Computing expectations with p-boxes : two views of the same problem

Lev Utkin[1] and Sebastien Destercke[2]

[1]Department of computer science, State Forest Technical Academy
St.Petersburg, Russia

[2]Institute of radioprotection and nuclear safety
Cadarache, France

ISIPTA, July 2007

Introducing Lev Utkin

Position

Prof. at computer science department, St.Petersburg

Main interests

- Reliability, uncertainty and risk analysis
- Use and aggregation of expert knowledge
- Decision theory

Collaborations

- Igor Kozine, Thomas Augustin

Introducing Sebastien Destercke (me)

Position

Phd student at the Institute of radiological protection and nuclear safety, under the supervision of Didier Dubois (IRIT) and Eric Chojnacki (IRSN)

Main interests

Treatment of information in uncertainty analysis, using imprecise models

- Information modeling
- Information fusion
- (In)dependence concepts
- Propagation of information

Why?

A p-box is a pair of lower/upper CDF $\underline{F}(x) \leq F(x) \leq \overline{F}(x),\ \forall x \in \mathbb{R}$

It is known that...

p-boxes have very low expressive power and, therefore, working with them usually give more imprecise and conservative results

... so, why bother about them?

- They're simple and easier to deal with
- They're very easy to explain
- If we can get an answer to our question by using them, why bother with more complex (and, likely, more expensive) models?

Different situations

Simple (and, still, common) cases

- Our model is simple (e.g. is a combination of monotonic operations like $\log, \exp, \times, /, +, -$)
- Guaranteed methods, although not giving best possible bounds, are satisfying

The worst case

- Big, huge model (i.e. computer codes) with lots of parameters (e.g. 51)
- Not a lot is known about the model
- Every single run or computation of the model takes a long time (and is therefore expensive)

The other cases (the one we're interested in)

- Model is partially known
- Rough tools not fine enough $\rightarrow$ we want to get finer answers

Problem statement

- P-box $\underline{F}(x) \leq F(x) \leq \overline{F}(x),\ \forall x \in \mathbb{R}$ describing our uncertainty on x
- We have a function h that is partially known
- We want to find lower ($\underline{\mathbb{E}}$) and upper expectations ($\overline{\mathbb{E}}$) of $h(x)$:

$$\underline{\mathbb{E}}h = \inf_{\underline{F} \leq F \leq \overline{F}} \int_{\mathbb{R}} h(x)\mathrm{d}F(x),\ \overline{\mathbb{E}}h = \sup_{\underline{F} \leq F \leq \overline{F}} \int_{\mathbb{R}} h(x)\mathrm{d}F(x).$$

- We're searching for the optimal distribution that will reach them, for some specific behavior of h
- h can be a contamination model, an utility function, or any characteristic (mean, probability of an event) about them.

General solutions to approximate ($\underline{\mathbb{E}}$),($\overline{\mathbb{E}}$)

Linear programming

Approximate solution by N points x_i:

$$\underline{\mathbb{E}}^* h = \inf \textstyle\sum_{k=1}^{N} h(x_k) z_k \text{ (lower)}$$
$$\text{or } \overline{\mathbb{E}}^* h = \sup \textstyle\sum_{k=1}^{N} h(x_k) z_k \text{ (upper)}$$

subject to

$$z_k \geq 0, \textstyle\sum_{k=1}^{N} z_k = 1, \; i = 1, \dots, N,$$

$$\textstyle\sum_{k=1}^{i} z_k \leq \overline{F}(x_i), \; \textstyle\sum_{k=1}^{i} z_k \geq \underline{F}(x_i)$$

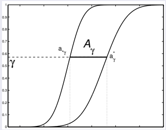
z_k: values of discretized F to optimize

- If N large: computational difficulties ($3N+1$ constraints)
- If N small: possible bad approximations

Random sets

P-box equivalent to multi-valued mapping $\Gamma(\gamma) = A_\gamma = [a_{*\gamma}, a^*_\gamma] \; \gamma \in [0,1]$,

$$a_{*\gamma} = \overline{F}^{-1}(\gamma) \quad a^*_\gamma = \underline{F}^{-1}(\gamma),$$



$$\underline{\mathbb{E}} h = \int_0^1 \inf_{x \in A_\gamma} h(x) \, d\gamma, \; \overline{\mathbb{E}} h = \int_0^1 \sup_{x \in A_\gamma} h(x) \, d\gamma.$$

- Solution : discretize the continuous random set in levels γ_i
- Difficulty: find sup, inf in A_{γ_i}
- If too few levels γ_i or poor heuristics : bad approximations

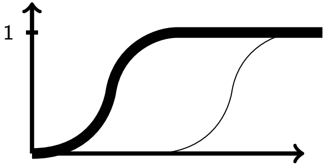
Simple case of monotonic functions

Non-decreasing

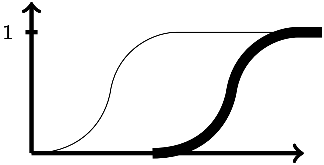
$\underline{\mathbb{E}}h=\int_{\mathbb{R}} h(x)\mathrm{d}\overline{F}(x)$, $\overline{\mathbb{E}}h=\int_{\mathbb{R}} h(x)\mathrm{d}\underline{F}(x)$,
$\underline{\mathbb{E}}h=\int_0^1 h(a_{*\gamma})d\gamma$, $\overline{\mathbb{E}}h=\int_0^1 h(a^*_\gamma)d\gamma$.

Non-increasing

$\underline{\mathbb{E}}h=\int_{\mathbb{R}} h(x)\mathrm{d}\underline{F}(x)$, $\overline{\mathbb{E}}h=\int_{\mathbb{R}} h(x)\mathrm{d}\overline{F}(x)$,
$\underline{\mathbb{E}}h=\int_0^1 h(a^*_\gamma)d\gamma$, $\overline{\mathbb{E}}h=\int_0^1 h(a_{*\gamma})d\gamma$.



Optimal F for $\underline{\mathbb{E}}h$ (non-decreasing h) or $\underline{\mathbb{E}}h$ (non-increasing h)



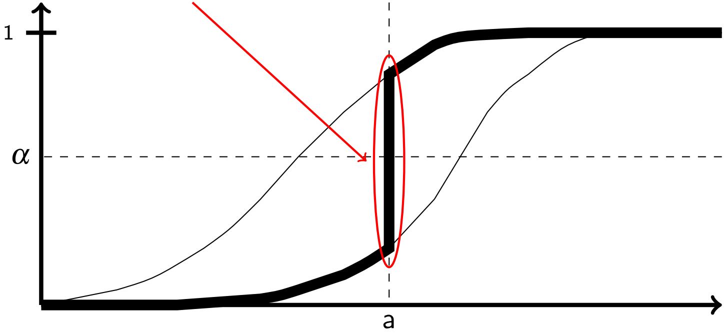
Optimal F for $\underline{\mathbb{E}}h$ (non-increasing h) or $\underline{\mathbb{E}}h$ (non-decreasing h)

One dimension, unconditional case ($\overline{\mathbb{E}}(h)$)

h has one maximum for $x = a$ and is decreasing in $[-\infty, a], [a, \infty]$

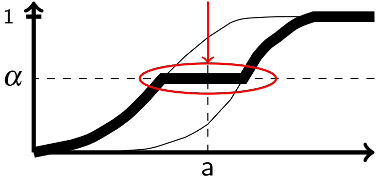
$$\overline{\mathbb{E}}(h) = \int_{-\infty}^{a} h(x)\mathrm{d}\underline{F} + h(a)\left[\overline{F}(a) - \underline{F}(a)\right] + \int_{a}^{\infty} h(x)\mathrm{d}\overline{F}$$

Probability mass concentrated on max.



One dimension, unconditional case ($\mathbb{E}(h)$)

Horizontal jump to "avoid" taking account of highest values



$$\mathbb{E}(h)=\int_{-\infty}^{\overline{F}^{-1}(\alpha)} h(x)\mathrm{d}\overline{F}+\int_{\underline{F}^{-1}(\alpha)}^{\infty} h(x)\mathrm{d}\underline{F}$$

with α solution of

$$h\left(\overline{F}^{-1}(\alpha)\right)=h\left(\underline{F}^{-1}(\alpha)\right)$$

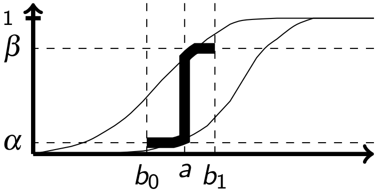
or, with random sets

$$\mathbb{E}h=\int_{0}^{\underline{F}(a)} h(a_{*\gamma})d\gamma+\int_{\underline{F}(a)}^{\overline{F}(a)} \min(h(a_{*\gamma}),h(a^{*}_{\gamma}))d\gamma+\int_{\overline{F}(a)}^{1} h(a^{*}_{\gamma})d\gamma$$

Algorithm to approximate the solution ?

- LP approach suggests (if we don't have analytical solution) to approximate level α by scanning range of values between $[\underline{F}(a),\overline{F}(a)]$
- RS approach suggests to discretize the p-box and to make **at most** two evaluations of h per level.

One dimension, conditional case



Optimal F for $\overline{\mathbb{E}}(h|B)$

Event $B = [b_0, b_1]$ is observed

$$\overline{\mathbb{E}}(h|B) = \sup_{\substack{\underline{F}(b_0) \leq \alpha \leq \overline{F}(b_0) \\ \underline{F}(b_1) \leq \beta \leq \overline{F}(b_1)}} \frac{1}{\beta - \alpha} \int_\alpha^\beta \sup_{x \in (A_\gamma \cap B)} h(x) \mathrm{d}\gamma,$$

$$\underline{\mathbb{E}}(h|B) = \inf_{\substack{\underline{F}(b_0) \leq \alpha \leq \overline{F}(b_0) \\ \underline{F}(b_1) \leq \beta \leq \overline{F}(b_1)}} \frac{1}{\beta - \alpha} \int_\alpha^\beta \inf_{x \in (A_\gamma \cap B)} h(x) \mathrm{d}\gamma,$$

Solution: need to find or approximate values (α, β) for which lower/upper expectations are reached with $\alpha \in [\underline{F}(b_0), \overline{F}(b_0)]$ and $\beta \in [\underline{F}(b_1), \overline{F}(b_1)]$

Multivariate Case

Problem introduction

1. $h : \mathbb{R}^2 \to \mathbb{R}$ is now a function of X and Y.
2. We assume our uncertainty on y is also described by a p-box
$$\underline{F}(y) \leq F(y) \leq \overline{F}(y),\ \forall x \in \mathbb{R}$$
3. h has one global maximum at point (x_0, y_0).
4. The marginal random set of variable Y is uniform mass density on sets $B_\kappa = [b_{*\kappa}, b^*_\kappa]$:
$$b_{*\kappa} := \sup\{y \in [b_{inf}, b_{sup}] : \overline{F}(y) < \kappa\} = \overline{F}^{-1}(\kappa),$$
$$b^*_\kappa := \inf\{y \in [b_{inf}, b_{sup}] : \underline{F}(y) > \kappa\} = \underline{F}^{-1}(\kappa).$$
5. Can $\underline{\mathbb{E}}h, \overline{\mathbb{E}}h$ be easily computed for various assumptions of independence (Couso et al., 2000)?

Multivariate case : summary

strong independence $\mathscr{P}_{XY} = \{p_X \times p_Y | p_X \in \mathscr{P}_X, p_Y \in \mathscr{P}_Y\}$

- For $\overline{\mathbb{E}}h$ probability mass again concentrated on the extremum (x_0, y_0).
- For $\underline{\mathbb{E}}h$, we have to find two "transition" levels instead of one $\rightarrow$ in n dimension, n such levels

RS ind. $\mathscr{P}_{XY} = \{m_{XY}(A,B) = m_X(A) \times m_Y(B) | m_X \equiv \mathscr{P}_X, m_Y \equiv \mathscr{P}_Y\}$

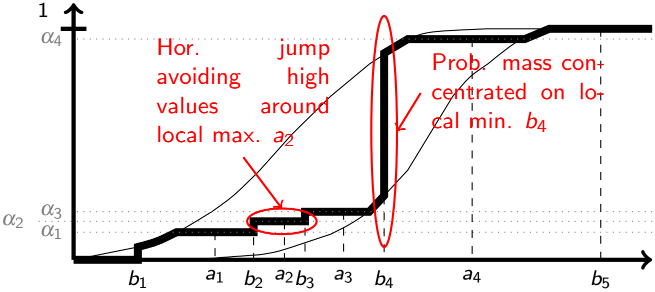
- $\underline{\mathbb{E}}(h) = \int_0^1 \int_0^1 \inf_{(x,y)\in[B_\kappa \times A_\gamma]} h(x,y)\mathrm{d}\kappa\mathrm{d}\gamma$, $\overline{\mathbb{E}}(h) = \int_0^1 \int_0^1 \sup_{(x,y)\in[B_\kappa \times A_\gamma]} h(x,y)\mathrm{d}\kappa\mathrm{d}\gamma$,
- In practice, approximate above equations by discretization

Unknown Interaction $\mathscr{P}_{XY} = \{P_{XY} | P_X \in \mathscr{P}_X, P_Y \in \mathscr{P}_Y\}$

- Using a result from (Fetz and Oberguggenberger, 2004), we can consider the set of all possible joint random sets having m_X, m_Y as marginals
- To approximate $\underline{\mathbb{E}}h, \overline{\mathbb{E}}h$, we need to solve an LP problem.

General case, lower expectation

h has alternate local maxima at points a_i and minima at points b_i, with $b_0 < a_1 < b_1 < a_2 < b_2 < \dots$



- Optimal F is a succession of horizontal and vertical jumps → probability masses concentrated on lower values
- Develop methods to efficiently evaluate vertical and horizontal (α_i) jumps

Conclusions and perspectives

Conclusions

Computing upper and lower expectations for models defined on reals is usually difficult, but we can greatly improve computational efficiency for various cases (i.e. reduce required computational times and/or evaluations of h).

Perspectives

- Extend various results (conditioning, multivariate case) to the more general case (alternate minima/maxima).
- Formalize and develop efficient algorithms to compute lower/upper expectations.
- Make similar work for other models of probability families (Possibility distributions, probability intervals, ...).