DE LA RECHERCHE À L'INDUSTRIE

www.cea.fr

Basics of VV-UQ

G. Perrin

CEA/DAM/DIF, France

Summer school CEA-EDF-INRIA |

July, 3rd 2017

Outline

An increasing role for simulation in our society

- Taking advantage of always increasing computational resources, the importance of simulation keeps increasing.
- It is now completely integrated in most of the decision making processes of our society.
- Thus, simulation has not only to be descriptive, but needs to be **predictive**.
- This implies that simulation is able to take into account the different existing sources of uncertainty.

Once uncertainties are considered, the information provided by simulation is richer. Instead of a unique deterministic value, simulation can :

- provide **confidence** on the prediction of quantities of interest,
- associate a **probability of occurrence** to different scenarii.

Examples

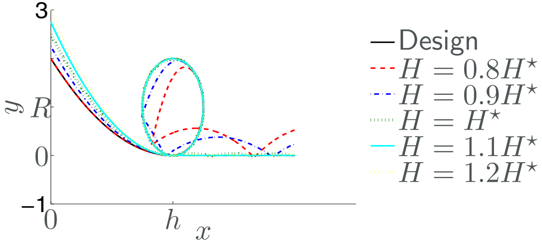
- confidence on weather forecasts,
- chance of success for a given medical operation,
- probability of failure for a structure...

The role of VV-UQ

In that context, VV-UQ approaches must be seen as efficient tools to construct such a confidence for the simulation results, in order to maximize the information provided to the decision maker, by:

- giving methods to integrate at best the available information (expert judgements / experimental measurements) to model the sources of uncertainty,
- proposing a (well-posed) probabilistic formulation of classical problems while taking into account uncertainties,
- providing methods to efficiently solve the probabilistic problem given models for the uncertainties.

A first example : looping design



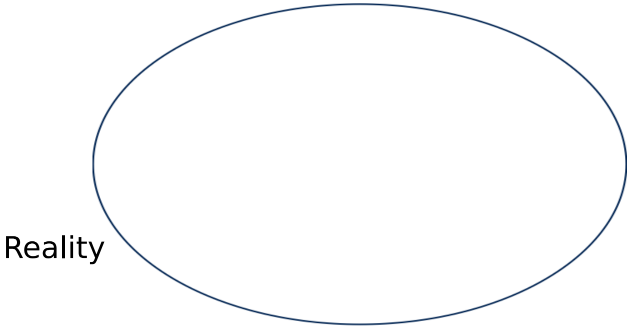
- Which parameters would you retain for the design of the looping?
- Neglecting the friction loss and the aerodynamic forces, it is possible to compute (by conservation of energy) a minimal initial height, $H^\star = 5R/2$, for the wagon to complete the looping. Comment and discuss.

Outline

Notations

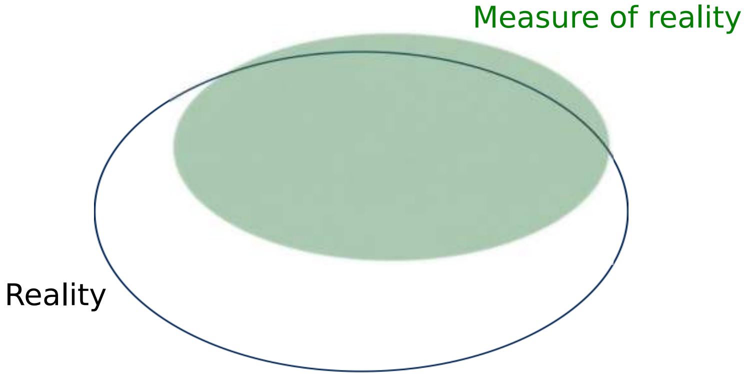
- Let $\mathcal{S}$ be a (physical, mechanical, economical,...) system of interest, whose behaviour is controlled by a vector of input parameters $\boldsymbol{x} \in \mathbb{X}$ (dimensions, material properties, boundary conditions,...).
- Let $y(\boldsymbol{x}) \in \mathbb{Y}$ be the output quantity of interest (QoI) that is used to monitor the system behaviour (one can think to a maximal acceleration of a maximal stress for instance).
- Application $\boldsymbol{x} \mapsto y(\boldsymbol{x})$ is *a priori* non-linear.
- $\boldsymbol{x}$ and y can refer to scalars, vectors, functions...
- Sets $\mathbb{X}$ and $\mathbb{Y}$ are most of the time subsets of $\mathbb{R}^d$ and $\mathbb{R}$.

Different versions of reality leading to different sources of uncertainty.



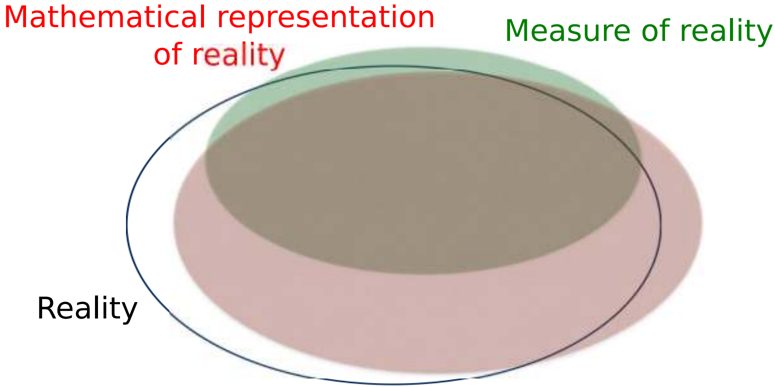
Notations

Different versions of reality leading to different sources of uncertainty.



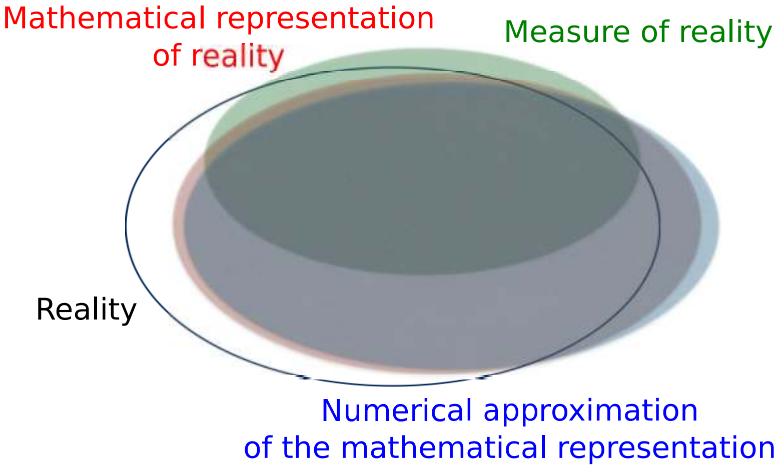
Notations

Different versions of reality leading to different sources of uncertainty.



Notations

Different versions of reality leading to different sources of uncertainty.



Output uncertainties

- Let y^{real} and y^{mes} be the real and the measured values of y.
- Let $y^{\text{mod}}(\cdot;\boldsymbol{\beta})$ and $y^{\text{sim}}(\cdot;\boldsymbol{\beta},\boldsymbol{\delta})$ be a parametric mathematical representation of y and its numerical approximation (from a discretization for instance).

Several sources of output uncertainty

- Controlling the measurement error, $\varepsilon^{\text{mes}} := y^{\text{mes}} - y^{\text{real}}$, is the objective of **metrology** (see the Guide to the expression of Uncertainty in Measurement (GUM) for more details).
- Choosing $\boldsymbol{\delta}$ to control the numerical error, $\varepsilon^{\text{num}} := y^{\text{mod}}(\cdot;\boldsymbol{\beta}) - y^{\text{sim}}(\cdot;\boldsymbol{\beta},\boldsymbol{\delta})$, is the objective of **verification**.
- Choosing the value of $\boldsymbol{\beta}$ is the objective of **calibration**.
- Controlling the simulation error, $\varepsilon^{\text{sim}} := y^{\text{real}} - y^{\text{sim}}(\cdot;\boldsymbol{\beta},\boldsymbol{\delta})$, is the objective of **validation**.

Verification

- Definition : the process of determining that a computational model accurately represents the underlying mathematical model and its solution $\leftrightarrow$ "solving the equations right".
- It refers to numerical analysis, and it is generally based on the use of reference solutions and the study of algorithms.

Validation

- Definition: the process of determining the degree to which a model is an accurate representation of reality from the perspective of the intended uses of the model $\leftrightarrow$ "solving the right equations".
- It is supposed to be based on the confrontation to experimental results and statistical tests (χ_2 test...):

$$y^{\text{mes}}(\boldsymbol{x}) = y^{\text{sim}}(\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\delta}) + \varepsilon^{\text{mes}}(\boldsymbol{x}) + \varepsilon^{\text{sim}}(\boldsymbol{x}).$$

Validation of complex systems

- If simulation is used to analyse high risk events when no/very few global experiments are available (train derailment, nuclear incident, car crashes...), validation is more complex.
- Classical validations are generally carried out on sub-systems, for which measurements are available, and the idea is then to accumulate evidence systematically to compute a credibility for the full-system simulation.
- To this end, several guides (mostly in the USA) have been proposed :
  - CSAU (Code Scaling, Applicability and Uncertainty) is used by the US Nuclear Regulatory Commission,
  - CAS (Credibility Assessment Scale) is developed by NASA
  - PMI (Predictive Maturity Index) is developed at Los Alamos.
  - PCMM (Predictive Capability Maturity Model) is developed at Sandia National Lab.

The value of x can also be uncertain. Two sources of uncertainty for x are generally distinguished:

- the **epistemic** uncertainty (reducible) : some deterministic components of x are not perfectly known,
- the **aleatory** uncertainty (irreducible) : some components of x are random by nature (manufacture dispersion, wind, natural radioactivity...).

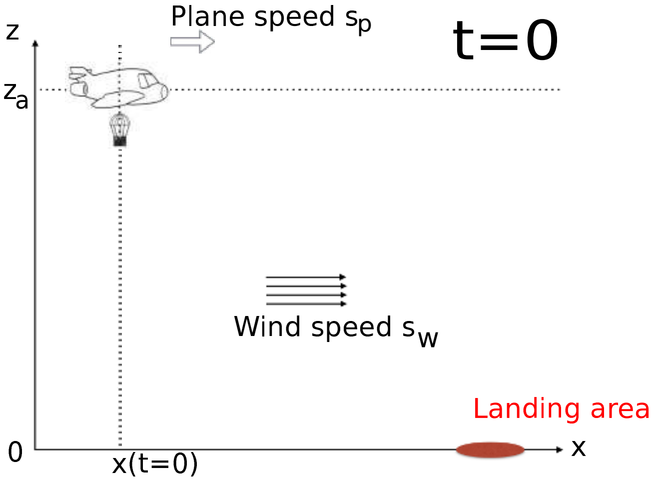
The **probability theory** is generally considered to model both input and output uncertainties.

This leads us to a hierarchical model for $y^{\mathsf{sim}}(x;\boldsymbol{\beta},\boldsymbol{\delta})$, as :

- x is random,
- $y^{\mathsf{sim}}(x;\boldsymbol{\beta},\boldsymbol{\delta})|x$ is also random.

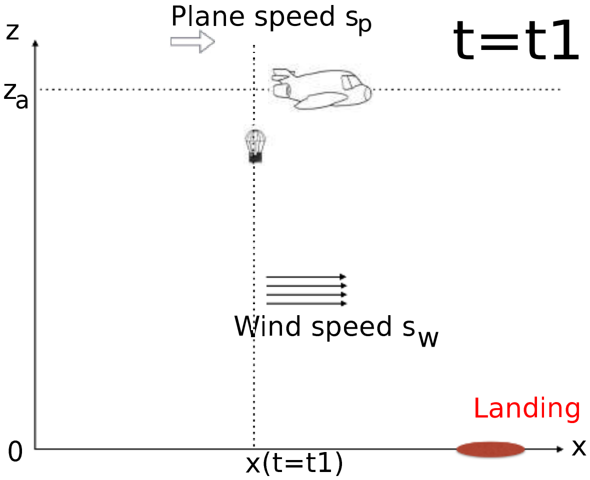
A second example

How simulation can help us to identify the landing area of a parachuted package?



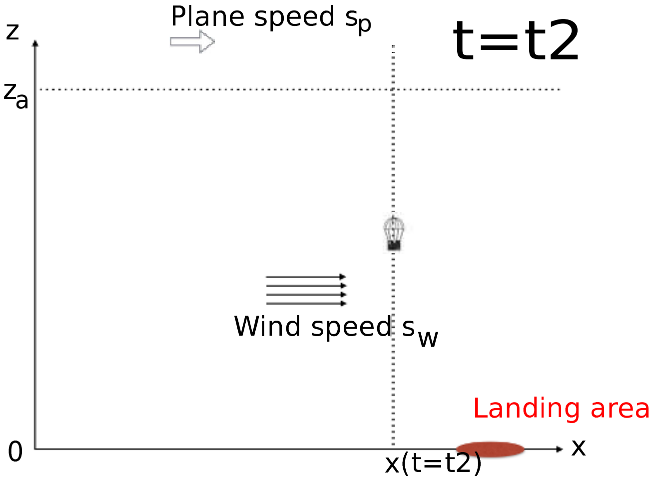
A second example

How simulation can help us to identify the landing area of a parachuted package?



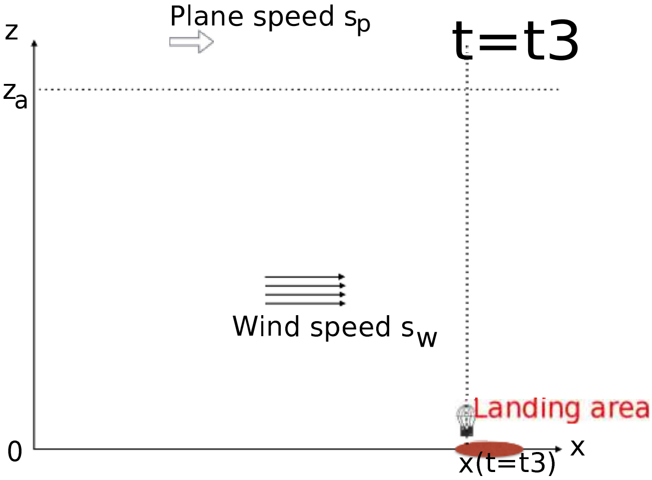
A second example

How simulation can help us to identify the landing area of a parachuted package?



A second example

How simulation can help us to identify the landing area of a parachuted package?



Outputs

- y^{real} is the true landing abscissa,
- With $\boldsymbol{\beta} = \boldsymbol{g}$, the modelled landing abscissa, $y^{\text{mod}}(\cdot;\boldsymbol{\beta})$, can be the horizontal abscissa of $\boldsymbol{x}(t)$ when $\langle \boldsymbol{x}(t), \boldsymbol{e}_z \rangle = 0$, where :

$$\begin{cases} m\ddot{\boldsymbol{x}} = m\boldsymbol{g} - \boldsymbol{A}(\dot{\boldsymbol{x}} - \boldsymbol{s}_w) \left\| \dot{\boldsymbol{x}} - \boldsymbol{s}_w \right\|, \\ \dot{\boldsymbol{x}}(t=0) = \boldsymbol{s}_p, \quad \boldsymbol{x}(t=0) = (x_0, z_a). \end{cases}$$

- With $\boldsymbol{\delta} = dt$, the simulated landing abscissa, $y^{\text{sim}}(\cdot;\boldsymbol{\beta},\boldsymbol{\delta})$, can be the horizontal abscissa of $\widehat{\boldsymbol{x}}(t)$ when $\langle \widehat{\boldsymbol{x}}(t), \boldsymbol{e}_z \rangle = 0$, with:

$$\begin{cases} \widehat{\boldsymbol{x}}(t+dt) = \widehat{\boldsymbol{x}}(t) + \dot{\widehat{\boldsymbol{x}}}(t)dt, \\ \dot{\widehat{\boldsymbol{x}}}(t+dt) = \dot{\widehat{\boldsymbol{x}}}(t) + \left(\boldsymbol{g} - \frac{1}{m}\boldsymbol{A}(\dot{\widehat{\boldsymbol{x}}}(t) - \boldsymbol{s}_w(t)) \left\| \dot{\widehat{\boldsymbol{x}}}(t) - \boldsymbol{s}_w(t) \right\| \right) dt, \\ \dot{\widehat{\boldsymbol{x}}}(t=0) = \boldsymbol{s}_p, \quad \widehat{\boldsymbol{x}}(t=0) = (x_0, z_a). \end{cases}$$

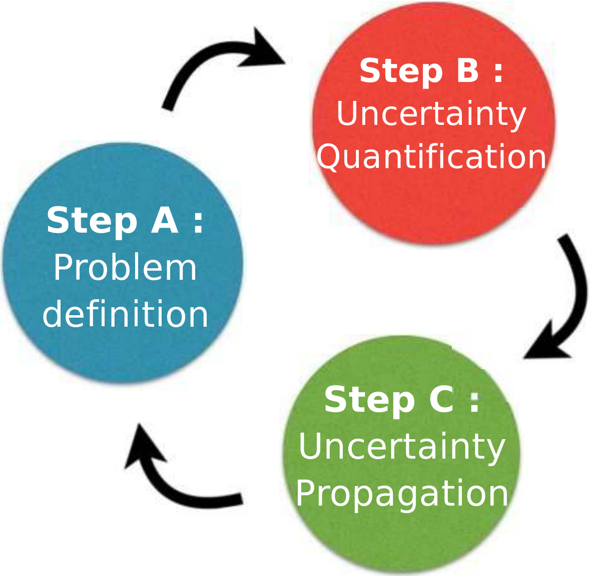
A second example

Inputs : $\boldsymbol{x} = (x_0, z_a, \boldsymbol{s}_p, \boldsymbol{s}_w, m, \boldsymbol{A})$

- (x_0, z_a), the dropping position (epistemic).
- $\boldsymbol{s}_p$, the plane speed at $t = 0$ (epistemic).
- $\boldsymbol{s}_w$, the wind speed, which can depend on time (aleatory).
- m, the mass of the package (epistemic).
- $\boldsymbol{A}$, the matrix that characterizes the aerodynamic forces of the package (epistemic).

- **Verification** $\leftrightarrow$ does the proposed numerical scheme allow $y^{\text{sim}}(\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\delta})$ to converge to $y^{\text{mod}}(\boldsymbol{x}; \boldsymbol{\beta})$ when dt tends to 0 / how to choose dt?
- **Validation** $\leftrightarrow$ is it necessary to consider numerical or model errors for y^{real} to belong to $\mathcal{I}_\alpha$, such that $\mathbb{P}(y^{\text{sim}}(\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\delta}) \in \mathcal{I}_\alpha) = 1 - \alpha$? / which structure can we propose for the model error?

Outline

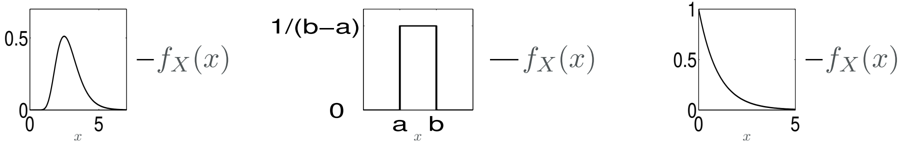

Step B :
Uncertainty
Quantification
Step A :
Problem
definition
Step C :
Uncertainty
Propagation

Step A : problem definition

1. What is the question?
2. What are the quantities of interest (QoI)?
3. Which models can be used to compute these QoI and to answer this question?
4. What are the inputs and the outputs of these models?
5. Which parameters have to be calibrated for these models to be run?
6. What is missing in the proposed models?
7. Do we have information about the structure of the model errors that could affect the prediction of the QoI?
8. Which method is used to solve these models?
9. On which numerical parameters are these solvers based?
10. Can we evaluate the numerical error associated with these solvers?

Step B : uncertainty quantification

1. What are the different sources of (input and output) uncertainty?
2. What is the available information about these uncertainties?
   - expert judgements (positive constraints, physical bounds,...),
   - direct and/or indirect experimental measurements?
3. What is the "most adapted" model to characterize these uncertainties, including their dependence structure (statistical inference)?
   - parametric models (Gaussian, uniform, Beta, ...),
   - non parametric models (kernel representations),
   - indirect representations (Polynomial Chaos expansion,...),
   - alternative theories (fuzzy sets, P-box, credibility functions,...).



Step C : uncertainty propagation

1. Do we have access to the equations on which the model is based?
   - → yes : intrusive and non-intrusive methods can be considered,
   - → no : only non-intrusive methods can be considered. In that case, the model is generally referred as a "black box".
2. Is the model linear (or slightly non-linear)? If yes, methods based on Taylor approximations can be used.
3. Is the model monotone?
4. Do we have information about the regularity of the model?

⇒ Sampling techniques (such as Monte Carlo / Markov Chain Monte Carlo approaches) are generally used to propagate the uncertainties.

⇒ The less regular is the model, and the more code evaluations are generally needed to achieve a given precision on the results.

$$\boldsymbol{\beta}, \boldsymbol{\delta} \qquad \qquad \boldsymbol{\theta}$$

$$\downarrow \qquad \qquad \downarrow$$

$$\text{CODE} \rightarrow y^{\text{sim}}(\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\delta}) + \varepsilon^{\text{sim}}(\boldsymbol{x}; \boldsymbol{\theta})$$

$$\boldsymbol{x} \nearrow \searrow \qquad \searrow \nearrow \; y^{\text{real}}(\boldsymbol{x})$$

$$\text{MEASURE} \rightarrow y^{\text{mes}}(\boldsymbol{x}) + \varepsilon^{\text{mes}}(\boldsymbol{x})$$

Direct problems

Given information on $\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\theta}$, we would like to evaluate some statistical quantities of y^{real}, such as:

- its mean and variance (**prediction**),
- probabilities of exceeding thresholds (**certification**),
- its full density (to be integrated in other models for instance...).

Step C : uncertainty propagation - two families of problems

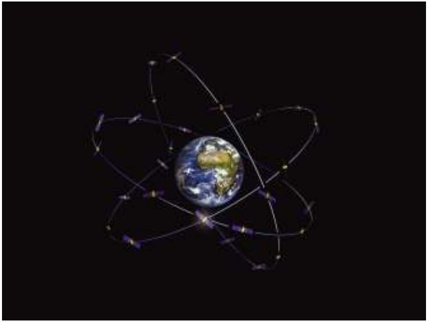
$$
\begin{array}{ccccccc}
 & & \boldsymbol{\beta}, \boldsymbol{\delta} & & \boldsymbol{\theta} & & \\
 & & \downarrow & & \downarrow & & \\
 & \nearrow & \text{CODE} & \rightarrow & y^{\text{sim}}(\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\delta}) + \varepsilon^{\text{sim}}(\boldsymbol{x}; \boldsymbol{\theta}) & \searrow & \\
\boldsymbol{x} & & & & & & y^{\text{real}}(\boldsymbol{x}) \\
 & \searrow & \text{MEASURE} & \rightarrow & y^{\text{mes}}(\boldsymbol{x}) + \varepsilon^{\text{mes}}(\boldsymbol{x}) & \nearrow &
\end{array}
$$

Inverse problems

Given information on $\boldsymbol{x}$ and y, we are interested in:

- the calibration of parameters $\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\theta}$ (**verification** and **validation**),
- classify the influence of each input (or group of inputs) on the variability of y (**sensitivity analysis**),
- find the best value of $\boldsymbol{x}$ with respect to given criteria on y (**optimization**, **conception**).

A third example : GPS positionning[1]



1. Based on location information given by 5 satellites (including uncertainty), how to compute the mean position of a GPS receiver $\boldsymbol{R}$?
2. What would be the radius of the sphere Ω centred on the mean position of $\boldsymbol{R}$ such that $\mathbb{P}(\boldsymbol{R} \in \Omega) = 90\%$?

1

[1] Inspired from the Mathematical Competitive Game 2014-2015 of the French Federation of Mathematical Games

A third example : GPS positionning

Step A : definition of the problem

- We have 5 satellites, and each satellite $\boldsymbol{S}_k$ sends a signal, which is analysed by a receiver $\boldsymbol{R}$.
- The clocks of the satellites are supposed to be perfect and synchronized (this is the "absolute time").
- The clock of the receiver may differ from the absolute time by a value τ, which is refered as "time shift" in the following.
- From the signal sent by each satellite, the receiver deduces the time $t_k = \tau + \|\boldsymbol{S}_k - \boldsymbol{R}\| / c$, or, equivalently, the distance $D_k = c\tau + \|\boldsymbol{S}_k - \boldsymbol{R}\|$, where c is the speed of light (c is supposed to perfectly known).

Remark: for this example, as it will be shown later, we don't need any numerical models.

A third example : GPS positionning

Step B: sources of uncertainty

- The time shift, τ, may be positive or negative, and is unknown. It is supposed to be uniformly distributed on an interval $\mathbb{T}$.
- We assume that each satellite returns a law of probability for D_k, $f_{\boldsymbol{D}} = f_{D_1} \times \cdots \times f_{D_5}$. In practice, such a law is usually given in a discrete form and is defined on an compact subset $\mathbb{D} \subset \mathbb{R}^5$.
- The positions of the satellites are also uncertain: the positions of the satellites $\boldsymbol{S}_k$ are supposed to be independent and distributed in spheres $\mathbb{S}_k$, whose centres and radii are known. Let $f_{\boldsymbol{S}_k}$ be the PDF of $\boldsymbol{S}_k$.

Remark: all these uncertainties are epistemic, there is no randomness here, but only lack of information...

A third example : GPS positionning

Step C: propagation of the uncertainty

By construction:

- $\mathbb{E}[\boldsymbol{R}] = \int_{\mathbb{R}^3} \boldsymbol{r} f_{\boldsymbol{R}}(\boldsymbol{r}) d\boldsymbol{r}$,
- $\mathbb{P}(\boldsymbol{R} \in \Omega) = \int_{\Omega} f_{\boldsymbol{R}}(\boldsymbol{r}) d\boldsymbol{r}$.

$\Rightarrow$ The key step is the computation of $f_{\boldsymbol{R}}$.

$\Rightarrow$ In that prospect, it is interesting to notice that for each value of $\boldsymbol{r}$:

$$
\begin{aligned}
f_{\boldsymbol{R}}(\boldsymbol{r}) &= \int_{\mathbb{T}\times\mathbb{S}_1\times\cdots\mathbb{S}_5} f_{\boldsymbol{R},\tau,\boldsymbol{S}_1,\ldots,\boldsymbol{S}_5}(\boldsymbol{r}, t, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_5) dt d\boldsymbol{s}_1 \ldots d\boldsymbol{s}_5, \\
&= \int_{\mathbb{T}\times\mathbb{S}_1\times\cdots\mathbb{S}_5} f_{\boldsymbol{D}}(ct + \|\boldsymbol{r} - \boldsymbol{s}_1\|, \ldots, ct + \|\boldsymbol{r} - \boldsymbol{s}_5\|) dt d\boldsymbol{s}_1 \ldots d\boldsymbol{s}_5, \\
&= \mathbb{E}\left[\frac{f_{\boldsymbol{D}}(c\tau + \|\boldsymbol{r} - \boldsymbol{S}_1\|, \ldots, c\tau + \|\boldsymbol{r} - \boldsymbol{S}_5\|)}{f_\tau(\tau) f_{\boldsymbol{S}_1}(\boldsymbol{S}_1) \ldots f_{\boldsymbol{S}_5}(\boldsymbol{S}_5)}\right].
\end{aligned}
$$

A third example : GPS positionning

Step C: propagation of the uncertainty - practical solving

- The support of the PDFs of $\boldsymbol{S}_k$, τ and $\boldsymbol{D}$ being compact, the support of $f_{\boldsymbol{R}}$ is also compact.
- Sampling techniques can therefore be used to compute this PDF on a discretized subspace of $\mathbb{R}^3$.
- Such a PDF can then be used to compute $\mathbb{E}\left[\boldsymbol{R}\right]$ and find Ω.

Numerical application

- The uncertainty on τ is generally $\pm 10^{-9}s$, the radii of the spheres for the satellite positions are commonly chosen equal to $2m$, whereas the uncertainty for the distance between the receiver and the satellite is often given by $\pm 10m$ (these distances being more than 20000km).
- Given these uncertainties, we find that the radius of sphere Ω, such that $\mathbb{P}(\boldsymbol{R} \in \Omega) = 90\%$, is around $2m$.

Outline

Deterministic case

From a very general point of view, optimization refers to the solving of :

$$\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}\in\mathbb{X}} y(\boldsymbol{x}), \qquad \text{where:}$$

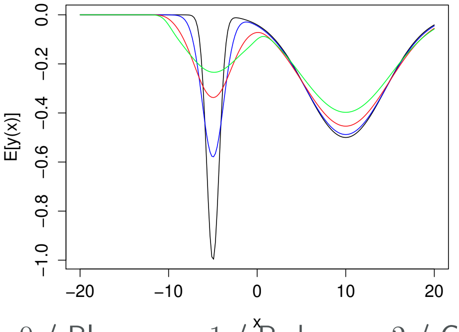
- $\mathbb{X}$ is a given subset of $\mathbb{R}^d$, which can include different kinds of (input and/or output) constraints,
- QoI y is a deterministic cost function that is adapted to the problem.

Stochastic case

- When uncertainties are considered, $y(\boldsymbol{x})$ becomes random.
- Instead of searching its minimum, we generally focus on some statistical quantities of y (its mean for instance).
- The choice of the new QoI has to be adapted to the application.

Example 1: uncertainty on the input only

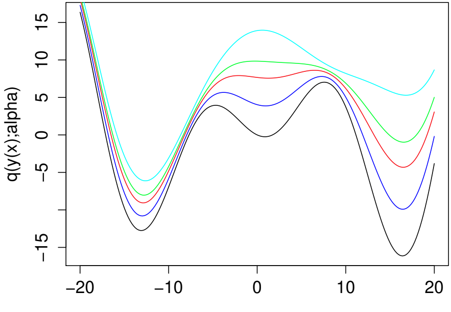
$$x^\star = \arg \min_{x \in [-20,20]} \mathbb{E}\left[y(x+\delta)\right], \quad \delta \sim \mathcal{N}(0, \sigma^2)$$



Black : $\sigma = 0$ / Blue : $\sigma = 1$ / Red : $\sigma = 2$ / Green : $\sigma = 3$.

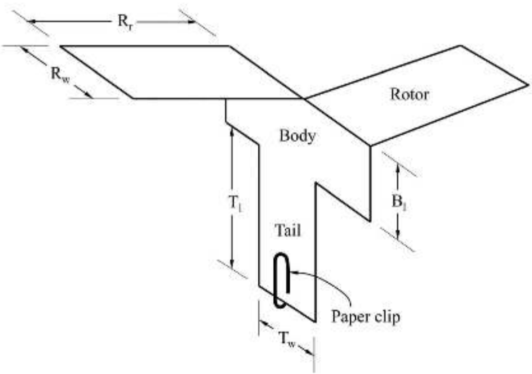
Example 2: uncertainty on the output only

$$x^\star = \arg \min_{x \in [-20,20]} q\left[y(x);\alpha\right],$$



Black : $\alpha = 50\%$ / Blue : $\alpha = 75\%$ / Red : $\alpha = 90\%$ / Green : $\alpha = 95\%$ / Cyan : $\alpha = 99\%$.

Example 3: optimization of a paper helicopter



Parameters:

- Dimensions R_r, R_w, T_l, B_l, T_w.
- Masses of the helicopter and of the paper clip.

Starting from a standard A4 paper, how to maximize the fall time of a paper helicopter dropped from a 5 meters height, which carries a paper clip?

Example 3: optimization of a paper helicopter

Some difficulties for the optimization:

- There are **testing uncertainties** due to differences in the way the helicopters are dropped, small variations in the height from which they are dropped, and inaccuracies in measuring the fall time $\Rightarrow$ each helicopter can be dropped several times / cameras can be used to videotape the fall and remove outliers.
- There are **construction uncertainty**, in that each constructed helicopter is different from other nominally identical helicopter $\Rightarrow$ for each choice of the dimensions, several helicopters can be built / electrical drawing tools and printers can be used to minimize these uncertainties.

Example 3: optimization of a paper helicopter

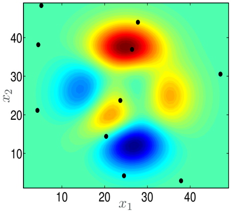
Some difficulties for the optimization:

- The two masses are difficult to measure as they are very small ⇒ it can be more efficient to assess the average paper mass per unit area and average mass of a clip by weighting multiple sheets of paper and a group of clips.
- A mechanical model based on a drag force that assumes quadratic dependence on the velocity can be used to predict the fall time ⇒ the drag coefficient has to be calibrated for this model to help the conception.

Outline

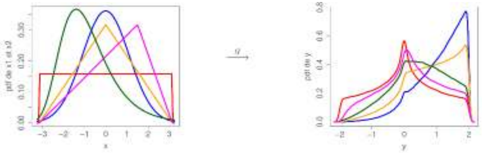
(1/4) Computational costs

- Most of the propagation methods are based on a huge number of code evaluations ($\sim 10^4 - 10^6$).
- When the computational cost associated with one evaluation is very high (for instance, some applications at CEA require around 100h on 100 computer cores in parallel for one evaluation), surrogate models are generally coupled to these approaches.
- Replacing the computer code by a surrogate introduces an other uncertainty, which also needs to be carefully controlled.



(2/4) Input / output characterization

- The uncertainty propagation completely depends on the uncertainty characterization.
- In most of the industrial applications, these uncertainties are relatively difficult to characterize. This is particularly true when very little information is available and/or when the input and output dimensions are high (high-dimensional random vectors, random fields...).
- This additional uncertainty can be taken into account by introducing uncertainties on the input distribution.



$$g(x) = cos(x_1) + cos(x_2)$$

Difficulties and works in progress

(3/4) Stochastic codes

- All the previous slides have been presented in the case when the computer code is deterministic, in the sense that running it twice leads to the exact two same values.
- This is not always the case (let us think to Monte Carlo solvers for instance), which can pose some additional difficulties.

(4/4) Data aggregation

In VV-UQ procedures, an other important difficulty comes from the aggregation of information coming from different experiments, from different experts, from different numerical models, which may be very different, and sometimes incoherent.

A last example : the SANDIA containment vessel benchmark

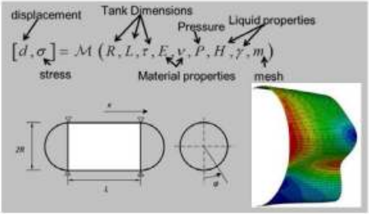
- There exists many storage tanks in the world, holding "Mystery Liquid" under pressure.
- No tank has actually failed, ever, but during standard safety testing, one tank's measurements exceeded a safety specification.

⇒ Are the tanks at risk of failure? Do we have to quickly remove all these tanks to prevent an accident?

⇒ How can evidence from experiments and simulations be integrated and used to support the final decision?

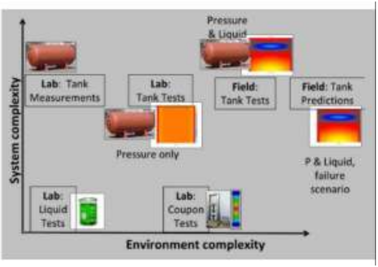
A last example : the SANDIA containment vessel benchmark

Numerical model



Four finite-element mesh resolutions were available, based on the same geometry.

Experimental data



Manufacture dispersions were also provided.

⇒ There is no "right" answer.

⇒ The method demonstration is sometimes more interesting than the final result.

Outline

Conclusions

- The formalism and the notations of VV-UQ based on the probability theory have been presented and illustrated on a series of examples.
- There exist alternative theories to assess prediction credibility.
- VV-UQ often provides more questions and discussions than clear-cut answers, and has to be seen as a tool that can help us to make a decision.
- VV-UQ approaches are not there to replace the physical / mechanical / economical models, but to optimize the information that can be extracted from them.
- Well characterizing the uncertainties does not necessarily mean that we reduce them.

Thank you for your attention.

Any questions?