Reliable detection of outliers with intervals method

Luc Jaulin, Lab-STICC, ENSTA-Bretagne
TOMS, Télécom-Bretagne, 7 février 2017

Interval analysis

Problem. Given $f : \mathbb{R}^n \to \mathbb{R}$ and a box $[\mathbf{x}] \subset \mathbb{R}^n$, prove that

$$\forall \mathbf{x} \in [\mathbf{x}], f(\mathbf{x}) \geq 0.$$

Interval arithmetic can solve efficiently this problem.

Example. Is the function

$$f(\mathbf{x}) = x_1 x_2 - (x_1 + x_2)\cos x_2 + \sin x_1 \cdot \sin x_2 + 2$$

always positive for $x_1, x_2 \in [-1, 1]$?

Interval arithmetic

$$\begin{aligned}
[-1,3]+[2,5] &= [1,8], \\
[-1,3]\cdot[2,5] &= [-5,15], \\
\mathrm{abs}([-7,1]) &= [0,7]
\end{aligned}$$

The interval extension of

$$f(x_1, x_2) = x_1 \cdot x_2 - (x_1 + x_2) \cdot \cos x_2 + \sin x_1 \cdot \sin x_2 + 2$$

is

$$\begin{aligned}[f]([x_1], [x_2]) \quad &= \quad [x_1] \cdot [x_2] - ([x_1] + [x_2]) \cdot \cos [x_2] \\ &\quad + \sin [x_1] \cdot \sin [x_2] + 2.\end{aligned}$$

Theorem (Moore, 1970)

$$[f]([\mathbf{x}]) \subset \mathbb{R}^{+} \Rightarrow \forall \mathbf{x} \in [\mathbf{x}], f(\mathbf{x}) \geq 0.$$

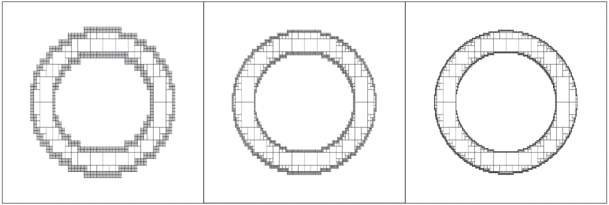
Set Inversion

A subpaving of $\mathbb{R}^n$ is a set of non-overlapping boxes of $\mathbb{R}^n$.
Compact sets $\mathbb{X}$ can be bracketed between inner and outer subpavings:

$$\mathbb{X}^- \subset \mathbb{X} \subset \mathbb{X}^+.$$

Example.

$$\mathbb{X} = \{(x_1, x_2) \mid x_1^2 + x_2^2 \in [1, 2]\}.$$



Let $\mathbf{f}:\mathbb{R}^n \to \mathbb{R}^m$ and let $\mathbb{Y}$ be a subset of $\mathbb{R}^m$. Set inversion is the characterization of

$$\mathbb{X} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{f}(\mathbf{x}) \in \mathbb{Y}\} = \mathbf{f}^{-1}(\mathbb{Y}).$$

We shall use the following tests.

$$\begin{array}{lllll} \text{(i)} & [\mathsf{f}]([\mathsf{x}]) \subset \mathbb{Y} & \Rightarrow & [\mathsf{x}] \subset \mathbb{X} \\ \text{(ii)} & [\mathsf{f}]([\mathsf{x}]) \cap \mathbb{Y} = \emptyset & \Rightarrow & [\mathsf{x}] \cap \mathbb{X} = \emptyset. \end{array}$$

Boxes for which these tests failed, will be bisected, except if they are too small.

Algorithm Sivia(in: $[\mathbf{x}](0), \mathbf{f}, \mathbb{Y}$)

1	$\mathscr{L} := \{[\mathbf{x}](0)\};$
2	pull $[\mathbf{x}]$ from $\mathscr{L}$;
3	if $[\mathbf{f}]([\mathbf{x}]) \subset \mathbb{Y}$, draw($[\mathbf{x}]$,'red');
4	elseif $[\mathbf{f}]([\mathbf{x}]) \cap \mathbb{Y} = \emptyset$, draw($[\mathbf{x}]$,'blue');
5	elseif $w([\mathbf{x}]) < \varepsilon$, {draw ($[\mathbf{x}]$,'yellow')};
6	else bisect $[\mathbf{x}]$ and push into $\mathscr{L}$;
7	if $\mathscr{L} \neq \emptyset$, go to 2

Set estimation

$$\mathbf{y} = \psi(\mathbf{p}) + \mathbf{e},$$

where
$\mathbf{e} \in \mathbb{E} \subset \mathbb{R}^m$ is the error vector,
$\mathbf{y} \in \mathbb{R}^m$ is the collected data vector,
$\mathbf{p} \in \mathbb{R}^n$ is the parameter vector to be estimated.

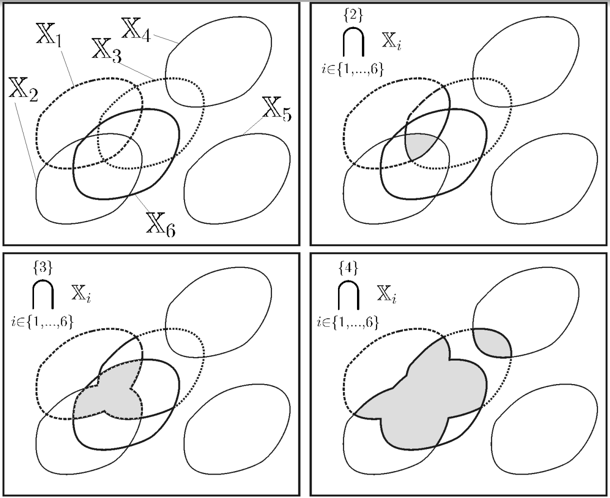
Or equivalently

$$\mathbf{e} = \mathbf{y} - \psi(\mathbf{p}) = \mathbf{f}_{\mathbf{y}}(\mathbf{p}),$$

The *posterior feasible set* for the parameters is

$$\mathbb{P} = \mathbf{f}_{\mathbf{y}}^{-1}(\mathbb{E}).$$

Relaxed intersection


$\mathbb{X}_1$
$\mathbb{X}_2$
$\mathbb{X}_3$
$\mathbb{X}_4$
$\mathbb{X}_5$
$\mathbb{X}_6$
$\bigcap^{\{2\}}_{i\in\{1,\dots,6\}} \mathbb{X}_i$
$\bigcap^{\{3\}}_{i\in\{1,\dots,6\}} \mathbb{X}_i$
$\bigcap^{\{4\}}_{i\in\{1,\dots,6\}} \mathbb{X}_i$

Exercise (IAMOOC)

Consider the bounded-error parameter estimation problem defined by

$$p_1 \cdot e^{p_2 \cdot t} \in [y](t)$$

with

t_i	0.2	1	2	4
$[y](i)$	$[1.5, 2]$	$[0.7, 0.8]$	$[0.1, 0.3]$	$[-0.1, 0.03]$

1) Draw an inner and an outer approximations for the set $\mathbb{P}_i$ of all $\mathbf{p}$ consistent with the ith data.
2) Compute the set all $\mathbf{p}$ consistent with all data except q of them, for $q \in \{0, 1, 2\}$.
3) Provide a method able to identify which data is an outlier.

Probabilistic-set approach

We decompose the error space into two subsets: $\mathbb{E}$ on which we bet $\mathbf{e}$ will belong and $\overline{\mathbb{E}}$. We set

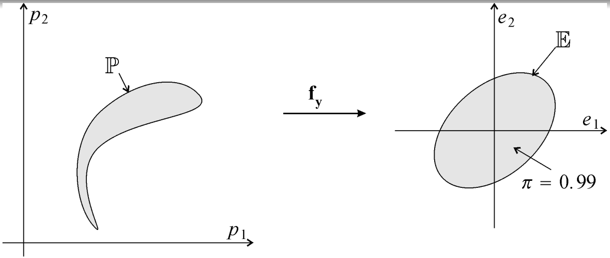
$$\pi = \Pr(\mathbf{e} \in \mathbb{E})$$

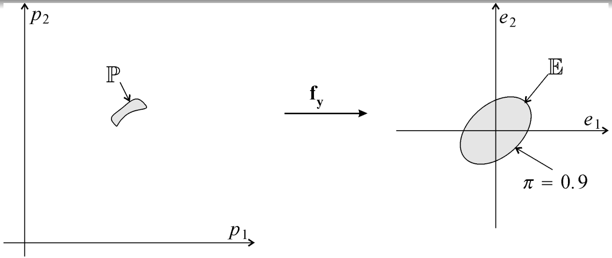
The event $\mathbf{e} \in \overline{\mathbb{E}}$ is considered as *rare*, i.e., $\pi \simeq 1$.

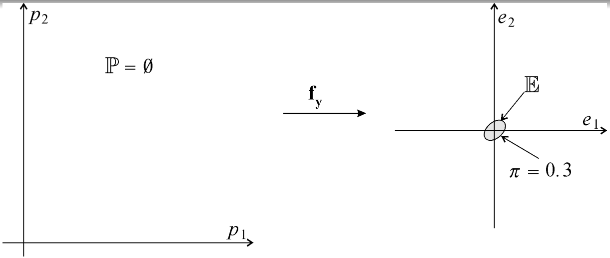
Once $\mathbf{y}$ is collected, we compute

$$\mathbb{P} = \mathbf{f}_{\mathbf{y}}^{-1}(\mathbb{E}).$$

If $\mathbb{P} \neq \emptyset$, we conclude that $\mathbf{p} \in \mathbb{P}$ with a prior probability of π.
If $\mathbb{P} = \emptyset$, than we conclude the rare event $\mathbf{e} \in \overline{\mathbb{E}}$ occurred.


p_2
$\mathbb{P}$
p_1
$\mathbf{f}_{\mathbf{y}}$
e_2
$\mathbb{E}$
e_1
$\pi = 0.99$


p_2
$\mathbb{P}$
p_1
$\mathbf{f}_{\mathbf{y}}$
e_2
$\mathbb{E}$
e_1
$\pi = 0.9$


p_2
$\mathbb{P} = \emptyset$
p_1
$\mathbf{f}_\mathbf{y}$
e_2
$\mathbb{E}$
e_1
$\pi = 0.3$

Consider the error model

$$\underbrace{\begin{pmatrix} e_1 \\ \vdots \\ e_m \end{pmatrix}}_{=\mathbf{e}} = \underbrace{\begin{pmatrix} y_1 - \psi_1(\mathbf{p}) \\ \vdots \\ y_m - \psi_m(\mathbf{p}) \end{pmatrix}}_{=\mathbf{f}_\mathbf{y}(\mathbf{p})}$$

The data y_i is an *inlier* if $e_i \in [e_i]$ and an *outlier* otherwise. We assume that

$$\forall i,\ \Pr(e_i \in [e_i]) = \pi$$

and that all e_i's are independent.

Equivalently,

$$\begin{cases} y_1 - \psi_1(\mathbf{p}) \in [e_1] & \text{with a probability } \pi \\ \vdots & \vdots \\ y_m - \psi_m(\mathbf{p}) \in [e_m] & \text{with a probability } \pi \end{cases}$$

The probability of having k inliers is

$$\frac{m!}{k!\,(m-k)!}\pi^k.(1-\pi)^{m-k}.$$

The probability of having strictly more than q outliers is thus

$$\gamma(q,m,\pi) = \sum_{k=0}^{m-q-1} \frac{m!}{k!\,(m-k)!}\pi^k.(1-\pi)^{m-k}.$$

Denote by $\mathbb{E}^{\{q\}}$ the set of all $\mathbf{e} \in \mathbb{R}^m$ consistent with at least $m - q$ error intervals $[e_i]$.
For $m = 3$, we have

$$\begin{array}{rcl}
\mathbb{E}^{\{0\}} & = & [e_1] \times [e_2] \times [e_3] \\
\mathbb{E}^{\{1\}} & = & ([e_1] \cap [e_2]) \cup ([e_2] \cap [e_3]) \cup ([e_1] \cap [e_3]) \\
\mathbb{E}^{\{2\}} & = & [e_1] \cup [e_2] \cup [e_3] \\
\mathbb{E}^{\{3\}} & = & \mathbb{R}^3.
\end{array}$$

$$\mathbb{P}^{\{q\}} = \mathbf{f}_{\mathbf{y}}^{-1}\left(\mathbb{E}^{\{q\}}\right) = \bigcap_{i\in\{1,\ldots,m\}}^{\{q\}} f_{y_i}^{-1}\left([e_i]\right).$$

Application to localization

A robot measures distances to three beacons.

i	x_i	y_i	$[d_i]$
1	1	3	$[1,2]$
2	3	1	$[2,3]$
3	-1	-1	$[3,4]$

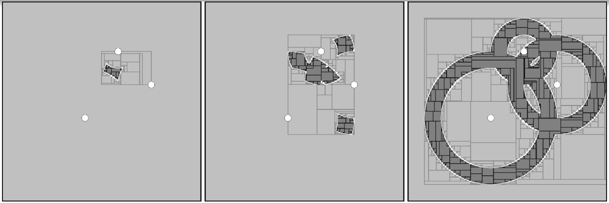
The intervals $[d_i]$ contain the true distance with a probability of $\pi = 0.9$.

The feasible sets associated to each data is

$$\mathbb{P}_i = \left\{ \mathbf{p} \in \mathbb{R}^2 \mid \sqrt{(p_1 - x_i)^2 + (p_2 - y_i)^2} - d_i \in [-0.5, 0.5] \right\},$$

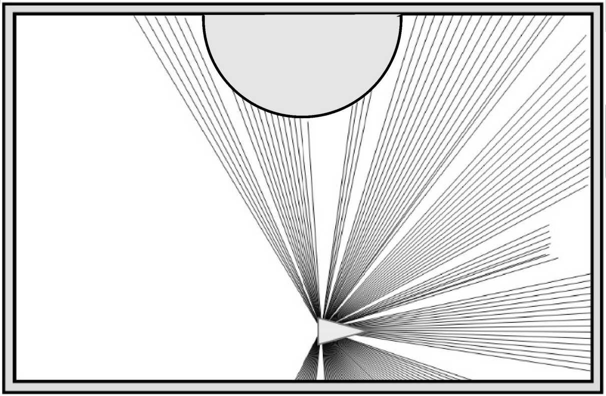
where $d_1 = 1.5, d_2 = 2.5, d_3 = 3.5.$

$$\begin{aligned}
\text{prob}\left(\mathbf{p} \in \mathbb{P}^{\{0\}}\right) &= 0.729 \\
\text{prob}\left(\mathbf{p} \in \mathbb{P}^{\{1\}}\right) &= 0.972 \\
\text{prob}\left(\mathbf{p} \in \mathbb{P}^{\{2\}}\right) &= 0.999
\end{aligned}$$



With real data


ENSTA
Bretagne



For $q = 16, m = 143, \pi = 0.95$, the probability of being wrong is

$$\alpha = \gamma(q, m, \pi) = 8.46 \times 10^{-4}.$$