f(x_1, \ldots, x_n) = f(\tilde{x}_1, \ldots, \tilde{x}_n) - f(\tilde{x}_1 - \Delta x_1, \ldots, \tilde{x}_n - \Delta x_n)$$

in Taylor series in $\Delta x_i$ and ignore quadratic and higher order terms in this expansion. As a result, we get the following expression:

$$\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i,$$  \hspace{1cm} (1)

where

$$c_i \stackrel{\text{def}}{=} \frac{\partial f}{\partial x_i} |_{x_i=\tilde{x}_i}.$$  

**Known result: in the linearized case, the main problem of interval computation is feasible to solve.** In the linearized case (1), the dependence of $\Delta y$ on $\Delta x_i$ is linear. Each variable $\Delta x_i$ takes all possible value from $-\Delta_i$ to $\Delta_i$.

It is easy to see that:

- when $c_i > 0$, then $\Delta y$ is an increasing function of $\Delta x_i$ and thus, its largest possible value is attained when $\Delta x_i$ takes its largest possible value $\Delta x_i = \Delta_i$;
• when \( c_i < 0 \), then \( \Delta y \) is a decreasing function of \( \Delta x_i \) and thus, its largest possible value is attained when \( \Delta x_i \) takes its smallest possible value \( \Delta x_i = -\Delta_i \).

In both cases, the optimizing value of \( \Delta x_i \) is equal to \( \text{sign}(c_i) \cdot \Delta_i \), where:

- \( \text{sign}(c_i) = 1 \) if \( c_i > 0 \), and
- \( \text{sign}(c_i) = -1 \) if \( c_i < 0 \).

For these optimizing values, each term \( c_i \cdot \Delta x_i \) takes the form \( c_i \cdot \text{sign}(c_i) \cdot \Delta_i \). One can easily see that \( c_i \cdot \text{sign}(c_i) = |c_i| \), so the largest possible value of \( \Delta y \) is equal to

\[
\Delta = \sum_{i=1}^{n} |c_i| \cdot \Delta_i. \tag{2}
\]

Similarly:

• when \( c_i > 0 \), then \( \Delta y \) is an increasing function of \( \Delta x_i \) and thus, its smallest possible value is attained when \( \Delta x_i \) takes its smallest possible value \( \Delta x_i = -\Delta_i \);

• when \( c_i < 0 \), then \( \Delta y \) is a decreasing function of \( \Delta x_i \) and thus, its smallest possible value is attained when \( \Delta x_i \) takes its largest possible value \( \Delta x_i = \Delta_i \).

In both cases, the optimizing value of \( \Delta x_i \) is equal to \( -\text{sign}(c_i) \cdot \Delta_i \). For this optimizing values, each term \( c_i \cdot \Delta x_i \) takes the form \( -|c_i| \cdot \Delta_i \), so the smallest possible value of \( \Delta y \) is equal to

\[
-\sum_{i=1}^{n} |c_i| \cdot \Delta_i = -\Delta.
\]

The value \( \Delta y \) thus ranges from \( -\Delta \) to \( \Delta \), so the range of possible values of \( y = \tilde{y} - \Delta y \) is the interval \([\tilde{y} - \Delta, \tilde{y} + \Delta]\).

Computing \( \Delta \) by using the expression (2) is definitely feasible: once we know \( c_i \) and \( \Delta_i \), this can be done by a simple linear-time algorithm.

**Need to consider constraints.** The above argument assumes that all possible combinations of values \( x_i \) are possible. In practice, we often have some dependence between these quantities.

In some cases, we have constraints like energy conservation, according to which some combination of the variables \( x_i \) is equal to a known value \( g(x_1, \ldots, x_n) = g_0 \). Substituting \( x_i = \bar{x}_i - \Delta x_i \) into this formula, we get

\[ g(\bar{x}_i - \Delta x_1, \ldots, \bar{x}_n - \Delta x_n) = g_0. \]

Since we assume that the measurement errors \( \Delta x_i \) are small – and thus, that their squares can be safely ignored – we can linearize this constraint and get a linear constraint of the type

\[ \sum_{i=1}^{n} g_i \cdot \Delta x_i = G_0, \tag{3} \]

where

\[ g_i \stackrel{\text{def}}{=} \frac{\partial g}{\partial x_i}_{x_i = \bar{x}_i} \quad \text{and} \quad G_0 = g_0 - g(\bar{x}_1, \ldots, \bar{x}_n). \]

In other cases, we have inequality constraints. For example, the second law of thermodynamics states that the entropy at the next moment of time must be smaller than or equal to the entropy at the previous moment of time. In such cases, we have constraints of the type \( h(x_1, \ldots, x_n) \geq h_0 \), which after linearization turn into linear inequalities

\[ \sum_{i=1}^{n} h_i \cdot \Delta x_i \geq H_0, \tag{4} \]

where

\[ h_i \stackrel{\text{def}}{=} \frac{\partial h}{\partial x_i}_{x_i = \bar{x}_i} \quad \text{and} \quad H_0 = h_0 - h(\bar{x}_1, \ldots, \bar{x}_n). \]

In such situations, instead of finding the range of all possible values of \( f(x_1, \ldots, x_n) \), we are interested in the range over all possible tuples \((x_1, \ldots, x_n)\) that satisfy the given constraints.
It turns out that in this case, we still have a feasible algorithm — although not as fast as in the no-constraints case.

**First new (simple) result: in the linearized case, computing the range under constraints is also feasible.** Indeed, in this case, we need to find the largest and the smallest possible values of a linear expression

\[ \Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i \]

under linear equalities and inequalities of the type (3) and (4).

The corresponding optimization problems are particular case of the general *linear programming* problem: optimize a linear function under linear constraints. Feasible algorithms are known for solving this problem; see, e.g., [15].

**Comment.** It should be noted, however, that while these algorithms are feasible, they are not as fast as the linear-time algorithm for computing the expression (2) — and there seems to be no hope of speeding them up by using, e.g., parallelization, since linear programming is known to be provably the most difficult to parallelize; see, e.g., [20].

**Second new result: in the linearized case, computing the range under sparsity constraint is also feasible.** Often, a natural constraint on the values is that the tuple \((x_1, \ldots, x_n)\) should be sparse, i.e., that no more than a certain number \((K)\) of values \(x_i\) are different from 0; see, e.g., [2–6, 10].

We are interested in finding the range for \(y = \tilde{y} - \Delta y\). In the linearized case, \(\Delta y\) is described by the formula (1), where \(\Delta x_i = \tilde{x}_i - x_i\). Substituting the definition of \(\Delta x_i\) into the formula (1) and substituting the resulting expression for \(\Delta y\) into the formula for \(y\), we get

\[ y = \tilde{y} - \sum_{i=1}^{n} c_i \cdot (\tilde{x}_i - x_i), \]

i.e.,

\[ y = Y + \sum_{i=1}^{n} c_i \cdot x_i, \]

where we denoted

\[ Y \overset{\text{def}}{=} \tilde{y} - \sum_{i=1}^{n} c_i \cdot \tilde{x}_i. \]

The values \(x_i\) for which \(0 \notin [\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]\) clearly cannot be equal to 0, so when looking for 0 values of \(x_i\) we should look among values for which \(0 \in [\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]\), i.e., for which \(\tilde{x}_i - \Delta_i \leq 0 \leq \tilde{x}_i + \Delta_i\). No more than \(K\) of these values are different from 0.

For each \(i\) for which \(c_i > 0\):  

- the largest value of the corresponding term \(c_i \cdot x_i\) is attained when \(x_i\) attains its largest possible value \(x_i = \tilde{x}_i + \Delta_i\), and  
- the smallest value of the corresponding term \(c_i \cdot x_i\) is attained when \(x_i\) attains its smallest possible value \(x_i = \tilde{x}_i - \Delta_i\).

For each \(i\) for which \(c_i < 0\):  

- the largest value of the corresponding term \(c_i \cdot x_i\) is attained when \(x_i\) attains its smallest possible value \(x_i = \tilde{x}_i - \Delta_i\), and  
- the smallest value of the corresponding term \(c_i \cdot x_i\) is attained when \(x_i\) attains its largest possible value \(x_i = \tilde{x}_i + \Delta_i\).

In both cases:

- the largest value \(y_i\) of the corresponding term \(c_i \cdot x_i\) is attained when \(x_i = \tilde{x}_i + \text{sign}(c_i) \cdot \Delta_i\), so that this largest value is equal to

\[ y_i = c_i \cdot x_i = c_i \cdot \tilde{x}_i + c_i \cdot \text{sign}(c_i) \cdot \Delta_i = c_i \cdot \tilde{x}_i + |c_i| \cdot \Delta_i; \]

and
the smallest value $y_i$ of the corresponding term $c_i \cdot x_i$ is attained when $x_i = \tilde{x}_i - \text{sign}(c_i) \cdot \Delta_i$, so that this smallest value is equal to

$$y_i = c_i \cdot x_i = c_i \cdot \tilde{x}_i - c_i \cdot \text{sign}(c_i) \cdot \Delta_i = c_i \cdot \tilde{x}_i - |c_i| \cdot \Delta_i.$$ 

For terms $i$ for which $0 \notin [\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]$, these are the terms we take to find the largest $\overline{y}$ and the smallest $y$ values of $y$.

For terms $i$ which could be 0:

- to find $\overline{y}$, we take the $K$ largest of these terms, and
- to find $y$, we take $K$ smallest of these terms.

Thus, we arrive at the following algorithm for computing $y$ and $\overline{y}$. In this algorithm, we separate indices $i$ for which $i \in [\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]$ as possible zeros and all other indices as definitely non-zeros.

The algorithm is as follows:

- first, for all indices $i$, we compute $\overline{y}_i = c_i \cdot \tilde{x}_i + |c_i| \cdot \Delta_i$ and $y_i = c_i \cdot \tilde{x}_i - |c_i| \cdot \Delta_i$;
- to compute $\overline{y}$ to $Y$, we add all the values $\overline{y}_i$ for all definitely non-zero $i$ and $K$ largest of the terms $\overline{y}_i$, corresponding to possible zeros;
- to compute $y$ to $Y$, we add all the values $y_i$ for all definitely non-zero $i$ and $K$ smallest of the terms $y_i$, corresponding to possible zeros.

This algorithm is clearly feasible: it takes linear time + time $O(n \cdot \log(n))$ for sorting; see, e.g., [8].

3 Linearized Case: Fuzzy Uncertainty

Towards a precise formulation of the problem. In fuzzy approach, instead of simply indicating which values are possible and which are not, we have experts’ degrees indicating to what extent each value is possible.

Once we know the corresponding degrees $\mu_i(\Delta x_i)$ for different possible values of the measurement errors $\Delta x_i$, we need to combine them into a degree $\mu(\Delta y)$ indicating to what extent each value $\Delta y$ is possible.

A value $\Delta y$ is possible if there exist values $\Delta x_1, \ldots, \Delta x_n$ all of which are possible and for which the formula (1) holds. In other words, $\Delta y$ is possible if and only if there exist $\Delta x_1, \ldots, \Delta x_n$ for which (1) is true and for which $\Delta x_i$ is possible and ... and $\Delta x_n$ is possible.

We know the degree $\mu_i(\Delta x_i)$ to which each value $\Delta x_i$ is possible. We get these degrees from the expert(s). Ideally, to find the degree to which $\Delta x_1$ is possible and $\Delta x_2$ is possible, etc., we should also ask experts – but there are so many combinations ($\Delta x_1, \ldots, \Delta x_n$) that asking about all of them is not realistic.

Since we cannot elicit these degree directly, we need to estimate them based on what we know – i.e., based on the degrees to which each $\Delta x_i$ is possible.

This is a typical situation in fuzzy reasoning:

- we know the degree $a$ and $b$ to which statements $A$ and $B$ are true, and
- we want to estimate to what extent their “and”-combination $A \& B$ is true.

Let $f_k(a, b)$ denote this estimate; the operation $f_k(a, b)$ is known as an “and”-operation, or, alternatively, a t-norm.

There are several reasonable properties that such an “and”-operation should satisfy. For example, since $A \& B$ and $B \& A$ means the same, we should have $f_k(a, b) = f_k(b, a)$, i.e., the “and”-operation should be commutative. Similarly, since $A \& (B \& C)$ means the same as $(A \& B) \& C$, the “and”-operation must be associative:

$$f_k(a, f_k(b, c)) = f_k(f_k(a, b), c).$$

All the operations with these properties are known. The simplest such operation is $f_k(a, b) = \min(a, b)$. This operation is widely used in fuzzy applications.

Other examples are operations of the type

$$f_k(a, b) = g^{-1}(g(a) + g(b)),$$

(5)
When is propagation of interval and fuzzy uncertainty feasible?

where \( g(x) \) is a decreasing function from \([0, 1]\) to non-negative numbers, and \( g^{-1}(x) \) is its inverse function. For example, for \( g(x) = -\ln(x) \), we get \( f_k(a, b) = a \cdot b \). This operation is also widely used in applications.

In general, operations of type (5) are universal approximators, in the sense that for every “and”-operation \( f_k(a, b) \) and for every \( \varepsilon > 0 \), there exists an operation of type (5) which is \( \varepsilon \)-close to \( f_k(a, b) \) for all \( a \) and \( b \); see, e.g., [18]. Thus, from the practical point, we can safely assume that the actually used “and”-operation is of type (5).

So, our estimate of the expert’s degree of confidence that \( \Delta x_1 \) is possible and \( \Delta x_2 \) is possible etc. is equal to

\[
\mu_k(\mu_1(\Delta x_1), \mu_2(\Delta x_2), \ldots).
\]

We have this degree for each tuple \((\Delta x_1, \ldots, \Delta x_n)\). We need to combine the degrees corresponding to different tuples that satisfy the property (1). The phrase “there exists” is, in essence, an “or”: it means that either this property holds for one of the tuples, or it holds for another tuple, etc. Similarly to “and”-operations, in fuzzy logic, we also have “or”-operations \( f_v(a, b) \) (also known as t-conorms). The simplest “or”-operation is \( f_v(a, b) = \max(a, b) \) which is widely used. Other operations include operations of the type (5) with increasing functions \( g(x) \). The mostly widely used example is \( f_v(a, b) = a + b - a \cdot b \) which corresponds to \( g(a) = -\ln(1 - a) \).

At first glance, it may seem that, similar to the fact that we can use generic “and”-operations, we can also use general “or”-operations. However, there is a big difference here:

- we use “and”-operations to combine a finite number of degrees, while
- the “or”-operation is used to combine degrees corresponding to infinitely many possible tuples.

For infinitely many degrees, “or”-operations of type (5) usually lead to the meaningless value 1, so the only meaningful choice is to use the max-operation \( f_v(a, b) = \max(a, b) \).

Thus, we arrive at the following formula for the desired membership function \( \mu(\Delta y) \):

\[
\mu(\Delta y) = \max \left\{ f_k(\mu_1(\Delta x_1), \ldots, \mu_n(\Delta x_n) : \sum_{i=1}^n c_i \cdot \Delta x_i = \Delta y \right\}.
\]

(6)

**Known result:** for min “and”-operation, the problem is feasible to solve. In the simplest case, when \( f_k(a, b) = \min(a, b) \), the formula (6) takes the form

\[
\mu(\Delta y) = \max \left\{ \min(\mu_1(\Delta x_1), \ldots, \mu_n(\Delta x_n) : \sum_{i=1}^n c_i \cdot \Delta x_i = \Delta y \right\}.
\]

(7)

This formula was first derived by Zadeh himself and is thus known as Zadeh’s extension principle.

It is known that to compute \( \mu(\Delta y) \), it is convenient to use \( \alpha \)-cuts, i.e., sets of the type \( \chi_\alpha(\Delta x_i) \equiv \{ \Delta x_i : \mu_i(\Delta x_i) \geq \alpha \} \) and \( y(\alpha) \equiv \{ \Delta y : \mu(\Delta y) \geq \alpha \} \).

Expert estimates for degrees \( \mu_i(\Delta x_i) \) usually decreases as we go further away from 0. In this case, the \( \alpha \)-cuts are simply intervals

\[
\chi_\alpha(\Delta x_i) = [\underline{\chi}_i(\alpha), \overline{\chi}_i(\alpha)]
\]

for appropriate endpoints \( \underline{\chi}_i(\alpha) \) and \( \overline{\chi}_i(\alpha) \).

By definition of \( \mu(\Delta y) \), we have \( \mu(\Delta y) \geq \alpha \) if and only there exist values \( \Delta x_1, \ldots, \Delta x_n \) for which (1) holds and for which min(\( \mu_1(\Delta x_1), \ldots, \mu_n(\Delta x_n) \)) \( \geq \alpha \), i.e., equivalently, for which \( \mu_1(\Delta x_1) \geq \alpha \) and \( \mu_2(\Delta x_2) \geq \alpha \), etc. In other words, \( \Delta y \in y(\alpha) \) if and only if there exist \( \Delta x_i \in \chi_\alpha(\Delta x_i) \) for which

\[
\Delta y = \sum_{i=1}^n c_i \cdot \Delta x_i.
\]

Thus, for each \( \alpha \), the interval \( y(\alpha) \) is simply the range of the linear expression (1) when each \( \Delta x_i \) is in the interval \( [\underline{\chi}_i(\alpha), \overline{\chi}_i(\alpha)] \).

To compute this range, we can represent each of these intervals as \( [\overline{\Delta x}_i - \delta_i, \overline{\Delta x}_i + \delta_i] \), where

\[
\overline{\Delta x}_i \equiv \frac{\underline{\chi}_i(\alpha) + \overline{\chi}_i(\alpha)}{2}
\]

and

\[
\delta_i = \frac{\overline{\chi}_i(\alpha) - \underline{\chi}_i(\alpha)}{2}.
\]
Then, the desired range is equal to
\[ [\hat{y} - \delta, \hat{y} + \delta], \]
where
\[ \hat{y} \equiv \sum_{i=1}^{n} c_i \cdot \Delta x_i \]
and
\[ \delta \equiv \sum_{i=1}^{n} |c_i| \cdot \delta_i. \]
All these explicit formulae are clearly feasible.

**Third new result:** the problem is feasible for any “and”-operation. Let us show that the problem remains feasible if instead of \( \min \), we use any “and”-operation of type (3).

For such an “and”-operation, the formula (6) takes the form
\[ \mu(\Delta y) = \max \left\{ g^{-1}(g(\mu_1(\Delta x_1)) + \ldots + g(\mu_n(\Delta x_n))) : \sum_{i=1}^{n} c_i \cdot \Delta x_i = \Delta y \right\}. \]
Since the function \( g(x) \) is decreasing, we have \( \mu(\Delta y) = g^{-1}(z) \), where
\[ z \equiv \min \left\{ g(\mu_1(\Delta x_1)) + \ldots + g(\mu_n(\Delta x_n)) : \sum_{i=1}^{n} c_i \cdot \Delta x_i = \Delta y \right\}. \]
Thus, if we can feasibly compute \( z \), we can also feasibly compute \( \mu(\Delta y) \) as \( g^{-1}(z) \).

To simplify our expression for \( z \), let us denote
\[ f_i(\Delta x_i) \equiv g(\mu_i(\Delta x_i)). \]
In these terms, the above formula for \( z \) takes the form
\[ z = \min \left\{ f_1(\Delta x_1) + \ldots + f_n(\Delta x_n) : \sum_{i=1}^{n} c_i \cdot \Delta x_i = \Delta y \right\}. \]
The problem of finding \( z \) is now a classical constraint optimization problem: minimize the sum
\[ f_1(\Delta x_1) + \ldots + f_n(\Delta x_n) \]
under the constraint
\[ \sum_{i=1}^{n} c_i \cdot \Delta x_i = \Delta y. \]
To solve this problem, we can use the usual Lagrange multiplier method to reduce it to the following unconstrained optimization problem: minimize the expression
\[ \sum_{i=1}^{n} f_i(\Delta x_i) + \lambda \cdot \left( \sum_{i=1}^{n} c_i \cdot \Delta x_i - \Delta y \right), \]
for an appropriate Lagrange multiplier \( \lambda \). Differentiating the above objective function with respect to \( \Delta x_i \) and equating the derivative to 0, we get
\[ f_i'(\Delta x_i) + \lambda \cdot c_i = 0, \]
i.e., that
\[ \Delta x_i = f_i^{-1}(\lambda \cdot c_i). \]
Thus, if we know \( \lambda \), we can feasibly compute all the values \( \Delta x_i \).

The only remaining problem is to find \( \lambda \). The value \( \lambda \) must be found from the condition (1). Thus, we can use, e.g., bisection to find the appropriate value \( \lambda \). Hence, the whole computation is feasible.
4 Beyond Linearized Case

For boxes, interval computation is NP-hard already for quadratic functions. For linear functions $f(x_1, \ldots, x_n)$, as we have mentioned, the problem of computing the interval range over a box

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

is feasible. However, already for quadratic functions $f(x_1, \ldots, x_n)$, this problem is NP-hard; see, e.g., [14].

Beyond boxes. But why concentrate on boxes? Boxes correspond to the case when we have no constraints. In practice, as we have mentioned, we often have constraints, so we have a subset of the box.

Known result: for ellipsoids, the problem is feasible for quadratic functions $f(x_1, \ldots, x_n)$. One of the typical sets of possible values is an ellipsoid, i.e., set of the type $Q(x_1, \ldots, x_n) \leq q_0$ for some quadratic function $Q(x_1, \ldots, x_n)$; see, e.g., [7, 13, 23] and references therein.

We want to find the range of a quadratic function $f(x_1, \ldots, x_n)$ over an ellipsoid. In other words, we want to find the maximum $y$ and the minimum $y$ of this function over an ellipsoid. The maximum or minimum occurs:

- either inside the ellipsoid
- or on its boundary $Q(x_1, \ldots, x_n) = q_0$.

If the maximum or minimum is attained inside, then all $n$ partial derivatives of the quadratic function $f(x_1, \ldots, x_n)$ must be equal to 0. Derivatives of a quadratic function are linear, so we get an easy-to-solve system of $n$ linear equations with $n$ unknowns, for which we can easily find the corresponding tuple $(x_1, \ldots, x_n)$ – and check that this tuple is indeed inside the given ellipsoid.

If the maximum or minimum is attained on the border, then the location of this maximum or minimum can be found by solving the following constraint optimization problem: optimize the function $f(x_1, \ldots, x_n)$ under the constraint $Q(x_1, \ldots, x_n) = q_0$.

To solve this problem, we can also use the Lagrange multiplier method and solve the corresponding unconstrained optimization problem of optimizing the expression

$$f(x_1, \ldots, x_n) + \lambda \cdot (Q(x_1, \ldots, x_n) - q_0).$$

For each $\lambda$, by differentiating this expression with respect to $x_i$ and equating each of these derivatives to 0, we get an easy-to-solve system of $n$ linear equations with $n$ unknowns. Thus we get the values $x_i(\lambda)$.

The only remaining problem is how to find $\lambda$ from the condition that $Q(x_1(\lambda), \ldots, x_n(\lambda)) = q_0$. But this is an equation with one unknown, and such equations are feasibly solvable. (The problem becomes computationally complex when the number of variables grows.)

Fourth new result: for quadratic function, the problem is feasible also for intersection of ellipsoids. Ellipsoids are sets of a very specific shape. Can we extend the above results to generic shapes?

Of course, since the problem is NP-hard for boxes, we do not expect a feasible algorithm for all the sets, but what we would like to have is a sequence of families of sets, with more and more parameters, for each of which we have a feasible algorithm – although the complexity of such algorithm may increase as we add more and more parameters to the family.

It turns out that as such a sequence of families, we can take intersections of ellipsoids:

- the first family is the family of ellipsoids,
- the next family is intersections of 2 ellipsoids,
- then intersections of 3 ellipsoids, etc.

Let us show that this way, we indeed get a universal approximation family, i.e., that any convex set can be thus approximated. Indeed, each convex set can be approximated, with any given accuracy, by a polyhedron: e.g., by the convex hull of sufficiently points on its surface; see, e.g. [22]. A convex polyhedron is an intersection of half-spaces that contain it, and a half-space can be approximated by an ellipsoid – since in an appropriate limit, an ellipsoid tends to a half-space.
For the intersection of $K$ ellipsoids $Q_k(x_1, \ldots, x_n) \leq q_k$, the corresponding Lagrange multiplier problem takes the form

$$f(x_1, \ldots, x_n) + \sum_{k=1}^{K} \lambda_k \cdot (Q_k(x_1, \ldots, x_n) - q_k) = 0.$$  

We also need to consider $2^K$ cases when the optimizing tuple $(x_1, \ldots, x_n)$ is inside some of the ellipsoids. In all these cases, we get a system of linear equations enabling us to find all the values $x_i$ as functions of $\lambda_1, \ldots, \lambda_K$: $x_i = x_i(\lambda_1, \ldots, \lambda_K)$. The values $\lambda_k$ must then be determined from the conditions that

$$Q_k(x_1(\lambda_1, \ldots, \lambda_K), \ldots, x_n(\lambda_1, \ldots, \lambda_K)) = q_k$$  

for all $k$ from 1 to $K$. We thus have a system of $K$ nonlinear equations with $K$ unknowns. When $K$ is fixed, this system is feasible – although the complexity of solving this system grows exponentially with $K$.

**Acknowledgments**

This work was supported in part by the US National Science Foundation grant HRD-1242122 (CyberSHARE Center of Excellence).

The authors are thankful to the anonymous referees for valuable suggestions.

**References**


When is propagation of interval and fuzzy uncertainty feasible?

Efficient training of neural networks with interval uncertainty

Jonathan C. Sadeghi, Marco De Angelis and Edoardo Patelli

Institute for Risk and Uncertainty, University of Liverpool, United Kingdom

Corresponding author: edoardo.patelli@liverpool.ac.uk

Abstract

In this paper we attempt to build upon past work on Interval Neural Networks, and provide a robust way to train and quantify the uncertainty of Deep Neural Networks. Specifically, we propose a back propagation algorithm for Neural Networks with constant width predictions. In order to maintain numerical stability we propose minimising the maximum of the batch of errors at each step. Our approach can accommodate incertitude in the training data, and therefore adversarial examples from a commonly used attack model can be trivially accounted for. We present preliminary results on a test function example. The reliability of the predictions of these networks are guaranteed by the non-convex Scenario approach to chance constrained optimisation. A key result is that, by using minibatches of size $M$, the complexity of our approach scales as $O(MN_{\text{iter}})$, and does not depend upon the number of training data points as with other Interval Predictor Model methods.

Keywords: Machine Learning, Imprecise Probability, Uncertainty Quantification, Neural Networks, Interval Predictor Models

1 Introduction

In recent years Deep Learning using Artificial Neural Networks has emerged as a generalised Machine Learning tool which has revolutionised supervised learning, reinforcement learning, as well as finding many applications in the field of Engineering, most often as efficient surrogates for large models [1]. In all fields, but particularly in safety critical engineering applications, it is essential to quantify the uncertainty of the Neural Network. The simplest and most widely used approaches attempt to quantify this uncertainty by analysing the mean squared error or explained variance ($r^2$) of the Neural Network on a test set. However, this approach does not robustly bound the predictions of the Neural Network. Subsampling techniques like bootstrapping can improve the uncertainty quantification, but the results are in general still unsatisfactory [2]. Bayesian Neural Networks (assisted by variational inference), where the weights are modelled as random variables, have emerged as the most popular tool for giving a prediction of the uncertainty of the Neural Network which is acceptable to statisticians [3]. Popular implementations in PyMC3 and Edward, amongst others, have made this approach viable in an industrial context [4]. However, many assumptions are necessary in order to apply this approach, for example the weights are commonly assumed to have a Gaussian prior and the likelihood function is also assumed to be Gaussian. In addition, Variational Inference is more complex to implement than the Backpropagation which is used by most Machine Learning engineers. Robust Neural Networks apply Bayesian model selection to discrete sets of Neural Networks, and whilst a prior is not required explicitly for the weights, other assumptions are made [5].

Interval Predictor Models are a recently developed machine learning technique for supervised learning which make interval predictions with guaranteed accuracy [6]. In other words, for every input, $x_i$, an Interval Predictor Model would predict $\bar{y}(x_i)$ and $\tilde{y}(x_i)$ instead of just $y_i$. The technique relies upon the solution of chance constrained convex optimisation programs by the scenario technique [7]. The scenario technique is a method of approximately solving optimisation problems with probabilistic constraints by considering a random sample. Scenario Optimisation is easier to use in practice than similar methods in statistical learning theory since no knowledge of the VC-dimension is required. The first published software implementation of Interval Predictor Models was made available in the open source OpenCossan software [8]. Crucially the scenario approach only applies when the data distribution of the input variables is stationary. This is an important limitation of the work currently. A key advantage over other machine learning techniques is that interval training data fits into the scenario optimisation framework coherently [9]. This can be seen as similar to the attack model of adversarial examples considered in [10], however our framework also permits robustness against uncertainty in training outputs.

Neural Networks with interval outputs were first proposed in [11], and further described in [12]. In
these papers the learning takes place by identifying the weights $W$, which solve the following program:

$$\arg\min_{W} \left[ \hat{y}(x_i) - y(x_i) : \hat{y}(x_i) > y_i > \bar{y}(x_i) \forall i \right],$$

(1)

where $\hat{y}(x)$ and $y(x)$ are obtained from two independent Neural Networks, such that $\hat{y}(x)$ and $y(x)$ are the output layers of networks, where layer $i$ is given by $f_i$

$$f_i = \tanh (W_i f_{i-1}),$$

(2)

where $f_i$ is a vector (which is the input vector when $i = 0$) and $W_i$ is the $i$th weight matrix. In practice this problem is solved by using a mean squared error loss function with a simple penalty function to model the constraints. Our experiments revealed that this penalty method requires careful choice of hyper-parameters to guarantee convergence, and hence we struggled to repeat the results from the paper. These Neural Networks act in a similar way to Interval Predictor Models, however the Interval Neural Networks do not include an assessment of the prediction accuracy (although this can be predicted to a reasonable degree of accuracy by using a test set). In [13] the Scenario Approach was extended to non-convex optimisation programs, and hence applied to a single layer Neural Network, with a constant width interval prediction, which was trained using the interior-point algorithm in Matlab. In other words the following program is solved:

$$\arg\min_{W,h} \left[ h : |y_i - \hat{y}(x_i)| < h \forall i \right],$$

(3)

where $h$ is a real number, and $\hat{y}$ represents the central line of the prediction obtained from the same network specified by Eqn. 2. In [14] the non-convex scenario approach is extended to problems solved with the sampling and discarding technique where the most restrictive constraints are optimally removed [15].

In this paper we propose a back propagation algorithm for Neural Networks with constant width predictions, and show how this can be efficiently used to create deep Interval Neural Networks. In order to obtain maintain numerical stability we avoid using the penalty method proposed by [11], and instead propose minimising the maximum squared error of the whole batch of gradients at each step. We present preliminary results on a test function example. On modern architectures using larger batches is desirable as this allows the step size to be increased, whilst maintaining a constant computation time for each step by making use of parallelisation. The predictions of these networks are guaranteed by the non-convex Scenario approach. We explain how interval training data can be accommodated in this paradigm [9].

2 Proposed Architecture

2.1 Overview

Our architecture is based on the model proposed in [13]. By solving

$$\arg\min_{W,h} \left[ h : |y_i - \hat{y}(x_i)| < h \forall i \right],$$

(4)

we are trying to find the Neural Network which minimises the maximum error rather than the mean squared error. In order to solve this problem efficiently we propose a modified backpropagation method for the Neural Network, i.e. gradient based optimisation. Eqn. 4 is chance constrained but is much easier to solve than Eqn. 1. To evaluate the gradient we compute the mean squared error for each data point in the training set, and then follow the gradient for the point with maximum error. Our algorithm is described in further detail in Algorithm 1.

It is immediately apparent that our algorithm is more costly than the standard back propagation method, since our method costs $O(NN_{iter})$, compared to a standard stochastic gradient descent cost of $O(N_{iter})$. Fortunately, the cost is not as high as it would initially seem as modern GPU architectures allow the some of the calculations to be parallelised. However, the largest GPU architectures have several thousand cores, so for data sets with millions of data points our approach would not be scalable. We should also add that the $N_{iter}$ required for convergence in both algorithms is not necessarily the same, as this depends on the variance of the gradient at each step.

In order to increase the speed of convergence the weights could be initialised to those obtained by training the network with a mean squared error loss function.
Algorithm 1 Maximum error backpropagation method

**Input:** Training data pairs \((x_i, y_i)\) for \(i = 1, ..., N\)

Randomly initialise weight tensor and \(h\).

for \(i = 1, ..., N_{\text{iter}}\) do

    for \(j = 1, ..., N\) do

        Compute Prediction \(\hat{y}(x_j)\)

        Compute error \(e_j = |y_j - \hat{y}(x_j)|\)

    end for

    Set \(k = \arg \max_j e_j\)

    Use gradient of loss function to update \(W\) and \(h\) \((W \leftarrow W + \eta \frac{\partial (y_k - \hat{y}(x_k))^2}{\partial W})\)

end for

for \(j = 1, ..., N\) do

    Compute Prediction \(\hat{y}(x_j)\)

    Compute error \(e_j = |y_j - \hat{y}(x_j)|\)

end for

Set \(h = \max_j e_j\)

**Output:** Weight tensor and \(h\)

---

2.2 Scalability Improvement

In order to reduce the computational cost of the algorithm we propose the use of minibatch stochastic gradient descent [16]. In other words, \(M\) training data points are randomly selected for each step. Our algorithm is described in Algorithm 2.

Algorithm 2 Maximum error backpropagation method, using minibatches

**Input:** Training data pairs \((x_i, y_i)\) for \(i = 1, ..., N\)

Randomly initialise weight tensor and \(h\).

for \(i = 1, ..., N_{\text{iter}}\) do

    Generate set, \(B\), of \(M\) random numbers, sampled without replacement between 1 and \(N\)

    Set \(k = \arg \max_{j \in B} |y_j - \hat{y}(x_j)|\)

    Use gradient of loss function to update \(W\) and \(h\) \((W \leftarrow W + \eta \frac{\partial (y_k - \hat{y}(x_k))^2}{\partial W})\)

end for

Generate set, \(B\), of \(M\) random numbers, sampled without replacement between 1 and \(N\)

Set \(h = \max_{j \in B} |y_j - \hat{y}(x_j)|\)

**Output:** Weight tensor and \(h\)

It is clear that using minibatches reduces the cost of our algorithm to \(O(MN_{\text{iter}})\), which is a potentially vast improvement when \(N >> M > 1\). However there is no guarantee that the minibatch selected at each step contains the true maximum of the error, and therefore we may not be following the true gradient, and hence may be solving the wrong optimisation program. However it turns out that this is not problematic, as small minibatch sizes can closely replicate the expectation of the true loss function (see Appendix A for proof). For large \(N\) we find that by using a minibatch of size \(M\), we solve a problem equivalent to minimising the \(\frac{1}{M}\)-th percentile of the empirical cumulative distribution function of the error for the whole training data set. This is not as problematic as it seems, due to developments which are presented later in the paper.

3 Reliability Assessment

3.1 Convex case

For the benefit of readers not familiar with the scenario approach to chance constrained optimisation we will first present an overview of the theory of scenario optimisation in the convex case.

A chance constrained optimisation program is an optimisation program of the following form

\[
\arg \min c^T x : P\{f(x, \delta) > 0\} \leq \epsilon,
\]

(5)

where \(\delta\) is a random variable, \(x\) is the design variable, \(c\) is a constant, and \(\epsilon\) is a parameter which constrains how often the constraints may be violated. This program can be approximately solved using
the scenario approach: the problem is solved for a random sample of the constraints, i.e. we solve
\[
\arg\min c^T x : f(x, \delta^i) \leq 0, i = 1, ..., N, \tag{6}
\]
where \(\delta^i\) represents the \(i\)th sampled value of the constraints \([7]\).

Intuition tells us that the solution will be most accurate when the dimensionality of the design variable is low and we take as many samples of the constraints as possible (in fact, an infinite number of sampled constraints would allow us to reliably estimate \(P\{f(x, \delta) > 0\}\), and hence solve the program exactly). However, in practice obtaining these samples is often an expensive process. Luckily the theory of scenario optimisation provides robust bounds on the robustness of the obtained solution. The bounds generally take the following form:
\[
P^N(V(\hat{x}_N)) > \epsilon) < \beta, \tag{7}
\]
This equation states that the probability of observing a bad set of data (i.e. a bad set of constraints) in future, such that our solution violates a proportion greater than \(\epsilon\) of the constraints \((V(\hat{x}_N) > \epsilon)\) where \(V(\hat{x}_N) = \frac{1}{N} \sum_{i}^N V_i\) and \(V_i = 1\) only if \(f(x, \delta^i) > 0\), is no greater than \(\beta\). The scenario approach gives a simple analytic form for the connection between \(\epsilon\) and \(\beta\) in the case that the optimisation program is convex:
\[
\beta = \frac{1}{\epsilon} \frac{n}{N + 1}, \tag{8}
\]
where \(N\) is the number of constraint samples in the training data set used to solve the scenario program, and \(n\) is the dimensionality of the design variable, \(x\). Other tighter bounds exist in the literature \([17]\). For a fixed \(N\) and \(n\) we obtain a plot as shown in Fig. 1. The plot demonstrates that by decreasing \(\epsilon\) slightly, \(1 - \beta\) can be made to be insignificantly small. Crucially our assessment of the solution is a-priori, although other techniques exist \([18]\). In the convex case the a-priori assessment is made possible by the fact that the number of support constraints (the number of constraints which if removed result in a more optimal solution) for a convex program is always less than the dimensionality of the design variable. For a non-convex program this is not necessarily the case, and therefore a new approach is required.

3.2 Non-convex case

Fortunately, another approach is possible. In \([13]\) the following bound is given for the non-convex case:
\[
P^N(V(\hat{x}_N) > \epsilon(s)) < \beta, \tag{9}
\]
where
\[
\epsilon(s) = \begin{cases} 
1, & \text{for } s = N, \\
1 - n^{-\frac{s}{N}}, & \text{otherwise},
\end{cases} \tag{10}
\]
and $s$ is the cardinality of the support set (in other words, the number of support constraints). The behaviour of this bound is similar to the convex case since in general increasing $n$ should increase the size of the support set.

### 3.3 Finding The Cardinality of the Support Set

So far we have shown how we can efficiently solve the scenario program which is required to train a Neural Network with interval predictions, provided that we know the number of support constraints. The only practical challenge which remains is to calculate the cardinality of the support set. This is in general a computationally expensive task since the scenario program must be solved $N$ times. In [13] an efficient algorithm is presented which only requires that the scenario problem is solved $s$ times. However in this section we propose a more efficient algorithm which can be derived from our approach.

If the maximum error is plotted with respect to iterations (see example in Fig. 2) then we will be able to spot when the Neural Network training process has converged by noticing a cyclic behaviour in the plot. This signifies that our optimiser is orbiting around a point in the design space. A simple modification to our algorithm allows us to record which points in the training data set contribute to this orbit, as they are the points with the maximum error at each step. If a point which is not a member of this set is removed then the orbit will not change. Therefore we propose that the cardinality of the support set is bounded by the cardinality of the set of points which contribute to the converged orbit.

When using minibatches the set of points which contribute to the orbit will be an over estimate due to statistical fluctuations. Therefore it will usually be necessary to run a number of refining iterations where the batch gradient is computed by finding the maximum error of the whole support set. This allows us to avoid the error resulting from using minibatches whilst avoiding a computational cost of order $O(NN_{test})$. The refining iterations also allow us to avoid unsatisfied constraints in the scenario program (as discussed in Section 2.2), which may occur due to choosing a minibatch size which is not sufficiently large.

### 3.4 Incertitude in Training Data

We have proposed an algorithm for use with crisp training data. However one of the main advantages of the Interval Predictor Model framework is that training data with incertitude (i.e. interval training data or fuzzy data) fits coherently into the paradigm [9]. An example of incertitude in training data is the adversarial attack model given in [10]. In fact, the proposed attack model places each training data point in an uncertain hyper-sphere ($\ell_2$ ball) rather than the intervals with which many uncertainty practitioners are accustomed to. However, both cases are convex sets and therefore the conceptual challenge of accommodating this training data is similar. Since our model is more complex than that proposed in [9] the computations required to accommodate interval data are also more complex.

For the case of interval imprecision in the output variables (i.e. we observe pairs $x_i$ and $[\bar{y}_i, \bar{y}_i]$) we can simply modify Eqn. 4 as follows:

$$\arg\min_{\tilde{W}, h} [h : \max(|\bar{y}_i - \tilde{g}(x_i)|, |\bar{y}_i - \tilde{g}(x_i)|) < h \forall i],$$  
(11)

which can be written in simplified form if the width of interval $[\bar{y}_i, \bar{y}_i]$ is constant for all data points.

For interval incertitude in the input training data the situation is more complex, and since the sum of squares approach used in [9] is not directly applicable our algorithm will be more costly. If we observe pairs $[\bar{x}_i, \tilde{x}_i]$ and $[\bar{y}_i, \tilde{y}_i]$ then we must solve

$$\arg\min_{\tilde{W}, h} [h : \max_{x \in [\bar{x}_i, \tilde{x}_i]} (|\bar{y}_i - \tilde{g}(x)|, |\bar{y}_i - \tilde{g}(x)|) < h \forall i],$$  
(12)

where the nested optimisation in the constraints is clearly to blame for the inefficiency of the algorithm. One approach to solving this problem would be to attempt to brute force the nested optimisation (i.e. discretise along the upper ‘edge’ of the incertitude box), however if the incertitude is large or the dimensionality of the training data is high then this becomes impractical. Another possibility is assuming the prediction of the Neural Network is approximately linear locally and using the gradient of the Neural Network with respect to the inputs (which is known analytically) to find an approximate solution to the nested optimisation problem. This is similar to the approaches proposed in [19] and [20], where the gradient is used to search within a set close to the original training data for points which maximise the loss function of the Neural Network. The crucial difference is that in our formulation we need only search on the surface of the set, since we aim to enclose the whole set in the Interval Neural Network.
Since neither of these methods is completely satisfactory we leave the problem of interval uncertainty ($\ell_\infty$ ball, $\ell_2$ ball and other cases) in the input training data as an unsolved problem.

4 Numerical Experiments

In order to illustrate the developed techniques we will demonstrate the Interval Neural Network on a modified version of a simple problem from [13]. We will train a Neural Network with 1 layer containing 10 neurons on 1250 samples from the following test function:

$$y = 0.3 \ast (15 \ast u \ast \exp(-3 \ast u) + w)$$

where $w$ is a normal distributed random variable with zero mean and standard deviation $\sigma = 0.025$. The input variable $u$ will be uniformly distributed between 0 and 1, since input data can be transformed into this range trivially. For clarity the algorithm used is described in Algorithm 3.

Algorithm 3 Maximum error backpropagation method as used in example

Input: Training data pairs $(x_i, y_i$ for $i = 1, ..., N)$

Randomly initialise weight tensor and $h$.

for $i = 1, ... , 5000$ do

Generate set, $B$, of $M$ random numbers, sampled without replacement between 1 and $N$

Set $k = \arg \max j \in B |y_j - \hat{y}(x_j)|$

Use gradient of loss function to update $W$ and $h$ ($W_i \leftarrow W_i + \eta \frac{\partial(y_k - \hat{y}(x_k))^2}{\partial W}$)

if $i > 4000$ then

$S_{i-4000} = k$
endif

end for

Set $B = S$

for $i = 1, ... , 100$ do

Set $k = \arg \max j \in B |y_j - \hat{y}(x_j)|$

Use gradient of loss function to update $W$ and $h$ ($W_i \leftarrow W_i + \eta \frac{\partial(y_k - \hat{y}(x_k))^2}{\partial W}$)

if $i > 4000$ then

$R_i = k$
endif

end for

Set $B = S$

Set $h = \arg \max j \in B |y_j - \hat{y}(x_j)|$

Use Eqn. 10 to calculate a bound on the violation probability of the identified Neural Network, with $\beta = 10^{-6}$ and $s$ equal to the cardinality of $R$.

Output: Weight tensor and $h$

The trained Neural Network is shown in Fig. 3. An annotated plot of the convergence (i.e. the maximum error at each step) is shown in Fig. 2. A minibatch size of 200 was used, and hence the support set (measured from assumed convergence after 4000 iterations) was an overestimate ($s = 181$). Running 100 refinement iterations allowed the size of the support set to be reduced to only 9 points.

Using Eqn. 10 we calculate

$$P^N(V(\tilde{x}_N) > 0.057) < 10^{-6}$$

for the trained Interval Neural Network.

5 Conclusions

We have shown how to create Neural Networks which quantify their uncertainty with interval predictions. Our approach converges reliably and allows some of the data to be ignored to reduce the width of the prediction interval. Our approach is not restricted to a specific architecture and makes use of many modern techniques in deep learning.

Our approach is limited by the fact that this interval is a constant width. In a future paper we will extend the approach to demonstrate how our work can be extended to a non-constant interval prediction width.
Figure 2: Plot of convergence of the Neural Network. The support set was determined by recording which training data points contribute to the change in weights after 4000 iterations (shown as a blue line).

Figure 3: Plot of trained Interval Neural Network. The support set is shown in yellow.

(a) Upper bound to support set, before refinement.  
(b) Upper bound to support set, after refinement.
References


A Proof of statements in Section 2.2

We will solve the Optimisation Problem in Eqn. 4 approximately using Algorithm 1. Explicitly we wish to minimise the loss function

\[
J = \max_{j \in [1,...,N]} (y_j - \hat{y}(x_j))^2.
\]  

Consider that the probability of selecting the true maximum of the squared error in a minibatch by random sampling without replacement is \( \frac{N}{N} \). The probability that the maximum point selected in the minibatch is the i-th largest in the training set is

\[
P(i) = \binom{N-i}{M-i} \binom{N}{M}.
\]  

Rec 2018 – J.C. Sadeghi and M. De Angelis and E. Patelli
Then to find the expectation of $i$ we calculate

$$E(i) = \sum_{i=1}^{i=N-M+1} i \frac{N-i}{M} = \frac{N + 1}{M + 1}. \quad (17)$$

In the case that $\frac{N}{M} >> 1$ we find that the expression for the expected percentile reduces to

$$\frac{E(i)}{N} \approx \frac{1}{M} \quad (18)$$

as promised. The variance of the percentile is

$$\text{Var} \left( \frac{i}{N} \right) = \frac{M(N-M)(1+N)}{N^2(1+M)^2(2+M)}, \quad (19)$$

which becomes

$$\text{Var} \left( \frac{i}{N} \right) \approx \frac{1}{M^2} \quad (20)$$

in the large $N$ limit. Therefore we see that the minibatch technique performs best when the size of the training set is large, but it is also necessary to increase the minibatch size to avoid the gradient having a large variance.
Iterative importance sampling for estimating expectation bounds under partial probability specifications

Matthias C. M. Troffaes,1* Thomas Fetz,2 and Michael Oberguggenberger2

1Department of Mathematical Sciences, Durham University, UK
2Arbeitsbereich für Technische Mathematik, Universität Innsbruck, Austria
*Corresponding author: matthias.troffaes@durham.ac.uk

Abstract
In this paper, we explore and enhance importance sampling techniques for calculating lower and upper expectations with respect to sets of probability distributions. We formalize an iterative algorithm that we proposed in earlier work, by formulating the algorithm as a procedure for identifying a fixed point. We show how the algorithm can break down under poor coverage of the sampling distribution, and explore simple methods to increase coverage and thereby improve the algorithm.

Keywords: importance sampling, reweighting, probability bounding, imprecise probability, lower prevision, Monte Carlo

1 Introduction
In many engineering problems, it can be hard to specify full probability densities for all parameters, due to lack of data and expert information. In such cases, we may prefer to work with partial probability specifications, or equivalently, sets of probability densities [1–3]. Typically, we may wish to estimate the lower expectation (lower prevision)

\[ \theta_s := \min_{t \in T} \int h(x) f_t(x) dx \]

of some function \( h \) with respect to some parametrized family of probability density functions \( f_t \), over all \( t \in T \). For example, in reliability engineering, \( h \) might be an indicator function of a failure region described by a limit state function \( g \) and then \( \theta_s \) is the lower probability of failure. Upper probabilities of failure can be treated in the same way.

In earlier work [4–8] we studied how to estimate \( \theta_s \) through importance sampling. In this paper, we formalize our work further by describing the sample as a parametrized function \( x_t(V) \) of a fixed random variable \( V \) (e.g. uniform in \([0,1]^k\)), and a function \( x_t \) of \( V \), such that \( x_t(V) \sim f_t \). This enables better control of the error, reduces the bias as shown in [8], and formalizes the technique of ‘fixing the seed’ across iterations of the optimisation steps [7]. Here, we study the convergence of the iterative importance sampling estimator developed in [6, 7] by formulating it as a fixed point of an operator. We contrast the iterative procedure with standard sampling, and we investigate how increased coverage of the sampling region can substantially improve the accuracy of the estimates. Examples demonstrate our approach.

2 Importance Sampling
Let \( f_t \) be a density parameterized by \( t \in T \). We are interested in estimating the lower and/or upper expectation of some function \( h \) with respect to \( f_t \) over all \( t \in T \):

\[ \theta(t) := \int h(x) f_t(x) dx, \]

\[ \theta_s := \min_{t \in T} \theta(t), \quad \theta^* := \max_{t \in T} \theta(t). \]

We assume that samples from \( f_t \) can be generated as follows. We start from a random variable \( V \) (e.g. uniform in \([0,1]^k\)), and a function \( x_t \) of \( V \), such that

\[ x_t(V) \sim f_t. \]

For example, if \( t = (\mu, \sigma) \) and \( f_t \) is \( N(\mu, \sigma^2) \), then \( V = (U_1, U_2) \) could be a standard bivariate uniform on \([0,1]^2\), and

\[ x_t(V) = \mu + \sigma \sqrt{-2 \ln U_1 \cos(2\pi U_2)} \]

will have the desired distribution [9]. Similar transformation methods are available for all standard distributions, and, in general, can be obtained for instance via inverse transform sampling. Specifically, if
Iterative importance sampling for expectation bounds under partial probability

Let \( F_t \) be the cumulative distribution function for the density \( f_t \), and \( V \) is a standard uniform random variable on \([0, 1]\), then

\[
x_t(V) = F_t^{-1}(V) \sim f_t.
\]

The reason for making the function \( x_t \) explicit is that we need to control the randomness throughout the algorithm that we will describe next. More specifically, we will need to describe the sample itself as a deterministic function of the parameter \( t \).

Imagine that we can start from an i.i.d. sample \( \Omega := (V_1, V_2, \ldots, V_n) \), to obtain an i.i.d. sample

\[
x_s(V_1), \ldots, x_s(V_n)
\]

from \( f_s \), for some fixed \( s \in T \). Now, because

\[
\int h(x)f_s(x) \, dx = \int f_s(x) h(x) f_s(x) \, dx,
\]

we can use this sample from \( f_s \) to estimate the expectation of \( h \) with respect to \( f_t \), for any \( t \in T \):

\[
\hat{\theta}_{\Omega,s}(t) := \frac{1}{n} \sum_{i=1}^{n} w_{st}(x_s(V_i)) h(x_s(V_i))
\]

where

\[
w_{st}(x) := \frac{f_t(x)}{f_s(x)}.
\]

Note that \( \hat{\theta}_{\Omega,s}(t) \) is an unbiased estimator for \( \theta(t) \):

\[
E(\hat{\theta}_{\Omega,s}(t)) = \theta(t).
\]

If this normalisation constant of the probability density \( f_t \) is expensive to calculate, then there is substantial gain by only evaluating it once for each desired \( t \) (or \( s \)), and to reuse it across all terms in the sum. Alternatively, self-normalised importance sampling can be used:

\[
\hat{\theta}_{\Omega,s}^*(t) := \frac{\sum_{i=1}^{n} w'_{st}(x_s(V_i)) h(x_s(V_i))}{\sum_{i=1}^{n} w'_{st}(x_s(V_i))}
\]

where \( w'_{st}(x) \) are defined in the same way as the weights \( w_{st}(x) \) but only up to a normalisation constant of the densities involved. This has the downside that the resulting estimate is only asymptotically unbiased [8].

A special case obtains when \( s = t \). In that case, we have standard sampling:

\[
\hat{\theta}_{\Omega}(t) := \hat{\theta}_{\Omega,t}(t) := \frac{1}{n} \sum_{i=1}^{n} h(x_t(V_i))
\]

because \( w_{tt}(x) = 1 \) for all \( x \). This leads to the following estimators for \( \theta_* \) and \( \theta^* \) [8]:

\[
\hat{\theta}_{s,t} := \hat{\theta}_{\Omega}(T_{s,t}) = \min_{t \in T} \hat{\theta}_{\Omega}(t) \quad \text{and} \quad \hat{\theta}_{s,t}^* := \hat{\theta}_{\Omega}(T_{s,t}^*) = \max_{t \in T} \hat{\theta}_{\Omega}(t),
\]

where

\[
T_{s,t} := \arg \min_{t \in T} \hat{\theta}_{\Omega}(t) \quad \text{and} \quad T_{s,t}^* := \arg \max_{t \in T} \hat{\theta}_{\Omega}(t).
\]

The estimates \( \hat{\theta}_{\Omega}(t) \) will be highly correlated for different values of \( t \), which helps reducing the bias, as shown in [8].

A difficulty with calculating \( T_{s,t} \) or \( T_{s,t}^* \) is that we need to evaluate \( h \) at points \( x_t(V_i) \), and these points will arbitrarily shift around as we optimize over \( t \). With importance sampling, however, for fixed \( s \), we only need to evaluate \( h \) for the points \( x_s(V_i) \), independently of \( t \). So if \( h \) is expensive to evaluate, then importance sampling is particularly useful, because we do not need to re-evaluate \( h \) for different \( t \) when optimizing over \( t \). In addition, we retain the benefit that the estimates \( \hat{\theta}_{\Omega,s}(t) \) will be highly correlated for different values of \( t \), helping to reduce the bias [8].

With importance sampling, for each \( s \), we have the following estimators for \( \theta_* \) and \( \theta^* \) [8]:

\[
\hat{\theta}_{s,t}(s) := \hat{\theta}_{\Omega,s}(\tau_{s,t}(s)) = \min_{t \in T} \hat{\theta}_{\Omega,s}(t) \quad \text{and} \quad \hat{\theta}_{s,t}^*(s) := \hat{\theta}_{\Omega,s}(\tau^*_{s,t}(s)) = \max_{t \in T} \hat{\theta}_{\Omega,s}(t),
\]

where

\[
\tau_{s,t}(s) := \arg \min_{t \in T} \hat{\theta}_{\Omega,s}(t) \quad \text{and} \quad \tau^*_{s,t}(s) := \arg \max_{t \in T} \hat{\theta}_{\Omega,s}(t).
\]

The quality of the importance sampling estimates can be verified in the standard way via confidence intervals that are constructed through repeated sampling [8].

REC 2018 – M.C.M. Troffaes and T. Fetz and M. Oberguggenberger
3 Iterative Importance Sampling

An issue with the importance sampling estimates is that their quality can be very poor if \( \tau_{\Omega}(s) \) is far from \( s \). A procedure for iteratively improving the choice of \( s \) was proposed in [6, 7]. The procedure essentially iteratively applies the operator \( \tau_{\Omega} \). Under the assumption that this iterative application

\[
s^{(k+1)} = \tau_{\Omega}(s^{(k)}), \quad k = 1, 2, \ldots
\]

of the operator \( \tau_{\Omega} \) reaches a unique fixed point, say \( S_{\Omega} \), our improved lower estimator is:

\[
\hat{\theta}^1 := \hat{\theta}_{\Omega,S_{\Omega}}(S_{\Omega}) = \frac{1}{n} \sum_{i=1}^{n} h(x_{S_{\Omega}}(V_i)).
\]

In numerical examples discussed in [6–8], normally, a fixed point is indeed obtained after few steps. As we shall see, however, \( \tau_{\Omega} \) is not necessarily continuous, and therefore it is not guaranteed that a fixed point exists. Even if it is continuous, \( \tau_{\Omega} \) is not necessarily contracting, and therefore it is not guaranteed that a fixed point can be found by repeated application of \( \tau_{\Omega} \) on itself.

However, asymptotically, as the sample size \( n \) increases, we are tempted to conjecture that, under suitable conditions, \( \tau_{\Omega} \) should have a fixed point \( S_{\Omega} \), and both \( T_{\Omega} \) and \( S_{\Omega} \) should converge, in probability, to

\[
t_s := \arg\min_t \theta(t).
\]

The intuition behind this conjecture is that \( \hat{\theta}_{\Omega,s} \) converges in probability to \( \theta \) as the sample size goes to infinity. In similar way we get the upper estimator \( \hat{\theta}^\dagger \).

Example 1 Here we consider the estimation of the upper probability of the event

\[
D = [-4, -3] \cup [-1, 1] \cup [3.1, 4.7],
\]

with respect to the set of normal distributions with mean \( t \in T = [-7, 7] \) and variance \( \sigma^2 = 2 \). The probability of \( D \), for each value of \( t \), is depicted in fig. 1. The values of \( t \) where local maxima are achieved are indicated by vertical dotted lines. Note that the upper probability of \( D \) is simply the upper expectation of

\[
h(x) := \begin{cases} 
1 & \text{if } x \in D, \\
0 & \text{otherwise}.
\end{cases}
\]

The probability \( h(t) = \int_D f(t) \, dx = \int_T h(x)f(t) \, dx \), is depicted in fig. 1 (left) for each value \( t \in T \). The exact maximum (upper probability) \( \theta^* \) is equal to 0.5488524 and is achieved for \( t = -0.001136 \).

Since the function \( h \) is not continuous, the approximation \( \hat{\theta}_\Omega \), eq. (12), of \( \theta \) is not continuous either. It is a step function with step sizes \( 1/n \). As we can see in fig. 1 (left) \( \hat{\theta}_\Omega \) is a quite good approximation (\( n = 1000 \)), but it is expensive to compute.

In fig. 2, we show the contour plots of \( \hat{\theta}_{\Omega,s} \) as a function of \( s \) and \( t \), for three different sample sizes \( n \). We also overlay the function \( \tau_{\Omega}^s(s) = \arg\max_t \hat{\theta}_{\Omega,s}(t) \) as a function of \( s \). Finally, we also show the result of iteratively applying \( \tau_{\Omega}^s \), starting from \( s^{(1)} = 6 \). For smaller sample sizes, in this case, we have cycling. For larger sample sizes, we no longer have cycling. We can see in the intermediate case that we only pick up the local maximum through the iterative procedure, although it is the global maximum of \( \hat{\theta}_{\Omega,S_{\Omega}} \) where \( S_{\Omega}^* \) is 3.6494. Table 1 provides the full numerical results of each iteration.

4 Increased Sampling Coverage

Unfortunately, in some already quite simple cases, the sample size required for \( \hat{\theta}_{\Omega,s} \) to converge to \( \theta \) can be excessively large. In particular, \( \hat{\theta}_{\Omega,s} \) may not reflect at all the shape of \( \theta \) especially when \( s \) is far from \( t \). The cause of this behaviour is that sampling from \( f_s \) may not cover regions where \( f_t \) is located. We can address this by modifying the sampling distribution \( f_s \) in order to increase this coverage. There are various ways of doing this: we can use a convex mixture of the original distribution and an additional distribution with large variance, or if possible we may also simply inflate the variance of the distribution directly.

For example, imagine that we wish to ensure that our sampling distribution covers the entire interval \([a, b]\) on the real line. For this purpose, we modify our sampling distribution to:

\[
f_s^R(x) := (1 - \alpha)f_s(x) + \frac{\alpha}{b-a}I_{x \in [a,b]}
\]
Figure 1: Exact probability $\theta$, standard Monte Carlo estimation $\hat{\theta}_\Omega$, and estimation $\hat{\theta}_{1,s=3.6494}$, for a fixed value of $s = 3.6494$ where $\theta$ has a local maximum, as a function of $t \in T = [-7, 7]$ (left). Contour plot of exact function $\vartheta(s, t) := \theta(t)$ and its maximum $\tau^*(s)$ (right), for comparison with the contour plots of the estimators in fig. 2.

Figure 2: Contour plots of $\hat{\vartheta}_\Omega(s, t) := \hat{\theta}_\Omega(s, t)$ and depiction of $\tau^*_\Omega(s) = \arg \max_{t \in T} \hat{\theta}_\Omega(s, t)$ (blue line) for three different sample sizes $n = 1000$, 10000, and 100000. The path of the iteration with starting value $s^{(1)} = 6$ is plotted as a red line. We have cycling (left), convergence to a local maximum (middle), and convergence to the global maximum (right).

<table>
<thead>
<tr>
<th>$N = 1000$, $s^{(1)} = 6$</th>
<th>$N = 10000$, $s^{(1)} = 6$</th>
<th>$N = 100000$, $s^{(1)} = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$s^{(k)}$</td>
<td>$\tau^*_\Omega(s^{(k)})$</td>
</tr>
<tr>
<td>1</td>
<td>6.0000</td>
<td>3.8500</td>
</tr>
<tr>
<td>2</td>
<td>3.8500</td>
<td>3.5560</td>
</tr>
<tr>
<td>3</td>
<td>3.5560</td>
<td>0.4760</td>
</tr>
<tr>
<td>4</td>
<td>0.4760</td>
<td>3.9900</td>
</tr>
<tr>
<td>5</td>
<td>3.9900</td>
<td>3.6540</td>
</tr>
<tr>
<td>6</td>
<td>3.6540</td>
<td>0.5880</td>
</tr>
<tr>
<td>7</td>
<td>0.5880</td>
<td>4.0740</td>
</tr>
<tr>
<td>8</td>
<td>4.0740</td>
<td>3.6960</td>
</tr>
<tr>
<td>9</td>
<td>3.6960</td>
<td>0.6160</td>
</tr>
<tr>
<td>10</td>
<td>0.6160</td>
<td>4.0880</td>
</tr>
</tbody>
</table>

Table 1: Iteration steps.
where \( I_{x \in [a, b]} = 1 \) if \( x \in [a, b] \) and 0 otherwise. To generate a sample from this distribution, we let

\[
x_s^R(U, V) := \begin{cases} a + (b - a)x & \text{if } U < \alpha \\ x_s(V) & \text{if } U \geq \alpha \end{cases}
\]

(23)

where \( U \) is a uniform \([0, 1]\) variable, and \( V \) is as before.

Now, we start from an i.i.d. sample \( \Omega := (U_1, V_1, \ldots, U_n, V_n) \) to obtain an i.i.d. sample

\[
x_s^R(U_1, V_1), \ldots, x_s^R(U_n, V_n)
\]

(24)

from \( f_s^R \), for some fixed \( s \in \mathcal{T} \). As before, we can use this sample from \( f_s \) to estimate the expectation of \( h \) with respect to \( f_t \), for any \( t \in \mathcal{T} \):

\[
\hat{\theta}_{\Omega, s}(t) := \frac{1}{n} \sum_{i=1}^{n} w_{st}(x_s^R(U_i, V_i))h(x_s^R(U_i, V_i))
\]

(25)

where

\[
w_{st}(x) := \frac{f_t(x)}{f_s^R(x)}.
\]

(26)

Here too, \( \hat{\theta}_{\Omega, s}(t) \) is an unbiased estimator for \( \theta(t) \):

\[
E(\hat{\theta}_{\Omega, s}(t)) = \theta(t).
\]

(27)

Obviously, we should choose \( a \) and \( b \) in a way that we are sure to generate samples that cover the range of \( f_t \) across all \( t \in \mathcal{T} \). For \( \alpha \), if we choose it too small, then we may not generate sufficient samples across the desired interval \([a, b]\). If we choose it too large, then \( f_s^R \) may be too far away from \( f_t \) regardless of our choice of \( s \), thereby removing the opportunity to increase the accuracy of the estimate through the iterative procedure, as the estimator will no longer depend on \( s \). A suggestion is to choose \( \alpha \) in a way that for every \( t \in \mathcal{T} \), there are at least a handful of samples that cover \( f_t \).

For more general cases, a convex mixture of target distributions \( f_t \) could also be used. For example, if \( \mathcal{T} = [0, 1] \), then we could take

\[
f_s^R(x) := (1 - \alpha)f_s(x) + \alpha \int_0^1 f_t(x) \, dt
\]

(28)

and

\[
x_s^R(U, V) := \begin{cases} x_{U/\alpha}(V) & \text{if } U < \alpha \\ x_s(V) & \text{if } U \geq \alpha \end{cases}
\]

(29)

where \( U \) is a uniform \([0, 1]\) variable, and \( V \) is as before. If need be, the integral can be approximated by a finite sum, and the sampling function can be adjusted accordingly. Theoretical arguments and further proposals for convex mixtures have been given in [10, §4.3].

The next example shows how we can improve coverage using variance inflation of the sampling distribution.

**Example 2** We consider the estimation of the lower probability of the event \( D = (\infty, -1] \cup [1, \infty) \) with respect to the set of normal distributions with mean \( t \in [-7, 7] \) and variance \( \sigma^2 = 2 \).

The probability \( \theta(t) \) is depicted in fig. 3 (left) for each value \( t \in \mathcal{T} \). The minimum (lower probability) \( \theta_* \) is equal to 0.4795 and is achieved for \( t_* = 0 \).

First, we consider \( f_s^R = f_s \), that is, no increased coverage from the sampling distribution. The cheaper function \( \hat{\theta}_{\Omega, s=0} \) using reweighting based on density \( f_s^R \) provides a good approximation near the solution \( t_* = 0 \) but not far away from \( t_* \). In this example, \( \hat{\theta}_{\Omega, s=0} \) leads to a completely wrong global minimum 0.0282 achieved at \( t = -7 \), as seen from the dashed curve in fig. 3 (left). The reason why this approximation is bad for \( t \) far away from \( t_* = 0 \) is that there are no or almost no sample points of the reweighting density \( f_s^R \) in areas \( D \) with high density of \( f_t \) for \( t \) going to \( \pm 7 \).

Figures 5 and 6 show what happens if instead we take \( f_t^R \) to be normally distributed with mean \( t \) but with increased variance \( \sigma^2 = 10 \). We observe that \( \hat{\theta}_{\Omega, s} \) provides a reasonably good approximation for \( \theta \) across a much wider range of values for \( s \) and \( t \). For the lowest sample size, 1000, we still observe cycling in the iterative method, as the variance is not sufficiently inflated to cover the more extreme ends of the range of the distributions that we are interested in. For larger sample sizes, the coverage becomes sufficient, and the iterative procedure produces a correct value.
Figure 3: Exact probability $\theta$, standard Monte Carlo estimation $\hat{\theta}_\Omega$ and estimation $\hat{\theta}_{\Omega,s}$ for a fixed value of $s$, as a function of $t \in T = [-7,7]$ (left). Contour plot of exact function $\vartheta(s,t) := \theta(t)$ and its minimum $\tau(s) := t^*$ (right), for comparison with the contour plots of the estimators in figs. 4 and 6.

Figure 4: Contour plots of $\hat{\vartheta}(s,t) := \hat{\theta}_{\Omega,s}(t)$ and depiction of $\tau(s) = \arg\min_{t \in T} \hat{\theta}_{\Omega,s}(t)$ (blue line) for three different sample sizes $n = 1000$, 10000, and 100000. The path of the iteration with starting value $s^{(1)} = 6$ is plotted as a red line. We have cycling in all cases, due to lack of coverage of the sampling distribution in regions where $t$ is far from $s$.

Figure 5: Exact probability $\theta$, standard Monte Carlo estimation $\hat{\theta}_\Omega$ and estimation $\hat{\theta}_{\Omega,s}$ for a fixed value of $s = 0$, as a function of $t \in T = [-7,7]$.
Figure 6: Contour plots of $\hat{\vartheta}(s,t) := \hat{\theta}_{\Omega,s}(t)$ and depiction of $\tau^*_{\Omega}(s) = \arg\min_{t \in T} \hat{\theta}_{\Omega,s}(t)$ (blue line) for three different sample sizes $n = 1000$, 10000, and 100000. The path of the iteration with starting value $s^{(1)} = 6$ is plotted as a red line. We still have cycling for sample size 1000, although $\hat{\theta}_{\Omega,s}(t)$ is clearly already a lot closer to $\theta(t)$ for a much wider range of values for $s$ and $t$. For the larger sample sizes, the iterative procedure converges quickly.

5 Conclusion

We set out to explore importance sampling techniques for calculating lower and upper expectations with respect to sets of probability distributions. We revisited the iterative algorithm proposed in [6, 7], and put it on a better mathematical foundation, by formulating it as a fixed point of an operator. We provided some intuition under which this operator has a fixed point, and thereby provides a good estimator.

We explored some numerical examples, and found that the procedure breaks down when the sampling distribution provides insufficient coverage. We proposed three simple methods to increase coverage. Nevertheless, quantifying the conditions under which importance sampling can provide a sufficiently accurate estimate under a wide range of importance sampling distributions remains an important open question.

Acknowledgments

This work is partially supported by the H2020 Marie Curie ITN, UTOPIAE, Grant Agreement No. 722734.

References


Global sensitivity analysis of uncertain linear structures subject to static load

Marcos A. Valdebenito,* Iván V. González, and Héctor A. Jensen

Civil Engineering Department, Santa Maria University, Chile

*Corresponding author: marcos.valdebenito[at]usm.cl

Abstract

This contribution presents a strategy for performing global sensitivity analysis of uncertain linear structures. Global sensitivity is expressed in terms of the so-called Sobol’ indices. These indices are calculated by means of Monte Carlo simulation involving substructuring and resampling. The application of the proposed framework is illustrated by means of a numerical example. The results obtained suggest that the proposed strategy can be highly efficient from a numerical viewpoint without compromising accuracy.

Keywords: Uncertain Linear Systems, Static Load, Global Sensitivity, Sobol’ Indices, Substructuring, Resampling.

1 Introduction

Sensitivity analysis can provide valuable insight on the behavior of uncertain mechanical and structural systems, pinpointing its most influential input parameters [1]. Such type of information is crucial for, e.g. risk analysis and decision making [2, 3]. One approach for performing sensitivity analysis is applying variance-based measures, such as Sobol’ indices [4]. These indices provide information on the fraction of the variance of the output response of a model that can be attributed to a particular random input variable or a group of input variables. The main advantages of Sobol’ indices when compared to other approaches for sensitivity analysis are its relatively ease of implementation and ability to capture higher-order effects [5, 6].

The estimation of Sobol’ indices is usually carried out via Monte Carlo simulation [7]. This is numerically demanding, as it comprises hundreds of thousands or even millions of deterministic structural analyses. Therefore, different approaches to circumvent this issue have been developed, including e.g. application of surrogate models [8], special sampling techniques [9], etc. In this context, this contribution presents an approach for calculating Sobol’ indices most efficiently for a particular class of problems, i.e. uncertain linear structures subject to static load. The approach is based on two concepts: substructuring [10] and resampling [11]. Substructuring is applied for creating a database of precomputed solutions for portions of the structure. Then, these precomputed solutions are resampled, producing a large number of samples of the structural response, but at a fraction of the cost associated with a full structural analysis. This is due to the fact that the substructures have already been analyzed separately (i.e. precomputed); therefore, a structural analysis at the interface level suffices for calculating the structural response. In this way, it is possible to estimate the sought sensitivity indices with high numerical efficiency. The application and advantages of the proposed scheme are demonstrated through a numerical example.

2 Formulation of the Problem

2.1 Structural Model

Consider a structural or mechanical system which is modeled as linear elastic and subject to static loading and which is represented using an appropriate framework such as, e.g. the finite element method [12]. The numerical model comprises a total of \( N_D \) degrees-of-freedom and it is assumed that some of its input parameters are not known precisely and are characterized as random variables. For the sake of simplicity, it is further assumed that the uncertain parameters are related to structural properties while loading is deterministic. The uncertain parameters of the model include a total of \( n_\theta \) random variables \( \theta_i, \ i = 1, \ldots, n_\theta \) (collected in vector \( \theta \)) whose probability distributions are \( p_i(\theta_i), \ i = 1, \ldots, n_\theta \). The equilibrium equation associated with the numerical model is:

\[
K(\theta)u(\theta) = w
\]
where $K(\theta)$ is the stiffness matrix, $u(\theta)$ is the displacement vector and $w$ is the loading vector. It is considered that the $l$-th component of the displacement vector is of particular interest and it is identified in the following as $y$, i.e. $y = y(\theta) = w_l(\theta)$.

### 2.2 Global Sensitivity Analysis: Sobol’ Indices

A possible means for characterizing the impact of the uncertain input parameters $\theta$ on the response of interest $y$ is carrying out a variance-based sensitivity analysis. This means that one is interested in determining the portion of the variance associated with $y$ that can be attributed to each uncertain parameter $\theta_i$ (or groups of them). This framework for carrying out sensitivity analysis has been termed in the literature as Sobol’ indices [7]. The framework consists of decomposing $y(\theta)$ into a summation of functions of increasing dimensionality with respect to $\theta$ such that the variance of $y$ is expressed as a summation of conditional variances.

Sobol’ indices allow determining the fraction of the response variance that can be attributed either to an input variable on its own (first-order indices) or to an input variable and its interaction with other variables (total indices). In the more general case, Sobol’ indices allow determining the fraction of the response variance that can be attributed to a group of random input variables. In this contribution, the focus is on first-order and total indices.

The first order Sobol’ index, denoted as $S_i$, is equal to the variance of the expected value of the response of interest conditioned on $\theta_i$ normalized by the total variance of the response. Mathematically, $S_i$ is defined as follows.

$$S_i = \frac{\mathbb{V}_\theta [\mathbb{E}_{\theta_{-i}} [y|\theta_i]]}{\mathbb{V}_\theta [y]}, \quad i = 1, \ldots, n_\theta$$

In the above equation, $\mathbb{E}_x$ and $\mathbb{V}_x$ denote expected value and variance operators calculated with respect to the set of random variables $x$, respectively, while $\theta_{-i}$ denotes the set of all random input variables associated with the model except for $\theta_i$. The total Sobol’ index, denoted as $S_{Ti}$, is defined as the complement of the variance of the expected value of the structural response conditioned on all random input variables of the model except $\theta_i$, normalized by the total variance.

$$S_{Ti} = 1 - \frac{\mathbb{V}_\theta [\mathbb{E}_{\theta_{-i}} [y|\theta_i]]}{\mathbb{V}_\theta [y]}, \quad i = 1, \ldots, n_\theta$$

### 2.3 Calculation of Sobol’ Indices

It can be noted from eqs. (2) and (3) that the evaluation of Sobol’ indices demands calculating expected values and variances of the structural response conditioned on different subsets of the uncertain input parameters. In turn, this means that a number of integrals involving the structural response and subsets of the uncertain input values and variances of the structural response conditioned on different subsets of the uncertain input parameters must be calculated. In general, the number of integrals to be calculated can be large. Furthermore, for cases of practical interest, the response of interest $y$ cannot be calculated analytically, as seen from eq. (1). The latter two issues favor the application of the Monte Carlo method for calculating Sobol’ indices [7]. The procedure to calculate the indices is the following (see, e.g. [1]).

1. Generate two independent sets $A$ and $B$ comprising $N$ independent samples of $\theta$ distributed according to $p_\theta(\theta_i)$, $i = 1, \ldots, n_\theta$. Each of these sets is actually a matrix of dimensions $n_\theta \times N$, i.e. each column of matrices $A$ and $B$ contains a sample of the input parameters of the structural model. The value $N$ should be chosen such that convergence of Monte Carlo integration is ensured [7], e.g. $N = 10^5$.

$$A = \begin{bmatrix} \theta_A^{(1)} & \theta_A^{(2)} & \ldots & \theta_A^{(j)} & \ldots & \theta_A^{(N)} \end{bmatrix}$$
$$B = \begin{bmatrix} \theta_B^{(1)} & \theta_B^{(2)} & \ldots & \theta_B^{(j)} & \ldots & \theta_B^{(N)} \end{bmatrix}$$

2. Define matrix $C_i$, $i = 1, \ldots, n_\theta$ equal to $B$ except for the $i$-th row, which is equal to the $i$-th row of matrix $A$. That is:

$$C_i = \begin{bmatrix} \theta_C^{(1)} & \theta_C^{(2)} & \ldots & \theta_C^{(N)} \end{bmatrix}, \quad i = 1, \ldots, n_\theta$$

$$\theta_C^{(j)} = \begin{bmatrix} \theta_1^{(j)} & \theta_2^{(j)} & \ldots & \theta_A^{(j)} & \ldots & \theta_B^{(j)} \end{bmatrix}^T, \quad j = 1, \ldots, N$$
3. Calculate the response of interest \( y \) for each set of samples (i.e. each column) of matrices \( A, B \) and \( C_i, i = 1, \ldots, n_\theta \).

\[
y^{(j)}_A = f(\theta^{(j)}_A) \quad y^{(j)}_B = f(\theta^{(j)}_B) \quad y^{(j)}_C = f(\theta^{(j)}_C), \quad i = 1, \ldots, n_\theta, \quad j = 1, \ldots, N
\]  

(8)

4. Calculate the following estimators.

\[
f_0 = \frac{1}{N} \sum_{j=1}^{N} y^{(j)}_A \quad D = \frac{1}{N} \sum_{j=1}^{N} y^{(j)}_A^2 - f_0^2
\]

\[
U_{S_i} = \frac{1}{N} \sum_{j=1}^{N} y^{(j)}_A y^{(j)}_C \quad U_{T_i} = \frac{1}{N} \sum_{j=1}^{N} y^{(j)}_B y^{(j)}_C
\]

(9)

(10)

5. The sought Sobol’ indices are estimated applying the following formulas.

\[
S_i = \frac{U_{S_i} - f_0^2}{D} \quad T_i = 1 - \frac{U_{T_i} - f_0^2}{D}, \quad i = 1, \ldots, n_\theta
\]

(11)

As noted from the procedure described above, the structural response must be evaluated a total of \( N_T = N(n_\theta + 2) \) times considering all samples of the input structural parameters contained in sets \( A, B \) and \( C_i, i = 1, \ldots, n_\theta \). Hence, the total number of samples of the structural response \( N_T \) is usually quite large in practical applications, e.g. \( N_T > 10^6 \). Undoubtedly, this imposes a huge numerical burden, as a prohibitive number of structural analyses must be carried out. Therefore, in the following, an approach for alleviating this numerical burden is discussed.

3 Proposed Approach

3.1 General Remarks

This section discusses an approach that allows to generate a large number of samples of the structural response most efficiently; then, these samples can be considered to calculate Sobol’ indices. The proposed approach combines substructuring and resampling. Each of these steps is discussed separately below.

3.2 Static Substructuring

Assume that each component of vector \( \theta \) affects a distinct portion of the structural system. Then, it is possible to separate the complete structure into \( n_\theta \) substructures connected between them by interfaces. Each substructure contains internal degrees-of-freedom, denoted by index \( I \), and boundary (or interface) degrees-of-freedom, denoted by index \( B \). Hence, the equilibrium equation associated with substructure \( i \) can be written using block matrix notation as follows [10].

\[
\begin{bmatrix}
K^{(i)}_{B,B} (\theta_i) & K^{(i)}_{B,I} (\theta_i) \\
K^{(i)}_{I,B} (\theta_i) & K^{(i)}_{I,I} (\theta_i)
\end{bmatrix}
\begin{bmatrix}
\mathbf{u}^{(i)}_B (\theta) \\
\mathbf{u}^{(i)}_I (\theta)
\end{bmatrix} =
\begin{bmatrix}
\mathbf{w}^{(i)}_B (\theta) \\
\mathbf{w}^{(i)}_I (\theta)
\end{bmatrix}
\]

(12)

The preceding equation can be expressed in terms of the boundary degrees-of-freedom applying static condensation [10, 13], yielding:

\[
\tilde{K}^{(i)} (\theta_i) \mathbf{u}^{(i)}_B (\theta) = \tilde{w}^{(i)} (\theta_i)
\]

(13)

where \( \tilde{K}^{(i)} (\theta_i) \) and \( \tilde{w}^{(i)} (\theta_i) \) are the condensed stiffness matrix and condensed load vector associated with substructure \( i \) and are defined as:

\[
\tilde{K}^{(i)} (\theta_i) = K^{(i)}_{B,B} (\theta_i) - K^{(i)}_{B,I} (\theta_i) \left(K^{(i)}_{I,I} (\theta_i)\right)^{-1} K^{(i)}_{I,B} (\theta_i)
\]

(14)

\[
\tilde{w}^{(i)} (\theta_i) = w^{(i)}_B - K^{(i)}_{B,I} (\theta_i) \left(K^{(i)}_{I,I} (\theta_i)\right)^{-1} w^{(i)}_I
\]

(15)

The condensed stiffness matrix and load vector of all substructures can be assembled in order to produce an equilibrium equation that involves all boundary degrees-of-freedom of the structure [10, 13]:

\[
\tilde{K} (\theta) \mathbf{u}_B (\theta) = \tilde{w} (\theta)
\]

(16)
Global sensitivity analysis of uncertain linear structures subject to static load

where \( \hat{K} \) denotes the stiffness matrix associated with the boundary degrees-of-freedom while \( \hat{w} \) and \( \hat{u}_B \) are the load and displacement vectors associated with the boundary degrees-of-freedom, respectively. Once the displacements \( \hat{u}_B \) have been determined by means of eq. (16), it is possible to determine the displacements at the internal degrees-of-freedom of the \( i \)-th substructure by means of the following expression:

\[
\hat{u}_i^{(i)}(\theta) = \left( K^{(i)}_{i,i}(\theta_i) \right)^{-1} \left( \hat{w}_i^{(i)} - K^{(i)}_{i,B}(\theta_i) \hat{u}_B^{(i)}(\theta) \right)
\]  

(17)

It should be noted that calculating the displacement vector applying either eq. (1) or eqs. (16) and (17) is completely equivalent, as it produces identical results and demands the same amount of numerical efforts. However, the application of substructuring may become advantageous in context with resampling, as described in the next section.

3.3 Resampling

A possible means for reducing the numerical costs associated with the calculation of \( N \) samples of the structural response is the application of resampling in combination with substructuring [11]. The main idea behind this procedure is the following. In a first stage, a relatively small number \( N_A \) of samples of the uncertain input parameters \( \theta^{(j)} \), \( j = 1, \ldots, N_A \) is generated, such that \( N_A < N \). For each of these samples, structural analysis is carried out at the substructure level. That is, the condensed stiffness and load vectors are generated in terms of the boundary degrees-of-freedom (see eqs. (14) and (15), respectively). Then, in a second stage, these stiffness matrices and load vectors associated with each substructure are assembled in order to calculate the displacements at both the boundary and internal degrees-of-freedom (see eqs. (16) and (17), respectively). Nonetheless, it is at this second stage where the concept of resampling comes into play: at the moment of assembling the stiffness matrix and load vector, the data required from the substructure level is chosen at random (with replacement) from the available set of \( N_A \) samples generated at the first stage. The advantage of applying resampling is that it is possible to generate a total number of samples of the structural response \( N \) which is larger than the total number of samples of each substructure \( N_A \) as a single sample of a substructure is used more than once. In fact, the application of this scheme would allow – in theory – to generate a total of \( N_A^n \) samples of the structural response. The whole idea of the resampling scheme is illustrated schematically in fig. 1, where it is assumed that a structure is modeled considering two substructures (\( n_\theta = 2 \)), while \( N_A = 2 \) and \( N = 3 \).

![Figure 1: Schematic representation of substructuring and resampling](image_url)

The main advantage of the scheme for generating samples of the structural response based on substructuring and resampling is the following. The calculation of \( N \) of the response demands performing structural analysis at the substructure level a total of \( N_A \) times plus additional \( N \) structural analysis at the boundary degrees-of-freedom. Recalling that \( N_A < N \) and that structural analysis at the level of the boundary degrees-of-freedom (see eq. (16)) can be numerically less demanding than performing a full system analysis (see eq. (1)), the proposed scheme can report important savings in numerical efforts. It should be noted that the resampling scheme described above produces samples of the structural response which possess a degree of correlation. This is due to the fact that one sample of a substructure is considered more than once. While this correlation does not introduce bias in the estimator of the Sobol’ indices, it may increase its variability [14]. At the moment, the quantification of the effect of this correlation remains an open challenge.

4 Example

The application of the proposed scheme for calculating Sobol’ indices that combines substructuring and resampling is illustrated by means of the following example. The example considered herein is a simplified model of a human tooth, as depicted in fig. 2, which has been adapted from [15]. The finite element
model of the tooth is developed under the assumption of a plane strain state and it involves a total of 15312 degrees-of-freedom. The anatomy of a human tooth can be modeled considering five different layers: bone, periodontal ligament, enamel, dentine and pulp. Due to several reasons (age, nutrition, dental hygiene, etc.), there is uncertainty concerning the elastic properties that characterize the behavior of the different layers. Hence, the Young’s moduli of each layer are characterized considering truncated Gaussian distributions with mean value and coefficient of variation as shown in Table 1; please note that the statistics reported in the Table correspond to the Gaussian distributions before truncation. The characterization of the uncertainty associated with the Young’s moduli is based on the data reported in [15–20]. The loading over the model consists of a deterministic horizontal point load of 1 [N] that simulates the effect of an orthodontics appliance. The response of interest is the horizontal displacement of the cusp of the tooth and hence, both first-order and total Sobol’ indices associated with this response are calculated.

![Figure 2: (a) Human tooth subject to static loading (b) finite element model](image)

Table 1: Parameters of the human tooth model

<table>
<thead>
<tr>
<th>Tooth layer</th>
<th>Poisson’s ratio</th>
<th>Young’s modulus Mean value $\mu_E$ [MPa]</th>
<th>Coefficient of Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Bone</td>
<td>0.30</td>
<td>14000</td>
<td>8%</td>
</tr>
<tr>
<td>2 Ligament</td>
<td>0.45</td>
<td>100</td>
<td>20%</td>
</tr>
<tr>
<td>3 Enamel</td>
<td>0.30</td>
<td>80000</td>
<td>10%</td>
</tr>
<tr>
<td>4 Dentine</td>
<td>0.31</td>
<td>17000</td>
<td>18%</td>
</tr>
<tr>
<td>5 Pulp</td>
<td>0.49</td>
<td>4</td>
<td>30%</td>
</tr>
</tbody>
</table>

The Sobol’ indices are estimated first applying Monte Carlo simulation considering a total of $N = 10^5$ samples. No substructuring or resampling schemes are applied and thus, the structural response is calculated by repeatedly solving eq. (1) for different samples of the input parameters. The results obtained are depicted in fig. 3 and are shown with dashed line. Please note that in this figure, the curves associated with the sensitivity indices of bone, enamel and pulp overlap, as their overall effect on the response under study is negligible.

In a second step, Sobol’ indices are estimated applying the substructuring and resampling scheme. Five substructures are considered for the analysis and each substructure matches the different layers of the tooth model. The number of samples of each substructure is set equal to $N_A = 10^3$; then, a total of $N = 10^5$ samples are generated by means of resampling. The results obtained are depicted in fig. 3 with solid line. This figure suggests an excellent agreement between the reference results and the results obtained with substructuring and resampling, as both lead to almost identical estimates of the first-order and total Sobol’ indices. Furthermore, the proposed approach for generating samples of the structural responses reduces the numerical costs considerably. In fact, a speedup factor of 2500 was observed when comparing the proposed approach with that where no substructuring and resampling are considered.

The results obtained reveal that the variability of the horizontal displacement of the tooth’s cusp is highly

---

REC 2018 – M.A. Valdebenito, I.V. González, H.A. Jensen
Global sensitivity analysis of uncertain linear structures subject to static load

Figure 3: (a) First-order and (b) total Sobol’ indices estimates. Solid line: proposed approach combining substructuring and resampling; dashed line: direct simulation (considering neither substructuring nor resampling).

influenced by the ligament. This make sense from a physical viewpoint, as both the bone and dentine are relatively rigid layers while the ligament that connects them is quite flexible.

5 Conclusions

The results presented in this contribution suggest that the application of substructuring and resampling for estimating Sobol’ indices can lead to substantial savings in numerical efforts without compromising accuracy. However, there are several practical issues that should be further investigated. For example, at the moment there is no criterion available for the selection of the number of substructures $N_A$ to be resampled. In addition, the application of resampling induces a correlation in the samples of the structural response that should be quantified. Furthermore, the possibility of accounting for more than one uncertain parameter per substructure should be addressed as well. All of these issues are currently under investigation.

Acknowledgments

This research is partially supported by CONICYT (National Commission for Scientific and Technological Research) under grant number 1180271. This support is gratefully acknowledged by the authors.

References

[9] G. Jia and A. Taflanidis, Efficient evaluation of Sobol’ indices utilizing samples from an aux-


An Efficient method for estimating conditional failure probabilities

Domenico Altieri,1 and Edoardo Patelli.1*

1Institute for Risk and Uncertainty, University of Liverpool, United Kingdom
*Corresponding author: edoardo.patelli@liverpool.ac.uk

Abstract

Conditional reliability measures provide a more detailed description of the performance of a system, being representative of different initial configurations. Commonly, since the failure region is characterized by a small probability of occurrence, advanced sampling techniques are required to reduce the computational effort of a simulation based approach. These techniques if on one hand decrease the number of samples needed to identify the failure domain, on the other hand do not generally allow a direct estimation of the conditional failure probability for different given inputs.

This study aims at providing an efficient and simple methodology to evaluate the conditional failure probability in the case of a static reliability analysis. In particular, under the assumption of probability density functions (PDFs) with a finite support, the failure region mapping process is carried out using surrogate PDFs associated with Sobol’ sequences in order to reduce as much as possible the model evaluations. Finally, the integration of the failure region in the standard normal space employs probabilistic weights instead of a classic indicator function to account for the uncertainty associated with the failure region definition.

The approach is verified by comparing the results against those obtained from a Latin Hypercube Sampling. The performance of the proposed method is evaluated in terms of computational costs and accuracy.

Keywords: Conditional failure probability, Reliability analysis, Failure region

1 Introduction

Despite the constant evolution of modern computers, numerical models require significant computational efforts to achieve the desired level of accuracy, in particular, for detailed and complex systems efficient simulation techniques result to be often of paramount importance to reach the required solution in a feasible time.

In the reliability analysis field, many advanced approaches [1–4] have been proposed in the last decades, having as main purpose a reduction in the number of simulations required for the failure probability estimation. Unfortunately, the techniques adopted by the aforementioned methodologies [5, 6] do not allow a direct estimation of reliability measures conditioned to specific sub-domains of the probabilistic space. Conditional failure probability represents indeed an important sensitivity indicator, since partial derivatives \( \frac{\partial P_f}{\partial v_i} \), that describe the effects of uncertain input variables \( v_i \) on the system reliability level, can be subsequently computed [7–9]. Moreover, an efficient approach is required especially when a reliability-based optimization is performed [7, 10, 11] and a full reliability analysis must be carried out for each sub-domain analyzed by the optimization algorithm.

The proposed approach aims at computing conditional failure probabilities starting from a failure region mapping based on a multinomial logistic regression [12]. The training of the classification algorithm employs quasi-random samples [13] that allow a further reduction in the model realizations required.

A simple case study that estimates the maximum displacement of a cantilever beam, with uncertain material and geometrical properties, subject to a random concentrated load is analyzed. Finally, a Latin Hypercube Sampling provides the reference results based on which a first comparison with the proposed method is carried out. The opensource Matlab toolbox OpenCossan [14, 15] is used for the probabilistic analysis.

2 Methodology

Given the failure event \( F \) and a general sub-domain \( B \) of the probabilistic space \( \Omega \), the conditional failure probability can be expressed following the Kolmogorov definition as:
An efficient method for estimating conditional failure probabilities

\[ P(F|B) = \frac{P(F \cap B)}{P(B)} \]

(1)

where \( P(F \cap B) \) represents the joint probability of the events \( F \) and \( B \), while the probability \( P(B) \) of the event \( B \) can be computed through the integral of the joint probability function \( \Phi(x) \) over the sub-domain \( B \), with \( x \) the vector of coordinates in \( \Omega \). In general, given a failure event \( F \), the associated probability of occurrence can be expressed as the integral of \( \Phi(x) \) over \( F \):

\[ P(F) = I = \int_{F} \Phi(x) dx = \int_{\Omega} \Phi(x) \cdot \hat{L} I dx \]

(2)

where \( I \) represents an indicator function equal to 1 when the system fails and 0 otherwise. Unfortunately, the multi-dimensional integration over the failure domain \( F \) may be often impossible to perform in a closed form [16], for this reason a simulation-based approach, that approximates the exact solution, is required.

The joint probability \( P(F \cap B) \), according to Equations 1 and 2, can be viewed as the integral of \( \Phi(x) \) in the standard normal space [17] over the sub-domain \( B \) and not \( \Omega \):

\[ P(F \cap B) = \int_{F \cap B} \Phi(x) dx = \int_{B} \Phi(x) \cdot \hat{L} I dx \]

(3)

The mapping from the physical space to the standard normal space can be easily obtained by assuming that the cumulative distribution functions of the random variables remain the same after the transformation in standard normal distributions [18]. Equation 1 can be now rewritten as:

\[ P(F|B) = \frac{P(F \cap B)}{P(B)} = \frac{\int_{B} \Phi(x) \cdot \hat{L} I dx}{\int_{B} \Phi(x) dx} \]

(4)

An approximated solution of \( \int_{B} \Phi(x) dx \) can be easily reached without additional computational efforts by means of deterministic samples \( \bar{x} \in B \) in the standard normal space [17]. For instance, to compute the failure probability \( P(F|V) \) conditioned to the variable \( V \), a sequence of sub-domains \( B_i = \{x_i \mid V = \text{constant} \} \) is explored by the samples \( \bar{x} \). More complex shapes \( s(x) \) of \( B \) may require a preliminary step to be defined, nonetheless \( s(x) \) just represents a combination of \( x \) depending on the specific problem.

An efficient solution of Equation 4 still needs an additional step to reduce the number of model realizations related to the indicator function \( I \).

### 2.1 Failure region mapping

A multinomial logistic regression (MLR) [12] is adopted to predict the probability that an initial system configuration reaches the failure region or not. The chosen classification algorithm allows the definition of intermediate classes in case conditional failure probabilities with respect other final performance levels are required (multiple failure regions). The indicator function in Equation 2 is therefore replaced by the probabilistic score \( w_i \) associated with each prevision \( Y_i \). In general, the probability that an observation \( i \) has \( K \) as predicted value is equal to:

\[ w_i = P(Y_i = K) = \frac{e^{\beta_K \cdot x_i}}{1 + \sum_{k=1}^{K} e^{\beta_k \cdot x_i}} \]

(5)

where \( x_i \) is the vector of attributes describing each observation \( i \), while \( \beta_K \) represents the vector of regression coefficients associated with the class \( K \).

The initial dataset for the MLR training is obtained by means of Sobol’ sequences from surrogate probability density functions in order to easily reach the failure region and getting at the same time a higher grade of coverage of the probabilistic space. Samples are generated from uniform distributions \( z(x) \) with upper and lower boundaries \((U, L)\) chosen according to the real distributions. In particular, given the original PDF \( f(x) \) and the associated cumulative function \( F(x) \), than \( F(L) \) and \( 1 - F(U) \) result to be lower than \( 10^{-7} \). An approximate solution of Equation 3 can now be obtained by sampling from the surrogate PDFs only in the sub-domain \( B \) and by evaluating the probabilistic score \( w_i \) associated with each realization:

\[ P(F \cap B) = \int_{B} \Phi(x) \cdot \hat{L} w_i dx \]

(6)
Despite the MLR algorithm is a linear classifier based on a separating hyperplane, more complex and general failure region shapes can be still mapped by adopting alternative kernels \[19\]. In Figure 1 a flowchart describes step-by-step the algorithm employed to estimate the failure probability corresponding to different \( B_i \).

### 3 Case study

A simple cantilever beam (Figure 2) subject to a random concentrated load, with uncertain material and geometric characteristics, has been analyzed as case study. The random variables are described in Table 1, with the associated uniform distributions employed for the failure region mapping. The maximum displacement \( R \) can be easily computed as:

\[
R = \frac{Q \cdot L^4}{8 \cdot E \cdot I} + \frac{P \cdot L^3}{3 \cdot E \cdot I} \tag{7}
\]

![Figure 2: Cantilever beam](image-url)
An efficient method for estimating conditional failure probabilities

Table 1: Random variables (R.V.), parameters and surrogate PDFs employed in the failure region mapping

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Type</th>
<th>μ</th>
<th>σ</th>
<th>z(x)</th>
<th>L</th>
<th>U</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load - P [N]</td>
<td>R.V.-Lognormal</td>
<td>5000</td>
<td>400</td>
<td>L.U</td>
<td>3900</td>
<td>6000</td>
<td>-</td>
</tr>
<tr>
<td>Elastic M.-E [MPa]</td>
<td>R.V.-Lognormal</td>
<td>10000</td>
<td>1600</td>
<td>R.V.-Uniform</td>
<td>1600</td>
<td>19000</td>
<td>-</td>
</tr>
<tr>
<td>Density - ρ [Kg/m^3]</td>
<td>R.V.-Lognormal</td>
<td>600</td>
<td>140</td>
<td>R.V.-Uniform</td>
<td>50</td>
<td>1350</td>
<td>-</td>
</tr>
<tr>
<td>Height - H [m]</td>
<td>R.V.-Normal</td>
<td>0.24</td>
<td>0.01</td>
<td>R.V.-Uniform</td>
<td>0.18</td>
<td>0.30</td>
<td>-</td>
</tr>
<tr>
<td>Length - L [m]</td>
<td>Parameter</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.8</td>
</tr>
<tr>
<td>Width - B [m]</td>
<td>Parameter</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.12</td>
</tr>
</tbody>
</table>

where \( Q = f(\rho) \) represents the uniformly distributed beam’s weight and \( I \) the moment of inertia.

4 Results

The system reliability has been analyzed assuming a maximum allowed displacement \( t = 8 \text{ mm} \) and the conditional failure probability \( P(R|E > t) \) has been evaluated for different given \( E \) values.

To assess the goodness of the result \( g(x) \) provided by the proposed method, a reference fragility curve \( fr(x) \), obtained through a Latin Hypercube Sampling, has been adopted. In particular, the comparison is carried out considering a different number of initial samples needed for the classification algorithm training.

Figure 3 shows the probabilistic score of each sample involved in the solution of Equation 4 for different fixed values of the elastic modulus \( E \). For illustrative purpose two only random variables have been considered.

Figure 4 reports the comparison with the reference curve obtained with a total of 7000 model realizations. The accuracy of the final match results to be dependent on the number of samples employed in the failure region mapping. In particular, by increasing the initial dataset the match improves reaching a good accuracy after only 250-350 samples.
Figure 4: Comparison for different datasets between the result $g(x)$ provided by the method and the target solution $f_r(x)$.

Figure 5 describes the evolution of the mean absolute error $D$ for different sizes of the dataset employed in the MLR training stage, where $D_N = \sum |g(x_N) - f_r(x_N)| / L$, with $N$ and $L$ equal to the number of initial samples and elements in $g(x_N)$ respectively. The variation of $D$ for different initial datasets shows a quick decrease of the error after only 300 samples, even if the results seem to not fully converge to the target solution, even increasing the training dataset, due to the limits in the MLR mapping capability.

Figure 5: Evolution of the mean absolute error $D$ for different training datasets.
5 Conclusions
A simplified approach to estimate conditional failure probabilities has been proposed. The reduction in the computational expense is reached by adopting a multinomial logistic regression trained over a predefined dataset generated with surrogate PDFs. The use of a quasi-random sampling technique allows obtaining a further improvement in the failure region mapping efficiency, even if the goodness of the final comparison with the target solution depends on the dimension of the dataset used for the training stage. Thanks to the MLR algorithm, the proposed approach provides even the possibility to define multiple thresholds in order to evaluate after one single training stage the system conditional reliability with respect different performance levels. According to the analyzed case study, the results show a good match starting from only 250 model realizations and a further reduction in the confidence bounds is obtained by increasing the initial samples. Additional analyses are required to check how the dimensionality affects the efficiency of the method and additional applications needed to be tested to generalize the effectiveness of the failure region mapping.

References
Why Taylor models and modified Taylor models are empirically successful: A symmetry-based explanation

Mioara Joldeş,1 Christoph Lauter,2 Martine Ceberio,3 Olga Kosheleva,4 and Vladik Kreinovich3*

1LAAS-CNRS, Toulouse, France
2LIP6, Sorbonne Université, France
3Department of Computer Science, University of Texas at El Paso, USA
4Department of Teacher Education, University of Texas at El Paso, USA
*Corresponding author: vladik[at]utep.edu

Abstract
In this paper, we show that symmetry-based ideas can explain the empirical success of Taylor models and modified Taylor models in representing uncertainty.
Keywords: Taylor models, Modified Taylor models, symmetries

1 Taylor Models and Modified Taylor Models: A Brief Reminder

General problem: reminder. In many practical situations:

• we know the equations that describe how desired quantities change with time (and space), and
• we want to use this knowledge to find the values of the desired quantities at different moments of time, at different locations, and for different values of the corresponding actions.

For example:

• we want to know how the location of a spaceship at time \( t \) depends on \( t \) and on the parameters of the launch: time, coordinates, orientation, etc.;
• another example is particle accelerators, in which particles move in a reasonable narrow tube with speeds close to the speed of light: we want to be able to predict the location of a particle beam at different moments of time for different values of the corresponding parameters.

In all these cases, we want to know the dependence of each desired quantity \( y \) on quantities \( x_1, \ldots, x_n \).

Often, we want guaranteed bounds. There are many numerical computation techniques for computing the desired dependence. However, numerical computations usually lead to an approximate description of the actual dependence: \( y_{\text{approx}}(x_1, \ldots, x_n) \approx y(x_1, \ldots, x_n) \).

In many real-life problems, it is important to provide not only an approximate value of \( y \) for given \( x_1, \ldots, x_n \), but guaranteed bounds on this value, i.e., values \( y_l(x_1, \ldots, x_n) \) and \( y_u(x_1, \ldots, x_n) \) for which

\[
y_l(x_1, \ldots, x_n) \leq y(x_1, \ldots, x_n) \leq y_u(x_1, \ldots, x_n).
\]

• This is important in space exploration: we want to make sure that the spaceship does not collide with anything and reaches its target within desired accuracy.
• This is important for particle accelerators: we want to make sure that the particle beam does not self-destruct by hitting the tube walls and instead, reaches the desired target.

Taylor models. In many real-life problems, the deviations from nominal values are small, so small that we can safely ignore terms which are quadratic or of higher order in terms of these deviations. In such situations, we can expand the desired dependence in Taylor series and keep only linear terms in this expansion. Hence linearization is one of the main tools in practical physics; see, e.g., [5].

To get a more accurate result, one can take into account quadratic terms. An even more accurate result emerges if we take into account cubic and higher order terms. In general, we take the sum of the first few terms in Taylor series and thus get a polynomial \( P(x_1, \ldots, x_n) \), i.e., a linear combination of monomials

\[
P(x_1, \ldots, x_n) = \sum c_{i_1 \ldots i_n} \cdot x_1^{i_1} \cdots x_n^{i_n}.
\]
A natural way to transform this approximate model into a guaranteed model is to supplement the approximate polynomial \( P(x_1, \ldots, x_n) \) with a guaranteed upper bound \( \Delta \) on the absolute value of the approximation error \( y(x_1, \ldots, x_n) - P(x_1, \ldots, x_n) \):

\[
|y(x_1, \ldots, x_n) - P(x_1, \ldots, x_n)| \leq \Delta.
\]

Once we know this upper bound, we can conclude that for each combination of values \( x_1, \ldots, x_n \), we have

\[
y(x_1, \ldots, x_n) \in P(x_1, \ldots, x_n) + [-\Delta, \Delta] = \sum c_{i_1 \ldots i_n} \cdot x_1^{i_1} \cdot \ldots \cdot x_n^{i_n} + [-\Delta, \Delta].
\]

The right-hand side of this inclusion is known as a Taylor model.

Taylor models has indeed been successfully used in many important applications; see, e.g., [1–3, 7, 8, 10].

**Comment.** The term “Taylor model” may be somewhat misleading: this term does not necessarily mean that the values \( c_{i_1 \ldots i_n} \) are the coefficients of the Taylor expansion of the original function \( y(x_1, \ldots, x_n) \), it simply means that to approximate the original function, we use a polynomial \( P(x_1, \ldots, x_n) \) – e.g., a Chebyshev polynomial – as opposed to approximating the function by, e.g., a rational expression.

**Taylor models and uncertainty.** In some cases, we know the exact dependence \( y(x_1, \ldots, x_n) \). This happens, e.g., in situations described by fundamental physics, when we know the exact equations. However, in many other cases, we only have an approximate model \( \tilde{y}(x_1, \ldots, x_n) \) for the desired dependence.

In many such cases, we know the upper bound \( \delta \) on the approximation error:

\[
|\tilde{y}(x_1, \ldots, x_n) - y(x_1, \ldots, x_n)| \leq \delta.
\]

In this case, if we have a polynomial \( P(x_1, \ldots, x_n) \) that approximates the model \( \tilde{y}(x_1, \ldots, x_n) \) with accuracy \( \delta' \), so that \( |\tilde{y}(x_1, \ldots, x_n) - P(x_1, \ldots, x_n)| \leq \delta' \), then we have

\[
|y(x_1, \ldots, x_n) - P(x_1, \ldots, x_n)| \leq \Delta \overset{\text{def}}{=} \delta + \delta',
\]

i.e., we have \( y(x_1, \ldots, x_n) \in P(x_1, \ldots, x_n) + [-\Delta, \Delta] \).

**Comment.** In addition to uncertainty about the dependence \( y(x_1, \ldots, x_n) \), we often also have uncertainty about the values of the quantities \( x_1, \ldots, x_n \). Indeed, these values usually come from measurements, and measurements are practically never absolutely accurate: the measurement results \( \tilde{x}_i \) are, in general, different from the actual (unknown) values of the corresponding quantities \( x_i \). In many cases, the only information that we have about the measurement errors \( \Delta x_i \overset{\text{def}}{=} \tilde{x}_i - x_i \) are the upper bounds \( \Delta_i \) on the measurement error’s absolute value: \( |\Delta x_i| \leq \Delta_i \). In such cases, once we have the measurement results \( \tilde{x}_i \), the only information that we have about each actual value \( x_i \) is that \( x_i \) belongs to the interval \( [\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i] \). In such situations, we need to find the range of the expression (1) when \( x_i \) takes values from these intervals.

**Modified Taylor models.** In the traditional Taylor model, we use a single upper bound \( \Delta \) to describe the approximation error for all possible combinations of the values \( x_i \). In many practical cases, however, the approximation error depends on the values \( x_i \). For example, when we predict a trajectory, usually,

- the predictions are very accurate for small \( t \),
- but become less and less accurate as the time \( t \) increases.

To get a better description of the model’s accuracy, it is therefore desirable to take into account that the approximation error may depend on \( x_i \).

One successful way to take this dependence into account was proposed in [4] under the name of modified Taylor models. In this description, each coefficient \( c_{i_1 \ldots i_n} \) is an interval:

\[
y(x_1, \ldots, x_n) \in \sum [c_{i_1 \ldots i_n}, \bar{c}_{i_1 \ldots i_n}] \cdot x_1^{i_1} \cdot \ldots \cdot x_n^{i_n}.
\]

**Comments.**

- The meaning of the formula (2) is that for each combination of the values \( x_1, \ldots, x_n \) there exist values \( c_{i_1 \ldots i_n} \in [c_{i_1 \ldots i_n}, \bar{c}_{i_1 \ldots i_n}] \) for which

\[
y(x_1, \ldots, x_n) = \sum c_{i_1 \ldots i_n} \cdot x_1^{i_1} \cdot \ldots \cdot x_n^{i_n}.
\]
• Models (2) also describe situations when we do not know the exact dependence $y(x_1, \ldots, x_n)$; in this case, different values of the coefficients $c_{i_1, \ldots, i_n}$ correspond to different possible dependency functions.

Why Taylor models? In the above text, we explained how Taylor series naturally appear if we consider more and more accurate polynomial approximations of the actual dependence $y(x_1, \ldots, x_n)$. The possibility to use polynomials comes from the known fact that any continuous function on a bounded domain can be, with any given accuracy, approximated by a polynomial.

However, many other families of functions have the same universal approximation property: rational functions, finite trigonometric series, functions corresponding to neural networks, etc. Several such approximate models have been used to approximate a real-life dependence, and in many practical cases, it turned out that Taylor models work the best – either they provide better accuracy within the same computation time, or faster computations for the same accuracy. For example, in particle accelerator design, Taylor modes turned to be the most efficient computational tool among the tools that have been tried.

This empirical fact natural leads to the first question that we will study in this paper: How can we explain this empirical fact? Why do Taylor series approximation turn out to be among the most empirically efficient?

Why modified Taylor models. The second question is related to the fact that, once we have selected the Taylor model, there are many ways to take into account uncertainty. In coming up with the formula (2), we represented the original polynomial as a linear combination of monomials, and allowed the coefficients at these monomials to vary within the appropriate intervals. Alternatively, we could select another non-monomial basis in the space of all polynomials, and allow coefficient corresponding to this basis to vary within some intervals. This will lead to a different family of approximating functions.

Empirical comparison [4] shows that in many practical situations, the formula (2) works better than formulas based on other bases: the formula (2) either leads to a better accuracy for the same computation time, or to faster computations if we fix the desired accuracy. So why is formula (2) empirically efficient?

What is known and what we do in this paper. A symmetry-based explanation of why Taylor models are efficient have been proposed in [9]. In this paper, we:

• first, use techniques from [6] to prove a stronger version of that result – thus providing a new justification for Taylor models, and

• then, extend this new result to a justification of the modified Taylor models.

This provides a theoretical explanation of why Taylor models work, and why modified Taylor models (2) are empirically successful.

2 Why Taylor Models: A New Justification

Towards formalization of the problem: we need to select a vector space. We want to select a family of functions $\mathcal{F}$, so that the results of our prediction have a form $y(x_1, \ldots, x_n) \in F(x_1, \ldots, x_n) + [-\Delta, \Delta]$ for some function $F \in \mathcal{F}$.

In the computer, the only directly hardware supported operations on real numbers are arithmetic operations. Programming languages also use elementary functions like sine, but inside the computer, all these functions are implemented via a sequence of arithmetic operations. In this sense, all computations on a computer reduce to a sequence of arithmetic operations.

Sometimes this sequence is explicit, sometimes it is hidden if we only consider the code in a high-level language, but it becomes clear if we take into account how the non-arithmetic mathematical functions are implemented.

Any function that is obtained by a sequence of arithmetic operations is analytical, i.e., it can be expanded into Taylor series. Thus, it is reasonable to restrict ourselves to analytical functions $F$.

We want to be able to represent functions from the class $\mathcal{F}$ inside a computer. Inside a computer, we can only represent a finite number of parameters, so it makes sense to consider finite-dimensional families of functions.

It would be useful to select the family $\mathcal{F}$ in such a way that an application of any arithmetic operation $\circ$ does not lead to additional approximation error. In other words, ideally, we would like to select $\mathcal{F}$ in such a way that, if two intermediate results $r$ and $s$ belong exactly to $\mathcal{F}$, then $r \circ s$ should also belong to $\mathcal{F}$. However, if we require that, then, since
we start with variables and
the family is closed under addition and multiplication,
we will end up with arbitrary polynomials, which contradicts to \( \mathcal{F} \) being finite-dimensional. So, we cannot require that the family \( \mathcal{F} \) be closed under all arithmetic operations:

- since we cannot require that for all operations,
- we should at least require it for the simplest ones: \(+, -\), and multiplication by a real number \( \lambda \).

In other words, we require that if \( F \in \mathcal{F} \) and \( G \in \mathcal{F} \), then \( F + G \in \mathcal{F} \) and \( \lambda \cdot F \in \mathcal{F} \). So, the family \( \mathcal{F} \) is a (finite-dimensional) vector space of functions.

**We should select the optimal vector space.** There are many possible vector spaces of functions. The question is: which of these vector spaces is the best (“optimal”) for our purpose?

When we say “the best”, we mean that on the set of all such spaces, there is a relation \( \succeq \) describing which family is better or equal in quality. This relation must be transitive (if \( \mathcal{F} \) is better than \( \mathcal{G} \), and \( \mathcal{G} \) is better than \( \mathcal{H} \), then \( \mathcal{F} \) is better than \( \mathcal{H} \)).

This relation must also clearly be reflexive: \( \mathcal{F} \succeq \mathcal{F} \) for every family \( \mathcal{F} \).

This relation is not necessarily asymmetric, because we can have two families of the same quality. However, we would like to require that this relation be final in the sense that it should define a unique best family \( \mathcal{F}_{\text{opt}} \), for which \( \forall \mathcal{G} \, (\mathcal{F}_{\text{opt}} \succeq \mathcal{G}) \). Indeed:

- if none of the families is the best,
- then this criterion is of no use.

So, there should be at least one optimal family. Similarly:

- if several different families are equally best,
- then we can use this ambiguity to optimize something else.

For example:

- if we have two families with the same approximating quality,
- then we can choose the one which is easier to compute.

As a result, the original criterion was not final: we obtain a new criterion: \( \mathcal{F} \succeq_{\text{new}} \mathcal{G} \), if:

- either \( \mathcal{F} \) gives a better approximation,
- or \( \mathcal{F} \sim_{\text{old}} \mathcal{G} \) and \( \mathcal{F} \) is easier to compute.

For this new optimality criterion, the class of optimal families is narrower.

We can repeat this procedure until we obtain a final criterion for which there is only one optimal family.

**Scale invariance.** The numerical value of each quantity \( x_i \) depends on the choice of the measuring unit. If instead of the original measuring unit, we choose a new one which is \( \lambda_i \) times smaller, then all numerical values are multiplied by \( \lambda_i \): \( x_i \rightarrow x'_i = \lambda_i \cdot x_i \). For example:

- if to measure height, we use centimeters instead of meters,
- then all numerical values of height are multiplied by 100: e.g., 2 m becomes 200 cm.

It is reasonable to require that the relative quality of two families should not change if we simply apply such re-scaling to one of the variables \( x_i \).

Thus, we arrive at the following definition.

**Definition 1.** Let \( n > 0 \) and \( N > 0 \) be integers.

- By a \( N \)-dimensional family, we mean a family \( \mathcal{F} \) of all functions of the type

\[ \{C_1 \cdot F_1(x_1, \ldots, x_n) + \ldots + C_N \cdot F_N(x_1, \ldots, x_n)\}, \]

where \( F_1, \ldots, F_N \) are given analytical functions, and \( C_1, \ldots, C_N \) are arbitrary (real) constants.
• By an optimality criterion, we mean a transitive reflexive relation \( \succeq \) on the set of all \( N \)-dimensional families.

• We say that a criterion is final if there exists one and only one optimal family, i.e., family \( \mathcal{F}_{\text{opt}} \) for which \( \forall \mathcal{G} (\mathcal{F}_{\text{opt}} \succeq \mathcal{G}) \).

• For every transformation \( T = \lambda \cdot x \ (\lambda > 0) \), and for every \( i \), we define
  \[
  (T_i(F))(x_1, \ldots, x_n) \overset{\text{def}}{=} F(x_1, \ldots, x_{i-1}, T(x_i), x_{i+1}, \ldots, x_n),
  \]
  and \( T_i(F) \overset{\text{def}}{=} \{ T_i(F) \mid F \in \mathcal{F} \} \).

• We say that a criterion \( \succeq \) is scale-invariant if for every two families \( \mathcal{F} \) and \( \mathcal{G} \), for every \( i \), and for every linear function \( T(x) = \lambda \cdot x \), \( \mathcal{F} \succeq \mathcal{G} \) implies \( T_i(\mathcal{F}) \succeq T_i(\mathcal{G}) \).

**Proposition 1.** Let \( \succeq \) be a final scale-invariant optimality criterion on the set of all families. Then, every function \( F \) from the optimal family \( \mathcal{F}_{\text{opt}} \) is a polynomial.

**Comments.**

• This result justifies the Taylor models.

• This result is stronger than the result from [9], since there, we also required that the optimality criterion be invariant if we change the starting point for measuring \( x_i \).

**Proof.**

1°. Let us first prove that the optimal family \( \mathcal{F}_{\text{opt}} \) is itself scale-invariant, i.e., that for every rescaling \( T \) and for every \( i \), we have \( T_i(\mathcal{F}_{\text{opt}}) = \mathcal{F}_{\text{opt}} \).

Indeed, let \( T \) and \( i \) be given. Since \( \mathcal{F}_{\text{opt}} \) is optimal, for every other family \( \mathcal{G} \), we have \( \mathcal{F}_{\text{opt}} \succeq T_i^{-1}(\mathcal{G}) \) (where \( T_i^{-1} \) means the inverse transformation). Since the optimality criterion \( \succeq \) is invariant, we conclude that \( T_i(\mathcal{F}_{\text{opt}}) \succeq T_i(T_i^{-1}(\mathcal{G})) = \mathcal{G} \). Since this is true for every family \( \mathcal{G} \), the family \( T_i(\mathcal{F}_{\text{opt}}) \) is also optimal. But since our criterion is final, there is only one optimal family and therefore, \( T_i(\mathcal{F}_{\text{opt}}) = \mathcal{F}_{\text{opt}} \).

2°. Since the family \( \mathcal{F}_{\text{opt}} \) is scale-invariant, in particular, it means that for every function \( F(x_1, \ldots, x_n) \) from this family, and for every \( \lambda > 0 \), the function

\[
F_{\lambda}(x_1, \ldots, x_n) \overset{\text{def}}{=} F(\lambda \cdot x_1, \ldots, \lambda \cdot x_n)
\]

also belongs to the optimal family.

3°. Let us now take any function \( F(x_1, \ldots, x_n) \) from the optimal family \( \mathcal{F}_{\text{opt}} \) and prove that this function is a polynomial.

The family consists of analytical functions, thus the selected function \( F(x_1, \ldots, x_n) \) is also analytical.

By definition, an analytical function \( f(x_1, \ldots, x_n) \) is an infinite series consisting of monomials \( m(x_1, \ldots, x_n) \) of the type

\[
a_{i_1 \ldots i_n} \cdot x_1^{i_1} \cdot \ldots \cdot x_n^{i_n}.
\]

For each such term, by its total order, we will understand the sum \( i_1 + \ldots + i_n \). The meaning of this total order is simple: if we multiply each input of this monomial by \( \lambda \), then the value of the monomial is multiplied by \( \lambda^k \):

\[
m(\lambda \cdot x_1, \ldots, \lambda \cdot x_n) = a_{i_1 \ldots i_n} \cdot (\lambda \cdot x_1)^{i_1} \cdot \ldots \cdot (\lambda \cdot x_n)^{i_n} =
\lambda^{i_1 + \ldots + i_n} \cdot a_{i_1 \ldots i_n} \cdot x_1^{i_1} \cdot \ldots \cdot x_n^{i_n} = \lambda^k \cdot m(x_1, \ldots, x_n).
\]

For each order \( k \), there are finitely many possible combinations of integers \( i_1, \ldots, i_n \) for which \( i_1 + \ldots + i_n = k \), so there are finitely many possible monomials of this order. Let \( P_k(x_1, \ldots, x_n) \) denote the sum of all the monomials of order \( k \) from the series describing the function \( F(x_1, \ldots, x_n) \). Then, we have

\[
F(x_1, \ldots, x_n) = P_0 + P_1(x_1, \ldots, x_n) + P_2(x_1, x_2, \ldots, x_n) + \ldots
\]

Some of these terms may be zeros – if the original expansion has no monomials of the corresponding order. Let \( k_0 \) be the first index for which the term \( P_{k_0}(x_1, \ldots, x_n) \) is not identically 0. Then,

\[
F(x_1, \ldots, x_n) = P_{k_0}(x_1, \ldots, x_n) + P_{k_0+1}(x_1, x_2, \ldots, x_n) + \ldots
\]
Since the family $\mathcal{F}_{\text{opt}}$ is scale-invariant, it also contains the function

$$F_\lambda(x_1, \ldots, x_n) = F(\lambda \cdot x_1, \ldots, \lambda \cdot x_n).$$

At this re-scaling, each term $P_k$ is multiplied by $\lambda^k$; thus, we get

$$F_\lambda(x_1, \ldots, x_n) = \lambda^{k_0} \cdot P_{k_0}(x_1, \ldots, x_n) + \lambda^{k_0+1} \cdot P_{k_0+1}(x_1, x_2, \ldots, x_n) + \ldots$$

Since $\mathcal{F}_{\text{opt}}$ is a linear space, it also contains a function

$$\lambda^{-k_0} \cdot F_\lambda(x_1, \ldots, x_n) = P_{k_0}(x_1, \ldots, x_n) + \lambda \cdot P_{k_0+1}(x_1, x_2, \ldots, x_n) + \ldots$$

Since $\mathcal{F}_{\text{opt}}$ is finite-dimensional, it is closed under turning to a limit. In the limit $\lambda \to 0$, we conclude that the term $P_{k_0}(x_1, \ldots, x_n)$ also belongs to the family $\mathcal{F}_{\text{opt}}$.

Since $\mathcal{F}_{\text{opt}}$ is a linear space, this means that the difference

$$F(x_1, \ldots, x_n) - P_{k_0}(x_1, \ldots, x_n) = P_{k_0+1}(x_1, x_2, \ldots, x_n) + P_{k_0+2}(x_1, x_2, \ldots, x_n) + \ldots$$

also belongs to $\mathcal{F}_{\text{opt}}$. If we denote, by $k_1$, the first index $k_1 > k_0$ for which the term $P_{k_1}(x_1, \ldots, x_n)$ is not identically 0, then we can similarly conclude that this term $P_{k_1}(x_1, \ldots, x_n)$ also belongs to the family $\mathcal{F}_{\text{opt}}$, etc.

We can therefore conclude that for every index $k$ for which term $P_k(x_1, \ldots, x_n)$ is not identically 0, this term $P_k(x_1, \ldots, x_n)$ also belongs to the family $\mathcal{F}_{\text{opt}}$.

Monomials of different total order are linearly independent. Thus, if there were infinitely many non-zero terms $P_k$ in the expansion of the function $F(x_1, \ldots, x_n)$, we would have infinitely many linearly independent functions in the family $\mathcal{F}_{\text{opt}}$—which contradicts to our assumption that the family $\mathcal{F}_{\text{opt}}$ is a finite-dimensional linear space.

So, in the expansion of the function $F(x_1, \ldots, x_n)$, there are only finitely many non-zero terms. Hence, the function $F(x_1, \ldots, x_n)$ is a sum of finitely many monomials—i.e., a polynomial.

The proposition is proven.

Comment. As we can see from the proof, to show that every function from the optimal family is a polynomial, we do not even need to use scale-invariance with respect to each of the variables: it is sufficient to require that the optimality criterion is invariant with respect to a simultaneous re-scaling of all the variables:

$$x_1 \to \lambda \cdot x_1, \ldots, x_n \to \lambda \cdot x_n.$$

### 3 Why Modified Taylor Models: A Justification

**Discussion.** In the original Taylor model, coefficients at the unknown functions were real numbers. The main idea behind modified Taylor models is that we can have interval-valued coefficients. Thus, we arrive at the following definition.

**Definition 2.** Let $n > 0$, $N > 0$, and $M > 0$ be integers.

- **By a $(N,M)$-family,** we mean a family $\mathcal{F}$ of all interval-valued functions of the type

  $$\{C_1 \cdot F_1(x_1, \ldots, x_n) + \ldots + C_N \cdot F_N(x_1, \ldots, x_n) + C_1 \cdot G_1(x_1, \ldots, x_n) + \ldots + C_M \cdot G_M(x_1, \ldots, x_n)\},$$

  where $F_1, \ldots, F_N$ and $G_1, \ldots, G_M$ are given analytical functions, $C_1, \ldots, C_N$ are arbitrary (real) constants, and $C_k = [\overline{C_k}, \underline{C_k}]$ are arbitrary intervals.

- **By an optimality criterion,** we mean a transitive reflexive relation $\succeq$ on the set of all $(N,M)$-families.

- **We say that a criterion is final if there exists one and only one optimal family,** i.e., family $\mathcal{F}_{\text{opt}}$ for which $\forall G (\mathcal{F}_{\text{opt}} \succeq G)$.

- **For every transformation** $T = \lambda \cdot x (\lambda > 0)$, and for every $i$, we define

  $$(T_i(F))(x_1, \ldots, x_n) \overset{\text{def}}{=} F(x_1, \ldots, x_{i-1}, T(x_i), x_{i+1}, \ldots, x_n),$$

  and $T_i(\mathcal{F}) \overset{\text{def}}{=} \{T_i(F) \mid F \in \mathcal{F}\}$. 

---

REC 2018 – M. Joldeş, C. Lauter, M. Ceberio, O. Kosheleva, V.Kreinovich
• We say that a criterion \( \succ \) is scale-invariant if for every two families \( \mathcal{F} \) and \( \mathcal{G} \), for every \( i \), and for every linear function \( T(x) = \lambda \cdot x \), \( \mathcal{F} \succ \mathcal{G} \) implies \( T_i(\mathcal{F}) \succ T_i(\mathcal{G}) \).

**Proposition 2.** Let \( \succ \) be a final scale-invariant optimality criterion on the set of all families. Then, every function \( F \) from the optimal family \( \mathcal{F}_{opt} \) is a sum of finitely many monomials with interval coefficients.

**Comment.** This result justifies the modified Taylor models.

**Proof.**

1°. First let us notice that each interval \( [\underline{C}_k, \overline{C}_k] \) can be represented as \( \underline{C}_k + [0, W_k] \), where \( W_k \defeq \overline{C}_k - \underline{C}_k \) is the width of the \( k \)-th interval. Substituting this expression into the general formula for a family, we conclude that each interval-valued function from the family is a linear combination of:

- real-valued functions \( F_j \) and \( G_k \) and
- functions \( G_k \) multiplied by an interval \([0, W_k]\).

Thus, without losing generality, we can safely assume that in each interval \( C_k \), the lower endpoint is 0, i.e., that each such interval has the form \([0, W_k]\) for some \( W_k \geq 0 \).

2°. Similarly to the proof of Proposition 1, we can prove that the optimal family is scale-invariant, i.e., remains unchanged if we re-scale each variable \( x_i \to \lambda_i \cdot x_i \).

In other words, for each interval-valued function from the optimal family,

- if we re-scale all the variables,
- we get an interval-valued function from the same family – but probably corresponding to different coefficients \( C_k' \) and \( W_k' \).

So, the re-scaled interval-valued function

\[
F'(x_1, \ldots, x_n) = [0, W_1] \cdot G_1(\lambda_1 \cdot x_1, \ldots, \lambda_n \cdot x_n) + \ldots + [0, W_M] \cdot G_M(\lambda_1 \cdot x_1, \ldots, \lambda_n \cdot x_n),
\]

where

\[
F'(x_1, \ldots, x_n) \defeq C_1', F_1(\lambda_1 \cdot x_1, \ldots, \lambda_n \cdot x_n) + \ldots + C_N', F_N(\lambda_1 \cdot x_1, \ldots, \lambda_n \cdot x_n),
\]

coincides with the interval-valued function

\[
F'(x_1, \ldots, x_n) + [0, W_1'] \cdot G_1(x_1, \ldots, x_n) + \ldots + [0, W_M'] \cdot G_M(x_1, \ldots, x_n),
\]

where

\[
F'(x_1, \ldots, x_n) \defeq C_1', F_1(x_1, \ldots, x_n) + \ldots + C_N', F_N(x_1, \ldots, x_n).
\]

Each interval-valued function is a convex set in the class of all functions, obtained by taking all possible values \( w_k \in [0, W_k] \).

Since the convex sets coincide, this means that their sets of extreme points should also coincide. These extreme points correspond to extreme values \( 0 \) and \( W_k \) of the parameters \( w_k \in [0, W_k] \). Thus, for the re-scaled family, they are:

\[
F'(x_1, \ldots, x_n), \quad F'(x_1, \ldots, x_n) + W_1 \cdot G_1(\lambda_1 \cdot x_1, \ldots, \lambda_n \cdot x_n), \ldots,
\]

\[
F'(x_1, \ldots, x_n), \quad F'(x_1, \ldots, x_n) + W_M \cdot G_M(\lambda_1 \cdot x_1, \ldots, \lambda_n \cdot x_n),
\]

\[
F'(x_1, \ldots, x_n) + W_1 \cdot G_1(\lambda_1 \cdot x_1, \ldots, \lambda_n \cdot x_n) + W_2 \cdot G_2(\lambda_1 \cdot x_1, \ldots, \lambda_n \cdot x_n), \ldots
\]

For the new family, the extreme points are:

\[
F'(x_1, \ldots, x_n), \quad F'(x_1, \ldots, x_n) + W_1' \cdot G_1(x_1, \ldots, x_n), \ldots,
\]

\[
F'(x_1, \ldots, x_n) + W_M' \cdot G_M(x_1, \ldots, x_n),
\]

\[
F'(x_1, \ldots, x_n) + W_1' \cdot G_1(x_1, \ldots, x_n) + W_2' \cdot G_2(x_1, \ldots, x_n), \ldots
\]

For \( \lambda_1 = \ldots = \lambda_n = 1 \), the first function in the first list coincides with the first function in the second list, etc. Since the dependence on \( \lambda_i \) is continuous, we cannot switch to different equalities, so always:
• the first extreme function from the first list must coincide with the first extreme function from the second list,
• the second extreme function from the first list must coincide with the second extreme function from the second list, etc.

Equality of the first terms means that, for every tuple \( C_1, \ldots, C_N \) and for every tuple \( \lambda_1, \ldots, \lambda_n \), there exist values \( C'_1, \ldots, C'_N \) for which

\[
F_\lambda(x_1, \ldots, x_n) = C_1 \cdot F_1(\lambda_1 \cdot x_1, \ldots, \lambda_n \cdot x_n) + \ldots + C_N \cdot F_N(\lambda_1 \cdot x_1, \ldots, \lambda_n \cdot x_n) =
F'(x_1, \ldots, x_n) = C'_1 \cdot F_1(x_1, \ldots, x_n) + \ldots + C'_N \cdot F_N(x_1, \ldots, x_n).
\]

This means that the class of all functions

\[
C_1 \cdot F_1(x_1, \ldots, x_n) + \ldots + C_N \cdot F_N(x_1, \ldots, x_n)
\]

corresponding to different values \( C_j \) is scale-invariant. Thus, based on Proposition 1, all the functions

\[
F_j(x_1, \ldots, x_n)
\]

are polynomials – i.e., a sum of finitely many monomials.

For each \( k \) from 1 to \( M \), since both the first and the \( (k+1) \)-st terms in the two lists are equal to each other, we conclude the differences between these term should also be equal. Thus, we conclude that

\[
W_k \cdot G_k(\lambda_1 \cdot x_1, \ldots, \lambda_n \cdot x_k) = W'_k \cdot G_k(x_1, \ldots, x_n).
\]

For \( W_1 = 1 \), this means that for every tuple \( \lambda_1, \ldots, \lambda_n \), there exists a value \( W'_k \) for which

\[
G_k(\lambda_1 \cdot x_1, \ldots, \lambda_n \cdot x_k) = W'_k \cdot G_k(x_1, \ldots, x_n).
\]

The function \( G_k(x_1, \ldots, x_n) \) is an analytical function and is, thus, the sum of monomials \( c_{i_1 \ldots i_n} x_1^{i_1} \cdots x_n^{i_n} \). Under re-scaling \( x_i \rightarrow \lambda_i \cdot x_i \), each monomial is multiplied by different coefficients \( \lambda_1^{i_1} \cdots \lambda_n^{i_n} \). So, the only case when the whole sum of monomials is multiplied by the same number \( W'_k \) is when the function \( G_k(x_1, \ldots, x_n) \) consists of a single monomial.

Thus, each interval-valued function from the optimal family is indeed a sum of finitely many monomials \( G_k(x_1, \ldots, x_n) \) with interval coefficients. The proposition is proven.

Acknowledgments
This work was supported in part by the US National Science Foundation grant HRD-1242122 (Cyber-Share Center of Excellence).

The authors are thankful to anonymous referees for valuable suggestions.

References
What is the economically optimal way to guarantee interval bounds on control?

Alfredo Vaccaro,1 Martine Ceberio,2 and Vladik Kreinovich2*

1 Department of Engineering, University of Sannio, Italy
2 Department of Computer Science, University of Texas at El Paso, USA
*Corresponding author: vladik[at]utep.edu

Abstract

For control under uncertainty, interval methods enable us to find a box $B = [u_1, \pi_1] \times \ldots \times [u_n, \pi_n]$ for which any control $u \in B$ has the desired properties – such as stability. Thus, in real-life control, we need to make sure that $u_i \in [u_i, \pi_i]$ for all parameters $u_i$ describing control. In this paper, we describe the economically optimal way of guaranteeing these bounds.

Keywords: Interval uncertainty, Actuators, Control, Economically optimal solution

1 Formulation of the Problem

1.1 Control problems: a very brief reminder

In control problems, we need to find the values of the control $u = (u_1, \ldots, u_n)$.

Usually, there are some requirements on the control: e.g., that under this control, the system should be stable, etc. These conditions are usually described by inequalities; see, e.g., [1].

1.2 From optimal control to constraint satisfaction

In general, there are many different controls that satisfy all the desired constraints.

In the ideal case, when:

- we know the exact initial state of the system and
- we know the equations that describe the system’s dynamics under different controls,

we can compute the exact consequences of each control. Thus, depending on what is our objective, we can select an appropriate objective functions and looks for the control that optimized this objective function.

The objective function depends on the task. For example, for selecting a plane trajectory, we can have different objective functions:

- In the situation of medical emergency, we need to find the trajectory of the plane that brings the medical team to the remote patient as soon as possible.
- For a regular passenger communications, we need to minimize expenses – and hence, instead of flying at the largest possible speed, we should fly at the speed that saves as much fuel as possible.
- For a private jet, a reasonable objective function is the ride’s smoothness.

In practice, we rarely know the exact initial state and the exact system’s dynamics. Often, for each of the corresponding parameters, we only know the lower and upper bounds on possible values, i.e., in other words, we only know the interval that contains all possible values of the corresponding parameter; see, e.g., [1–4]. In such cases, for each control, instead of the exact value of the objective function, we get the range of possible values $[v, \pi]$. Computing this range under interval uncertainty is a particular case of the main problem of interval computations [1–3].

In such situations, it make sense, e.g., to describe all the control $u$ which are possibly optimal, i.e., for which $\pi(u) \geq \max_{u'} v(u')$.

1.3 What we get from interval computations

In situations of interval uncertainty, interval methods enable us to find a box $B = [u_1, \pi_1] \times \ldots \times [u_n, \pi_n]$ for which any control $u \in B$ has the desired properties – such as stability or possible optimality.
1.4 Resulting problem
Thus, in real-life control, we need to make sure that \( u_i \in [\underline{u}_i, \overline{u}_i] \) for all parameters \( u_i \) describing control.

What is the most economical way to guarantee these bounds?

2 Towards Formulating the Problem in Precise Terms

2.1 Analysis of the problem

Actuators are never precise, so we can only set up the control value \( u_i \) with some accuracy \( a_i \). Thus, if we aim for the midpoint \( \overline{um}_i \equiv \frac{\underline{u}_i + \overline{u}_i}{2} \), we will get the actual value \( u_i \) within the interval \([\overline{um}_i - a_i, \overline{um}_i + a_i]\).

The only way to guarantee that the control value is indeed within the desired interval is to measure it. Measurement are also never absolutely precise. Let us assume that we use a measuring instrument with accuracy \( \varepsilon_i \). This means that for each actual value \( u_i \) of the corresponding parameter, the measured value \( \tilde{u}_i \) is somewhere within the interval \([u_i - \varepsilon_i, u_i + \varepsilon_i]\). Based on the measurements, the only thing we can conclude about the actual (unknown) value \( u_i \) is that it belongs to the interval \([\tilde{u}_i - \varepsilon_i, \tilde{u}_i + \varepsilon_i]\). We want to make sure that all the values from this interval are within the desired interval \([\underline{u}_i, \overline{u}_i]\), i.e., that

\[
\underline{u}_i \leq \tilde{u}_i - \varepsilon_i \tag{1}
\]

and

\[
\tilde{u}_i + \varepsilon_i \leq \overline{u}_i. \tag{2}
\]

These inequalities must hold for all possible values \( \tilde{u}_i \in [u_i - \varepsilon_i, u_i + \varepsilon_i] \). For the inequality (1) to hold for all these values, it is sufficient to require that this inequality holds for the smallest possible value \( \tilde{u}_i = u_i - \varepsilon_i \), i.e., that we have

\[
\underline{u}_i \leq (u_i - \varepsilon_i) - \varepsilon_i = u_i - 2\varepsilon_i. \tag{3}
\]

Similarly, for the inequality (2) to hold for all the values \( \tilde{u}_i \in [u_i - \varepsilon_i, u_i + \varepsilon_i] \), it is sufficient to require that this inequality holds for the largest possible value \( \tilde{u}_i = u_i + \varepsilon_i \), i.e., that we have

\[
(u_i + \varepsilon_i) + \varepsilon_i = u_i + 2\varepsilon_i \leq \overline{u}_i. \tag{4}
\]

The inequalities (3) and (4) must hold for all possible values \( u_i \in [\overline{um}_i - a_i, \overline{um}_i + a_i] \). For the inequality (3) to hold for all these values, it is sufficient to require that this inequality holds for the smallest possible value \( u_i = \overline{um}_i - a_i \), i.e., that we have

\[
\underline{u}_i \leq \overline{um}_i - 2\varepsilon_i - a_i. \tag{5}
\]

Similarly, for the inequality (4) to hold for all the values \( u_i \in [\overline{um}_i - a_i, \overline{um}_i + a_i] \), it is sufficient to require that this inequality holds for the largest possible value \( u_i = \overline{um}_i + a_i \), i.e., that we have

\[
\overline{um}_i + 2\varepsilon_i + a_i \leq \overline{u}_i. \tag{6}
\]

Let us denote the half-width of the interval \([\underline{u}_i, \overline{u}_i]\) by \( \Delta_i \equiv \frac{\overline{u}_i - \underline{u}_i}{2} \). In terms of the half-width, \( \overline{u}_i - \underline{u}_i = \overline{u}_i - \underline{u}_i = \Delta_i \). Thus, the inequalities (5) and (6) are equivalent to the inequality

\[
2\varepsilon_i + a_i \leq \Delta_i. \tag{7}
\]

2.2 In the optimal solution, we have equality

In general, the more accuracy we want, the more expensive will be the corresponding measurements and actuators. From this viewpoint, if \( 2\varepsilon_i + a_i < \Delta_i \), then we can use slightly less accurate actuators and/or measuring instruments and still guarantee the desired inequality (7). Thus, in the most economical solution, in the formula (7), we should have the exact equality:

\[
2\varepsilon_i + a_i = \Delta_i, \tag{8}
\]

i.e., equivalently,

\[
a_i = \Delta_i - 2\varepsilon_i. \tag{9}
\]

2.3 Towards resulting formulation of the problem

So, the problem is to find, among all the values \( a_i \) and \( \varepsilon_i \) that satisfy the equality (9), the values for which the overall expenses are the smallest possible.

To solve this problem, we need to know how the cost of actuators and measurements depends on accuracy. To analyze this dependence, we start with a 1-D case, when we only have a single control parameter \( u_1 \).
3 1-D Case, When We Have a Single Control Parameter $u_1$

3.1 Cost of actuators: main idea

Actuators – such as robotic arms – are usually rather crude, so we may not be able to properly orient the robot after the first attempt. A natural way to provide a better accuracy is to repeat the attempts until we get the desired location (or, in general, the desired value of the parameter $u_i$).

3.2 Cost of actuators: from idea to a formula

Let us assume that the given actuator can provide the value $u_i$ with some accuracy $A_i$. This means that if we aim for the midpoint $u_{mi}$, we will get values from the interval

$$[u_{mi} - A_i, u_{mi} + A_i].$$

(10)

We do not know the relative frequency of different values within this interval. Since we have no reasons to assume that some of these values are more probable and some are less probable, it is therefore reasonable to assume that all the values from the interval (10) are equally probable, i.e., that we have a uniform probability distribution on this interval.

By repeatedly trying, we want to get the value $u_i$ within the interval

$$[u_{mi} - a_i, u_{mi} + a_i]$$

for some $a_i < A_i$. For the uniform distribution, the probability to be within a subinterval is proportional to the width of this subinterval – namely, it is equal to the ratio between the width of the subinterval and the width of the entire interval. In particular, the probability that at each try, the value $u_i$ is within the interval is equal to the ratio $p = \frac{2a_i}{2A_i} = \frac{a_i}{A_i}$. So, on average, we need $\frac{1}{p} = \frac{A_i}{a_i}$ iterations to get into the desired interval (11).

3.3 Cost of measurements

How to gauge the cost of accurate measurements? It is known that if we start with a measuring instrument with 0 mean and standard deviation $\sigma_i$, then, by performing $M$ independent measurements and averaging the results, we get a $\sqrt{M}$ times smaller standard deviation.

Indeed, the variance of the sum of independent random variables is equal to the sum of the variances, so for the sum of $M$ measurement errors, the variance is $m \cdot \sigma_i^2$, and thus, the standard deviation is $\sqrt{m} \cdot \sigma_i$. The arithmetic average is obtained by dividing the sum by $M$, so its standard deviation is

$$\frac{\sqrt{M} \cdot \sigma_i}{M} = \frac{\sigma_i}{\sqrt{M}}$$

So, if we start with a measuring instrument with accuracy $\sigma_i$, and we want accuracy $\varepsilon_i$, we need to repeat each measurement $M_i$ times, where $\frac{\sigma_i}{\sqrt{M_i}} = \varepsilon_i$, i.e., $M_i = \frac{\sigma_i^2}{\varepsilon_i^2}$. The cost of measurement is proportional to the number $M_i$ of such measurements, so it is equal to

$$m_i \cdot M_i = m_i \cdot \frac{\sigma_i^2}{\varepsilon_i^2},$$

where by $m_i$, we denoted the cost of a single measurement.

3.4 Overall cost

Let $T_i$ denote the cost of a single actuator try.

The overall cost of one try of an actuator is equal to the actuator trying cost $T_i$ plus the measurement cost $m_i \cdot \frac{\sigma_i^2}{\varepsilon_i^2}$, i.e., it is equal to $T_i + m_i \cdot \frac{\sigma_i^2}{\varepsilon_i^2}$. To achieve the actuator accuracy $a_i = \Delta_i - 2\varepsilon_i$, we need to perform $\frac{A_i}{a_i} = \frac{A_i}{\Delta_i - 2\varepsilon_i}$ tries. Thus, the overall cost $C$ of all the tries is equal to

$$C = \frac{A_i}{\Delta_i - 2\varepsilon_i} \cdot \left(T_i + m_i \cdot \frac{\sigma_i^2}{\varepsilon_i^2}\right).$$

(12)

3.5 Resulting optimization problem

In the 1-D case, we need to find the value $\varepsilon_i$ for which the cost (12) is the smallest possible.
3.6 Discussion

• When $\varepsilon_i$ is close to 0, the cost of measurement tends to infinity, so we have a very large overall cost.

• Similarly, when $a_i$ is close to 0, i.e., when $\varepsilon_i \approx \Delta_i/2$, the actuator cost becomes very large, so the overall cost is also very large.

Thus, there should be values $\varepsilon_i$ between 0 and $\Delta_i/2$ for which the cost of maintaining $u_i$ within the desired interval $[u_i, u_i]$ is the smallest possible.

3.7 How to solve this optimization problem

In the 1-D case, where we have a single unknown $\varepsilon_i$, to find the optimal value of this unknown, we can simply differentiate the objective function (1) with respect to $\varepsilon_i$ and equate the derivative to 0.

Minimizing the expression (12) is equivalent to minimizing its logarithm

$$
\ln(A_i) - \ln(\Delta_i - 2\varepsilon_i) + \ln \left( T_i + m_i \cdot \sigma_i^2 \cdot \varepsilon_i \right).
$$

Differentiating this expression with respect to $\varepsilon_i$ and equating the derivative to 0, we get

$$
\frac{2}{\Delta_i - 2\varepsilon_i} - 2 \cdot m_i \cdot \sigma_i^2 \cdot \varepsilon_i^{-2} \cdot \frac{1}{T_i + m_i \cdot \sigma_i^2 \cdot \varepsilon_i} = 0.
$$

Dividing both sides by 2, moving the negative term to the right-hand side, and explicitly multiplying the expressions in the right-hand side, we get

$$
\frac{1}{\Delta_i - 2\varepsilon_i} = \frac{m_i \cdot \sigma_i^2}{\varepsilon_i^3 \cdot T_i + m_i \cdot \sigma_i^2 \cdot \varepsilon_i}.
$$

Bringing both fractions to the common denominator, we get a cubic equation

$$
\varepsilon_i^3 \cdot T_i + m_i \cdot \sigma_i^2 \cdot \varepsilon_i = m_i \cdot \sigma_i^2 \cdot (\Delta_i - 2\varepsilon_i),
$$

i.e., equivalently,

$$
T_i \cdot \varepsilon_i^3 + 3m_i \cdot \sigma_i^2 \cdot \varepsilon_i - m_i \cdot \sigma_i^2 \cdot \Delta_i = 0.
$$

3.8 Resulting algorithm

To find the optimal accuracy $\varepsilon_i$ of the measuring instrument, we can use one of the standard methods (e.g., Newton’s method) so solve the cubic equation (13).

Then, we can find the optimal value $a_i$ of the actuator accuracy as $a_i = \Delta_i - 2\varepsilon_i$.

4 General Case

4.1 Notations

In general, we may have several actuators. Let us denote the number of actuators by $A$. For each actuator $a = 1, \ldots, A$, let us denote:

- the cost of one try by $T_a$, and
- the number of the corresponding control parameters by $n_a$.

For each actuator $a$, and for each $i$ from 1 to $n_a$, let us denote:

- the $i$-th control parameter by $u_{ai}$;
- the bounds of the corresponding control parameter by $\underline{u}_{ai}$ and $\overline{u}_{ai}$;
- the midpoint of the resulting interval by $u_{mai} \overset{\text{def}}{=} \frac{\underline{u}_{ai} + \overline{u}_{ai}}{2}$;
- the half-width of the corresponding interval by $\Delta_{ai} \overset{\text{def}}{=} \frac{\overline{u}_{ai} - \underline{u}_{ai}}{2}$;
- the bounds achievable on one try by $A_{ai}$,
the desired actuator accuracy by $a_{ai}$,

- the desired measurement accuracy by $\varepsilon_{ai}$,

- the accuracy of the corresponding measuring instrument by $\sigma_{ai}$, and

- the cost of a single measurement with that accuracy by $m_{ai}$.

4.2 Relation between accuracies of actuator and measurement

Similarly to the 1-D case, we can conclude that in the general case, for each $a$ and $i$, we have $2\varepsilon_{ai} + a_{ai} = \Delta_{ai}$, i.e., $a_{ai} = \Delta_{ai} - 2\varepsilon_{ai}$.

4.3 Cost of actuators

We assume that the actuator $a$ can provide the value $u_{ai}$ with some accuracy $A_{ai}$. This means that if we aim for the midpoint $u_{mai}$, we will get values from the interval $[u_{mai} - A_{ai}, u_{mai} + A_{ai}]$. Similar to the 1-D case, it is reasonable to assume that all the combinations $(u_{a1}, \ldots, u_{an})$ from the corresponding box

$$[u_{ma1} - A_{a1}, u_{ma1} + A_{a1}] \times \cdots \times [u_{man} - A_{ana}, u_{man} + A_{ana}]$$

are equally probable, i.e., that we have a uniform probability distribution on this box.

By repeatedly trying, we want to get the value $(u_{a1}, \ldots, u_{an})$ within the smaller box

$$[u_{ma1} - a_{a1}, u_{ma1} + a_{a1}] \times \cdots \times [u_{man} - a_{ana}, u_{man} + a_{ana}].$$

For the uniform distribution, the probability to be within a sub-box is proportional to the volume of this sub-box – namely, it is equal to the ratio between the volume of the sub-box and the volume of the entire box. In particular, the probability that at each try, the values $a_{ai}$ are within the desired box is equal to the ratio

$$p = \frac{(2a_{a1}) \cdot \cdots \cdot (2a_{ana})}{(2A_{a1}) \cdot \cdots \cdot (2A_{ana})} = \frac{a_{a1} \cdot \cdots \cdot a_{ana}}{A_{a1} \cdot \cdots \cdot A_{ana}}.$$

So, on average, we need

$$\frac{1}{p} = \frac{A_{a1} \cdot \cdots \cdot A_{ana}}{a_{a1} \cdot \cdots \cdot a_{ana}} = \prod_{i=1}^{n_a} \frac{A_{ai}}{a_{ai}}$$

iterations to get into the desired box

$$[u_{ma1} - a_{a1}, u_{ma1} + a_{a1}] \times \cdots \times [u_{man} - a_{ana}, u_{man} + a_{ana}].$$

4.4 Overall cost for each actuator

Similarly to the 1-D case, for each $i$, the cost of measuring the value $u_{ai}$ with accuracy $\varepsilon_{ai}$ is equal to $m_{ai} \cdot \frac{\sigma^2_{ai}}{\varepsilon^2_{ai}}$. Thus, the overall cost of measuring all these values is equal to the sum

$$\sum_{i=1}^{n_a} m_{ai} \cdot \frac{\sigma^2_{ai}}{\varepsilon^2_{ai}}.$$

The overall cost of one try of an actuator is equal to the actuator trying cost $T_a$ plus the measurement cost:

$$T_a + \sum_{i=1}^{n_a} m_{ai} \cdot \frac{\sigma^2_{ai}}{\varepsilon^2_{ai}}.$$

To achieve the actuator accuracy $a_{ai} = \Delta_{ai} - 2\varepsilon_{ai}$, we need to perform

$$\prod_{i=1}^{n_a} \frac{A_{ai}}{a_{ai}} = \prod_{i=1}^{n_a} \frac{A_{ai}}{\Delta_{ai} - 2\varepsilon_{ai}}$$

tries. Thus, the overall cost $C_a$ of all the tries is equal to

$$C_a = \sum_{i=1}^{n_a} m_{ai} \cdot \frac{\sigma^2_{ai}}{\varepsilon^2_{ai}} \cdot \left(T_a + \sum_{i=1}^{n_a} m_{ai} \cdot \frac{\sigma^2_{ai}}{\varepsilon^2_{ai}} \right). \quad (14)$$
4.5 The overall cost of all the actuators
The overall cost $C$ of all the actuators can be obtained by adding up all the costs of all the actuators:

$$C = \sum_{a=1}^{A} C_a = \sum_{a=1}^{A} \left( \prod_{i=1}^{n_a} \frac{A_{ai}}{\Delta_{ai} - 2\varepsilon_{ai}} \cdot \left( T_a + \sum_{i=1}^{n_a} m_{ai} \cdot \frac{\sigma_{ai}^2}{\varepsilon_{ai}^2} \right) \right). \tag{15}$$

4.6 Resulting optimization problem
In the general case, we need to find the values $\varepsilon_{ai}$ for which the cost (15) is the smallest possible.

4.7 Towards solving the optimization problem
First, one can notice that each cost $C_a$ depends only on the parameters corresponding to this actuator. Thus, to optimize the overall cost $C$, it is sufficient to optimize the cost $C_a$ for each actuator $a$.

For each $a$, minimizing $C_a$ is equivalent to minimizing its logarithm

$$\sum_{i=1}^{n_a} \ln(A_{ai}) - \sum_{i=1}^{n_a} \ln(\Delta_{ai} - 2\varepsilon_{ai}) + \ln \left( T_a + \sum_{i=1}^{n_a} m_{ai} \cdot \frac{\sigma_{ai}^2}{\varepsilon_{ai}^2} \right).$$

Differentiating this expression with respect to $\varepsilon_{ai}$ and equating the derivative to 0, we get

$$\frac{2}{\Delta_{ai} - 2\varepsilon_{ai}} - 2 \cdot m_{ai} \cdot \sigma_{ai}^2 \cdot \frac{1}{\varepsilon_{ai}^3} \cdot \frac{1}{m_{a}} = 0, \tag{16}$$

where we denoted

$$m_{a} \overset{\text{def}}{=} T_a + \sum_{j=1}^{n_a} m_{aj} \cdot \frac{\sigma_{aj}^2}{\varepsilon_{aj}^2}. \tag{17}$$

Dividing both sides of the formula (16) by 2, moving the negative term to the right-hand side, and explicitly multiplying the expressions in the right-hand side, we get

$$\frac{1}{\Delta_{ai} - 2\varepsilon_{ai}} = m_{ai} \cdot \frac{\sigma_{ai}^2}{\varepsilon_{ai}^3} \cdot m_{a}. \tag{18}$$

Bringing both fractions to the common denominator, we get a cubic equation

$$\varepsilon_{ai}^3 \cdot m_{a} = m_{ai} \cdot \sigma_{ai}^2 \cdot (\Delta_{ai} - 2\varepsilon_{ai}),$$

i.e., equivalently,

$$m_{a} \cdot \varepsilon_{ai}^3 + 2m_{ai} \cdot \sigma_{ai}^2 \cdot \varepsilon_{ai} - m_{ai} \cdot \sigma_{ai}^2 \cdot \Delta_{ai} = 0. \tag{18}$$

4.8 Resulting algorithm
For each actuator $a$, once we fixed the value $m_{a}$, we can find each value $\varepsilon_{ai}$ ($i = 1, \ldots, n_a$) by solving the cubic equation (18).

We can then check whether our guess was correct by checking whether the formula (17) is satisfied for the resulting values $\varepsilon_{ai}$. By using bisection, we can find the value $m_{a}$ for which the equality (17) is satisfied.

Then, we can find the optimal value $a_{ai}$ of the actuator accuracy as $a_{ai} = \Delta_{ai} - 2\varepsilon_{ai}$.

Acknowledgments
This work was supported in part by the US National Science Foundation grant HRD-1242122 (Cyber-ShARE Center of Excellence). This work was partly performed during Alfredo Vaccaro’s visit to El Paso.

We are greatly thankful to the anonymous referees for valuable suggestions.

References

Worst-case solution spaces for systems design under uncertainties

Marco Daub,1* and Fabian Duddeck1,2

1Technische Universität München, Germany, 2Queen Mary University of London, UK
*Corresponding author: marco.daub[at]tum.de

Abstract
A general goal in systems design is to find a design that guarantees an optimal performance of the system. In practice, this cannot be reached as uncertainties occur which are responsible for a crucial reduction of the performance. Therefore, uncertainties must be taken into account in order to reach optimal or nearly optimal performance. In robust design optimization this is often addressed by considering uncertainties in uncontrollable parameters and controllable design variables that serve both as an input for the system performance. However, a mathematical quantification of the uncertainties is required which is in general unavailable at an early design phase of development due to a lack-of-knowledge situation. In this work, these uncertainties are regarded by assigning intervals to the uncontrollable parameters and a maximal worst-case solution space, also in the form of intervals, to the design variables. This worst-case solution space contains only designs which satisfy a performance requirement for all possible uncontrollable parameters. Furthermore, the center of the solution space is the design which tolerates the largest uncertainties in its variables.

Keywords: Systems design, interval uncertainties, lack of knowledge, solution spaces, optimization

1 Introduction
The method of using maximal box-shaped solution spaces for systems design was introduced in [9]. It is a set-based design approach (see [8] for an introduction) which provides a set of designs that fulfill all crucial requirements on the systems responses. Thus, the decision for a particular design can be postponed. As permissible intervals for design variables are made available with this approach, the decision for a design can also be decoupled between the single design variables. Furthermore, variations in design variables can be handled with these solution spaces. Beneath the uncertainties in the design variables, also referred to as controllable variables, uncertainties in uncontrollable parameters are of special interest in systems design, [6]. Both serve as an input of a system model that is assumed to be certain in this work. In comparison to uncertainties concerning uncontrollable parameters, uncertainties in design variables often cannot be quantified mathematically. This is due to imprecise or incomplete knowledge which especially occurs during early phase of development. In order to tolerate maximal deviations in design variables, a robust design method was proposed in [4], and extended for example by [3] and [7]. Here, a design is sought that is the center of a maximal permissible neighborhood, which can be box-shaped, similar to the approach using solution spaces.

In this paper, worst-case solution spaces are introduced that also incorporate interval uncertainties in uncontrollable parameters. Moreover, the general idea of how solution spaces help to handle uncertainties in design variables is motivated and the connection to the robust design method mentioned above is illustrated. In chapter 2, basic definitions of systems design are given and occurring uncertainties are investigated. Chapter 3 discusses methods for systems design concerning uncertainties in design variables and leads to worst-case solution spaces. In chapter 4, standard and worst-case solution spaces are compared at a crash design problem.

2 Basics
2.1 Systems Design
A goal in complex systems or systems design is to find a design for which the system performs optimally. Here, a design can be expressed as a d-dimensional vector $\mathbf{x}$. The entries of $\mathbf{x}$ are the design variables $x_i$, $i = 1, \ldots, d$. Each $x_i$ is technically lower bounded by $x_{\text{ls},i}$ and upper bounded by $x_{\text{us},i}$. Thus, it holds componentwise $x_{\text{ls}} \leq \mathbf{x} \leq x_{\text{us}}$ and $\Omega_{\text{ds}} = [x_{\text{ls}}, x_{\text{us}}]$, a multi-dimensional interval, can be defined as the design space. It is remarked that the same notations are used for all vector inequalities and multi-dimensional intervals in the following. Because the design can be chosen within $\Omega_{\text{ds}}$ by a decision maker,
the design variables are also called controllable variables. Besides controllable variables, there are uncontrollable parameters, see [6]. Those are collected in a $q$-dimensional vector $p$ and their values can neither be influenced by the decision maker anymore nor be influenced in general. Typical uncontrollable parameters include a priori chosen design variables, and system-specific or environmental parameters. Together, the design variables and the uncontrollable parameters form the inputs of the system. The outputs of the system are the responses which depend on the inputs and can be collected in a $m$-dimensional vector $z$. A diagram for a system like this is visualized in Figure 1.

![Diagram of a system with inputs and outputs](image)

**Figure 1: Diagram of a system with inputs and outputs**

The mapping of the inputs to the outputs can be described by the system performance functions $f$, where

$$z = f(x, p)$$

holds. In order to reach an optimal performance, some of the system responses must be minimized or maximized. Given these goals, an optimal design can be sought by using standard optimization methods which can also be multi-objective. However, as uncertainties can occur during systems development, it is possible that the resulting design causes a bad system performance. Therefore, uncertainty consideration has to be incorporated in the decision process for selecting a design.

### 2.2 Uncertainties

There are various ways to classify uncertainties, [1]. One is to differentiate from a system point of view. Uncertainties can be found in design variables, in uncontrollable parameters, and in the system model including its responses. Uncertainties in the design variables mainly occur because they cannot be realized accurately. This is for example due to manufacturing tolerances, or design variables are outputs of lower level design variables indeed. The same holds for uncontrollable parameters if they are a priori chosen design variables. Furthermore, there are also variations which follow for example from changing environmental or operating conditions. Uncertainties in the system model result from the approximation of real physical objects in order to use models. This includes also measuring errors when building the model. Other classifications of uncertainties can be done by means of their mathematical quantification or by distinguishing between aleatoric or epistemic type.

In early phase of system design, there is often not much information available on the uncertainties. This is mainly due to imprecise or incomplete knowledge which corresponds to the epistemic type. Therefore, a proper mathematical quantification is not possible in most cases. However, if it can be ensured that the realized values are bounded, the uncertainties can be at least modeled as interval uncertainties. In the following, the focus is put on uncertainties in design variables and in uncontrollable parameters. Similar to [6], it is assumed that the bounds of the uncontrollable parameters are known, i.e.

$$p_l \in [p_l^l, p_l^u]$$

with fixed $p_l^l$ and $p_l^u$ for $l = 1, \ldots, q$. The bounds of the design variables are unknown. Though, for given nominal values $\bar{x}_i$, $i = 1, \ldots, d$ which correspond to the design values chosen by the designer, the real values are assumed to be symmetrically bounded around these nominal values, i.e.

$$x_i \in [\bar{x}_i - \delta_i, \bar{x}_i + \delta_i]$$
with $\delta_i \geq 0$, $\bar{x}_i - \delta_i \geq x_{i,d_{i}}^l$, and $\bar{x}_i + \delta_i \leq x_{i,d_{i}}^u$ for $i = 1, \ldots, d$. As stated above, the values of $\delta_i$, $i = 1, \ldots, d$ are typically unknown. However, they can be also specified as maximal tolerances for the nominal values.

3 Solution Spaces

3.1 Motivation

In this section, the use of solution spaces for systems design and for handling the prescribed uncertain situation is motivated. Therefore, uncertainties in $p$ are neglected at first. Furthermore, $\delta_i = \bar{\delta}$ for $i = 1, \ldots, d$ is assumed and written as $\delta = (\bar{\delta}, \ldots, \bar{\delta})$.

If only one performance function is given that shall be minimized, i.e. $m = 1$, the robust regularization $f_{\bar{\delta}}$ can be defined according to [5]. For given $\delta$, it is

$$f_{\bar{\delta}}(\bar{x}, p) = \sup \{ f(x, p) : x \in [\bar{x} - \delta, \bar{x} + \delta] \},$$

which is a function of $\bar{x}$. Thus, by incorporating all possible realizations of $x$, the worst case for the system performance is always regarded. Then, a minimization of the robust regularization yields a design $\bar{x}$ which can be considered as robust design. This is because the worst value of the system response for uncertain design variables is minimal.

In systems design it is not always necessary to minimize systems responses. Often, it is also sufficient if they do not exceed or fall below particular threshold values, [9]. If this is applicable, the problem can be tackled from the opposite side. As $\delta$ is typically unknown, it can also be maximized under given constraints. In the following, multiple performance functions are regarded again, i.e. $m \geq 1$. Furthermore, thresholds $f_{c}(p)$ are given which upper bound the systems responses and depend on uncontrollable parameters. Hence, the set of all permissible designs, i.e. the designs within the design space which fulfill the constraints on their responses, are

$$\Omega_c = \{ x \in \Omega_{ds} : f(x, p) \leq f_{c}(p) \}.$$  

The set $\Omega_c$ is also called complete solution space. It shall be remarked that lower-bound thresholds can be considered in the same way. Therefore, they and their corresponding systems responses has to be multiplied by $-1$. In order to get nominal values for the design with maximal tolerances, the optimization problem reads

$$\begin{align*}
\max_{\bar{x}, \delta} \quad & \bar{\delta} \\
\text{subject to} \quad & x \in \Omega_{ds}, \ f(x, p) \leq f_{c}(p) \quad \forall x \in [\bar{x} - \delta, \bar{x} + \delta] 
\end{align*}$$

with $\bar{\delta} \geq 0$ which was for example considered in [4]. Like before, its solution $(\bar{x}^*, \delta^*)$ contains nominal values that can be regarded as robust design. Here, this design is forced into the center of $\Omega_c$. Therefore, problem (6) is also called a design centering problem, [3].

As any $x \in [\bar{x}^* - \delta^*, \bar{x}^* + \delta^*]$ with $\delta^* = (\bar{\delta}^*, \ldots, \bar{\delta}^*)$ is a permissible design, the set

$$\Omega^* = [\bar{x}^* - \delta^*, \bar{x}^* + \delta^*]$$

is a box-shaped solution space within the complete solution space. Further, the solution space $\Omega^*$ can be used for systems design itself while offering some advantages which are discussed in the following. The method was first introduced by [9].

3.2 The standard and the worst case

In Figure 2, the solution space $\Omega^*$ is visualized.
The solution space $\Omega^*$, can be used for systems design without having to regard the design space or system responses separately. Given $\Omega^*$, the decision maker can choose a design in form of nominal values, i.e. $\tilde{x}$ within $\Omega^*$, by regarding the following situations: If the real values of $\delta$ are unknown, the decision maker should always select $\tilde{x}$, which is the center of the box-shaped solution space, in order to tolerate maximal variations of that design. If the real values of $\delta$ are known however and $\delta \leq \delta^*$ holds, he can select $\tilde{x}$ from $[\tilde{x}^* - (\delta^* - \delta), \tilde{x}^* + (\delta^* - \delta)]$ and all variations from that design are guaranteed to be permissible. This can be done for example by minimizing the robust regularization or by solving another optimization problem. Otherwise, there is no design where all variations from that design can be guaranteed to be permissible.

Nevertheless, there are no benefits from choosing a design from the box-shaped solution space compared to the complete solution space so far. Instead, some permissible design possibilities from the complete solution space are lost. Though, the major advantages of using $\Omega^*$ is that decoupled decisions for choosing design variables are enabled. Each $\tilde{x}_i$ can be chosen within $[\tilde{x}^*_i - (\delta^* - \delta), \tilde{x}^*_i + (\delta^* - \delta)]$ independently without having to incorporate the other decisions regarding the design variables for $i = 1, \ldots, d$. Thus, the decisions also do not have to be met simultaneously and single decisions can be easily adapted at a later stage. Together with the property that the design which tolerates the largest variations of all designs in $\Omega_c$ is the center of $\Omega^*$, providing box-shaped solution spaces can be a useful tool for systems design.

However, from the perspective of uncertainties, solution spaces only account for uncertainties in design variables. In this paper, uncertainties in the uncontrollable parameters, like described in (2), are also incorporated. In doing so, the optimization problem (6) becomes

$$
\begin{align*}
\text{maximize} & \quad \delta \\
\text{subject to} & \quad x \in \Omega_{ds}, \quad f(x, p) \leq f_c(p) \quad \forall x \in [\tilde{x} - \delta, \tilde{x} + \delta], \quad \forall p \in [p^l, p^u] \quad (8)
\end{align*}
$$

where $\delta \geq 0$ must be guaranteed. The solution of this problem $(\tilde{x}^1, \delta^1)$ builds up the worst case box-shaped solution space $\Omega^1$. Here, the constraints $f(x, p) \leq f_c(p)$ can be considered as hard constraints because the designs in $\Omega^1$ must fulfill them for every realization of $p \in [p^l, p^u]$. Selecting a design $\tilde{x}$ within $\Omega^1$ can be done similarly to the case which does not consider uncertainties in $p$. As this procedure is further restricted due to uncertainties in the design variables, the overall design process becomes very conservative.

### 3.3 Reformulation of the problem

The optimization problem (8) is not easy to solve for arbitrary system performance functions $f$. In the following, the case where $f_j$ is a linear or rather affine function of $x$ with

$$
\begin{align*}
f_j(x, p) - f_{c,j}(p) &= a_j^T(p)x - b_j(p)
\end{align*}
$$

for $j = 1, \ldots, m$ is considered. Here, $a_j(p)$ is a $d$-dimensional vector with $a_j(p) = (a_{j,1}(p), \ldots, a_{j,d}(p))$, and $b_j(p)$ is a scalar. In total, the inequalities $f_j(x, p) \leq f_{c,j}(p)$ are a system of linear inequalities in the form $A(p)x \leq b(p)$ for fixed $p$, where the rows of $A(p)$ and $b(p)$ are formed by $a_j(p)$ and $b_j(p)$.
respectively. With equation (9), problem (8) becomes

\[
\begin{align*}
\text{maximize} & \quad \tilde{\delta} \\
\text{subject to} & \quad x \in \Omega_{ds}, \quad a_j^T(p)x \leq b_j(p) \quad \forall x \in [\bar{x} - \tilde{\delta}, \bar{x} + \tilde{\delta}], \quad \forall p \in [p^1, p^u], \ j = 1, \ldots, m, \\
\end{align*}
\]  

(10)

where \(\tilde{\delta} \geq 0\) has to be ensured. Here, the inequality \(a_j^T(p)x \leq b_j(p)\) is fulfilled for all \(x \in [\bar{x} - \tilde{\delta}, \bar{x} + \tilde{\delta}]\) and all \(p \in [p^1, p^u]\) if and only if it is \(\sup\{a_j^T(p)x : x \in [\bar{x} - \tilde{\delta}, \bar{x} + \tilde{\delta}], \ p \in [p^1, p^u]\} \leq \inf\{b_j(p) : p \in [p^1, p^u]\}\). Given that \(x_{ds,i}^1 \geq 0, \ i = 1, \ldots, d, \) holds, then it is

\[
\begin{align*}
\sup\{a_{j,i}(p)x_i : p \in [p^1, p^u], \ x_i \in [\bar{x}_i - \tilde{\delta}, \bar{x}_i + \tilde{\delta}]\} = \sup\{a_{j,i}(p) : p \in [p^1, p^u]\}(\bar{x}_i + w_{j,i}\tilde{\delta}) \\
\end{align*}
\]

(11)

with

\[
\begin{align*}
w_{j,i} = \begin{cases} 
-1 & \text{for } \sup\{a_{j,i}(p) : p \in [p^1, p^u]\} \leq 0, \\
1 & \text{for } \sup\{a_{j,i}(p) : p \in [p^1, p^u]\} > 0
\end{cases}
\]

(12)

for \(i = 1, \ldots, d, \ j = 1, \ldots, m\). Moreover, if there is a \(p^{*j} \in [p^1, p^u]\) for each \(j\) that maximizes \(a_{j,i}(p)\) for \(i = 1, \ldots, d\) and minimizes \(b_j(p)\) simultaneously, then problem (10) can be reformulated to

\[
\begin{align*}
\text{maximize} & \quad \tilde{\delta} \\
\text{subject to} & \quad -\tilde{\delta} \leq 0, \ -\bar{x} + \tilde{\delta} \leq -x_{ds,i}, \ \bar{x} + \tilde{\delta} \leq x_{ds}, \\
& \quad a_j^T(p^{*j})\bar{x} + a_j^T(p^{*j})W_j\tilde{\delta} \leq b_j(p^{*j}), \ j = 1, \ldots, m,
\end{align*}
\]

(13)

where \(W_j\) is a diagonal \(d \times d\)-matrix for that the \(i\)th entry on the diagonal is given by \(w_{j,i}\) from equation (12). A proof for this can be done similar to [3]. Problem (13) is a linear optimization problem and can be solved numerically by standard linear optimization techniques.

4 Application to Crash Design

4.1 Systems Specifications

A vehicle that has to perform in a frontal crash can be considered as a systems design problem. Here, significant performance measures which are responses of the system can be the maximal acceleration, the energy absorption, and the order of deformation of the components. For a given vehicle structure and a deformation in one direction, all of them can be calculated from the corresponding force-deformation characteristics of its components. Here, it is assumed that the force-deformation characteristics can be designed directly where their degrees of freedom are used as design variables. In general however, they are only responses of lower level design variables, like material properties. Therefore, an appropriate calculation of the lower level design variables must follow in a second step which is not considered in this paper.

In the following, the focus is put on a scenario where the vehicle is driven against a rigid wall at full overlap with an initial velocity \(v_0\), like done in crash tests. Instead of optimizing the system responses, requirements in the form of threshold values are set. These requirements contain the bounding of the maximal acceleration by a critical \(a_c\), the complete absorption of the impact energy, and that the deformation of the components starts at the front of the vehicle.

As an example for a crash design problem, a front structure of a vehicle, taken from [2], is shown in Figure 3 and considered. This structure was obtained from the geometry of a system by assuming a discrete vehicle mass distribution, given by \(m_1, m_2,\) and \(m_3\), and by mapping the coordinates of the geometry to coordinates of simultaneous deformation. In addition, it is assumed here that every component will deform only partially, given as a combined sum of \(s_1, s_2,\) and \(s_3,\) before it behaves rigid. These rigid parts were also removed from the geometry in Figure 3.

Figure 3: Front structure of a vehicle with four components, three masses, two load paths, and three sections, compare [2]
There are four components and therefore four force-deformation characteristics. In every section, they are modeled as piece-wise constant. Thus, there are six degrees of freedom in total, meaning six design variables. One design variable belongs to each the first \( (F_1) \) and the forth \( (F_6) \) and two to each the second \( (F_2 \) and \( F_3 \)) and the third \( (F_4 \) and \( F_5 \)) component. Every \( F_i \) is lower bounded by \( F_{0,i}^{\text{ds}} = 0 \text{kN} \) and upper bounded by \( F_{0,i}^{\text{ds}} = 300 \text{kN} \) for \( i = 1, \ldots, 6 \) what forms the design space \( \Omega_{\text{ds}} \).

In order to keep the maximal acceleration below the critical \( a_c \), the sum of deformation forces of simultaneously deforming parts must be bounded, which is stated by three linear inequalities for the given structure. These inequalities are composed by the columns 1-3 of \( A(p) \) in equation (15) and of \( b(p) \) in equation (16). To absorb the complete impact energy, the sum of integrals of the component’s deformation force over their deformation length must be greater than the impact energy itself. As the deformation forces are constant in every section this constraint is also a linear inequality which is composed by the columns 4 of \( A(p) \) in equation (15) and of \( b(p) \) in equation (16). To ensure that the deformation of the components starts at the front of the vehicle, the deformation force in every section of a previous component must be smaller than the force which is necessary to start the deformation of a subsequent component. Again, these requirements can be expressed as linear inequalities which are composed by the columns 5-7 of \( A(p) \) in equation (15) and of \( b(p) \) in equation (16).

In total, constraints on the system responses are given by

\[
A(p)(F_1, \ldots, F_6)^T \leq b(p)
\]

with

\[
A(p) = \begin{pmatrix}
1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 \\
-\frac{s_1}{m_1 + m_2 + m_3} & -\frac{s_2}{m_2 + m_3} & -\frac{s_3}{m_2 + m_3} & -\frac{s_1}{m_1 + m_2 + m_3} & -\frac{s_2}{m_2 + m_3} & -\frac{s_3}{m_2 + m_3} \\
1 - \frac{s_1}{m_1 + m_2 + m_3} & 0 & 0 & 1 - \frac{s_1}{m_1 + m_2 + m_3} & 0 & 0 \\
-\frac{s_1}{m_1 + m_2 + m_3} & 0 & 0 & -\frac{s_1}{m_1 + m_2 + m_3} & 0 & 0 \\
0 & -\frac{s_1}{m_1 + m_2 + m_3} & 0 & 0 & 1 - \frac{s_1}{m_1 + m_2 + m_3} & -1
\end{pmatrix}
\]

and

\[
b(p) = \begin{pmatrix}
(m_1 + m_2 + m_3)a_c & (m_2 + m_3)a_c & m_3a_c & -\frac{1}{2}(v_0)^2 & 0 & 0 & 0
\end{pmatrix}^T,
\]

see [2]. Therefore, the system performance functions together with their threshold values can be also written in the form of equation (9). The uncontrollable parameters are collected in the vector \( p = (s_1, s_2, s_3, m_1, m_2, m_3, v_0, a_c) \). These are the section lengths \( s_1, s_2, s_3 \), the masses \( m_1, m_2, m_3 \), the initial velocity of the vehicle \( v_0 \), and the critical acceleration \( a_c \). Here, the intervals where their real values can be found are given in table 1.

<table>
<thead>
<tr>
<th>Uncontrollable parameter</th>
<th>Interval of the real values</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_1 )</td>
<td>([0.19 \text{ m}, 0.21 \text{ m}])</td>
</tr>
<tr>
<td>( s_2 )</td>
<td>([0.14 \text{ m}, 0.16 \text{ m}])</td>
</tr>
<tr>
<td>( s_3 )</td>
<td>([0.19 \text{ m}, 0.21 \text{ m}])</td>
</tr>
<tr>
<td>( m_1 )</td>
<td>([95 \text{ kg}, 105 \text{ kg}])</td>
</tr>
<tr>
<td>( m_2 )</td>
<td>([145 \text{ kg}, 155 \text{ kg}])</td>
</tr>
<tr>
<td>( m_3 )</td>
<td>([1150 \text{ kg}, 1250 \text{ kg}])</td>
</tr>
<tr>
<td>( v_0 )</td>
<td>([15.5 \text{ m/s}, 15.6 \text{ m/s}])</td>
</tr>
<tr>
<td>( a_c )</td>
<td>([290 \text{ m/s}^2, 310 \text{ m/s}^2])</td>
</tr>
</tbody>
</table>

Table 1: Uncertainty modelling of the uncontrollable parameters

As there exists a \( \mathbf{p}^{*, \odot} \in \mathbb{P}^l \times \mathbb{P}^u \) that maximizes the entries of the \( j \)th column of \( A(p) \) and minimizes the \( j \)th entry of \( b(p) \) for \( j = 1, \ldots, m \), the crash design problem to find worst-case box-shaped solution space can be formulated in the form of problem (13) and be solved accordingly.

### 4.2 Solution Spaces for Crash Design

In Figure 4, the worst case box-shaped solution space \( \Omega \) for the crash design problem is visualized.
In addition to the worst case solution space, also the standard case with $p = \frac{1}{2}(p^l + p^u)$ is shown. As $\frac{1}{2}(p^l + p^u)$ is contained in $[p^l, p^u]$, the complete solution space that accounts for all $p \in [p^l, p^u]$ is contained in the one that accounts only for $p = \frac{1}{2}(p^l + p^u)$. Therefore, the sizes of the intervals that form $\Omega^\dagger$ are smaller or equal to the ones that form $\Omega^\star$. For the crash design problem, these intervals are contained in each other where the interior intervals are oriented towards the upper bounds of the exterior intervals, see Figure 4. This can be explained by the influence of the uncertainties in uncontrollable parameters on the constraints that build the solution space. Here, the influence on the energy absorption which is related to small design variable values is greater than the influence on the maximal acceleration which concerns large design variable values.

5 Conclusions

In this paper, worst-case box-shaped solution spaces were introduced as a method for systems design. These solution spaces consist of maximal intervals for the design variables within these, all requirements on the system responses are satisfied. Furthermore, a decoupled selection of the design variables is enabled by using solution spaces. Compared to standard solution spaces not only uncertainties in controllable design variable are considered in this method, but also uncertainties in uncontrollable parameters which were given as intervals here.

Additionally, it was shown how worst-case solution spaces can be calculated for affine performance functions. Applied to a crash design problem, the influence of uncertainties in uncontrollable parameters on the solution spaces could be illustrated. For further research, the focus should not be only limited to affine performance functions and interval uncertainties for uncontrollable parameters in order to extend the methods applicability to systems design.

Acknowledgments

This work was supported by the SPP 1886 “Polymorphic uncertainty modelling for the numerical design of structures” of the German Research Foundation, DFG.

References


Digital twin of nuclear waste management facilities

M Bankhead, K Fritsch J MacDonald-Taylor, D Trivedi, G Ecclestone, and W. J. O’Brien

Abstract

We describe the functions and features of a prototype digital twin developed to support effluent management from a facility undergoing decommissioning. The concept of a digital twin is fairly recent and there are relatively few descriptions of digital twin frameworks in the open literature. Moreover the application to decommissioning is novel and challenging due to the dynamic nature of the process and the large uncertainty. It has been recognized that digitalization of knowledge across the life-cycle could deliver significant benefits to the nuclear industry; Successful decommissioning projects require improved process reliability therefore reducing risks associated with schedule and cost whilst ensuring operations are undertaken safely. The digital twin consists of the integration of numerical models that cover the operation of the asset. These models simulate from a system level description of scheduled operations and process flows of waste material extending down to chemical models that can account for changes in speciation and solubility as a function of process conditions. Statistical tools have been developed to calculate the uncertainty in plant data and this can be propagated through to the predictions. The digital twin has been developed within a modern enterprise server bus framework with a web based front end. Back-end services are scalable and have been designed to take advantage of high performance distributed computing to accelerate model solution.

Keywords: Digital Twin, Uncertainty quantification, chemical process engineering.

1 Introduction

Nuclear decommissioning is a strategic goal of the United Kingdom to achieve the end state of all historic nuclear sites by 2125. These sites include the 1st Generation Magnox fleet of nuclear reactors and the associated recycling facilities at Sellafield site. These 1st generation facilities were not built with decommissioning in mind and this contributes to the relatively high cost relative to today’s operating 2nd and 3rd generation plant where lessons have been learned.

The Magnox Swarf Storage Silos (MSSS) are a facility that is currently operating on Sellafield site and is undergoing a process of decommissioning. The facility was established as a temporary storage facility for waste materials, principally fuel cladding and miscellaneous solids wastes, from military and civilian reprocessing operations. These materials are stored under water to minimise the hazard associated with oxidation and secondary benefit of radiation shielding. The building is aging and the structural integrity represents a significant risk; as such the facility is a priority for decommissioning. A simplified description of the process will involve the removal and safe storage of solid wastes. Radioactivity has leached from the solid wastes into the cover water. Furthermore further liquid wastes will be generated through evaporative top ups and washing of waste grabs. These liquid wastes will also be removed and they will be discharged through a treatment plant which will remove radioactivity before discharge to sea. The first operations to remove and replace radioactive liquors, liquor activity reduction (LAR), have already started, with the first phase completed. From next year, solids wastes retrievals will begin. Over the next two decades, significant quantities of waste will be retrieved from these silos by mechanical grab and transferred to waste boxes. Liquors within the waste silos and additional liquors generated in operations containing potentially hazardous levels of radioactive species will be treated.

Planning the retrievals process is challenging because no plans were drawn up for how waste would be taken out of the building when it was built in the 1960s. Inventory records were kept but it is unknown how the waste has changed after several decades of storage and so it is difficult to predict what will arise when waste is retrieved. The strategy adopted by the programme is to undertake early phase operations in a cautious manner (leading phase) to gather information during early retrievals (learning phase), and this knowledge allows acceleration of the decommissioning in later phases. Digitalization of the information management has the potential to extract as much useful information as possible from retrievals to improve predictions of how the plant will operate in later years. A digital twin representation of the real world asset is required to capture this learning to lead future operations.
2 Introduction to digital twins
The origin of the concept of a digital twin is uncertain. Early references to the concept can be found in NASA and aerospace technology roadmaps, [2], [3]. The role of digital twins in the nuclear industry has been argued by [4]. A digital twin is a framework capable of predicting the most important variables of a physical asset. It is built from a number of components which are integrated into a single powerful tool that enables the end user to predict the outcome of scenarios based on changing process variable. At the core of the digital twin are phenomenological relationships, which can be mechanistic models or data driven models that predict the values of outputs as a function of the inputs. As the digital twin is a representation of a real world operating plant, vehicle or building it is implicit that there will be some coupling of the framework to data obtained from the system. The tight integration of data and model allows the twin to learn from the behaviour of the operating plant. The final part of the picture is data required to validate the phenomenological models. This data can be acquired from commissioning or laboratory experiments and is used to establish the credibility of the models. Credibility is obtained from quantifying the uncertainty and therefore is useful to couple within the digital twin mathematical models to quantify uncertainty in both the model (epistemic or systematic uncertainty in the phenomenological model) and the plant data (aleatoric or statistical uncertainty).

2.1 Origin of a digital Twins for MSSS liquor waste management
To plan decommissioning operations we need to know how much waste is to be treated, what the waste consists of, when it needs to be treated and what other wastes need treating at the same time that will compete for the finite capacity in the treatment plants. To assist planning, models are developed to understand the uncertainty associated with each aspect. Models are developed in computer software and simulations are carried out where the inputs are varied to quantify the uncertainties. Data is gathered from both the real world asset and laboratory experiments to first parametrise and then validate these models. A key part of the solution is to ensure that this information is then made available to those who need it to make decisions.

In decommissioning we have the problem that the process is inherently not steady state as over the lifetime of the operation we want to achieve an end state of zero output and the waste arising varies with the type of operation being carried out. Therefore we must use tools for modelling the dynamics of the system as we progress from an operational (where outputs reflect normal operations) to fully decommissioned state (with zero or close to zero outputs). Process operations can be both discrete (grabbing some waste and transferring it to a box) and quasi-continuous (batched transfer of radioactive liquor from the silos to treatment plant and continuous processes such a liquor evaporation). It can also be important to consider additional variability due to the availability of resources and failure rates of plant equipment. We can add to this complexity the dynamic changes in the system chemistry which in turn impacts on our ability to process the effluent. All of these processes require different types of model and the usual practice is to develop independent models in software specifically optimally designed for the purpose of solving a particular type of real-world problem.

When considering the performance of the decommissioning process as a whole, we need to understand how the outputs of the various models link together. Standard practice within the industry is to exchange data via a quality assured process of written reports which have been assessed technical peer review. However this process is inefficient in that only a limited amount of data can be interpreted and passed to the next model. This limits the quantity and fidelity of the solution that can be interpreted by the decision maker. Once the need to exchange data is determined there is an obvious next step to consider how the models themselves can be integrated into a digital twin to improve performance and enhance the quality assurance on the exchange of data. The adoption of a digital twin widens the stakeholders that need to interact with each model establishing a need for a simple, easy to use interface so that those who need to access the inputs and outputs can do so.

3 Methodology
3.1 Existing and planned components of the digital twin framework
The software architecture is based on a Service Oriented Architecture where information (and models) is broken down into discrete blocks such that one service only deals with one particular type of data. A central service register tracks where services and versions are located. At present an event management system has not been implemented as there is currently only the need to implement a limited number of pre-defined operations. The components of this architecture are described in more detail in the following sections. The services are presented to the end user via a web interface with industry standard security protocols. Each of the services is implemented using micro-server architecture with dynamic load.
balancing and parallel distributed computing enabled. The service is deployed on a high performance distributed computing architecture, which includes scheduler execution services to manage simulation throughput of sensitivity and uncertainty quantification calculations.

3.2 Source data and data management
The first task is to identify the temporal nature of the data which reflects its behaviour in-service through to decommissioning. Historic data refers to the liquor inventory data from tipping records and sample data acquired during early phase of the operations. We have the highest degree of confidence in the data acquired in the more recent time periods (within the last ten years) as we can be sure of the quality of the data and analytic records are stored in a digital format. Finally, there is considerable potential to acquire new data during retrievals and capturing this data is a key part of the learning process that is to be followed.

We have developed a data management system to both store the liquor data which is based on a SQL database for storing the analytical data and MongoDB database for storing documents, which includes simulation outputs. When connecting different physical and temporal scales, we need to consider how data from one scale can be extrapolated (coarse-grained or interpolated) to another. For a decreasing granularity of the data, we can apply statistical methods to either spatially or temporally average the data. For an increasing granularity of data we need to apply interpolation to fill in the gaps between ‘coarse-grained’ data measured over a long time scale and data required at a ‘fine grained’ level. If this data was not measured, it must be artificially generated (interpolated) in a realistic and representative manner. A statistical analysis functionality developed in R-shiny provides these interpolation functions. We have analysed the data from the liquor activity reduction operations and have determined its effectiveness. A linear trend was fitted to the data with functions in ggplot2. The confidence interval was determined from the t-distribution with P=95%. The effectiveness of the operation can then be determined by comparing the slopes of the fit to an extrapolation using radioactive decay (eq. 2). The results from a typical analysis are shown in figure 3, overlaid with the results of a simulation from the current development version of the digital twin.

Implementation of extrapolation and interpolation are described in the following section and are limited to specific cases where codes are coupled: extrapolation is only currently only applied to the experimental data trending and interpolation is currently only used to reduce the number of calls to PHREEQC from the process model.

3.3 Models and outputs
Going beyond the data management the design of the digital twin is that of a hierarchy of coupled multi-scale and multi-physics models that together capture the key phenomenological properties of the asset. Individual models can be placed into categories based on the spatial and temporal resolution of the underlying physical models:

3.3.1 Chemical and physical underpinning
Thermodynamic and kinetic models representing the changes in chemical species within the effluent. Physical (chemical) models represent particulate interactions at very small spatial and temporal resolutions (lengths in nm, time steps hours and timescales in years with temporal resolution of fractions of seconds). Also represents mechanistic models of physical processes like settling of colloidal suspensions and bulk flow of liquids through pipes and transfer lines. The key phenomenological properties of the effluent arising from decommissioning operations are the solubility of species that can challenge downstream treatment plants and the potentially in future we can model the formation of mobile slow-settling or colloidal materials that can similarly be transported via pumps and transfer lines. Chemical reactions between the remaining solids and the liquors could result in the release of activity which in future will be captured in the phenomenological models.

The development of a purely mechanistic model to predict speciation and solubility of legacy wastes is an unsurmountable challenge due to the chemical complexity of the system and a paucity of information of the source terms (detailed chemical knowledge of the waste inventory). Data is readily available for the concentration of species in solution obtained from sampling and chemical analysis on up to a daily basis, and extending back to the 1970’s (though less frequently) when waste was still being deposited. Trending of this data can be used to predict the change in concentrations over time. We have integrated a tool set developed in the R statistical language within the digital twin framework to undertake trending analysis of the input data. This can compute the trends over time to compute the release rates of activity
from current operations. As decommissioning progresses we can analyse the emerging data to update the
the correlations as required.

The challenge comes when identifying limiting cases, for example where the concentration of a species
assumes a nominally constant value. We have therefore derived a hybrid approach where mechanistic
models developed in PHREEQC are used to capture the solubility/ speciation behaviour at the point
where a step change occurs, for example the limit of solubility. The solubility of species is determined
by the solubility product; this is an equilibrium reaction that expresses the concentration of the ions in
solution and in the solid product in terms of an equilibrium coefficient. The effective concentration (or
activity) of species in solution is corrected for the excess free energy of mixing via Pitzers method,[5].

The materials in the silos are continuously slowly changing due to reactions such as corrosion. However
we can use this approach by assuming that the principle of local thermodynamic equilibrium applies due
to the large difference between the silo heterogeneity (which is macroscopic) with the thermodynamic
principle of heterogeneity (which is microscopic). Fundamental chemical data is managed in a database
and has been critically assessed to determine its suitability for this specific application.

The PHREEQC model can be used to determine whether a measurement exceeds the limit of solubility
of the species. If the measured value exceeds the solubility limit, then either excess quantity of a species
is present as a third phase (colloidal); or the measurement is suspect. At equilibrium, the concentration
of a substance in solution is given by the distribution constant, the ratio of the concentrations in the
solid and aqueous phases (eq. 1). Currently we can only measure the concentration in solution however
we can assume that the solids concentration is effectively constant as it is in large excess and is based
on the inventory. To obtain $k_D$, we used the solution data where we can assume the concentration is
constant (at equilibrium) then apply eq. 1.

$$k_D = \frac{[A]_{solid}}{[A]_{aq}}$$  \hspace{1cm} (1)

A further mechanistic process that can be captured is radioactive decay. Many of the important
radioactive species have short half-lives and the change in concentration of these over time can be accounted
for. The half-life is computed from a simple 1st order exponential decay expression for the amount of
substance $N$ remaining, the integral form being:

$$\log \frac{N}{N_0} = -kt$$  \hspace{1cm} (2)

The effectiveness of liquor dilution on the radioactivity of the waste can be calculated as the ratio
of the slope of the measure radioactivity data over 1st order decay constant, $k$, which is related to the
half-life of the radionuclide.

3.3.2 Unit operations/ component models

Conservation models for mass, energy and phase changes in chemical species evaluated over the physical
scales of filtration, pipes, settling tanks, etc. over length scales equal to the real physical dimensions of
these vessels and time-scales up to those of the typical residence times within the vessels (lengths from
centimetres to meters, time steps in minutes and timescales from seconds to months).

For the MSSS process unit operations are the individual silos, transfer lines and discharge tank. The
models are developed on a compartment by compartment basis which reflects the known inventory and
chemical analysis of each compartment. Each compartment group typically has hydraulic connectivity
and these results in common liquor chemistry until such time as retrievals have progressed to the point
where liquor levels fall below the connectors. Both the compartment model and the chemical process
dynamic model are developed in the gPROMS software and language. As the chemistry of the waste
is linked to the compartment, we have directly integrated the chemical model developed in PHREEQC
with the compartment model. This coupling is implemented in a C++ object interface to deliver high
performance. However, calling PHREEQC results in a computational overhead that could be problematic
if running a simulation over the whole lifetime of planned retrievals (25+ years.) and so the model has
been optimised in this respect with interpolation used to minimise external calls.

3.3.3 Process models

Conservation models for mass, evaluated over the physical scales of connected unit operations (sequences
of vessels, pipes etc. in either series or parallel) over length scales representing the components of a
chemical plant or series of plants (Lengths from metres to 100s of metres, time steps in hours and
timescales from days to decades). The process model is set up to represent the transfer operations
associated with the planned decommissioning choreography.

The process model computes dynamically the changes in silo inventory as a function of planned re-
trieval operations. Each of the silo models is close-coupled within the process flowsheet and the gPROMS
model describes mathematically the mass flow of materials between them. Mass balance equations are
computed to calculate the volume change and the activity change as liquor is transferred between the
compartments and to the discharge. The general form for activity \( A \) of soluble species \( i \) is

\[
\frac{dA_{(aq)i}}{dt} = \sum \text{Ain}_{(aq)i} + L_{(aq)i} + R_{(aq)i} - \sum \text{Aout}_{(aq)i}
\]

(3)

For radioactive species an additional decay term (eq.2) is added to the RHS of eq. 3. The reaction
term covers the rate of change of species that are known to undergo chemical reactions. For reactions
we can assume the reaction is limited by \( k_D \) or by solubility, determined from the PHREEQC model
and the limiting value is chosen. Release fractions, which define the quantities released during retrieval
operations, are also determined from \( k_D \). As these cannot be known for future operations, they will need
to be fitted to plant data once these operations actually start. The leach rates, \( L \), are determined by
the rate of reaction, which are assumed to be instantaneous in most cases. For some species, we adjust
the rate constant minimising the sum of squares of the error until the model prediction closely matches
the measured trend. Currently the activity balance for 29 chemical species is tracked in the model and 8
reactions are modelled. We also apply balance equations of a form similar to eq. 3 for the mass of solids
in the silos and volume of liquors, both of which change over time depending on the operations being
carried out (for example a waste grab removes solid waste, and a liquor top-up adds water to the silos to
increase the volume).

Interpolation between model scales is currently only applied to the linking of the PHREEQC chemical
model and the gPROMS process model. Calling the PHREEQC model at every timestep would be
prohibitively expensive over the run-time of the simulation, and so it is called based on schedule and
linked to specific operations. The log form of a 1st order rate equation, where the rate constant is either
assumed to be constant and large or is back fitted as discussed earlier, is used to numerically scale the
error in the process model to bring the prediction in line with the PHREEQC prediction.

3.3.4 Operational Research (OR)

Conservation models for mass and usually other physical constraints (availability of manpower etc.)
evaluated over a series of connected plant flowsheets (length and time-scales identical to the process
models). Some processes are simplified to increase the range of process variables that can be modelled.

This models the retrieval operations as a function of resource constraints and ultimately provides
the input to the process model. The system model is developed in FlexSim from Saker Solutions. The
process model is developed to match the operations modelling in the OR model, with the additional
overlay of the chemical changes as included through the silo model. Ultimately it is the OR model that
provides decision makers with information on what decommissioning schedule is achievable. The two way
exchange with the process chemical models assures that this is not limited by chemistry.

3.3.5 Uncertainty Quantification

Deterministic simulation of the impact of variable uncertainty on the simulation outcome computed for
an individual campaign (months) or over the decommissioning schedule as a whole (years) in order to
obtain a probabilistic outcome for the discharge profile for liquid effluent.

To understand the importance of the sensitivity of the performance of downstream plants we need
to evaluate the propagation of uncertainty of the feed liquors and eventually also to understand the
impact of model uncertainty. The distribution of variable uncertainty can be obtained from a top-down
analysis of the liquor chemical analysis using the R functionality that has been built into the digital
twin. To undertake sensitivity analysis and eventually uncertainty quantification we plan to integrate
the COSSAN-X software with the gPROMS model. We have already demonstrated the potential of this
technique for effluent treatment in published work [6]. This will enable multi-variate sensitivity studies
to quantify the uncertainty in the outputs.
Digital twin of nuclear waste management facilities

Figure 1: 3D Render of the Magnox Swarf Storage Silos (Source: Sellafield Ltd, original video on ITV.com). The OR model described in 3.4.4 uses 3D visualisation to demonstrate potential bottlenecks in decommissioning activities. The image shows three of the silos which contain a mix of solid debris, settled solids and liquid wastes. The digital twin has been designed to support operations for the removal and treatment of the liquid waste.

4 Conclusions

We have presented a short technical description of a framework for a digital twin of silo decommissioning. The development process of the system architecture is being delivered in several phases, with the database and statistical functionality already implemented and the current work programme integrating the process model service into the architecture. This work will continue in line with the commencement of the next phase of retrieval operations with the approach evolving as the availability of data increases. Value has already been demonstrated with the individual twin components already used to support key decision milestones and the capability we will develop will mitigate against schedule delays, contributing to the saving millions over the lifetime of the plant decommissioning. One of the ways this will be achieved will be to put the tool in the hands of decision makers allowing for real-time analysis of current and planned operations.

Key challenges that have been overcome in this project include collaborative working between teams, across different organisations, which have traditionally focused on solving problems within their own discipline. Further collaboration will be necessary to implement features on the current roadmap, particularly uncertainty quantification. The tools produced to date have been used to support key decisions and as the technology evolves, widening the stakeholder interaction then we will gain further tangible benefits against the current decommissioning challenge. One of the key learning points from this programme has been in the value of data. From an analysis of the quantity of historic data available from which to construct a digital twin, it is clear that increased knowledge of past operations would greatly assist decision making today. It is important for industry to re-evaluate the value and need for data given the continued rise of digital technologies.

References

Figure 2: Schematic of the service oriented architecture of the digital twin framework. Showing the high level components (left) and the service Oriented Architecture (right). The R service is integrated as a client and the process model is integrated as a web service (NNL Modelling Web) as part of the High Level Architecture. A more detailed description of the architecture is presented in section 3.3 and 3.4.

Figure 3: Results of the trending analysis showing the decreasing silo radioactivity with the commencement of decommissioning operations. Sample points are shown as filled circles and the linear fits to the trend and the radioactive decay prediction are shown. Due to the fact that these results represent a work in progress other specific details have been omitted. The simulation (light blue) is the result of the combination of the PHREEQC chemical model coupled with the process simulator (as described in sections 3.4.1-3.4.3). The model prediction falls within the confidence interval and represents a step towards validation and acceptance of the model.
Human involvement the unforeseen and robustness

Franz M. Knoll

1Vice-president, NCK Inc, Montreal, QC, Canada
*Corresponding author: franz.knoll[at]nck.ca

Abstract

Human error is present in all human endeavours and has proven to be quite intractable scientifically due to the great variety of causes and circumstances associated with it. To prevent it from taking effect altogether has turned out to be equally out of reach as checking and verification techniques are limited timewise and cost wise. Safety margins are intended to protect against expected variation of building parameters; they are tuned to probabilistic assessments of these variations, with no consideration of those caused by humans. An alternative strategy to mitigate the effect of human shortcomings in planning and execution of construction projects is proposed to be equipping structures to essentially resist unforeseen circumstances by resilience and reserves of strength and deformability, i.e. robustness. The art of robustness as required by modern building codes is not intended to compensate for the "unforeseen" caused by human error and shortcomings but in fact it does. Nevertheless, it should not be interpreted as a panacea compensating for sloppy work, or for a relaxation of checking and verification.

Keywords: Human error, Robustness

1 Introduction

Human made systems being what they are, infested with the shortcomings of human cognition, reliability and attitude, they can be seen as the product of an incomplete optimisation. Depending on the situation that optimisation varies from a well-thought out and long process such as the development of an airplane model, to the momentary snap decision made on a construction site where the progress of the work was stopped due to a “site” problem. In all such situations insight and devotion of the human participants is of decisive importance and the effective mobilisation of these is a primary task of management. It has been posited that every error that occurs and is allowed to persist, is also and always an error of management. Management implies the organization of the process, i.e to put appropriate human resources in the right positions.

Check lists and prescribed protocols have been put in place and work effectively in many contexts but not in others. When looking back on more than fifty years of practice as structural engineer, in commercial and institutional construction, this author has learned that much of the “hunting down” of errors happens in more or less informal ways where “well seasoned”, i.e. experienced people view and review the production, communication and execution of the information that constitutes the construction process from conceptual design to physical expression.

We know for quite some time that this “filtering” reduces an initial rate of erroneous information and decisions by one or two orders of magnitude (1,3). If we wish to make it yet more effective, the conclusion seems to be that this can only be done by enhancing the circumspection, insight and commitment of the humans involved, and to increase the time and financial resources devoted to control and review. Limitations to both are often found among the main causes of something going wrong.

Codification, standardization and regulation were meant to help with this but have now become so all-encompassing, convoluted and voluminous as to be counterproductive. The response of industry and practice to this has been to delegate it all to computer processing which introduces an entirely new source of error that we are presently forced to recognize: The computer glitch. Once again, the only means to control this newly created problem, is to engage the insight of experienced humans.

In spite of all the progress of technology and codification the real rate of failure in construction does not seem to be abating, for reasons that cannot be discussed in this paper (increased complexity, sharpened pencils for economic gain, etc). It is clear however that it would be desirable to find other means to do something about the accidents and losses that still happen, complementing the “filtering” and optimization the construction process provides.

If little can be done about the frequency of errors, something may be possible to mitigate and control the consequences of those errors. Strategy doing this is the design for robustness, making the structures...
more “resilient”. Its essence and limitations are discussed in section 3 hereafter.

2 The human mind, its strength and weakness

The cognitive processes involved in “error hunting” have not been researched in any depth to this author’s knowledge although their manifestations for example in the form of apprehension, are rather familiar. They frequently occur in sleepless early morning hours when, medically speaking, the blood temperature in our bodies is slightly lower than during the active daytime, making things appear more sombre or outright scary, leading the mind to “ruminate” on what is bothering. On numerous occasions this has led to better perception and to timely and pertinent correction of decisions and information. It’s pertinence is very much related to the emotional involvement one has with the professional activity. If those emotions are of a negative nature, or non-existent, for example because of personal conflicts at the work place, or personal problem situations, unease of body or mind, attention and cognitive digestion will suffer. On the other hand, an engineer too confident and sure of himself will tend to be less circumspect and attentive to potentially adverse situations than his worried colleague.

Between the formalized protocol and the informal review by experienced professionals every variation can be found for the activity of error control, all of them depending for their effect on the emotional involvement of the humans doing it. It will be difficult to produce research that stands up to the rigours of modern science when it comes to emotional involvement of the human organism with the work it is doing. It is this author’s experience however with observing colleagues in the engineering enterprise I have worked for more than fifty years, that individuals who made their work a part of their life rather than just an exercise to “make a living” have been much more successful in producing satisfactory and reliable results. People whose priorities lie elsewhere tended to quit the company sooner, leaving often a ‘mess’ behind that needed to be cleaned up, meaning to get rid of erroneous information such as ‘bad details’, ill thought through decisions, contradictions etc.

It is a tempting conclusion to believe that the power of ’lateral thinking’, i.e. imagination, insight and circumspection, are directly related to how completely the mind is attending to the task. ’What I like doing I am doing well and what I am doing well I like doing’. It may be that all this is having to do with our need for a sense of accomplishment, as basic as the daily bread, one might say.

Some types of potential sources of errors are better dealt with formally such as inconsistencies of data and communication – is the intent of the conceptual engineering drawings truly reflected on the manufacturers shop drawings? Others such as modifications along the process, must be reviewed for all their ramifications on other elements, something that involves the contribution of people familiar with those. Conceptual and fundamental decisions must be revisited for all their consequences which may reverberate throughout the entire destiny of a building project, including use and, eventually, deconstruction: why was the penny-wise decision made for the more “economic” construction method which causes future modifications to the structure to become so much more onerous, complicated or risky?

There are fundamental differences between large and complex endeavours such as offshore oil platforms, nuclear power stations on one hand, and everyday simple projects like high rise residential constructions, road bridges, warehouses, parking garages etc., on the other.

In the first case, vast amounts of regulation and protocols are installed and extensive studies conducted in order to detect potentially dangerous scenarios – with mixed success as we now know, following such recent accidents as Fukushima or BP’s platform in the Golf of Mexico. On the commercial construction site time pressure reigns and things must be decided, adjusted, corrected or adapted quickly which does not permit any lengthy testing, study or reflection.

Experience shows that the accumulation of several harmless looking circumstances can cause great and unanticipated damage. The anticipation of such scenarios may sometimes resemble clairvoyance and is the product of experience, imagination and dedication; it is often treated in brainstorming sessions, involving participants from several disciplines, with some success. However, this still resembles a “shot gun” approach where a good portion of the issues is reviewed in what may be called “guided random” manner, depending on what crosses the mind of the participants, i.e. their power of imagination. To make the review process more exhaustive, a more systematic approach will be needed. It has been proposed that this should be given to Artificial Intelligence, and it certainly looks as though this will come - we are not there yet.

In any case, the human input will be needed for a while to come, and the role of random style lateral thinking will always exist, including its emotional basis.
3 The unforeseen, safety margins and robustness

Every building process and its product are carrying a certain measure of uncertainty that can traditionally be represented by some probabilistic concept such as a “bell curve”. This uncertainty has numerous sources, to begin with variations of material properties, imprecisions of geometry, underestimated exposure, etc.

Codified safety margins are intended to compensate for this. They are limited to what the Code Committee was able and willing to recognize as “legitimate” or “expected” variations. It has been customary since the time design rules became based on probabilistic concepts to calculate numerical values of the safety margins from “\(\beta\)” factors of 3.5 or so, meaning that variations from “nominal” values up to 3.5 standard deviations should be compensated by factors reflecting pessimistic expectation about reality:

\[
\text{Prob (real resistance > exposure)} \geq 1 - (10^{-5} \text{ to } 10^{-6})
\]

and

\[\text{factor } x \text{ nominal resistance } \geq \text{factor } x \text{ expected exposure}\]

Theoretically then, a probability of \((10^{-5} \text{ to } 10^{-6})\) would be the result, far away from reality which is \(10^{-2} \text{ to } 10^{-3}\) or so depending on the circumstances and how one counts. These probabilities, real or theoretical, correspond to a sort of consensus in society that this is by and large as it should be since the cost of increased margins to produce a lower probability of failure, would not be acceptable. Of course, this does not prevent society from hunting down and finding a guilty party if something goes wrong.

As a consequence, safety margins are and have been essentially unchanged for several decades and the quest for better safety has recently taken a different tack; it focusses on the description of the failure and the goal has become to mitigate its consequences, excluding catastrophic events and limiting the cost associated to what one did not succeed to catch and correct in time. This can be summarized with the term “robustness”. A precise and all-inclusive definition of this concept has been elusive but the intent is clearly to deal with the shortcoming of probabilistic theory at the tail end of the “bell curve” where the discrepancy between theory and reality manifests itself.

Where probability theory described the behavior of events of a random nature, the product of building construction is not so random, thanks to the involvement of human agents as we now know. It makes the tail end of the bell curve “thicker”, with more frequent “outlayers” than the theory would have it.

The sources and manifestations of that involvement are varied and manifold. To control it through the organization of checking and review processes has always been the goal of the builders, and is recently being made the object of research. At best, this will be partially successful, leaving a residue of unhappy circumstances as “facts of life”.

This is where robustness comes in, mostly in the physical sense, i.e. making structures respond in less unfavorable ways to the unforeseen which now includes the faults produced by human agents.

Robustness was always implied in the work of thoughtful builders; it has recently been approached in a more systemic way mostly by researchers with experience in real life construction (2,4). In its practical application it is not a quality lending itself to be measured in the traditional sense by units of some sort. It must be adapted to particular circumstances which may themselves include elements that cannot be quantified meaningfully, e.g. the omission or misplacement of physical parts of a construction, errors of calculation or interpretation, lack of attention, ignorance, the unforeseen exposure to destructive agents, bad workmanship etc.

Robustness will help to attenuate the consequences of all this. However, as an answer to the problem of the unforeseen, it cannot be a final and all-encompassing solution, for two principal reasons:

- Robust structures are not meant to “survive” the unforeseen unaffected but their state “after the event” must only satisfy certain criteria, e.g. for buildings to remain standing after a strong earthquake, permitting the occupants to leave. Other examples include structures which have been “massacred” by thoughtless destructive alteration, for instance in connection with the installation of new piping or electrical and mechanical systems, or have simply been overloaded. Local deformations, cracks or slight damage may be acceptable, collapse is not.
- The scope and variety of human “action” and its effects are limitless and cannot be compensated entirely by ever so thoughtful and targeted measures and strategies. In this context the wilful destruction by war or sabotage, or the degradation through rot or corrosion over time due to lack of maintenance are not events that can be compensated in all circumstances by measures of robustness.
Robustness is a concept that has its limits, at least within “reason”, and must be seen as complementary to the elimination of errors and flaws through control processes. It could be brought to extreme degrees if it must, at commensurably high cost. Some fortifications of the Second World War still exist – they were built so massive that their destruction became too onerous and expensive. Some of the Egyptian pyramids have survived more or less unscathed – some historians say that their construction was instrumental in ruining Egypt’s economy of the period.

Robustness has been included in modern building codes as a requirement for new construction (most codes are intended principally for this, at the exclusion of existing structure being altered, modified or extended.) How to do this is mostly left to the insight of the engineer – only recently has research on structural robustness been undertaken in academia. It was recognized as a rather vague concept with less than well defined parameters and aspects that does not lend itself to the abstraction and quantification most research calling itself scientific, needs as a prerequisite.

Mostly, it is presently treated, in the form of hypothesized scenarios, "what will happen if..." the selection of which is very much a matter of insight and thoughtfulness of the people involved. Obviously, this methodology has failed in the case of Fukushima. It may be said however, that we now command the tools to analyze what happens to structures when they are taken beyond the traditional limits of stress and deformation that were mostly based on the classic elastic theory. Forgetting about minute precision, behaviour of materials in the non-linear, inelastic range can be evaluated and applied to structural systems, predicting their response to loads or deformations exceeding codified values. Research on seismic response has spearheaded this and it’s findings can be transposed to other scenarios. There will always be limits to this, however.

An example from real life may illustrate a number of things discussed above: The case involves a small bridge with two pylons and stays made from steel rods. Everything went well at first and the bridge was opened for use by normal road traffic. However, the cumulative effects of a number of faults ultimately led to an accident scenario where the structure suffered severe damage, the bridge had to be closed for traffic and repaired at a cost comparable to the initial cost of construction, including the expense for all the legal and procedural action that accompanies such events.

- The tight schedule had caused problems of procurement, and among numerous other concessions, the substitution of an inferior class of steel had to be accepted. Instead of the specified quality which would have had a favourable behaviour at cold temperature,

- A steel with a transition temperature well above the winter weather was used for the guys. With temperatures to be expected in the low -30’s, this turned out to be a serious problem. Forensic investigation, following the failure, turned up the fact that the steel for the bridge had been procured from 24 different sources, all with impeccable certificates...

- The guys were made from 3\(\frac{1}{2}\) inch (89 mm) diameter solid rods, each guy consisting of four units arranged in blocks of four in section, with a spacing of 8” (200 mm). (The use of solid rods for this purpose has recently been discouraged or prohibited by some authorities, presumably due to an accumulation of bad experience).

- The mode of attachment at both ends of the rods did not permit rotation without secondary bending moments.

- The damping associated with vibration of the guys was nearly nonexistent, and no tying of the four rods had been provided because no reason had been identified to do this.

In a cold winter night (-20° or so) a blizzard with winds of moderate speed blowing principally across the bridge apparently caused the leeward members of each guy to vibrate vertically (wake buffeting). Because of the very low damping, the movement in and out of the wind shade of the windward rods, was excited to a damaging amplitude for a number of cycles of the order of 20000 during this particular storm. This led to low cycle fatigue failure of 4 rods in 4 different places, at the anchorage points, due to the secondary moments. The bridge did not collapse however, since the three units remaining at each guy provided sufficient reserve resistance. The cumulative effect of several errors caused this accident. Its consequences were limited, thanks to robustness which had been provided through redundancy (four rods for each guy).
4 Conclusion

Human shortcomings will always be with us and the research into the circumstances that bring it about and the ways and means how it can be counteracted is an exercise aimed at minimizing injury, loss of life or value and the cost of “making good”.

One of the ways to deal with the effects of human error in the construction industry is to try and mitigate these by designing structures to be robust. Robustness can loosely be described as the ability of structures to withstand the unforeseen without loosing their function where that function may change with the unforeseen event, to be reduced to, e.g. merely standing up rather than collapse, while other functions are permitted to be lost. In this discussion, the effects of human error or shortcoming are treated as “unforeseen” events, as are natural catastrophes surpassing the limits of the hypothetically expected. This may be justified by the fact that the distinction of what exactly was the human ingredient in an unfortunate outcome is sometimes difficult.

In any case, the effect of unforeseen circumstances is no different whether caused by human agents or not, or to which degree, and to deal with them usefully through structural design measures, no distinction is needed. Robustness is a strategy which has always been useful for this.

References

Optimising cargo loading and ship scheduling subject to
uncertain sea levels

N. Le Carrer,* S. Ferson, and P. L. Green

Institute for Risk and Uncertainty, Department of Engineering, University of Liverpool, United Kingdom
*Corresponding author: nlc[at]liverpool.ac.uk

Abstract

Until now, works in the field of tide routing (i.e. optimisation of cargo loading and ship scheduling decisions in tidal ports and shallow seas) have omitted the uncertainty of sea level predictions. However, from the widely used harmonic tide forecasts to the more performant hydrodynamic models, sea level predictions are not perfectly reliable. Consequences for the maritime industry are significant: current solutions to tide routing may be robust through the introduction of arbitrary slack but they are not optimal. Given the financial implications at stake for every additional centimetre of draft and the dramatic effects of a grounding, an investigation of tide routing from the perspective of risk analysis seems necessary.

Considering the journey of a bulk carrier between two ports, a shipping decision model is designed to compute optimal cargo loading and scheduling decisions, given the time series of the sea level point forecasts in these ports. The objective function is chosen from well-know risk-averse models. In a realistic case study between two British ports with 54-hour-ahead sea level predictions, four risk metrics are compared, as well as the effect of the underlying sea level stochastic representation. Results show the economic added-value of adopting a risk perspective in cargo loading and ship scheduling optimisation.

Keywords: Operational research in maritime industry; Robust optimisation; Decision-making under uncertainty

1 Introduction and literature review

1.1 Ship scheduling in tidal areas

A ship’s draft is the distance between the waterline and the bottom of the hull. It is a fundamental characteristic of a ship and forms a major constraint in terms of scheduling or cargo loading decisions because a poor choice can lead to grounding in tidal areas or shallow waters. Yet the research on ship loading has mostly focused on operations safety and logistic aspects (see for instance a review in ? ). The question of scheduling with time-varying draft was not tackled until recently, when Kelareva and colleagues developed a deterministic procedure to optimise ship scheduling and cargo loading decisions of multiple vessels at a single port [? ? ]. Their procedure is based on the short-term predictions of under-keel clearance provided by the DUKC® software (OMC International, 1993, described in ? ]. Similar operational tools have been developed since then, e.g. the MetOceanView solution, by the commercial brand of the New Zealand Meteorological Service). Specifically, from real-time environmental measurements, the physical responses to the ship moving in a dynamic environment (squat, heel and wave) are quantified in terms of their effect on the under-keel clearance. Recent and actual water depth measurements in navigation channels are integrated to tide predictions. All these factors are combined to determine the optimal cargo loading and ship scheduling decisions given estimations of the under-keel clearance. In such a deterministic optimisation approach, safety margins have to be introduced as the under-keel clearance is only estimated a priori. One can ask whether taking into account the stochastic nature of sea levels (and, consequently, the under-keel clearance) could reduce this safety margin to some theoretical minimum - this is one of the aspects investigated in the current paper.

The work of ? ] was extended to a shipping cost optimisation problem for a fleet considering time-varying draft restrictions at waypoints, variable ship speed and cargo loads as well as flow control through busy waterways [? ]. The specific waterway ship scheduling problem was later formulated by ? ], who integrated tide as a constraint in their approach to optimally schedule the flow of incoming and outgoing ships through different shipping channels (so that the waiting times were globally minimised).

Similarly, researchers focusing on the berth allocation problem, which aims at scheduling berth and crane allocation to optimise port throughput, introduced tide as a constraint only quite recently. While
early works [? ? ] were more concerned with the quantification of the economic impact of tides on port operations, recent studies developed practical models and solutions for berth scheduling optimisation [? ? ] or quay crane allocation [? ] in tidal ports.

1.2 Shipping optimisation in stochastic environments

Maritime transportation is an activity particularly subject to risk, i.e. the possibility of a loss. From the weather at sea to port variables (berth availability, loading/unloading works), including the volatility of bunker fuel prices, a range of uncertain factors condition the outputs of a shipping operation. In spite of its significant impacts on shipping productivity, the issue of uncertainty has remained marginal in the research on maritime transportation until recently. Indeed, as stressed by [? ], due to the complexity and untractability of some shipping problems, authors often introduce simplifications (constant speed, single cargo type, basic weather model, etc.) that are different from one study to another, making comparison difficult. The introduction of stochasticity is often limited to the modelling of a single or a very limited number of factors (e.g. weather [? ], market demand [? ], weather and berth occupation [? ]).

Water depth is also a significant uncertain factor. Although tide forecasts used to predict the water depths in shallow seas are traditionally given by harmonic analysis from past observations, a range of causes can modulate the observed water levels. These encompass weather influence, river discharge, the interaction between currents, shallow water seabed and ship traffic [? ] and lead to significant deviations between astronomical tides and actual water level observations (called residuals hereafter: the difference between observations and predictions). [? ] estimate that the root mean square error on the high tide predictions in UK tide stations is typically 10 cm and rises to 29 cm for high tidal range stations. [? ] note that sea level residuals can amount to 30% of the total measured sea level in Hillarys Boat Harbour, Western Australia.

Uncertainty about future water depths has considerable impacts on shipping optimisation. First, as shown in the case study presented in Section ??, even for a small-sized carrier of horizontal dimensions 85 m ×15 m, one additional centimetre of under-keel clearance can be turned into an extra freight of about 13 metric tons (mt) whose value ranges from US$ 2,500 for a single hold of malting barley1 to more than US$ 223,000 for a single hold of tin2 with little increase in operational costs in short journeys3. Secondly, when it costs thousands of dollars a day to operate the same vessel, missing a tide-window because of a negative anomaly in the water depth is significantly costly to the shipper, to say nothing about the cost of grounding and its potential environmental consequences.

1.3 Robustness in shipping optimisation

In all the approaches mentioned in Section ??, water depths are considered as perfectly predictable variables. Although [? ] introduced a conservative 15-minute departure window for each departure/arrival decision, the authors justified the slack as a way to take into account the inertia of large ships in port operations rather than to account for sea level uncertainties. Unfortunately, as actual water levels are often different from predicted, such a deterministic assumption is either not robust or requires the introduction of safety margins to make it robust, at the cost of optimality [? ]. Large operational costs of ships tend to prevent the shippers from adding significant slack in their schedule [? ], a ship being productive only when it is sailing.

An original approach to robustness in ship routing and scheduling is found in [? ], who introduced the concept of “risky arrival”. A penalty cost proportional to the risk of a given schedule is integrated when optimising the transportation cost of a fleet. The work of [? ] should also be mentioned as it questions the applicability of mathematical optimisation in a real port context. The authors especially highlight the situation where a small change in the model inputs leads to a radically different optimal solution. The concept of persistence is introduced as a new feature of the optimisation model, so that small changes in the input values do not drastically change the nature of the optimal solution.

1.4 Objective and contribution

The present work aims at filling a gap in the field of tide routing by considering the uncertainty associated with water depths. A robust analysis of cargo loading and ship scheduling decisions in tidal areas is drawn through a realistic case study. The question at hand is: how can we optimise the cargo loading and ship

---

1Agriculture & Horticulture Development Board, UK Prices, January 2018.
3To paraphrase [? ]: if the fuel consumption of the empty ship is 20% less than that of the laden ship [? ] and if the increase in fuel consumption is linear in draft difference, then if the vessel at hand shows a difference of 2.47m between laden and empty draft, 1 cm of extra draft equals to 0.08% of fuel consumption increase. For short sea shipping journeys and small extra load draft, this can be neglected.
scheduling decisions, given imperfect sea level forecasts, without foregoing safety? To this purpose, a simple 1 ship - 1 leg maritime shipping decision model is introduced. The model assumes that an industrial operator has the sea level forecasts at two ports at a given time $t_0$, over a prediction horizon $T$. On this basis, the operator has to decide the amount of a given commodity to convey from one port to the other and when to depart. This model does not address the uncertainty associated with the under-keel clearance arising from dynamical responses to the sea state (heeling, heaving, squat effect), nor from imperfect or changing bathymetry. We limit our point to the uncertainty about still water levels, resulting from deviations to the tide predictions. The dynamical sources of uncertainty could, however, be integrated into a similar approach in order to address the open water problem.

In the following, our model allows us to demonstrate the economic potential of robust under-keel clearance optimisation. Beyond the application to industrial shipping, for which the bulk cargo load is quite flexible, this work wants to raise awareness of the economic potential for small vessels (mini-bulkers), cheap commodities (grains) and small ports strongly affected by tidal effects (i.e. limited dredging). In the current context of transportation greening, we expect this to be an important area for future applications.

Section 2 has introduced the motivations for the investigation of robust cargo loading and scheduling optimisation in tidal areas and outlined the state of the art around this issue. Section 3 introduces the case study and sets up the economic shipping decision model. In Section 4, the uncertainty on port sea level forecasts is discussed and a selection of four risk metrics, providing so-called robust alternatives to the deterministic decision-making process are presented. The implementation setting is summarised and the statistical modelling of sea level residuals is introduced. Section 5 discusses the results of these approaches and assesses their distributional robustness over a range of sea level models. Finally, the findings are summarised in Section 6 and perspectives are opened.

2 Shipping decision model

2.1 Case study

To illustrate the approach in this paper, a case study is presented. This gives the reader context for the model development that is detailed later. We consider a farm cooperative that owns a small-size bulk carrier of design draft 5.2 m, dimensions 85 m × 15 m and carrying capacity 5,170 mt. The company uses it to carry various farm commodities between ports along the United Kingdom coast. The malting barley produced in the fields of Eastern England is much desired by breweries further south in the UK. Consequently a single hold of malting barley is regularly shipped from Lowestoft to Portsmouth.

Lowestoft inner harbour offers quays with a maintained depth of at least 5.7 m regardless of the tide \(^8\). This allows the bulk carrier to arrive and berth at any time, provided it is not fully laden. However, in order to facilitate sustainable transport connections, the arrival and unloading take place at the Harbour Railway Jetty of Portsmouth port. The latter is a limiting factor for time scheduling and cargo loading because nautical charts indicate a depth of 2.3 m below Chart Datum \(^7\). Given the range of tides in the Portsmouth site, this means that the water depth varies between about 2.4 and 7.6 m which restricts the access of the vessel to the jetty, even with empty tanks. Given the vessel size and a freight unit value of US$ 195.61 per metric ton\(^4\), 1 cm of additional draft equals an extra freight of 13.05 mt which conveys an extra profit of US$ 2,556. As described before, although a heavier ship will consume more fuel, for small vessels and short sea voyages, it remains much more profitable for the company to increase the overall cargo loading if possible.

On November 19th 2016 at 16:30 UTC, we assume that the cooperative has to decide how much barley will be freighted and when the vessel will depart. To this purpose, they use the long term harmonic tide forecasts as sea level predictions. Hence the problem at hand: given the economic, vessel and port parameters summarised in Table 7, given the tide predictions in the tide gauge stations of Portsmouth Harbour and Lowestoft Port\(^5\):

1. What is the optimal decision in terms of cargo load and departure time, if the harmonic tide forecasts were considered as perfect?

2. Is this decision robust to actual port and sea level conditions?

3. What is an optimal and robust shipping decision if the uncertainty on tide forecasts is taken into account?

4. What shipping benefit can be guaranteed, given an error level, from the robust solution?

\(^{8}\)Agriculture & Horticulture Development Board, UK Prices, January 2018

\(^{5}\)The tide predictions are relative to the Harbour Master’s Office gauge station in Lowestoft and to the Victory Jetty gauge station in Portsmouth. Both are maintained and provided by the UK Environmental Agency.
Table 1: Model parameters.

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Lowestoft</th>
<th>Portsmouth</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journey</td>
<td>Decision time</td>
<td>19-Nov-2016, 16:30:00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journey</td>
<td>Mean distance between departure and arrival ports</td>
<td>195</td>
<td></td>
<td>Nautical miles</td>
</tr>
<tr>
<td></td>
<td>Sea water mean density</td>
<td>1.025</td>
<td></td>
<td>Kilogram per cubic meter</td>
</tr>
<tr>
<td>Ship design</td>
<td>Mean operational sailing speed</td>
<td>13</td>
<td>Knot</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ship horizontal surface</td>
<td>15 \times 85</td>
<td></td>
<td>Meter\times Meter</td>
</tr>
<tr>
<td>Ship design</td>
<td>Minimum cargo load (ballast)</td>
<td>1,870</td>
<td></td>
<td>Metric ton</td>
</tr>
<tr>
<td>Ship design</td>
<td>Deadweight tonnage (carrying capacity)</td>
<td>5,170</td>
<td></td>
<td>Metric ton</td>
</tr>
<tr>
<td>Ship design</td>
<td>Half-laden ship draft</td>
<td>5.2</td>
<td></td>
<td>Meter</td>
</tr>
<tr>
<td>Ship design</td>
<td>Fuel consumption rate of the laden ship at sea</td>
<td>8</td>
<td>Ton per day</td>
<td></td>
</tr>
<tr>
<td>Ship design</td>
<td>Fuel consumption rate of the ship at port</td>
<td>1</td>
<td>Ton per day</td>
<td></td>
</tr>
<tr>
<td>Monetary</td>
<td>Fuel cost</td>
<td>387</td>
<td></td>
<td>US$ per ton</td>
</tr>
<tr>
<td>Monetary</td>
<td>Other operational costs (staff, maintenance)</td>
<td>2,500</td>
<td></td>
<td>US$ per day</td>
</tr>
<tr>
<td>Monetary</td>
<td>Average bulk cargo value</td>
<td>195.6</td>
<td></td>
<td>US$ per ton</td>
</tr>
<tr>
<td>Monetary</td>
<td>Berthing and loading/unloading operation cost within normal opening times</td>
<td>1,486</td>
<td>1,239</td>
<td>US$ per hour</td>
</tr>
<tr>
<td>Monetary</td>
<td>Berthing and loading/unloading operation cost outside of normal opening times</td>
<td>1,858</td>
<td>1,548</td>
<td>US$ per hour</td>
</tr>
<tr>
<td>Monetary</td>
<td>Daily port fee</td>
<td>1,363</td>
<td>1,115</td>
<td>US$ per day</td>
</tr>
<tr>
<td>Port</td>
<td>Bulk material (un)loading rate</td>
<td>1,000</td>
<td>1,200</td>
<td>Ton per hour</td>
</tr>
<tr>
<td>Port</td>
<td>Normal port opening time</td>
<td>[7:00, 19:00]</td>
<td>[7:00, 19:00]</td>
<td>-</td>
</tr>
<tr>
<td>Port</td>
<td>Minimum allowed under-keel clearance to navigate in port still waters</td>
<td>10% static draft</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Forecast</td>
<td>Sea level forecast time step</td>
<td>15</td>
<td></td>
<td>Minute</td>
</tr>
<tr>
<td>Forecast</td>
<td>Horizon of the sea level predictions</td>
<td>54</td>
<td></td>
<td>Hour</td>
</tr>
</tbody>
</table>

5. How do the robust solution and guaranteed benefit depend on the residuals modelling and risk metric under consideration?

2.2 Model overview

The model used here is a simplified low-dimensional representation of the maritime inventory routing problem. A material is produced at a given rate in a production site (called loading port) and consumed at other sites (called unloading ports) at specified rates. Given storage capacities in the production and consumption places, what is the optimal design of routes and fleet schedule to minimise the shipping costs (sailing and port costs) without interrupting any of the production or the consumption in the aforementioned sites? The optimisation is made on an industrial shipping basis. In other words, the shipper owns the material to be shipped and wants to maximise the net benefit of the shipment (the value of the cargo loaded minus the shipping costs). The fleet consists of a single bulk carrier or general cargo ship and the study is restricted to a leg between the loading (departure, \( p_1 \)) and unloading (arrival, \( p_2 \)) ports. A constant ship speed is assumed (provided by the ship specifications). From this, the goal is to optimise the decision vector \( d \), consisting of the departure time \( t_d \) and the cargo \( m \) to be loaded given the sea level predictions available at time \( t_0 \), spanning the horizon \( T \) in the entrance channels of both ports and given constraints from the ship design (carrying capacity), safety at sea (minimum acceptable water under keel), port management (opening times and price bands for port labour). For now, unlimited
storage capacities in both ports are assumed. The question of rate of production in the departure port (i.e. offer) and rate of consumption in the arrival port (i.e. demand) is not taken into account.

In this simplified problem, the ship is assumed to be in the departure port at time $t_0$ with empty tanks. The most recent predictions $\hat{X}_p(t)$ for the sea levels in both ports $p = \{p_1, p_2\}$ over the horizon $T$ are available. Time is discretised with the time step $\Delta t$ (following the precision in the sea level prediction and observation time series). Here and in the following, in order to simplify the notations, $t_d$ will be relative to the origin of our time axis, $t_0$.

### 2.3 Model description

The model takes time series of sea level point-forecasts in both departure and arrival ports as inputs. Given contextual parameters regarding the journey, including ship characteristics, freight and port management, generic constraints about acceptable under-keel clearance, latest arrival time and cargo load and finally the net return computation rule for a journey, it computes the optimal cargo loading and departure time by means of a particle swarm optimisation (PSO) solver.

#### 2.3.1 Journey parameters

Table ?? defines the model’s input parameters. A few comments and justifications are provided here.

The operational speed is assumed to be fixed and constant over the journey (as it is often the case in maritime shipping models). Operational port costs are subject to price bands. Although most often docks and loading / unloading operations are accessible 24 hours a day 7 days a week, the cost of such operations depends on the local port schedule. For example, midweek vs weekend periods for Liverpool port are shown in [? ? ]. The simple price band framework allows us to simulate a range of situations: nights vs days, week days vs weekends, bank holidays. Finally, the safety margin coefficient, $\alpha$, in terms of legally required under-keel clearance to use the confined navigation channel of port $p$, is set to 10% of the laden ship draft as this is usual practice at limited speeds [? ? ]. The open sea version would require adding a 30% margin to the dynamical draft.

#### 2.3.2 Sea level input variables

The sea level point predictions in each port are harmonic tide forecasts, available online through the British Oceanographic Data Center portal. The time step, $\Delta t = 15$ minutes, sets a minimum bound on the resolution of our departure time solution.

#### 2.3.3 Model variables

The ship draft, a key element in shipping planning and realisation, is a function of the cargo load as well as the fuel mass in the tanks at the time of interest. The fuel mass is estimated from the fuel consumption rates at sea and at port, the time already spent at sea and at port respectively, as well as the total fuel load necessary to move the ship from one port to another and (un)load material. Considering Archimedes’ principle and the equilibrium of forces in a gravitational field, the draft can be estimated from the equality between ship’s weight and water displacement. The latter is a function of the half laden ship’s draft, the ship’s horizontal area, its carrying capacity and the water density. Dynamical effects such as the squat effect or the heel due to the wind and the wave responses can reduce the under-keel clearance temporarily. They are not taken into account here beyond the safety margins because, as stated previously, we consider the still water problem.

#### 2.3.4 Constraints

The ship’s cargo and scheduling have to satisfy a few constraints. First, the cargo load $m$ cannot exceed the tank capacity and must fit with the requirements for safe structural behaviour of the hull (i.e. equal minimum ballast). $m$ includes the fuel load necessary to carry the laden ship on the mean distance between the two ports at specified speed and load/unload the freight at specified rates in each port. Second, to enter/leave a port at a given time, the water depth must be greater than the ship draft plus the safety margin. Third, the ship cannot leave port $p_1$ before the cargo is loaded and must arrive in $p_2$ before the horizon $T$ is reached. Finally the time of arrival (i.e. berthing at port $p_2$) must not exceed the prediction horizon $T$. 

---

REC 2018 – N. Le Carrer, S. Ferson, P. L. Green
2.3.5 Shipping return

The problem is to find the optimal combination of decisions \( d^* = (t^*_d, m^*) \) that maximises the net benefit \( B \) (hereafter called objective function or objective), where:

\[
B(t_d, m; \hat{X}_{p_1}(t), \hat{X}_{p_2}(t)) = \begin{cases} Z & \text{if delivered on time,} \\ V - (O + P + U) & \text{otherwise.} \end{cases}
\]

The gross value \( V \) is the merchant value of the cargo \( m \) unloaded in \( p_2 \). It depends on the unit value of the freight. From there, we subtract the operational costs of the journey, starting from \( t_0 \) (time of decision) with an empty ship and finishing after unloading the material in port \( p_2 \). These charges encompass the propulsion costs \( O \), computed as a function of the total time spent at sea and at port respectively, the fuel unit price and the fuel consumption rates at sea and at port respectively. Operational charges also include usage costs \( U \), depending on the total time of the journey, loading and unloading steps included, as well as on the hourly usage cost (staff) of the ship. Finally, port costs \( P \) have to be included. They are calculated on the basis of the daily port fee, hourly handling prices in normal hours and outside normal hours in each port as well as the time spent in each port within and outside normal hours respectively.

\( Z \) is the cost of not making the delivery in time (i.e within the horizon \( T \)). Depending on the aim of the user, \( Z \) can also take into account the negative externalities on the environment and society of a grounding \((Z \to -\infty)\) or simply the loss for the shipper \((Z = -V - (O + P + U))\).

3 A probabilistic approach to decision making

Using the model described above, one can choose an optimisation technique (e.g. particle swarm optimisation or simulated annealing) to compute the optimal decision to take at time \( t_0 \), according to the sea level forecast time series \( \hat{X}_p(t) \) for the two ports \( p = \{p_1, p_2\} \). Such a calculation does not consider the actual stochastic behaviour of the water depth. Mean sea levels are locally influenced by a range of factors, including weather. A residual \( e_p(t) = X_p(t) - \hat{X}_p(t) \) between the predictions and the observations can lead to either a regret \((e_p > 0): \text{the shipper could have loaded more or departed earlier}) or a loss \((e_p < 0): \text{in order to adjust to the actual water level the journey is delayed, or a grounding can happen})\.

In other words, the resulting solution is risky as it does not tolerate a negative deviation to prediction nor port delays. In order to account for the uncertainty on the outcome of a given decision and its potentially dramatic consequences for the shipping company, it is sensible to work in the frame of risk averse optimisation.

From the classical mean-risk [? ] and chance-constrained perspectives [? ] to the more recent so-called robust optimisation models (e.g. worst-case, minimax regret, uncertainty sets, see [? ] for an historical overview and [? ] for an extensive presentation), operational research has developed a range of approaches to address the notion of uncertain decision-making. In these problems, the questions at stake are: a) Are all the scenarios acceptable, or feasible, whatever their probability of occurrence? (e.g. is a ship grounding acceptable?) b) How much does the decision-maker give way to objective optimality in order to guarantee feasibility? Any solution to stochastic optimisation is a trade-off between feasibility and performance, or said otherwise, between variance and guaranteed value of the objective function.

A (robust) optimisation approach must thus define the attitude of the decision-maker towards risk and the specificities of her optimisation problem before computing any solution. Between the two extreme approaches that are worst-case (always feasible) and deterministic optimisation (best performance e.g. for the most probable scenario, no uncertainty taken into account), lie a range of models depending on the decision-maker’s requests as regards performance and feasibility. We introduce in the following a representative selection of them, before comparing their outputs in Section ??.

3.1 Risk models

3.1.1 Regret

In decision-making under uncertainty, it is common to adopt the gain shortfall perspective. In this case, risk takes the meaning of the loss in profit due to the fact that decision \( d \in D \) is taken at time \( t_0 \) based on imperfect forecasts \( \hat{X}_p \in X \) of the environment state \( X_p \in X \). Let \( F_p \) be the cumulative distribution function over \( X_p \), which is conditional on information on the prior values of \( X_p \) and possible other information. Let \( \hat{F}_p \) be a predictive distribution of \( X_p \) (that is a distribution over \( \hat{X}_p \)) provided by the forecaster at \( t_0 \). Let \( \hat{X}_p(t) \) be a point forecast time series of \( X_p(t) \) over time \([t_0, t_0 + T]\), \( B(\cdot, \cdot) : D \times X \to \mathbb{R} \) the utility function (namely the net benefit of the journey based on decision \( d \)) and \( y(\cdot) : X \to D \) an

REC 2018 – N. Le Carrer, S. Ferson, P. L. Green
optimal action function defined by:

$$y(\hat{F}_p) = \arg \max_{d \in D} \left( \mathbb{E}[B(d, \hat{X}_p)]_{\hat{F}_p} \right) = \arg \max_{d \in D} \int_{\mathcal{X}} B(d, \hat{X}_p) d\hat{F}_p$$

(2)

The loss function $L(.,..) : \mathcal{D} \times [0, 1] \to \mathbb{R}$ is then defined by as:

$$L\left(y(\hat{F}_p), F_p\right) = B\left(y(X_p), X_p\right) - B\left(y(\hat{F}_p), X_p\right)$$

(3)

for all $\hat{X}_p, X_p \in \mathcal{X}$. In other words, the utility of the decision made under uncertainty $B\left(y(\hat{F}_p), X_p\right)$ is compared to the utility resulting from the decision made under perfect knowledge of the future $B\left(y(X_p), X_p\right)$.

With an absolute robust approach, each possible shipping decision $d$ is mapped to the maximum loss it can generate, whatever its probability of occurrence. The optimal decision minimises:

$$d^* = \min_{d \in \mathcal{D}} \left\{ \max_{F_p} \left\{ L\left(d, X_p\right) \right\} \right\}$$

(4)

Its less conservative counterpart involves mapping each decision $d$ to the regret it generates in average:

$$d^* = \min_{d \in \mathcal{D}} \left\{ \mathbb{E} \left[ L\left(d, X_p\right) \right]_{F_p} \right\}$$

(5)

Looking more closely at the definition of the loss which we aim to minimise (the expectation over the space of sea level residuals), one can notice that minimising $\mathbb{E} \left[ L\left(d, X_p\right) \right]_{F_p}$ is equivalent to finding the decision $d^*$ that maximises the expected benefit $\mathbb{E} \left[ B\left(d, X_p\right) \right]_{F_p}$.

### 3.1.2 Mean-risk

What appears to be the first risk model developed in operational research involves adding a penalty known as the risk functional to the expected objective outcome of a given decision, and thus setting:

$$d^* = \max_{d \in \mathcal{D}} \left\{ \mathbb{E} \left[ B\left(d, X_p\right) \right]_{F_p} - \beta \mathbb{D}[B]_{F_p} \right\}$$

(6)

where the parameter $\beta \geq 0$ allows to quantify the price of risk.

In the simplest case, the risk functional is proportional to the standard deviation of the objective:

$$\mathbb{D}[B] = \left( \mathbb{E} \left[ (B - \mathbb{E}[B])^2 \right]_{F_p} \right)^{1/2}$$

(7)

Negative and positive deviations to the mean do not have the same implications in terms of risk. In the case of maximising the shipping benefit, positive deviations to the expected benefit are welcome, contrary to negative ones. The standard deviation cannot fully describe such asymmetrical behaviour of the utility function. The lower semi-deviation of order $\gamma$ is consequently introduced as:

$$\mathbb{D}[B] = \left( \mathbb{E} \left[ (B - \mathbb{E}[B])^\gamma \right]_{F_p} \right)^{1/\gamma}$$

(8)

Note that, in the following, we use $\gamma = 2$ and $\beta = 1$.

### 3.1.3 Worst-case

The absolute robust way of optimising the shipping net benefit is to prevent any unfeasible scenario and maximise the outcome in the worst possible scenario. In other words, finding the decision:

$$d^* = \max_{d \in \mathcal{D}} \left\{ \min_{F_p} \left\{ B\left(d, X_p\right) \right\} \right\}$$

(9)
3.1.4 Chance-constrained

Although strictly speaking robust in terms of feasibility, the worst-case approach is often criticised for being too conservative in practical implementations.

The chance-constrained perspective allows more flexibility. Given a level of guarantee $\zeta$, it computes the decision maximising the ensured benefit at this level, in other words:

$$d^* = \max_{d \in D} \left\{ \inf_b \left\{ P \left( B \left( d, X_p \right) \leq b \right) \leq 1 - \zeta \right\} \right\}$$

(10)

In our experiments, we use $\zeta = 0.99$, that is we look for the maximal benefit allowing an error rate less than or equal to 1%.

3.2 Implementation

The problem of deterministic shipping optimisation was defined in Section ??.. It consists of finding the decision $d^* = \left( t^*_d, m^* \right)$ maximising the net benefit of the shipping given sea level forecasts $\hat{X}_p, p = \{p_1, p_2\}$. Similarly, the risk minimisation problem consists of finding the decision maximising one of the objective functions (or risk functions) defined in Section ??.

Both are constrained 2-dimensional optimisation tasks whose objective functions are not continuous nor differentiable. As a result, classical analytical optimisation techniques cannot be used. Hence the call to derivative-free algorithms such as particle swarm optimisation to estimate $d^*$.

A range of other computational methods could have been implemented as well. However, PSO was chosen because it generally demonstrates good convergence and execution speed properties in addition to its simplicity of implementation. A review and comparison of the derivative-free approaches is provided in ??]. PSO is an iterative stochastic optimisation technique that imitates the natural swarm behaviour of a bird flock ??]. At each iteration, the elements (particles) of the flock explore the search space in a semi-random way and evaluate the fitness (value of the function to optimise) of their positions. They share the information so that their next move is influenced by both their own findings and the findings of the other members of the swarm. The algorithm stops when the desired number of iterations is reached and the position with optimum fitness is returned. Algorithm ?? describes the procedure and our implementation choices.

Because the risk functions defined in Section ?? cannot be written in closed forms due to the definition of the net benefit $B$, it is natural to turn to Monte Carlo simulations to estimate them, within the PSO procedure. Algorithm ?? shows the general approach, now referred to as $R_{PSO}$. $B_{PSO}$ refers to the “deterministic” optimisation of the shipping benefit (by means of Algorithm ??), that is without taking into account any uncertainty on the sea level forecasts (although technically PSO is a stochastic technique). Hereafter we name nominal state the forecasted sea-level state.

3.3 Distributional robustness

Mathematically, the risk functions defined above are based on the distribution of the economic output of a given decision when the sea levels in both departure and arrival ports vary around their nominal state. Such a definition implies that the risk evaluation is model-dependent: its accuracy depends on the quality of the modelling of sea level residual distributions. In this section, we summarise the results of an analysis of the residuals distribution in Portsmouth and Lowestoft ports.

The dataset used for the modelling and then the testing consists of sea level residuals sampled every 15 minutes between 00:15-01/01/2006 and 23:45-31/12/2016 UTC, in each port. We split it into two parts: even years (dataset $D_e$) and uneven years ($D_u$). The former is used for modelling the sea level residuals by means of best-fit distributions. It is then used for the shipping optimisation procedure per se. Finally, $D_u$ is used as validation set, to perform simulations and gather statistics on the performance of the optimisation outputs.

Three distributions are tested on $D_e$: Gaussian, Logistic and Gaussian mixture model (GMM). The number of components in the Gaussian mixture models (3 for Portsmouth, 5 for Lowestoft) were chosen so as to minimise the Akaike Information Criterion [??]. This criterion assesses the “goodness of fit” of a model to a dataset while introducing a penalty that increases with the number of free parameters requiring estimation. The aim is to find the optimal trade-off between model complexity and loss of information.

Kolmogorov-Smirnov statistics were computed to quantify the “goodness-of-fit” of each model. They reject at the 1% significance level the null hypothesis that the residuals follow a Gaussian or Logistic distribution for both ports. A graphical analysis of the three models shows that the Gaussian distribution significantly under-represents the small deviations of sea level observations with respect to tide.
Algorithm 1 Particle Swarm Optimisation procedure

1. Initialise randomly the position \( \mathbf{d}_i \) of each particle \( i \) in the search space \( \mathcal{D} \) and set their initial velocity vector to \( \mathbf{0} \).
2. For each step \( j \):
   (a) For each particle \( i \):
      i. Compute the objective function \( f(\mathbf{d}_i) \) (e.g. the net benefit \( B(\mathbf{d}_i, X_{p_1}(t), X_{p_2}(t)) \) resulting from the shipping decision \( \mathbf{d}_i \) with actual sea levels \( X_p(t), p = \{p_1, p_2\} \)). This is the fitness of position \( \mathbf{d}_i \).
      ii. Update the personal best \( \mathbf{b}_i(j) \) of each particle i.e its position with optimal fitness among the set of previous iterations. Similarly, identify and update the global best \( \mathbf{g}(j) \), that is the best solution among the positions visited by the whole swarm so far.
      iii. Move each particle according to the following equation of motion:
\[
\mathbf{d}_i(j + 1) = \mathbf{d}_i(j) + \mathbf{v}_i(j + 1),
\]
   where the velocity is defined by:
\[
\mathbf{v}_i(j + 1) = \omega(j + 1)\mathbf{v}_i(j) + c_1 R_1(\mathbf{b}_i(j) - \mathbf{x}_i(j)) + c_2 R_2(\mathbf{g}(j) - \mathbf{x}_i(j)).
\]
The cognitive \( c_1 \) and social \( c_2 \) coefficients are set up so as to optimise the ratio between individual exploitation and social interaction while the linearly decreasing inertia weight \( \omega(j) \) limits ‘velocity explosion’. The diagonal matrices \( R_1 \) and \( R_2 \) introduce stochasticity in the walk of the particles. Their diagonal elements are generated from a Lozi 2-dimensional chaotic map [?].
3. Stop when the desired number of steps is reached.

Algorithm 2 Procedure \( R_{PSO} \)

1. Initialise randomly the position \( \mathbf{d}_i \) of each particle \( i \) in the search space \( \mathcal{D} \) and set their initial velocity vector to \( \mathbf{0} \).
2. For each step \( j \):
   (a) For each particle \( i \):
      i. Sample time series of tide residuals \( \epsilon_p(t) \) in ports \( p = \{p_1, p_2\} \) from a distribution model. In this first study, we assume the residuals to be independent from port to port at a given time, and between any two given times at a given port. The spatial independence is not a strong assumption as long as ports are not too close. On the other side, the time independence can be discussed as on short durations the residuals show correlation.
      ii. Compute the net benefit \( B(\mathbf{d}_i, X_{p_1}(t), X_{p_2}(t)) \) for the simulated sea level conditions. These are given by the nominal state modified by the tide residuals, namely at port \( p \): \( X_p(t) = \hat{X}_p(t) + \epsilon_p(t) \).
      iii. Repeat steps ?? to ?? until the number \( N_s \) of simulated environments requested to compute the empirical risk function is reached.
      iv. Estimate the latter from the \( N_s \) outputs of step ??, e.g. for the chance-constrained approach at error level \( 1 - \zeta \), find the quantile of level \( 1 - \zeta \) (that is the cutoff benefit \( b \) such as \( P(B < b) = 1 - \zeta \) i.e. \( P(B > b) = \zeta \)).
   (b) Move particles according to the general PSO procedure described in Algorithm ??, in the search space, step by step. Here the objective function to be maximised are the risk functions computed in step ??, e.g. the cutoff benefit \( b \).
   (c) Stop when the desired number of steps is reached and return the position with optimal fitness.
predictions. Hence the introduction of the Gaussian mixture model, that globally represents the original residual distribution with greater fidelity. Besides, the GMM is able to capture the long tails that the single Gaussian or Logistic cannot. This could be important, as extreme events are usually in the tails.

This analysis will first be used to assess the distributional robustness of the optimisation procedures in Section ??, by analysing the effect of the residual modelling on the optimisation results. Besides, on a standard desktop computer running Linux, sampling from a Logistic distribution is about 10 times quicker than from a Gaussian distribution and 15 times quicker than from a 5 component mixture distribution. In the following sections we check whether the difference in risk outputs and its implication in real-world decision making justify the added complexity of the GMM input model.

4 Results & Discussion

All the results in terms of benefit $B$ will be expressed as multiples of the value of the minimum cargo load, $B_0 = \text{US$363,550}$. We also set the cost of not making the delivery in time to $(Z = -V - (O + P + U))$. Negative benefits would thus imply a grounding or the impossibility to reach the arrival port within the specified time horizon.

4.1 Deterministic case

The $B_{PSO}$ procedure recommends the ship to leave Lowestoft Harbour at 23:00 UTC on November 19th 2016 with an overall barley freight of 3,835.0 mt. The standard deviations of these recommendations are estimated to be 0.5 mt in freight and less than 15 minutes in time (from 1,000 independent runs).

Figure ?? presents a mapping of the final shipping benefit over the decision search space $D$, given the forecast \textit{a priori} at hand and given perfect forecasts, i.e. the \textit{a posteriori} exact observations of the sea level depths. The optimal decision according to $B_{PSO}$ in each scenario differ by 1 hour and 30 minutes in time and 527 mt in cargo load. In other words, the deterministic solution under imperfect harmonic predictions is far away from optimality in the real-world of non-zero residuals. Besides, it is quite straightforward to see on these maps that both solutions are very sensitive to perturbations. A 15 min departure/arrival shift or a negative error in the actual sea levels both shift the expected benefit from its maximum to the negative area.

One way to get over the second limitation is to improve the accuracy of sea level forecasts. This is currently achieved by means of storm surge models. To take into account the local weather perturbations, these models use atmospheric forecasts as forcing in shallow-water hydrodynamic simulations e.g. the CS3 storm surge model covering the sea of the northwest European continental shelf [? ]. Nevertheless, whatever the accuracy reached, these forecasts cannot prevent the issue of port perturbations and delays. Hence it seems reasonable to develop a robust solution instead of a single deterministic optimisation.
Table 2: Statistics over 50 runs of the outputs in terms of decision-making. The optimal cargo load $m$, departure time $t_d$ and guaranteed benefit $B_{98}$ at the level of 2% (over 100,000 simulations) are expressed in metric tons, UTC and fraction of $B_0$ respectively. The uncertainty is computed as the standard deviation of the results.

<table>
<thead>
<tr>
<th>Distribution Risk metric</th>
<th>GMM</th>
<th>Logistic</th>
<th>Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean-Regret</td>
<td>$m = 3,947 \pm 20$</td>
<td>$m = 3,940 \pm 10$</td>
<td>$m = 3,957 \pm 9$</td>
</tr>
<tr>
<td></td>
<td>$t_d = 00:45 \pm 30\text{mm}$</td>
<td>$t_d = 01:00 \pm 15\text{mm}$</td>
<td>$t_d = 00:30 \pm 15\text{mm}$</td>
</tr>
<tr>
<td></td>
<td>$B_{98} = 1.901$</td>
<td>$B_{98} = 1.894$</td>
<td>$B_{98} = 1.909$</td>
</tr>
<tr>
<td>Worst-Case</td>
<td>$m = 3,943 \pm 15$</td>
<td>$m = 3,935 \pm 10$</td>
<td>$m = 3,961 \pm 11$</td>
</tr>
<tr>
<td></td>
<td>$t_d = 00:45 \pm 15\text{mm}$</td>
<td>$t_d = 01:00 \pm 30\text{mm}$</td>
<td>$t_d = 00:45 \pm 30\text{mm}$</td>
</tr>
<tr>
<td></td>
<td>$B_{98} = 1.899$</td>
<td>$B_{98} = 1.891$</td>
<td>$B_{98} = 1.908$</td>
</tr>
<tr>
<td>Mean-Risk</td>
<td>$m = 3,946 \pm 18$</td>
<td>$m = 3,933 \pm 15$</td>
<td>$m = 3,963 \pm 16$</td>
</tr>
<tr>
<td></td>
<td>$t_d = 00:45 \pm 30\text{mm}$</td>
<td>$t_d = 00:45 \pm 15\text{mm}$</td>
<td>$t_d = 00:30 \pm 30\text{mm}$</td>
</tr>
<tr>
<td></td>
<td>$B_{98} = 1.901$</td>
<td>$B_{98} = 1.895$</td>
<td>$B_{98} = 1.905$</td>
</tr>
<tr>
<td>Chance-Constrained</td>
<td>$m = 3,959 \pm 11$</td>
<td>$m = 3,956 \pm 9$</td>
<td>$m = 3,976 \pm 6$</td>
</tr>
<tr>
<td></td>
<td>$t_d = 00:30 \pm 15\text{mm}$</td>
<td>$t_d = 00:45 \pm 15\text{mm}$</td>
<td>$t_d = 00:45 \pm 15\text{mm}$</td>
</tr>
<tr>
<td></td>
<td>$B_{98} = -2.239$</td>
<td>$B_{98} = 1.905$</td>
<td>$B_{98} = -2.253$</td>
</tr>
</tbody>
</table>

4.2 Risk models

We now use $R_{PSO}$ to compute the optimal shipping decision under uncertain sea levels. Each of the four risk metrics presented in Section ?? is combined with one of the three sea level residuals distribution models under consideration. Table ?? reports the statistical results of each combination as regards the optimal cargo load, departure time and the resulting guaranteed benefit at the error rate of 2%, that is the 2% percentile $B_{98}$. The latter is estimated from 100,000 Monte Carlo simulations. In order to prevent a methodological bias, these simulations sample the sea level by means of bootstrapping (over dataset $D_a$, c.f. Section ??).

As the purpose of the $R_{PSO}$ procedure is to support decision-making, it is necessary to analyse the consequences of the above results as regards their translation in terms of practical shipping decision. The overall majority of the computed departure times are located within a 30 mn time slot centered on 00:45. Taking into account the relative inertia of large vessels and generally slow port dynamics (from decision to subsequent actions), this range of uncertainty can be seen as a buffer to consider in the decision-making schedule. Trying to increase the precision on $t_d$ would be meaningless considering the real world context of a maritime shipping problem.

As expected, the worst-case approach is the more conservative and generally computes the lowest loads ($m \approx 3,945$ mt overall). The mean-risk model with a penalty equal to one standard lower deviation behaves very similarly. The chance-constrained approach returns the highest loads ($m \approx 3,966$ mt overall) and the mean-regret (or expected-benefit) approach is intermediary. This is a quite general observation, whatever the sampling distribution. As regards the distribution impact, Logistic sampling produces more conservative loads than the GMM approach and further again, than the Gaussian one. The difference between the maximal and minimal loads abovementioned is in the range of 30 mt, that is in our case study less than 3 centimetres of draft. This invariably leads to quite similar guaranteed benefits $B_{98}$ between the worst-case, mean-risk and mean-regret for a given distribution. On the contrary, the guaranteed benefit of the chance-constrained solution is much less stable: either maximum or minimum (and) negative. This illustrates the concept of distributional (non-) robustness: according to the sea level modelling (Logistic versus Gaussian and GMM), the solution computed by $R_{PSO}$ leads to either very satisfying outcomes overall or to a very likely failure.

Figure ?? summarises most of the information discussed above: whatever the risk metric, a Logistic sampling will produce more stable (smaller variance) outcomes than the other models. It also shows that, strictly speaking, only the mean-risk approach could be said to be distributionally robust. Indeed, the ranges of the reduction in standard deviation and in guaranteed benefit when the underlying distribution varies (2 and 0.6 % respectively) are much smaller than for the other metrics (closer to 8 and 0.9 % respectively). Considering the money at stake, even variations of 0.1% $B_{98}$ are worth a few thousand dollars, so should not be neglected. Three observations can be highlighted as well. First, in this particular case study, the stochastic optimisation based on risk metrics allows the owner to (in most of the configurations) save money as the guaranteed benefit is above the expected benefit of the deterministic
Figure 2: Performance of each optimisation approach (a risk metric combined with a sea level residuals distribution) from the perspective of the reduction of the guaranteed benefit at the error level of 2% and the standard deviation of the actual shipping benefit, with respect to the performances of the “deterministic” solution based on sea level forecasts alone. 100,000 Monte Carlo simulations are used to compute these statistics, with bootstrap sampling. The chance constrained and GMM or Gaussian sampling are not represented here as the reduction in guaranteed benefit is out of scope, reaching 200%.

decision in real conditions. Second, the spatial organisation of the points underlines a general pattern in robust optimisation: the guaranteed benefit increases at the cost of the increase in variance $\sigma^2$. Finally, as noted by [? ], the variation in actual benefit is about one order of magnitude smaller than the reduction in its standard deviation.

5 Conclusion

Figure ?? summarises some of the above considerations in a 3-dimensional view of the optimisation problem. A map of the standard deviation is estimated with bootstrap sampling for each couple $(t_d, m)$ of the search space. On top of the map, we report the decision suggested by the net benefit optimisation from sea level forecasts, perfect forecasts (i.e. perfect knowledge of the future) and by four optimisation approaches. Figure ?? gives a good overview of the set of solutions returned by all the approaches and presented above.

As the owner of the company, you could use the benefit optimisation decision that is based on the deterministic harmonic forecasts, load 3,835 mt of barley and cast off at 23:00. However the outcome of this decision, given the actual observations of sea levels is $-2.15B_0$. This is much less desirable than the benefit $2.12B_0$ that you could make if you knew the future perfectly and left Lowestoft port at 00:30 with 4,362 mt on board. Using the stochastic optimisation method developed in this paper, you could load cargo between 3,935 and 3,959 mt, raise anchor between 00:30 and 01:00 and get a net benefit from $1.89B_0$ to $1.91B_0$. If these decisions were reported in Figure ?? (mapping based on actual sea level conditions), one could notice that a port re-scheduling of up to 2 hours (earlier) or 4 hours (delay) would not substantially change the benefit, nor a variation (in standard limits) in sea level conditions. Besides, Figure ?? reminds that the variance in the actual benefit is substantially reduced for our solutions, contrary to the variance of the deterministic proposition. In other words, the approach $R_{PSO}$ proposes a robust solution. This is true for any risk metric introduced here apart from the chance-constrained, and true for any sampling distribution although a Gaussian generally leads to solutions with less predictable economic outcomes. Recalling the questions raised in the motivation of the problem (Section ??), in this case study, our stochastic approach demonstrated to be economically valuable with respect to the standard (deterministic) approach. Besides, a simple Logistic modelling of the residuals is enough to produce quality results, similar to those gained by means of a GMM.

One can note that the cargo load output $m^*$ can be turned into a safety margin $\Delta r$ to be deducted
Points of interest discussed in the text are also reported. The mapping use Monte Carlo simulations of 1,000 journeys by means of bootstrap re-sampling.

from the maximum draft that would have been allowed given the sea level tide forecasts at hand at $t_0$ (procedure $B_{PSO}$). For future works, it would be interesting to compare $\Delta r$ with what “non-stochastic” commercial softwares would suggest on a similar problem, so as to assess the quality and potential added value of our model.

Avenues of research on the problem raised in this paper include defining sounder uncertainty sets on which the risk metrics would then be applied. A finer modelling of the sea level residuals would also be judicious, exploiting the cyclic character of data.

Acknowledgements

This research is funded by the Engineering & Physical Sciences Research Council (EPSRC) and the Economic & Social Research Council (ESRC) with grant no. EP/L015927/1. EPSRC and ESRC are greatly acknowledged for their funding and support. We use data from the National Tidal and Sea Level Facility, provided by the British Oceanographic Data Centre and funded by the Environment Agency. Finally, we would like to thank Kevin Horsburgh from the British National Oceanographic Centre, for his introduction to the field of sea level observation and prediction.

References


M. Christiansen and K. Fagerholt, Robust ship scheduling with multiple time windows, Naval Research Logistics (NRL) 49, 611 (2002).


Peel Ports Group, Port charges: Port of Liverpool & Port of Manchester (2017).


A. Shapiro, D. Dentcheva, and A. Ruszczyński, Lectures on stochastic programming: modeling and theory (SIAM, 2009).


J.-y. Gotoh, M. J. Kim, and A. Lim, Robust empirical optimization is almost the same as mean-
A reinforcement learning framework for optimisation of power grid operations and maintenance

R. Rocchetta 1,* M. Compare 1,2 E. Patelli 1 and E. Zio 2,3,4,5

1 Institute for Risk and Uncertainty, Liverpool University, United Kingdom
2 Aramis s.r.l., Milano, Italy
3 Energy Department, Politecnico di Milano, Italy
4 Systems Science and the Energetic challenge, Foundation Electricite’ de France at Centrale Supelec
5 Eminent Scholar, Department of Nuclear Engineering, Kyung Hee University, Republic of Korea

*Corresponding author: roberto.rocchetta@liverpool.ac.uk

Abstract

In this work, we investigate Reinforcement Learning (RL) for managing Operation and Maintenance (O&M) of power grids equipped with Prognostic and Health Management (PHM) capabilities, which allow tracking the health state of the grid components. RL exploits this information to select optimal O&M actions on the grid components giving rise to state-action-reward trajectories maximising the expected profit. A scaled-down case study is solved for a power grid, and strengths and weaknesses of the framework are discussed.

Keywords: Reinforcement Learning, Prognostic and Health Management, Operation and Maintenance, Degradation, Power Grid, Uncertainty

1 Introduction

Modern power grids are complex systems, including many highly interconnected components. Maximising the grid productivity while ensuring a safe and reliable delivery of power is of uttermost importance for grid operators. This requires developing robust decision-making frameworks, which give account to both the complexity of the asset and the uncertainties on its operational conditions, component degradation, failure behaviours, external environment, etc.

Nowadays, the grid management issue is further challenged by the possibility of equipping grid elements with Prognostics and Health Management (PHM) capabilities, which allow tracking the health state evolution. This information can be exploited by grid operators to further increase the profitability of their assets [1–6].

Reinforcement Learning (RL) [7, 8] has been used in the last decades to solve a variety of realistic control and decision-making issues in the presence of uncertainty, including power grid management. In the RL paradigm, a controller (i.e. the decision maker) learns from the interaction with the environment (e.g. the grid) by observing states, collecting rewards and selecting actions to maximise the future revenues, considering the aleatory uncertainties in the environment behavior. The state-action-reward trajectories [9] can be gathered from direct interaction with the real system (e.g. [10]), or from its realistic simulation [7]. This makes RL suitable to power grid management optimization, as it can cope with both the complexity of the asset and the unavoidable uncertainties related to its operation.

In [6], an RL framework based on Q-learning is proposed to solve constrained load flow and reactive power control problems in power grids. Kuznetsova et al. [5] develop an optimisation scheme for consumers actions management in the microgrid contest and accounting for renewable volatility and environmental uncertainty. In [9], a comparison between RL and a predictive control model is presented for a power grid damping problem. In [4], the authors review recent advancements in intelligent control of micro grids including few attempts using RL methods. However, none of the revised works employs RL to find optimal combined Operation and Maintenance (O&M) policies for power grids with degrading elements.

We present an RL framework to support O&M decisions for power grids equipped with PHM systems, which seeks for the settings of the generator power outputs and the scheduling of preventive maintenance actions that maximize the grid load balance and expected profit over an infinite time horizon, while considering the uncertainty of power production from Renewable Energy Sources (RES), power loads and component failure behaviors.

The rest of this paper is organized as follows: Section 2 presents the RL framework for optimal decision
A reinforcement learning framework for optimisation of power grid making under uncertainty. A scaled-down power grid application is proposed in Section 3, whereas the results and limitations of RL are discussed in Sections 4 and 5, respectively. Section 6 closes the paper.

2 Modelling framework for optimal decision making under uncertainty

As anticipated above, developing a RL framework for power grid O&M management requires defining the environment, the actions that the agent can take in every state of the environment, the state transitions the actions lead to and, finally, the rewards associated to each state-action-transition step.

2.1 State space

Consider a power grid made up of elements \( C = \{1, ..., N\} \), physically and/or functionally interconnected, according to the given grid structure. Similarly to [11], the features of the grid elements defining the environment are the \( N_d \) degradation mechanisms affecting the degrading components \( d \in D \subseteq C \) and the \( N_p \) setting variables of power sources \( p \in P \subseteq C \). For simplicity, we assume \( D = \{1, ..., |D|\} \), \( P = \{|D| + 1, ..., |D| + |P|\} \) and \( |D| + |P| \leq N \).

The degradation processes evolve independently on each other according to a Markov process defining the transition probability from state \( s_i^t(t) \) at time \( t \) to the next state \( s_i^t(t+1) \), where \( s_i^t(t) \in \{1, ..., S_i^t\} \quad \forall t, d \in D, i = 1, ..., N_d \). Similarly, for the power sources production, a Markov process defines the probabilistic dynamic of power setting variables from \( s_p^p(t) \) at time \( t \) to the next state \( s_p^p(t+1) \), where \( s_p^p(t) \in \{1, ..., S_p^p\} \quad \forall t, p \in P, j = 1, ..., N_p \). Then, system state vector \( S \in \mathcal{S} \) at time \( t \) reads:

\[
S(t) = \left[ s_1^1(t), s_1^2(t), ..., s_N^{P+|D|}(t) \right] \in \mathcal{S}
\]

(1)

2.2 Actions

Actions can be performed on the grid components \( g \in G \subseteq C \) at each \( t \). The system action vector \( a \in \mathcal{A} \) at time \( t \) is:

\[
a(t) = \left[ a_{g_1}(t), ..., a_{g_g}(t), ..., a_{|G|-1}(t) \right] \in \mathcal{A}
\]

(2)

were action \( a_{g_g}(t) \) is selected for component \( g_g \in G \) among a set of mutually exclusive actions \( a_{g_g} \in \mathcal{A}_{g_g} \). The action set \( \mathcal{A}_{g_g} \) can include operational actions (e.g. closure of a valve, generator power ramp up, etc.) and maintenance actions (e.g. preventive and corrective). Constraints can be defined for reducing \( \mathcal{A}_{g_g} \) to a subset \( \mathcal{A}_{g_g} \subseteq \mathcal{A}_{g_g} \). For example, Corrective Maintenance (CM), cannot be taken on As-Good-As-New (AGAN) components and, similarly, it is mandatory action for failed components. In an optimistic view [11], both Preventive Maintenance (PM) and CM actions are assumed to restore the AGAN state for each component. An example of Markov process for a 4 degradation state component is presented in Fig.1, where circle markers indicate maintenance actions and squared markers indicate other actions, i.e. operational actions.

2.3 Transition probabilities

Transition probability matrices are associated to each feature \( f \) of each component \( c \in P \cup D \) and to each action \( a \in \mathcal{A} \), where \( f \in \{1, ..., N_d\} \) if \( c \in D \) and \( f \in \{1, ..., N_p\} \) otherwise, as follows:

\[
\mathcal{P}_{c,f}^a = \begin{bmatrix}
p_{1,1} & p_{1,2} & \cdots & p_{1,S_f^c} \\
p_{2,1} & p_{2,2} & \cdots & p_{2,S_f^c} \\
\vdots & \vdots & \ddots & \vdots \\
p_{S_f^c,1} & p_{S_f^c,2} & \cdots & p_{S_f^c,S_f^c}
\end{bmatrix}^a
\]

(3)

where \( p_{i,j} \) represents the probability of transition from state \( i \) to state \( j \) of feature \( f \) of component \( c \) and conditional to the action \( a \) in a time varying setting, i.e. \( \mathcal{P}_{c,f}^a(s_j | a, s_i) \). The normalization propriety holds, i.e. \( \sum_{j=1}^n p_{i,j} = 1 \). In practice, element \( p_{i,j} \) of the transition probability matrix \( \mathcal{P}_{c,f}^a \) can be estimated as the relative frequency of the measured component state to fall into the \( j^{th} \) state at time \( t+1 \) provided that it was at the \( i^{th} \) state in the previous time step when the action \( a \) was taken.

REC 2018 – R. Rocchetta, M. Compare, E. Patelli, E. Zio
Figure 1: The Markov Decision Process associated to the health state of a degrading component.

2.4 Rewards

Numerical rewards are case-specific and obtained by solving a physico-economic model of the system, which evaluates how good is the transition from one state to another given that action $a$ is taken:

$$R(t) = \mathfrak{R}(S(t+1), a(t), S(t)) \in \mathbb{R}$$

2.5 Reinforcement Learning and SARSA($\lambda$) method

Generally speaking, the goal of RL methods for optimal control is to find the optimal action-value function $Q_{\pi^*}(S,a)$, which provides an estimation of future revenues when an action $a$ is taken in state $S$, following the optimal policy $\pi^*$:

$$Q_{\pi^*}(S,a) = \mathbb{E}_{\pi^*} \left[ \sum_{t=0}^{\infty} R(t) | S(t), a(t) \right]$$ (4)

Among the wide range of RL algorithms, we adopt SARSA($\lambda$), which is a temporal difference learning methods (i.e. it changes an earlier estimate of $Q$ based on how it differs from a later estimate) employing eligibility traces to carry out backups over n-steps and not just over one step [7]. Details on SARSA($\lambda$) are provided in Algorithm 1 in the Appendix.

3 Case study

A scaled-down power grid case study is used to test the RL decision making framework. The grid includes: 2 controllable generators; 5 cables for the power transmission; 2 non-controllable RES which are connected to 2 loads and provide them electric power depending on random weather conditions (Fig. 2). Then, $|\mathcal{C}|=11$. Two traditional generators (Gen1 and Gen2) are installed as displayed in Fig. 2 and controlled to minimize power unbalances on the grid. We assume that the 2 controllable generators and links 3 and 4 are affected by degradation and, thus, are equipped with PHM capabilities to inform the decision-maker on their degradation states, then $\mathcal{D} = \{1,2,3,4\}$. The two loads and the two renewable generators define the grid power setting, $\mathcal{P} = \{5,6,7,8\}$

3.1 States and Actions

In the case study, we consider $N^d = 1$ degradation features, $d = 1, \ldots, 4$ and $N^p = 1$ power features $p = 1, \ldots, 4$. We consider 4 degradation states for the generators, $s^d_i = \{1, \ldots, S^d_i = 4\}$ for $d = 1, 2$, whereas three states are associated to the power lines $s^p_i = \{1, \ldots, S^p_i = 3\}$, $d = 3, 4$. State 1 refers to the AGAN conditions, state $S^d_i$ to the failure state and states $1 < s^d_i < S^d_i$ to degraded states in ascending order. For each load, we consider 3 states of increasing power demand $s^p_i = \{1, \ldots, S^p_i = 3\}$ for $p = 5, 6$ and three states of increasing power production are associated to renewable sources, $s^p_i = \{1, \ldots, S^p_i = 3\}$ for
Figure 2: The power grid structure and the position of the 4 PHM capabilities, 2 renewable sources, 2 loads and 2 controllable generators.

$p = 7, 8$. Then, the total number of state vectors combinations is 11664 and the grid state vector at time $t$ is defined as follows:

$$S(t) = \{s_1^1, s_2^2, s_3^3, s_4^4, s_5^5, s_6^6, s_7^7, s_8^8\}$$

The agent can operate both generators with the aim to maximise the system revenue by minimizing unbalance between demand and production, while preserving the structural and functional integrity of the system, $g \in G = \{1, 2\}$. Other actions can be performed by other agents on other components (e.g. transmission lines), but being outside from the control domain of the first agent those are assumed included in the environment. Then, the action vector reads $a = [a_1, a_2]$. Five O&M actions can be performed on each controllable generator, for a total of 25 combinations, thus giving rise to a 291600 state-action pairs. The action set for each generator is the following:

$$A_g = \{1, \ldots, 5\} \quad g \in \{1, 2\}$$

where the first 3 (operational) actions affect the power output of the generator, changing it to one of the 3 allowed power levels. The last 2 actions are preventive and corrective maintenance actions, respectively. It is assumed that CM is mandatory for failed generators. Furthermore, highly degraded generators (i.e. $S_d^g = 3$, $d = 1, 2$) are assumed degraded in their operational performance and only the lower power output can be obtained (only $a_g = 1$ action is allowed). Tables 1-3 display the costs for each action and the corresponding power output of the generator, the line electric parameters and the relation between state indices $s_p^1$ and the power variable settings, respectively.

<table>
<thead>
<tr>
<th>Action:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{g=1}$ [MW]</td>
<td>40</td>
<td>50</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$P_{g=2}$ [MW]</td>
<td>50</td>
<td>60</td>
<td>120</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$C_{a,g}$ [m.u.]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>500</td>
</tr>
</tbody>
</table>

Table 1: The power output of the 2 generators in [MW] associated to the 5 available actions and action costs in monetary unit [m.u.].

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>Am [A]</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen 1</td>
<td>Load 1</td>
<td>125</td>
<td>0.0845</td>
</tr>
<tr>
<td>Gen 1</td>
<td>Load 2</td>
<td>135</td>
<td>0.0719</td>
</tr>
<tr>
<td>Gen 1</td>
<td>Gen 2</td>
<td>135</td>
<td>0.0507</td>
</tr>
<tr>
<td>Load 1</td>
<td>Gen 2</td>
<td>115</td>
<td>0.2260</td>
</tr>
<tr>
<td>Load 2</td>
<td>Gen 2</td>
<td>115</td>
<td>0.2260</td>
</tr>
</tbody>
</table>

Table 2: The transmission lines proprieties.

<table>
<thead>
<tr>
<th>State index $s_p^1$</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 5$ Demanded [MW]</td>
<td>60</td>
<td>100</td>
<td>140</td>
</tr>
<tr>
<td>$p = 6$ Demanded [MW]</td>
<td>20</td>
<td>50</td>
<td>110</td>
</tr>
<tr>
<td>$p = 7$ Produced [MW]</td>
<td>0</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>$p = 8$ Produced [MW]</td>
<td>0</td>
<td>20</td>
<td>60</td>
</tr>
</tbody>
</table>

Table 3: The physical values of the power settings in [MW] associated to each state $S_p^1$ of component $p \in P$. 
3.2 Probabilistic Model

State transitions may occur from time \( t \) to the next time step \( t+1 \) and are specifically defined for each feature of each component. The 2 loads have identical transition probability matrices and also the degradation of the transmission cables and generators are described by the same Markov process. Thus, for ease of notation, the component subscripts have been dropped. Each action \( a \in A \) is associated to a specific transition probability matrix \( P_{a}^{g} \) describing the evolution of the generator health state conditioned to its operative state or maintenance action. It can be noticed that probabilities associated to operational actions, namely \( a_{g} = 1, 2, 3 \), affect differently the degradation of the component. For those actions, the bottom row corresponding to the failed state has only zero entries. This is to indicate that operational actions cannot be taken for failed generators, but only CM is allowed. The transition matrices for the considered features are defined as follows:

\[
P_{a}^{d=1} = \begin{pmatrix} 0.98 & 0.02 & 0 & 0 \\ 0 & 0.95 & 0.05 & 0 \\ 0 & 0 & 0.97 & 0.1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{for } d = 1, 2
\]

\[
P_{a}^{d=2} = \begin{pmatrix} 0.97 & 0.03 & 0 & 0 \\ 0 & 0.95 & 0.05 & 0 \\ 0 & 0 & 0.9 & 0.1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{for } d = 1, 2
\]

\[
P_{a}^{d=3} = \begin{pmatrix} 0.95 & 0.04 & 0.01 & 0 \\ 0 & 0.95 & 0.04 & 0.01 \\ 0 & 0 & 0.97 & 0.03 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{for } d = 1, 2
\]

\[
P_{a}^{d=4} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{for } d = 1, 2
\]

\[
P_{a}^{d=5} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.15 & 0 & 0 & 0.85 \end{pmatrix} \quad \text{for } d = 1, 2
\]

\[
P_{a}^{4} = \begin{pmatrix} 0.9 & 0.08 & 0.02 \\ 0.1 & 0.97 & 0.03 \\ 0 & 0 & 0.9 \end{pmatrix} \quad \forall a, d = 3, 4
\]

\[
P_{a}^{5} = \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.4 \\ 0.2 & 0.4 & 0.4 \end{pmatrix} \quad \forall a, p = 5, 6
\]

\[
P_{a}^{7} = \begin{pmatrix} 0.5 & 0.1 & 0.4 \\ 0.3 & 0.3 & 0.4 \\ 0.1 & 0.4 & 0.5 \end{pmatrix} \quad \forall a
\]

3.3 Reward Model

When the agent performs an action at time \( t \); the environment provides a reward and leads the system to its state at time \( t+1 \). The reward is calculated as the sum of 4 different terms: (1) the revenue from selling electric power, (2) the cost of producing electric power by traditional generators, (3) the cost associated to the performed actions and (4) the cost of not serving energy to the customers. Mathematically, the reward reads:

\[
R(t) = \sum_{p=5}^{6} (L_p(t) - ENS_p(t)/\Delta_t) \cdot C_{el} - \sum_{g=1}^{2} P_g \cdot C_g - \sum_{g=1}^{2} C_{a,g} - \sum_{p=5}^{6} ENS_p(T) \cdot C_{ENS}
\]

where \( L_p \) is the power demanded by component \( p \), \( C_{el} \) is the price paid by the loads for per-unit of electric power, \( P_g \) is the power produced by the generators, \( C_g \) is the cost of producing the unit of power, \( C_{a,g} \) is the cost of the action \( a \) on the generator \( g \), \( \Delta_t \) is the time difference between the present and the next system state and it is assumed to be 1 h, \( ENS_p \) is the energy not supplied to the load \( p \) and is a function of the grid state vector and lines and generators electrical proprieties and availability, i.e. \( ENS(t) = \mathcal{G}(S, A, M, X) \) where \( \mathcal{G} \) defines the constrained DC power flow solver [12]. \( C_{ENS} \) is the cost of the energy not supplied. The costs \( C_{ENS}, C_g \) and \( C_{el} \) are set to 5, 4 and 0.145 monetary unit (m.u.) per-unit of energy or power, respectively.

4 Results and Discussion

The SARSA(\( \lambda \)) algorithm (Algorithm 1 in the Appendix) has been used to provide an approximate solution to the decision problem. The stochastic grid model is used to sample control trajectories only, i.e. it provides a reward and a new state when an action and the old state is provided as input. The SARSA method has been run changing parameters setting and accumulating eligibility traces. The initial state...
A reinforcement learning framework for optimisation of power grids has been selected for the episodic loop randomly, using a degradation-weighted probability mass function $f_S(s) \propto \sum_{d=1}^{[D]} s_d^t$. This sampling scheme is used to better estimate action-value functions in rarely visited states (i.e., low-probability states with many failed/highly degraded components), which speeds up the convergence of the SARSA method. For validation, Bellman’s optimality [13]-[14] has been solved to provide a reference optimal action-value function. The Bellman’s results are in good agreement with the SARSA results, as it can be seen from Fig. 3.

The SARSA($\lambda$) results are summarised in Fig. 4, where the curves provide a compact visualization of the distribution of $Q_\pi^*(s,a)$ over the states for the available 25 combinations of actions. Three clusters can be identified: on the far left, we find the set of states from which CM on both generators is performed; being CM a costly action, this leads to a negative expectation of the discounted reward. The second cluster ($C_2$) corresponds to the 8 combination of one CM and any other action on the operating generator. The final cluster ($C_1$) of 16 combinations of actions includes only PM and operational actions. If corrective maintenance is not performed, higher rewards are expected.

In Fig. 5, each sub-plot shows the the highest expected discounted power grid return, $Q_\pi^*(s,a)$, adopting the optimal policy, conditional to a specific degradation states of the generators and for increasing electric load demand. It can be noticed that if the generators are both healthy or slightly degraded (i.e. $\sum_{d=1}^{[D]} s_d^t = 2, 3, 4$) an increment in the overall load demand leads to an increment in the expected reward, due to the larger revenues from selling more electric energy to the customers. On the other hand,
if the generators are highly degraded or failed (i.e. $\sum_{d=1}^{2} s_d^1 = 7, 8$), an increment in the load demand leads to a drop in the expected revenue. This is due to the increasing risk of load curtailments and associated cost (i.e. cost of energy not supplied), and to the impacting PM and CM actions costs.

Figure 5: The maximum $Q_\pi^*(S,a)$ (i.e. maximum expected discounted cumulative reward) for increasing total load and different degrading condition of the generators.

4.1 Policies comparison

We have empirically found that SARSA(0.5) policies outperform SARSA($\lambda$), $\lambda = 0$ and $\lambda = 1$. Thus, two SARSA(0.5) have been further investigated, by setting the truncation windows $T$ to 50 and 250 time steps for each episode, respectively. Table 4 shows the results of the SARSA($\lambda$) algorithms (columns 3 and 4, respectively) and compares them with the MDP (Bellman’s optimality) solution (column 2) and 2 artificial suboptimal policies: $Q_{50rnd}$ (column 5), which is artificially obtained randomizing the action to be selected in 50 % of the states and selecting the MDP optimal action for the remaining states and $Q_{100rnd}$ (column 6), where all states have a random action associated with. Three representative system states $S_1 = [1, 1, 1, 1, 1, 1, 1, 1]$ and $S_2 = [4, 1, 1, 1, 1, 1, 1, 1]$ and $S_3 = [4, 4, 3, 3, 3, 3, 3, 3]$ are used to compare the expected discounted return $Q$. The 3 states are associated with substantially different rewards as they have been selected from the 3 clusters $C_1$, $C_2$ and $C_3$, respectively (see Fig. 4): $S_1$ has both generators in the AGAN state, $S_2$ has on generator out of service whilst $S_3$ has both generators failed. $Act$ is defined as the portion of actions taken from the SARSA($\lambda$) policies that are equal to those taken using the reference MDP optimal policy in the corresponding states; $E[R(t)]$ is the expected averaged non-discounted return, i.e.

$$E \left[ \sum_{t=1}^{T} R(t) \right],$$

independent from the initial state of the system. It is interesting to notice that SARSA(0.5) provides better policies (i.e. higher expected discounted and non-discounted returns) compared to $Q_{50rnd}$ and $Q_{100rnd}$. This is true even if $Q_{50rnd}$ has higher $Act$ compared to the SARSA policies, i.e. more than 60 % of the $Q_{50rnd}$ actions are equal to the MDP actions whilst less than 50 % for the SARSA. This points out that the optimal policy is very sensitive to some of the state-action combinations and less to others. In other words, taking the wrong action in some states can lead to a catastrophic drop in the expected return, whilst in other cases a sub-optimal action affects less the expected revenue (e.g. making generator 1 produce power rather than generator 2 or vice versa).

Fig. 6 presents in details 2 control trajectories obtained selecting greedily actions with the MDP Bellman’s policy (top plot) and the SARSA($\lambda$) policy (bottom plot), rewards are displayed on the y-axis and actions and states (see Table 5) are associated to each time step. It is interesting to observe that by following an optimal policy, PM actions are sometimes recommended even if the generators are As-Good-As-New. This might seem counter intuitive, but it can be explained considering the degradation model settings. A PM action taken in an AGAN degradation state will assure a transition to the AGAN state. In this sense, the MDP policy is ready to accept a slightly lower revenue (due to PM costs), but with the advantage of suspending the degradation process, especially when the power produced by RES can be used to minimise imbalances between power production and the 2 loads are small.
Table 4: The MDP Bellman’s optimality and the RL results compared with suboptimal policies.

<table>
<thead>
<tr>
<th></th>
<th>MDP</th>
<th>SARSA(0.5)</th>
<th>Q50 rnd</th>
<th>Q100 rnd</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_{s_1}$</td>
<td>5719</td>
<td>5511</td>
<td>5555</td>
<td>4191</td>
</tr>
<tr>
<td>$Q_{s_2}$</td>
<td>2898</td>
<td>2577</td>
<td>2664</td>
<td>1297</td>
</tr>
<tr>
<td>$Q_{s_3}$</td>
<td>-1721</td>
<td>-1816</td>
<td>-1813</td>
<td>-2956</td>
</tr>
<tr>
<td>Act top1</td>
<td>100 %</td>
<td>48.8 %</td>
<td>49.1 %</td>
<td>62.1 %</td>
</tr>
<tr>
<td>Act top3</td>
<td>100 %</td>
<td>66.5 %</td>
<td>66.5 %</td>
<td>71.4 %</td>
</tr>
<tr>
<td>$E[R(t)]$</td>
<td>529.8</td>
<td>478.8</td>
<td>488.1</td>
<td>370.3</td>
</tr>
<tr>
<td>$N_e$</td>
<td>-</td>
<td>5e5</td>
<td>5e5</td>
<td>-</td>
</tr>
<tr>
<td>$T$</td>
<td>-</td>
<td>50</td>
<td>250</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 6: Actions taken in 2 separate control trajectories using MDP and SARSA policies. Initial state $s_1$ and next states are randomly generated by the underlying probabilistic model (see Table 5).

Table 5: The state vectors for the MDP and SARSA control trajectories in Figure 6.

<table>
<thead>
<tr>
<th>Gens</th>
<th>Loads</th>
<th>RES</th>
<th>Lines</th>
<th>MDP states trajectory</th>
<th>SARSA states trajectory</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>$s_2$</td>
<td>$s_3$</td>
<td>$s_1$</td>
<td>$s_1$</td>
<td>$s_1$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

5 Discussion on Limitation

While RL, like stochastic dynamic programming (DP), has in principle a very broad scope of application, it has to face computational issues when the state-action spaces of the control problem are very large. In such a case, RL has to be combined with regression techniques capable of interpolating over the state-action space the data obtained from (relatively) few control trajectories [9]. Most of the research in this context has focused on parametric function approximators, representing either some (state-action) value functions or parameterized policies, together with some stochastic gradient descent algorithms (see e.g. [8] or [15]). Further research work will focus on the development of enhanced RL algorithms, capable of dealing with
imprecise rewards (e.g. due to unavailable/unreliable models), partial observability and issues related to scarcity of samples due to low-probability of specific state-action pairs.

6 Conclusion
A framework based on Reinforcement Learning for optimal decision making of power grid systems affected by uncertain operations and degradation mechanisms has been investigated. Power grid models can include PHM devices, which are used to inform the agent about the health state of the system components. This information helps to select optimal O&M actions on the system components.
The SARSAR(\(\lambda\)) method was used to solve a control problem for a scale down power grid with renewable and PHM capabilities. The RL results have been compared to the reference Bellman’s optimality solution and are in good agreement, although inevitable approximation errors have been observed. The framework proved to be flexible and effective in tackling a small but representative case study and future works will test its applicability to more realistic (larger) state-action spaces. To this aim, artificial neural networks will be used in future research work for state-action space regression to scale up to larger grids. This necessary verification for a possible future applicability of the method.

Acknowledgments
The authors would like to acknowledge the gracious support of this work through the EPSRC and ESRC Centre for Doctoral Training on Quantification and Management of Risk & Uncertainty in Complex Systems & Environments Grant number (EP/L015927/1)

References
Appendix

The SARSAL(λ) algorithm starts initializing the action-value function \( Q \) and eligibility traces \( Z \) tables. Then, the values for the learning rate \( \alpha \), the discount factor \( \gamma \), the decay rate of the traces \( \lambda \in [0, 1] \) and the greediness factor \( \epsilon \) (or a policy \( \pi \) to be evaluated) are selected. After this initialization, the episodic loop starts with a random sample (or selection) of an initial state \( s_t \), then, an action \( a_t \) is selected based on the adopted policy, e.g. \( \epsilon \)-greedy or \( \pi(|s_t) \). A \( \epsilon \)-greedy policy consists of random actions, taken with probability \( \epsilon \), or greedy actions taken with probability \( 1-\epsilon \) (i.e. actions for which \( Q \) is maximised). Once the initial state-action pair is obtained, the episode \( e \) is evaluated (i.e. a sequence of action-rewards-state-actions). Temporal difference errors \( \delta_t \) at the time step \( t \) are calculated, traces replaced or accumulated and \( Q \) updated. The episode terminates when a predefined truncation horizon \( T \) is reached (i.e. maximum time length of the episode). The procedure is iterated until a predefined number of events \( N_E \) is obtained. The SARSAL(0) is guaranteed to convergence to an optimal action-value function for a Robbins-Monro sequence of step-sizes \( \alpha_t \), for further details regarding stopping criteria and convergence the reader is referred to [16].

RL approaches can tackle control problems with infinite optimisation horizon by approximating the solution with a T-stage approach. In this sense, windows of \( T \) time steps are used to truncating the time horizon, thus reducing the computational burdens [9]. The SARSAL(λ) algorithm works as follows [7]:

**Data:** Set \( e = 1 \), \( N_E \), \( \epsilon \) (or a policy \( \pi \) to be evaluated), \( \alpha \), \( \gamma \), \( \lambda \);
Initialize \( Q(s,a) \), for all \( s \in S \) and \( a \in A \), arbitrarily (e.g. \( Q = 0 \));
Initialize traces \( Z(s,a) = 0 \), for all \( s \in S \) and \( a \in A \);

**while** \( e < N_E \) (Episodic Loop) **do**

Set \( t = 1 \);
Initialize starting state \( s_t \) e.g. randomly;
Select action \( a_t \in A(s_t) \) using policy derived from \( Q \) (e.g. \( \epsilon \)-greedy) or \( \pi(|s_t) \);

**while** \( t < T \) (run an episode) **do**

Take action \( a_t \), observe \( s_{t+1} \) and reward \( R_t \);
Select action \( a_{t+1} \in A(s_{t+1}) \) using policy derived from \( Q \) (e.g. \( \epsilon \)-greedy) or \( \pi(|s_{t+1}) \);
Compute temporal difference \( \delta_t \) and update traces: \( \delta_t = R_t + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t) \);
\( Z(s_t,a_t) = Z(s_t,a_t) + 1 \) (accumulate traces) or;
\( Z(s_t,a_t) = 1 \) (replace traces);
Update \( Q \) and \( Z \) for each \( s \) and \( a \): \( Q(s,a) = Q(s,a) + \alpha \delta_t Z(s,a) \);
Set \( t = t + 1 \);

end

go to next episode \( e = e + 1 \);
end

**Algorithm 1:** The SARSAL(λ) algorithm adopting replacing or accumulating eligibility traces settings.
Index of Authors

Ahrens, A., 21
Altieri, D., 163
Andrews, J., 103
Auer, E., 21
Bankhead, 193
Beer, M., 87
Belina, A., 125
Bellani, L., 223
Benavente-Peces, C., 21
Betancourt, D., 41
Broggi, M., 87
Brown, D.L., 51
Ceberio, M., 169, 177
Chiachío-Ruano, J., 103
Chiachío-Ruano, M., 103
Compare, M., 223
Daub, M., 185
De Angelis, M., 137
Duddeck, F., 185
Ecclestone, G., 193
Edler, P., 31
Faes, M., 79
Ferson, S., 207
Fetz, T., 147
Freitag, S., 31
Fritsch, 193
Gonzalez, I.V., 155
Graf, W., 63
Green, P.L., 207
Hanss, M., 13
He, L., 87
Hose, D., 13
Imholz, M., 79
Jensen, H.A., 155
Joldes, M., 169
Kaliske, M., 63
Knoll, F., 201
Kosheleva, O. M., 125, 169
Kreinovich, V., 125, 169, 177
Kremer, K., 31
Lauter, C., 169
Le Carrer, 207
MacDonald-Taylor, 193
Meschke, G., 31
Modares, M., 111, 117
Moens, D., 79
Mohammadi, J., 111
Muhanna, R. L., 41, 117
Mullen, R. L., 41, 117
Neves, L.C., 51
O’Brien, W.J., 193
Oberguggenberger, M., 1, 147
Patelli, E., 137, 163, 223
Pierce-Brown, J., 51
Pownuk, A.M., 125
Rocchetta, R., 223
Sadeghi, J., 137
Schietzold, N., 63
Schwarz, M., 1
Topa Gomes, A., 87
Troffaes, M., 147
Vaccaro, A., 177
Valdebenito, M., 155
Van Doninck, B., 79
Zio, E., 223