# (Meta)Modeling with Gaussian processes: An overview

O. Roustant[a]

with contributions from: M. Binois, Y. Deville, N. Durrande, D. Ginsbourger, R. Le Riche, A. Lopez Lopéra, E. Padonou

[a] Mines Saint-Étienne

—

SFdS, Paris Saclay, 2018 May
50èmes journées de statistique

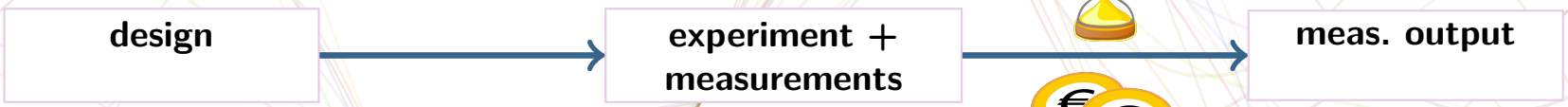
# Outline

# Outline

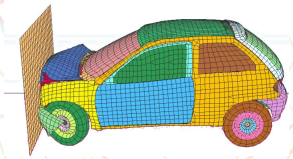
# Metamodeling – Computer experiments

**design**

# Metamodeling – Computer experiments



# Metamodeling – Computer experiments



**design**

**experiment + measurements**

**meas. output**

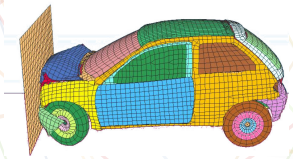
**simulator inputs**
$\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$

**simulator**
*physical model, numerical methods*

**simulator output**
$f_{\text{sim}}(\mathbf{x}^1), f_{\text{sim}}(\mathbf{x}^2), \ldots, f_{\text{sim}}(\mathbf{x}^n)$

# Metamodeling – Computer experiments

**design** → **experiment + measurements** → **meas. output**



**simulator inputs** $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ → **simulator** *physical model, numerical methods* → **simulator output** $f_{\text{sim}}(\mathbf{x}^1), f_{\text{sim}}(\mathbf{x}^2), \ldots, f_{\text{sim}}(\mathbf{x}^n)$

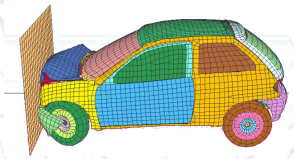
**metamodel inputs** $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ → **metamodel** $y = \mathbf{g}(\mathbf{x})^\top \beta + Z(\mathbf{x})$ → **metamodel output** $\widehat{y}(\mathbf{x}), \ldots$

# Metamodeling – Computer experiments

**design**

**experiment + measurements**

**meas. output**



**simulator inputs**
$\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$

**simulator**
*physical model, numerical methods*

**simulator output**
$f_{\text{sim}}(\mathbf{x}^1), f_{\text{sim}}(\mathbf{x}^2), \ldots, f_{\text{sim}}(\mathbf{x}^n)$

param.

design

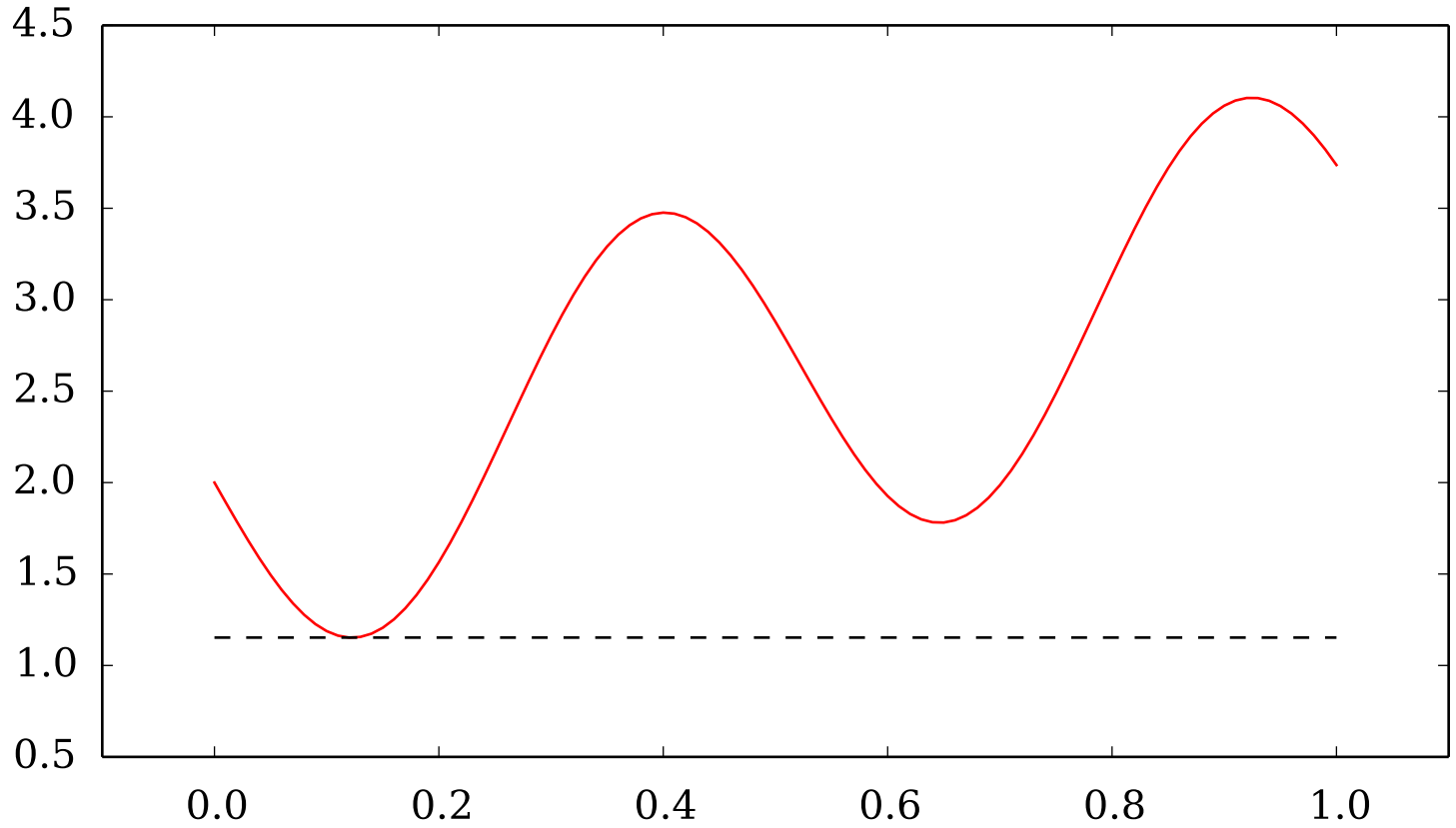
**metamodel inputs**
$\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$

**metamodel**
$y = \mathbf{g}(\mathbf{x})^\top \beta + Z(\mathbf{x})$

**metamodel output**
$\widehat{y}(\mathbf{x}), \ldots$

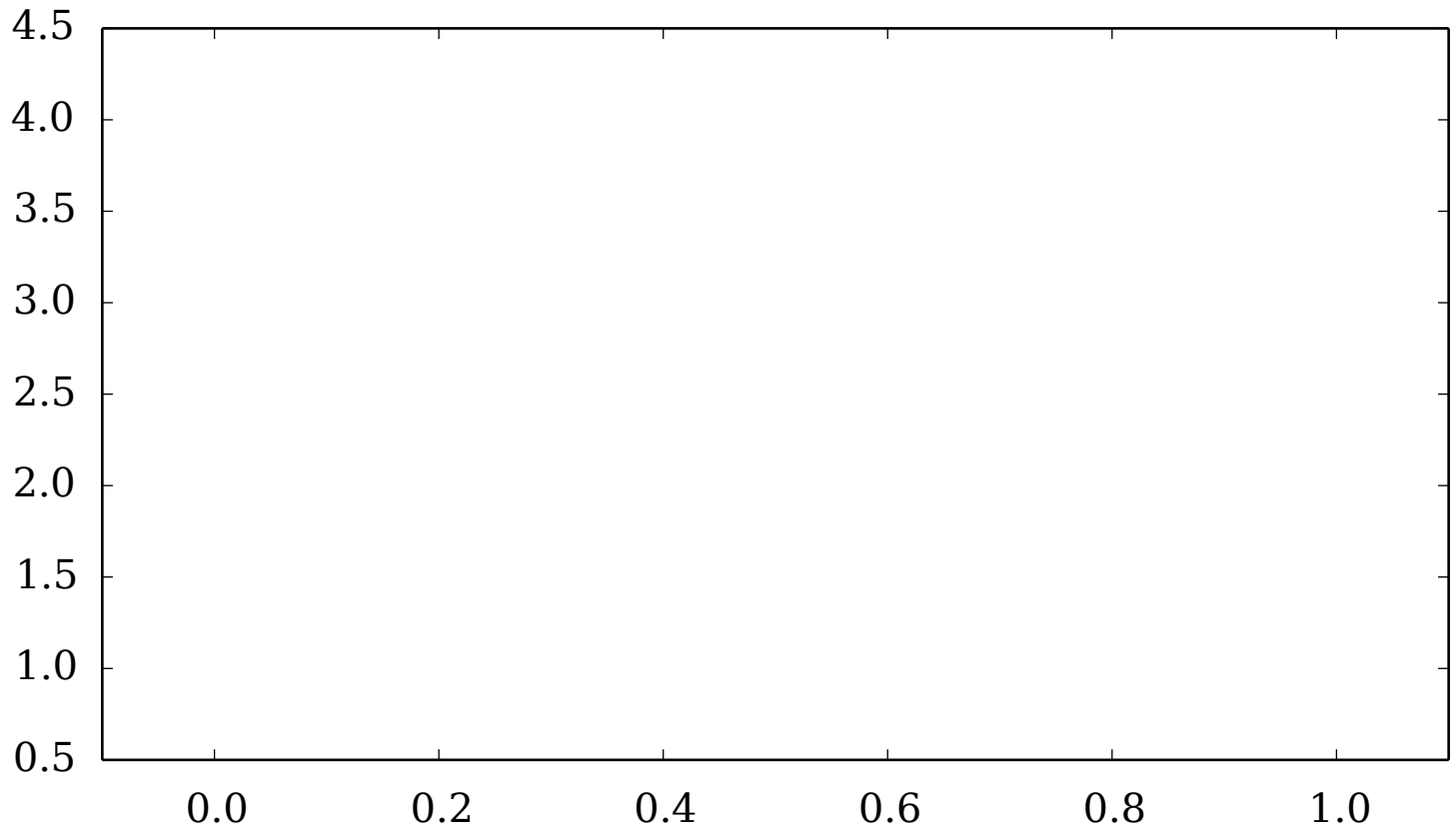
# GP-based optimization

How to find the global minimum of a function... when each evaluation is costly ?



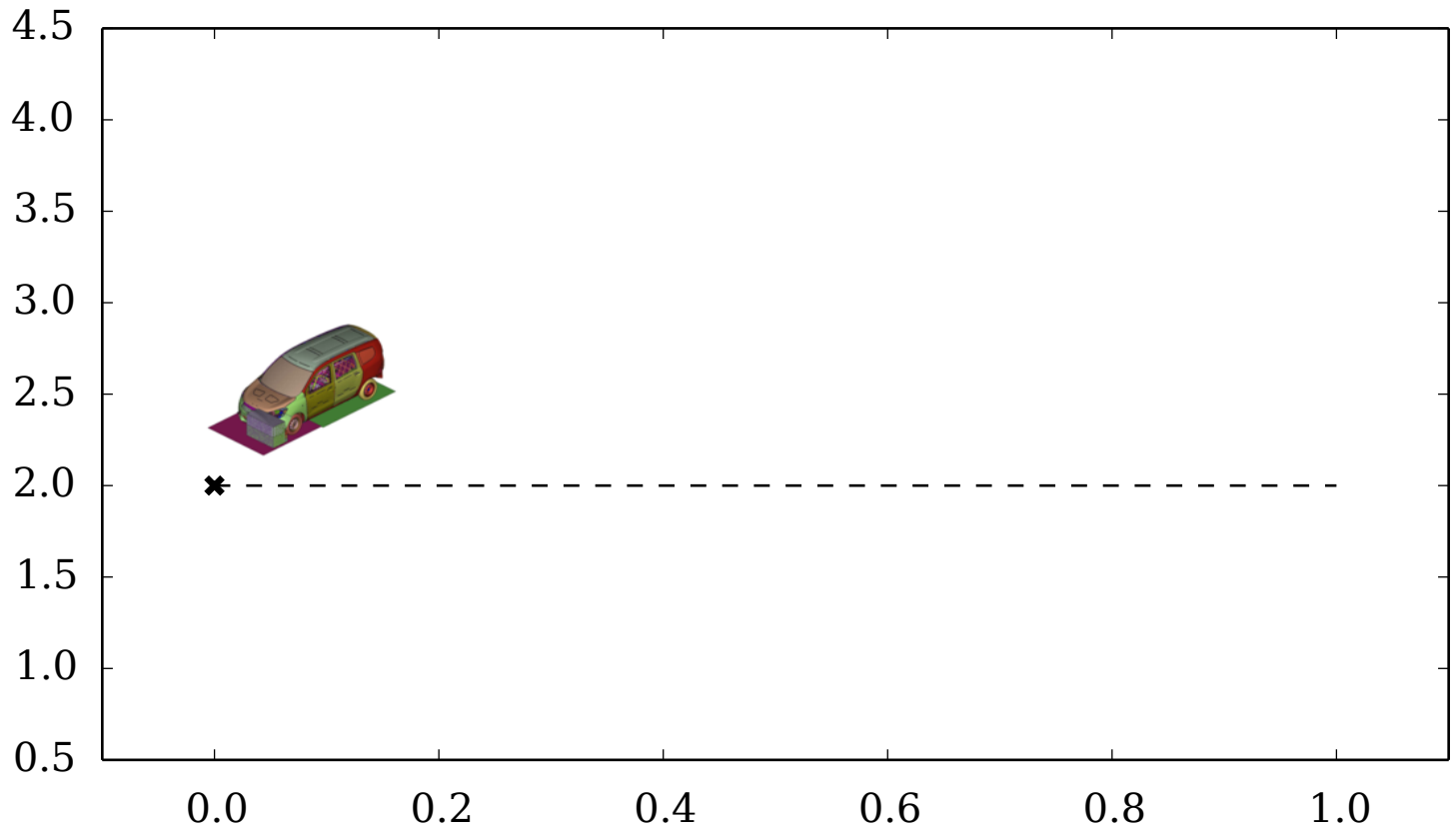
# GP-based optimization

How to find the global minimum of a function... when each evaluation is costly ?



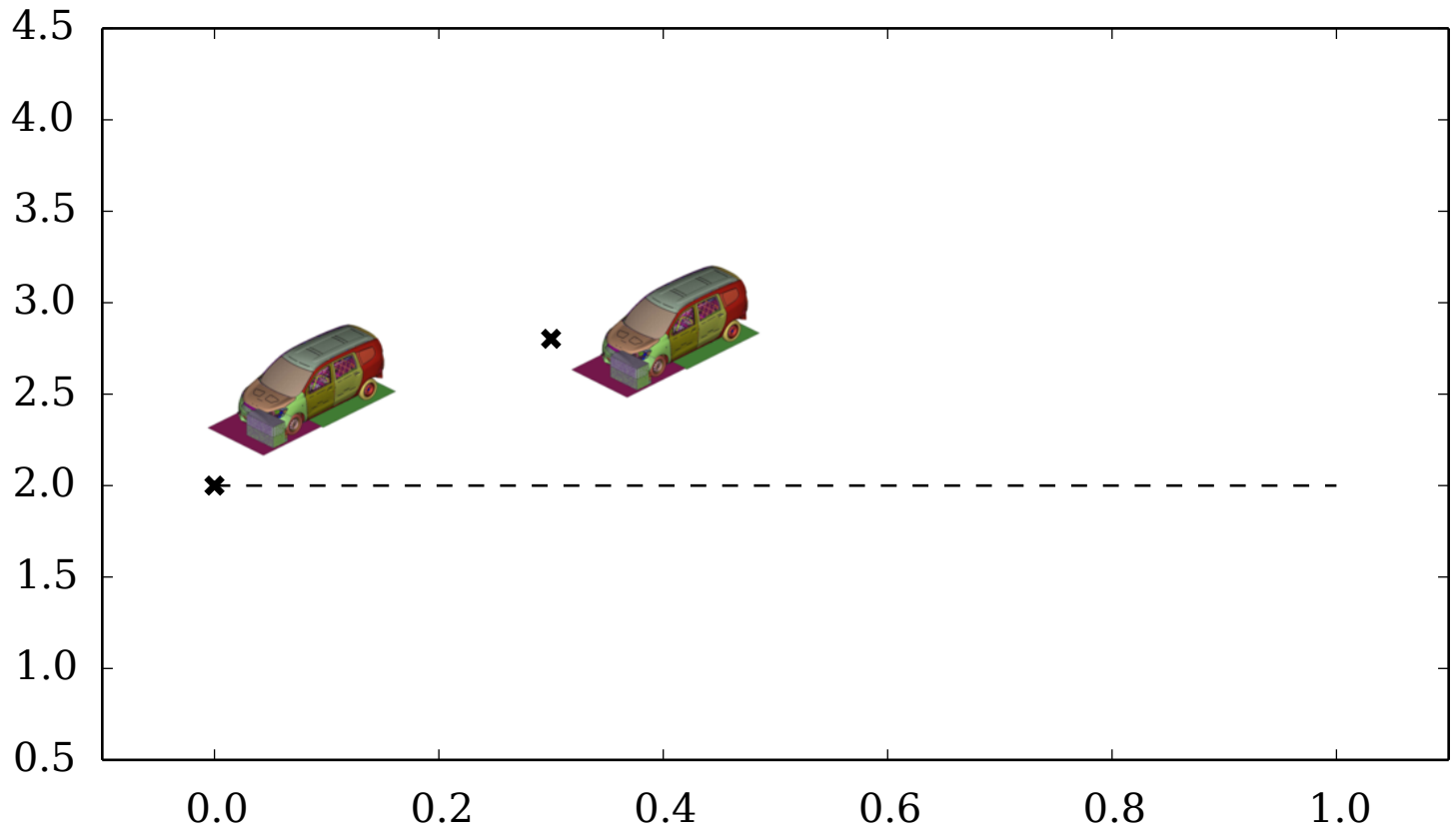
# GP-based optimization

How to find the global minimum of a function... when each evaluation is costly ?



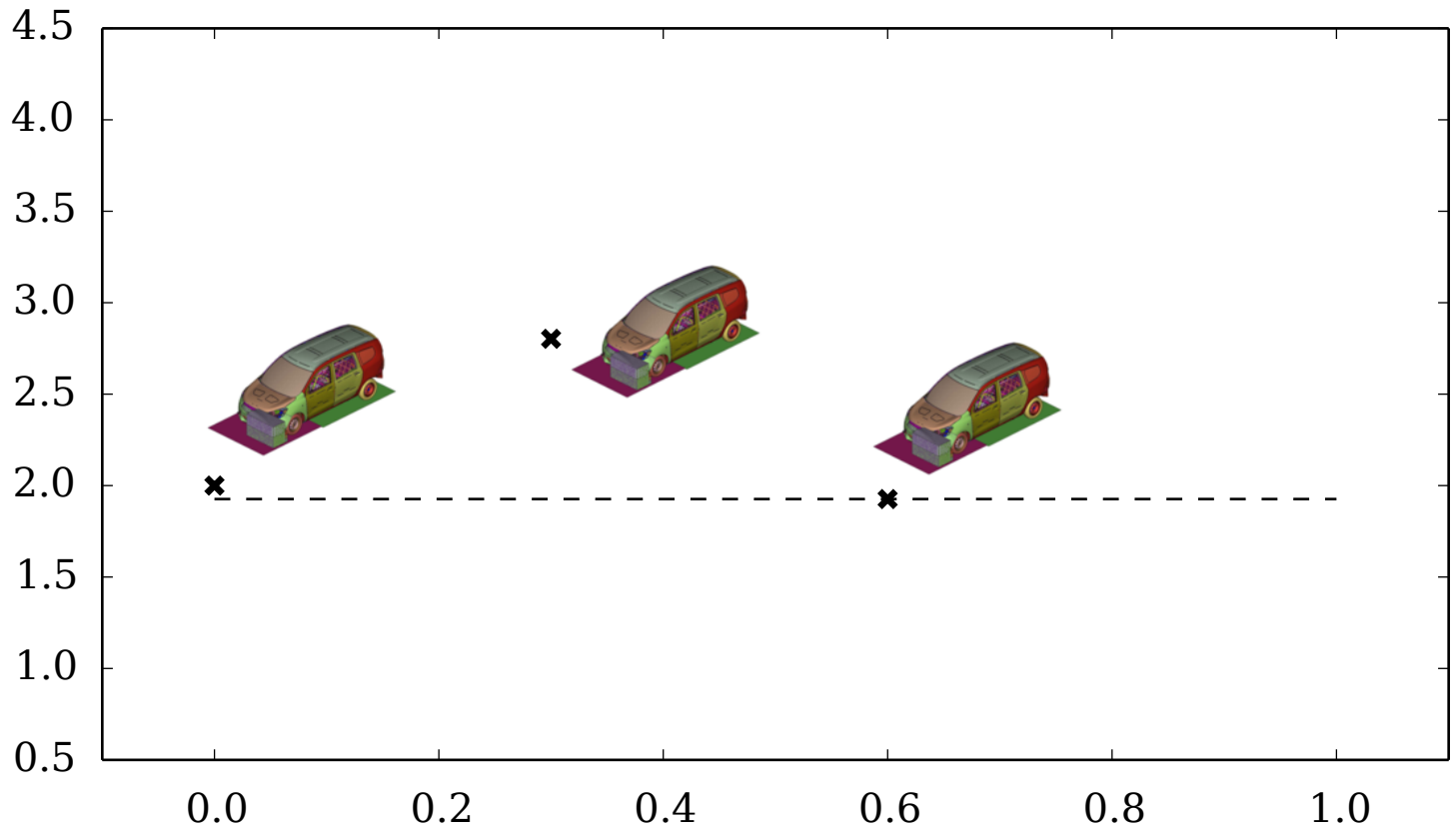
# GP-based optimization

How to find the global minimum of a function... when each evaluation is costly ?



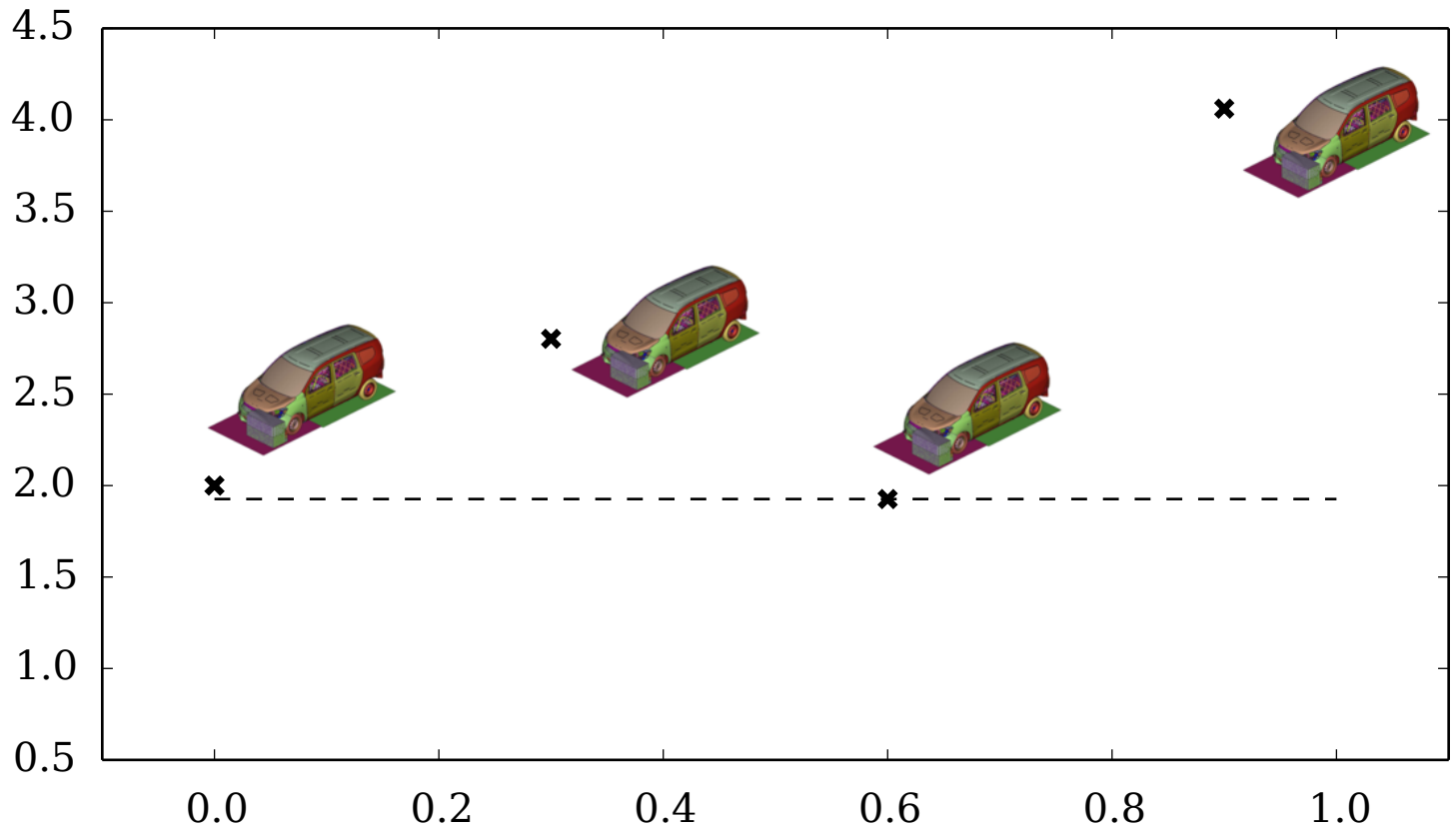
# GP-based optimization

How to find the global minimum of a function... when each evaluation is costly ?



# GP-based optimization

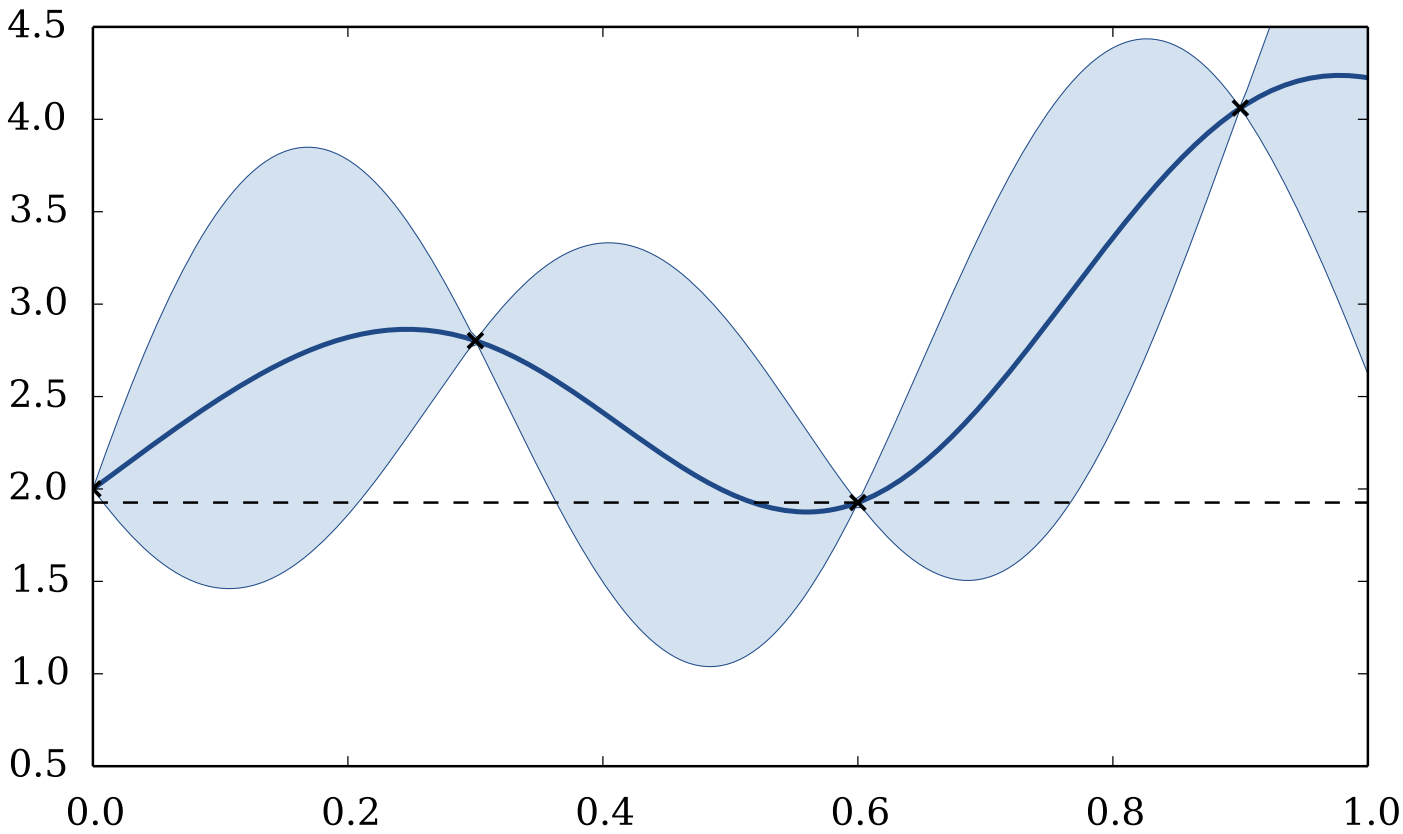
How to find the global minimum of a function... when each evaluation is costly ?



# GP-based optimization

A solution : **GP-based (or "Bayesian") optimization** [Močkus, 1975, Jones et al., 1998]

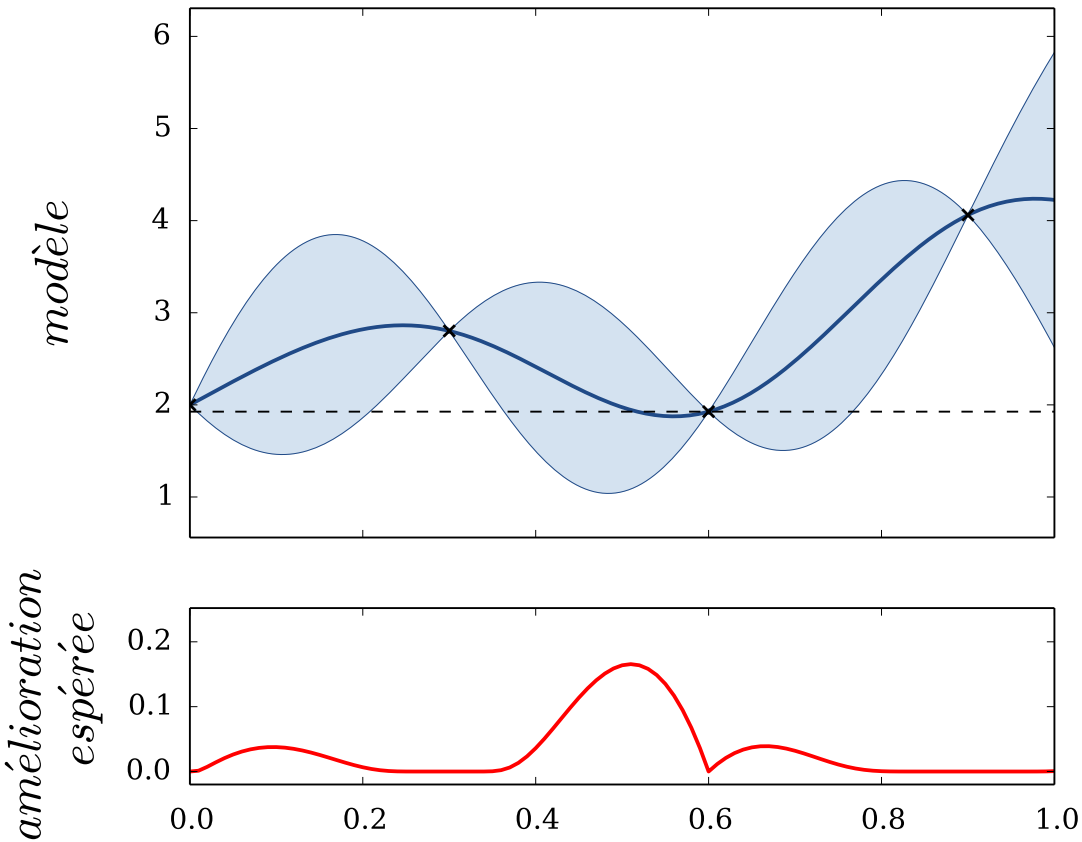
First ingredient : a GP model $Y$



# GP-based optimization

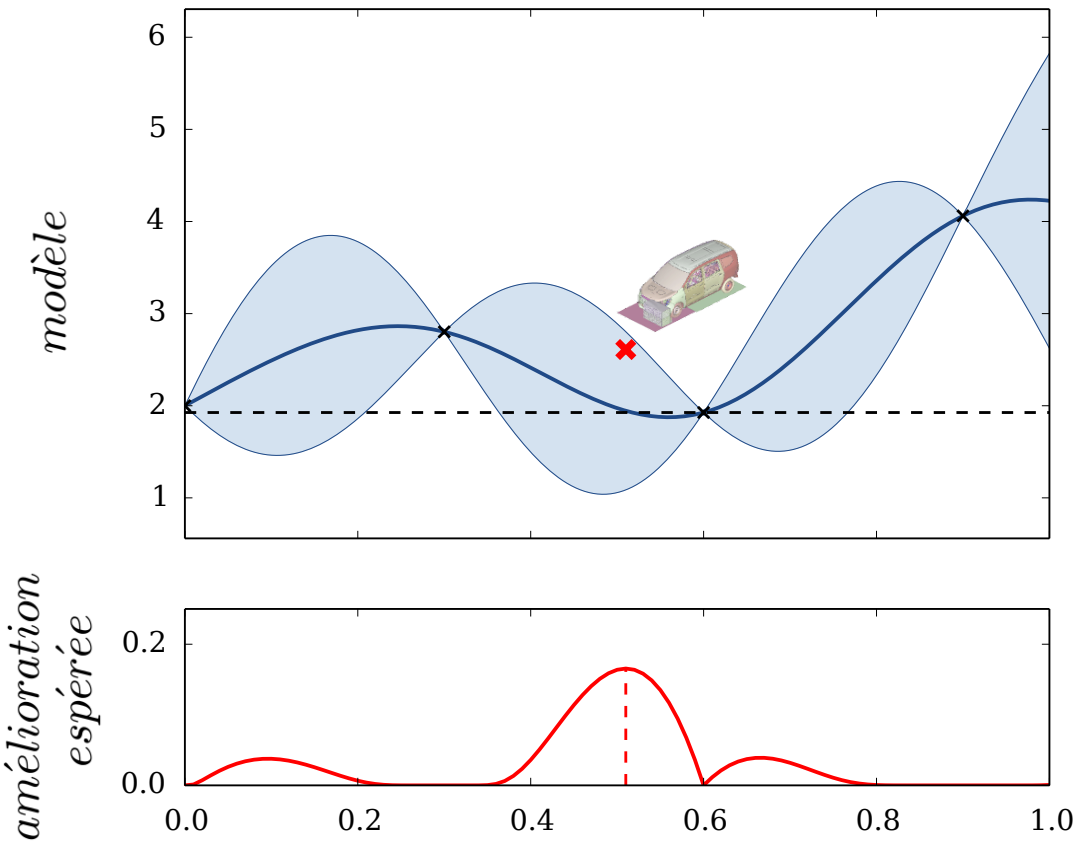
Second ingredient : an **easy-to-compute** criterion **accounting for uncertainty at unknown regions**, e.g. here "expected improvement"

$$EI(x) = \mathrm{E}([f_0 - Y(x)]^+ | Y(x_1), \ldots, Y(x_n)) \qquad f_0 : \textit{ current minimum}$$



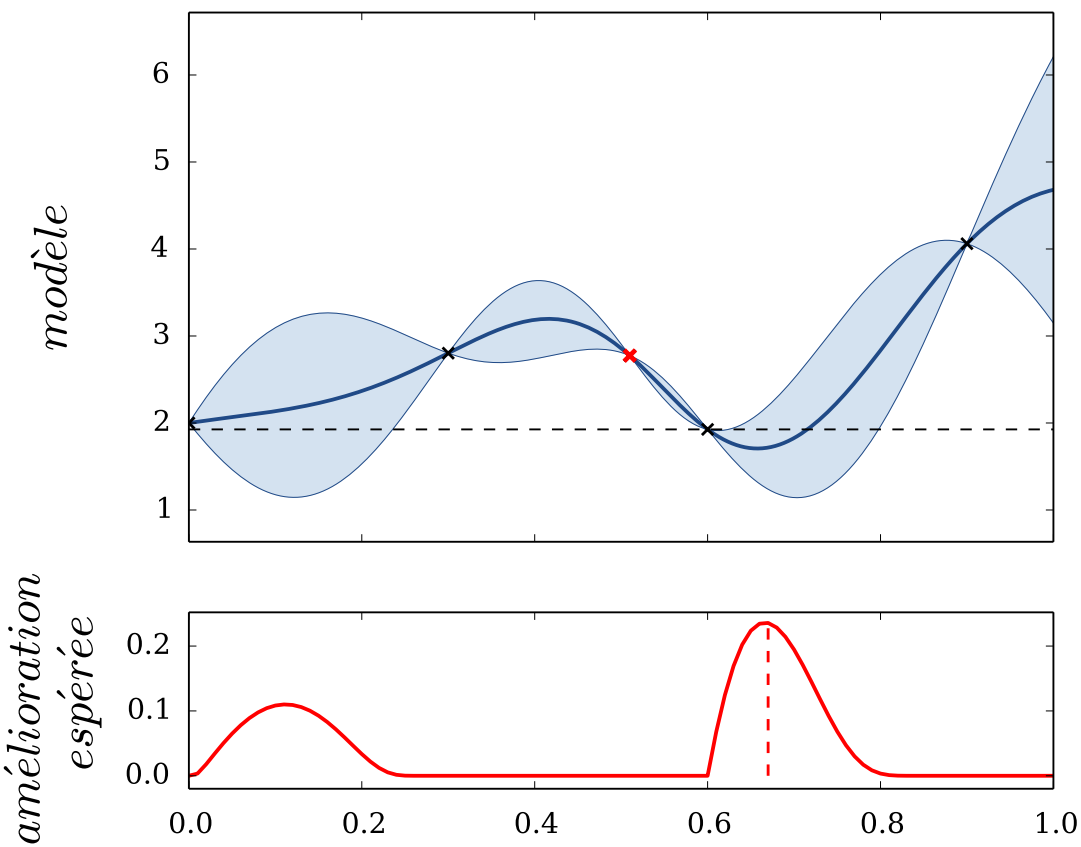
# GP-based optimization

The algorithm (here "EGO") : (1) Find the next point by maximizing the criterion $\rightarrow$ (2) Evaluate the function $\rightarrow$ (3) Update the GP model $\uparrow$



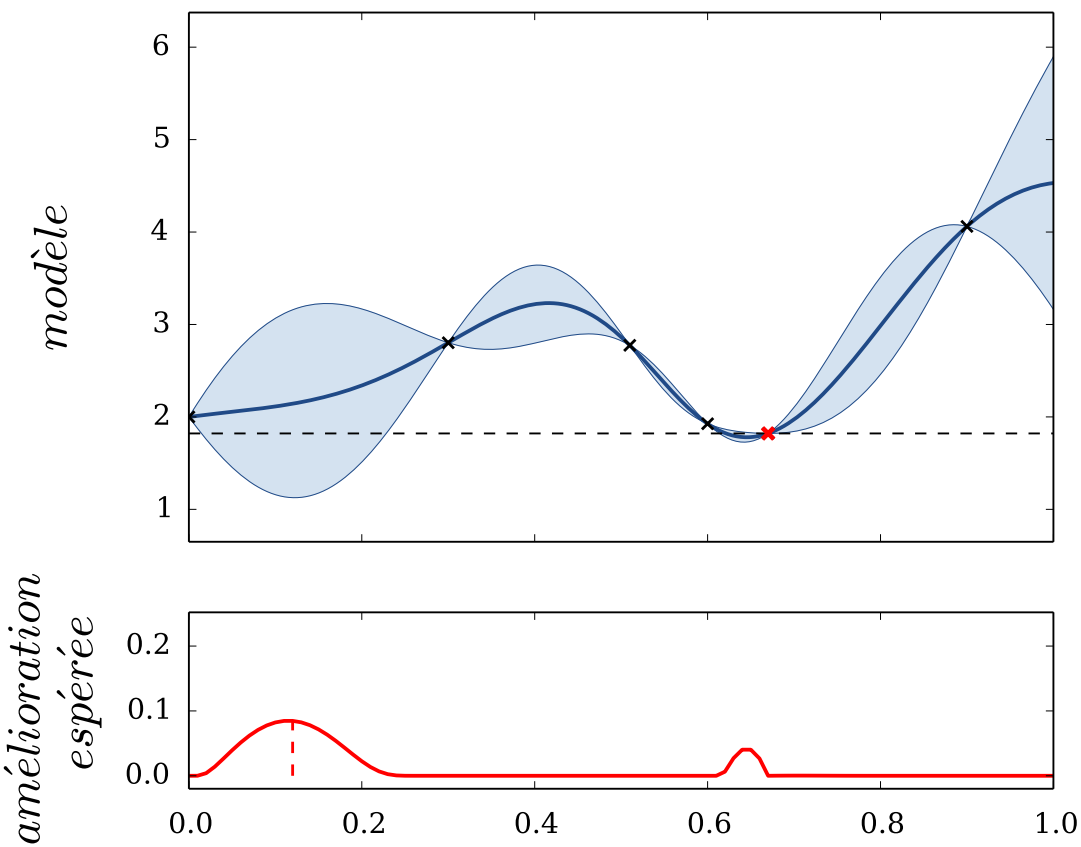
# GP-based optimization

Iteration 2 :



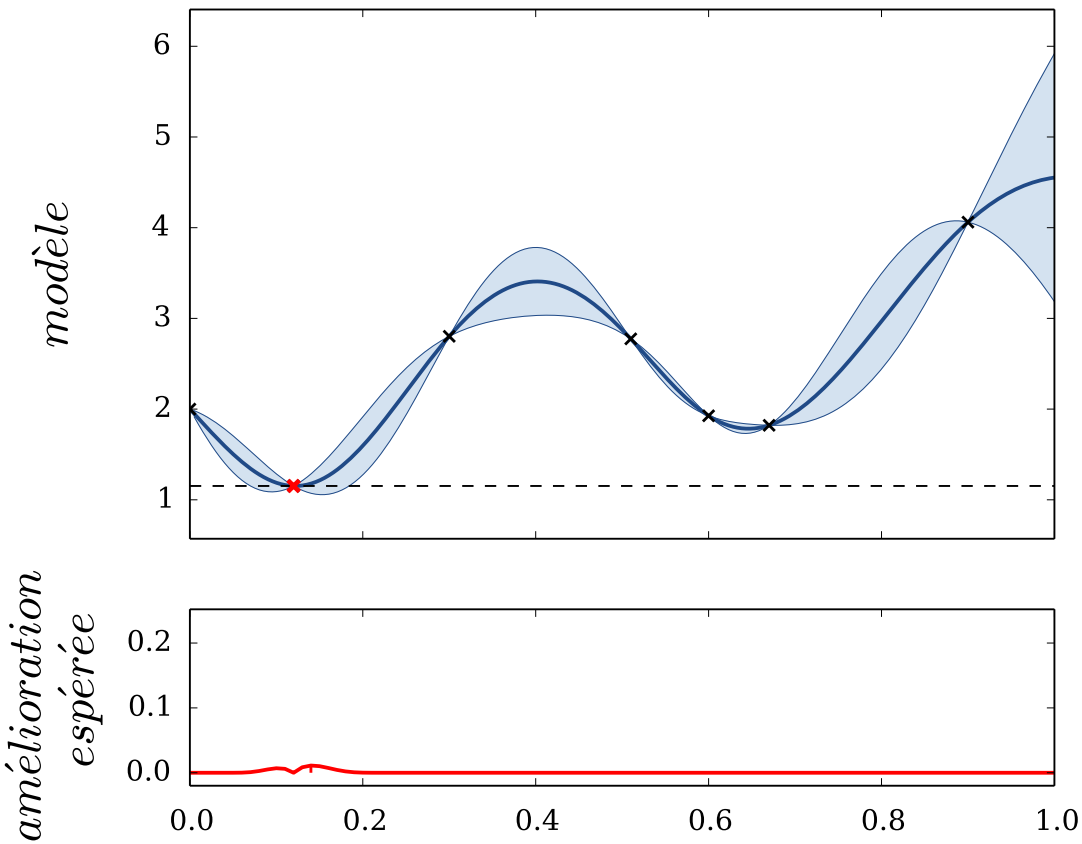
# GP-based optimization

Iteration 3 :



# GP-based optimization

Theory shows that **EGO algorithm provides a dense sequence of points**, up to a slight condition on the kernel used for GPs [Vazquez and Bect, 2010] .

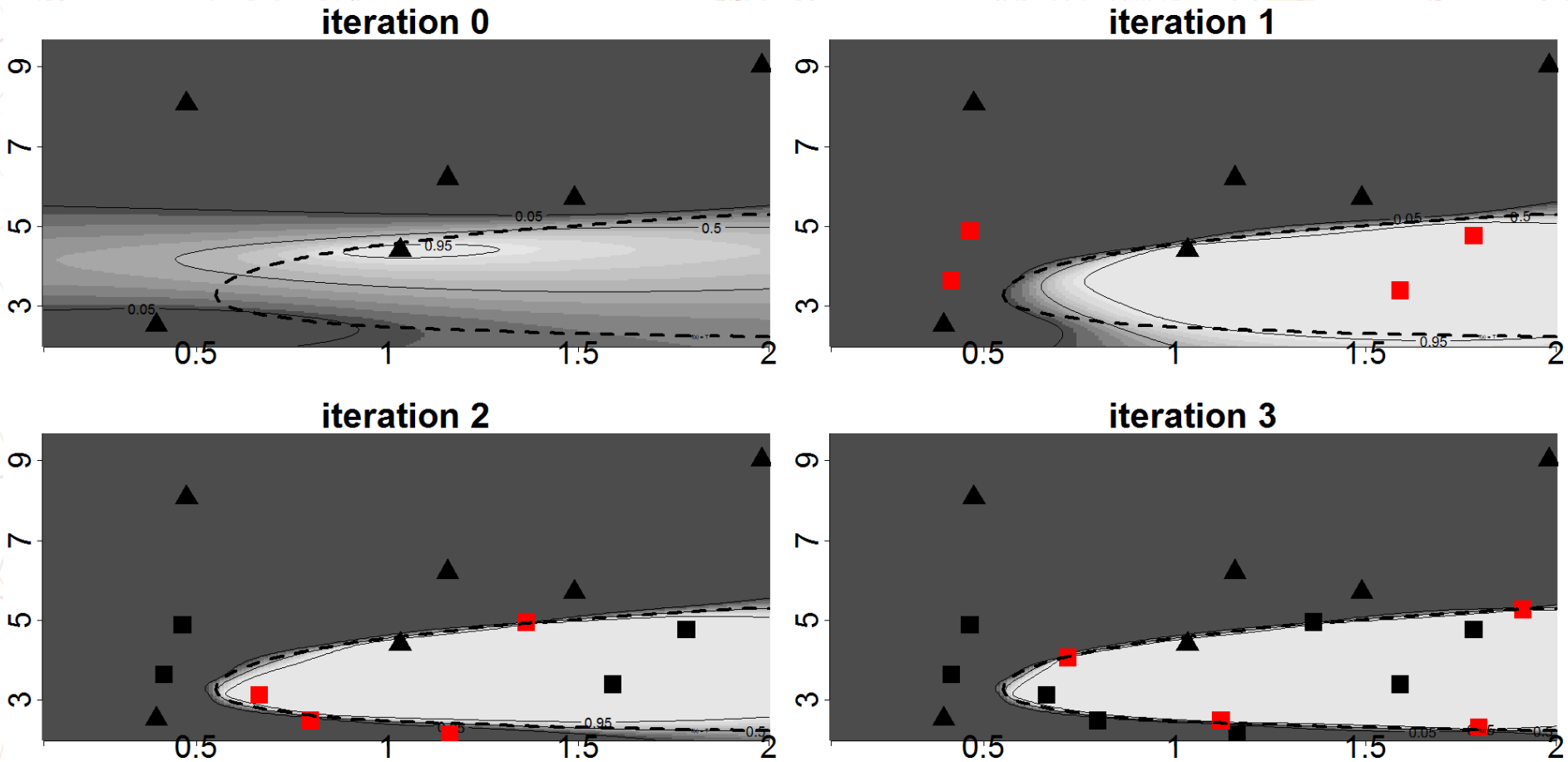


## GP-based inversion

Same receipt for estimating a probability of failure ("SUR" strategy).

See [Chevalier et al., 2014] for details and [Bect et al., 2017] for a convergence analysis with supermartingales.

Illustration : Estimation of the nuclear criticity region $k_{\text{eff}} > 0.95$



# From geostatistics to metamodels

## Time line

- 1951 : Spatial interpolation in geosciences [Krige, 1951]
  $\rightarrow$ "Kriging"
- 1963 : Foundations of geostatistics [Matheron, 1963]
- 1989 : Computer experiments, metamodelling [Sacks et al., 1989]
  $\rightarrow$ Application to dimensions $\geq 4$

## Some books about metamodelling with GPs

- "The Design and Analysis of Computer Experiments" [Santner et al., 2003]
- "Design and Modeling for Computer Experiments" [Fang et al., 2006]
- "Gaussian process for machine learning" [Rasmussen and Williams, 2006]

## Gaussian processes

Gaussian processes are stochastic processes (or random fields) s.t. every finite dimensional distribution is Gaussian. $\rightarrow$ Parameterized by two functions

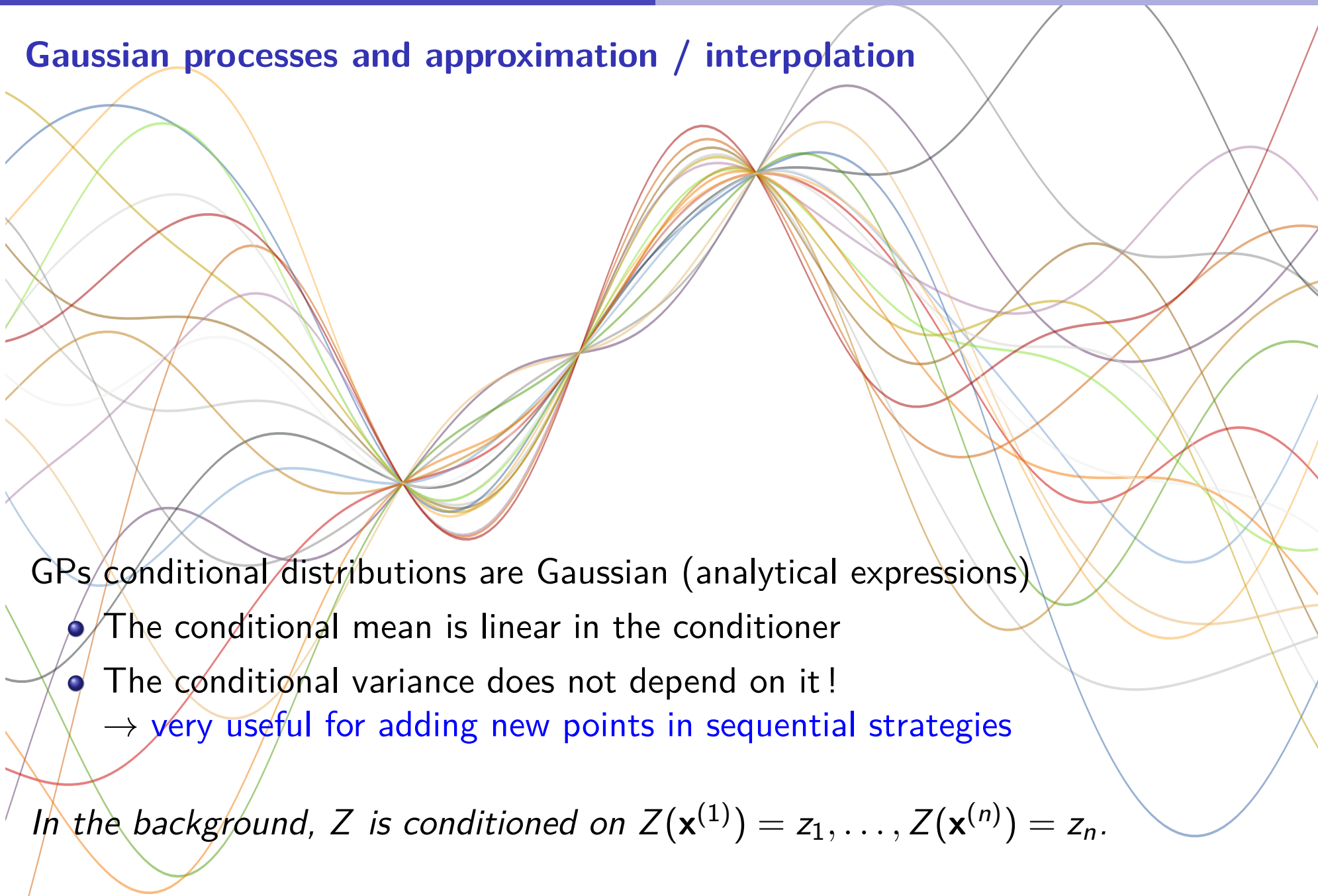
$$Z_{\mathbf{x}} \sim GP(\underbrace{m(\mathbf{x})}_{trend}, \underbrace{k(\mathbf{x}, \mathbf{x}')}_{kernel})$$

- The trend can be any function.
- The kernel is positive semidefinite :

$$\forall n, \alpha_1, \ldots, \alpha_n, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}, \qquad \sum_{i=1}^{n} \alpha_i \alpha_j k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \geq 0.$$

It contains the spatial dependence.

# Gaussian processes and approximation / interpolation



GPs conditional distributions are Gaussian (analytical expressions)

- The conditional mean is linear in the conditioner
- The conditional variance does not depend on it !
  $\rightarrow$ very useful for adding new points in sequential strategies

*In the background, $Z$ is conditioned on $Z(\mathbf{x}^{(1)}) = z_1, \ldots, Z(\mathbf{x}^{(n)}) = z_n$.*

# Gaussian processes, splines and RKHS

## The 3 faces of a kernel

$GP(0, k(\mathbf{x}, \mathbf{x}')) \quad \Leftrightarrow \quad$ p.s.d. functions $k \quad \Leftrightarrow \quad$ RKHS : $\mathcal{H} = \overline{\text{span}\{k(., \mathbf{x}), \mathbf{x} \in D\}}$

where $\mathcal{H}$ is a "Reproducing Kernel" Hilbert Space with dot product :

$$\langle k(\mathbf{x}, .), k(\mathbf{x}', .) \rangle = k(\mathbf{x}, \mathbf{x}') \qquad (*)$$

*RKHS can be also defined as Hilbert spaces of functions such that evaluations $f \to f(\mathbf{x})$ are continuous : By Riesz theorem, there exists a unique $k(., \mathbf{x})$ s.t.*

$$f(\mathbf{x}) = \langle f, k(., \mathbf{x}) \rangle$$

*Choosing $f = k(., \mathbf{x}')$ gives the reproducing identity (*).*

*Ref : [Aronszajn, 1950], [Berlinet and Thomas-Agnan, 2011].*

# Gaussian processes, splines and RKHS

## Correspondence between interpolation spline and GP conditional mean

**[Kimeldorf and Wahba, 1971]**

The interpolation spline is defined by the functional problem

$$(*) \qquad \min_{h \in \mathcal{H}} \|h\| \qquad s.t. \quad h(\mathbf{x}^{(i)}) = z_i,\ i = 1, \ldots, n$$

If $\mathcal{H}$ is the RKHS of kernel $k$, and if $K = (k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}))_{1 \leq i,j \leq n}$ is invertible, then (*) has a unique solution in the finite dimensional space spanned by the $k(., \mathbf{x}^{(i)})$ :

$$h_{\text{opt}}(\mathbf{x}) = \mathrm{E}\left[ Z_{\mathbf{x}} \,\middle|\, Z(\mathbf{x}^{(i)}) = z_i,\ i = 1, \ldots, n \right]$$

$\rightarrow$ In this sense, GPs are generalizing interpolation splines.

*The first part (reduction to finite dimension) is know as Representer theorem.*

# Playing with kernels

A lot of flexibility can be obtained with kernels !

## Building a kernel from other ones (basic examples)

| | |
|---|---|
| Sum, tensor sum | $k_1 + k_2,\ k_1 \oplus k_2$ |
| Product, tensor product | $k_1 \times k_2,\ k_1 \otimes k_2$ |
| ANOVA | $(1 + k_1) \otimes (1 + k_2)$ |
| Warping | $k(\mathbf{x}, \mathbf{x}') = k_1(f(\mathbf{x}), f(\mathbf{x}'))$ |
| ... | ... |

*See examples in [Rasmussen and Williams, 2006]*... and in this talk !

# Why Gaussian processes are popular in (geo)statistics / machine learning ?

## GP strengths

- Probabilistic models that **traduce spatial dependence**
  $\rightarrow$ provide realistic uncertainty in unvisited area
- **Conditional distributions are analytical**, and the cond. var. is constant
  $\rightarrow$ Useful for prediction and sequential strategies
- **Parameterized by functions** : mean and covariance **(kernel)**
  $\rightarrow$ Flexibility
- At the crossing between **rich mathematical theories**
  $\rightarrow$ Stochastic proc., Reproducing Kernel Hilbert Spaces, Positive Definite Functions

## Aim of this talk

*To illustrate the flexibility of GPs in various (meta)modeling situations.*

# Outline

# Various information... same framework

Question : What is the common point between GPs whose sample paths are...

| | |
|---|---|
| Centered | $\int Z_{\mathbf{x}}\, \mu(d\mathbf{x}) = 0$ |
| Harmonic | $\frac{\partial^2 Z_{\mathbf{x}}}{\partial x_1^2} + \frac{\partial^2 Z_{\mathbf{x}}}{\partial x_2^2} = 0$ |
| Symmetric | $Z_{g.\mathbf{x}} = Z_{\mathbf{x}} \quad \forall g \in G$ |
| Additive | $Z_{\mathbf{x}} = \sum_{j=1}^{d} Z_{x_j}^j$ |
| ... | |

$G$ : Finite group of symmetries.

# Various information... same framework

Answer : The information can be reformulated$^\star$ with a linear operator $T$

| | | |
|---|---|---|
| Centered | $\int Z_{\mathbf{x}}\, \mu(d\mathbf{x}) = 0$ | $T(f) = \int f(\mathbf{x})\, \mu(d\mathbf{x})$ |
| Harmonic | $\frac{\partial^2 Z_{\mathbf{x}}}{\partial x_1^2} + \frac{\partial^2 Z_{\mathbf{x}}}{\partial x_2^2} = 0$ | |

# Various information... same framework

Answer : The information can be reformulated$^\star$ with a linear operator $T$

| | | |
|---|---|---|
| Centered | $\int Z_{\mathbf{x}}\,\mu(d\mathbf{x}) = 0$ | $T(f) = \int f(\mathbf{x})\,\mu(d\mathbf{x})$ |
| Harmonic | $\frac{\partial^2 Z_{\mathbf{x}}}{\partial x_1^2} + \frac{\partial^2 Z_{\mathbf{x}}}{\partial x_2^2} = 0$ | $T(f) = \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) + \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x})$ |
| Symmetric | $Z_{g.\mathbf{x}} = Z_{\mathbf{x}} \quad \forall g \in G$ | |

# Various information... same framework

Answer : The information can be reformulated$^\star$ with a linear operator $T$

| | | |
|---|---|---|
| Centered | $\int Z_{\mathbf{x}}\,\mu(d\mathbf{x}) = 0$ | $T(f) = \int f(\mathbf{x})\,\mu(d\mathbf{x})$ |
| Harmonic | $\frac{\partial^2 Z_{\mathbf{x}}}{\partial x_1^2} + \frac{\partial^2 Z_{\mathbf{x}}}{\partial x_2^2} = 0$ | $T(f) = \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) + \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x})$ |
| Symmetric | $Z_{g.\mathbf{x}} = Z_{\mathbf{x}} \quad \forall g \in G$ | $T(f) = f(\mathbf{x}) - \frac{1}{\lvert G\rvert}\sum_{g\in G} f(g.\mathbf{x})$ |
| Additive | $Z_{\mathbf{x}} = \sum_{j=1}^{d} Z^j_{x_j}$ | |

# Various information... same framework

Answer : The information can be reformulated$^\star$ with a linear operator $T$

| | | |
|---|---|---|
| Centered | $\int Z_{\mathbf{x}}\,\mu(d\mathbf{x}) = 0$ | $T(f) = \int f(\mathbf{x})\,\mu(d\mathbf{x})$ |
| Harmonic | $\frac{\partial^2 Z_{\mathbf{x}}}{\partial x_1^2} + \frac{\partial^2 Z_{\mathbf{x}}}{\partial x_2^2} = 0$ | $T(f) = \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) + \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x})$ |
| Symmetric | $Z_{g.\mathbf{x}} = Z_{\mathbf{x}} \quad \forall g \in G$ | $T(f) = f(\mathbf{x}) - \frac{1}{\lvert G\rvert}\sum_{g\in G} f(g.\mathbf{x})$ |
| Additive | $Z_{\mathbf{x}} = \sum_{j=1}^{d} Z^j_{x_j}$ | $T(f) = f(\mathbf{x}) - \left(m + \sum_{j=1}^{d}(E[f(\mathbf{X})\lvert X_j = x_j] - m)\right)$ |
| ... | | with $m = \mathrm{E}[f(\mathbf{X})]$ and $X_1, \ldots, X_n$ ind. r.v. |

$^\star$ The reformulation is partial for the symmetric condition.

## Two results for GPs constrained by linear equalities [Ginsbourger et al., 2016]

- $Z : GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$
- $T$ : linear operator on $Y$ such that $T(m) = 0$ ($+$ integrability condition)

### Argumentwise property for kernels

$T$ can be defined in a unique way on the RKHS corr. to $k$ and

$$\forall \mathbf{x} : T(Z)_{\mathbf{x}} = 0 \quad \Leftrightarrow \quad \forall \mathbf{x}' : T(k(., \mathbf{x}')) = 0$$

*Proof idea :* $Cov(T(Z)_{\mathbf{x}}, Z_{\mathbf{x}'}) = T(Cov(Z_{\mathbf{x}}, Z_{\mathbf{x}'}))$

## Two results for GPs constrained by linear equalities [Ginsbourger et al., 2016]

- $Z : GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$
- $T$ : linear operator on $Y$ such that $T(m) = 0$ ($+$ integrability condition)

### Argumentwise property for kernels

$T$ can be defined in a unique way on the RKHS corr. to $k$ and

$$\forall \mathbf{x} : T(Z)_{\mathbf{x}} = 0 \quad \Leftrightarrow \quad \forall \mathbf{x}' : T(k(., \mathbf{x}')) = 0$$

*Proof idea :* $Cov(T(Z)_{\mathbf{x}}, Z_{\mathbf{x}'}) = T(Cov(Z_{\mathbf{x}}, Z_{\mathbf{x}'}))$

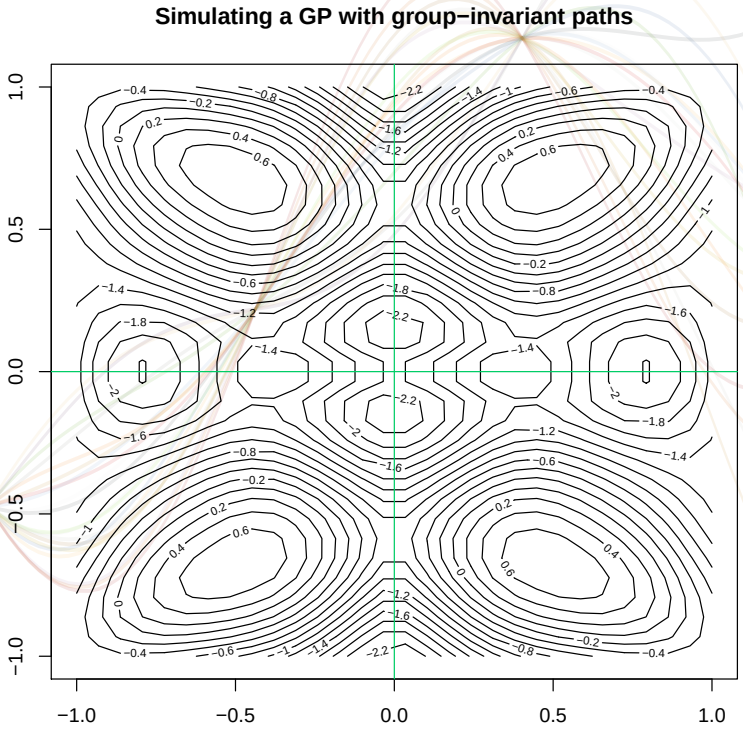
### Inheritance to conditional distributions

Let $Z^c$ the GP $Z$ conditional on $Z_{\mathbf{x}^{(i)}} = z_i$, $i = 1, \ldots, n$. Then

$$\forall \mathbf{x} : T(Z)_{\mathbf{x}} = 0 \quad \Rightarrow \quad \forall \mathbf{x} : T(Z^c)_{\mathbf{x}} = 0$$

*Proof idea : The cond. mean and cond. cov. are linear functions of* $k(., \mathbf{x}')$.

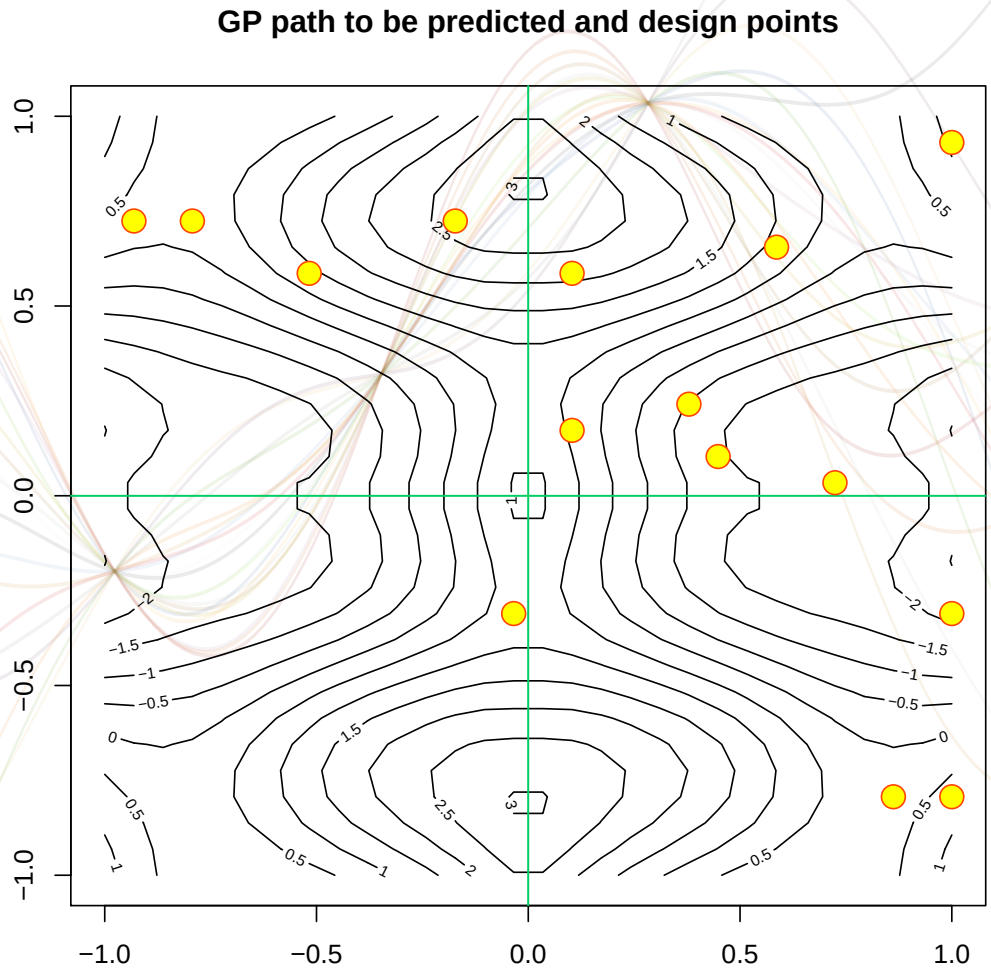
# Illustration : Symmetric GPs



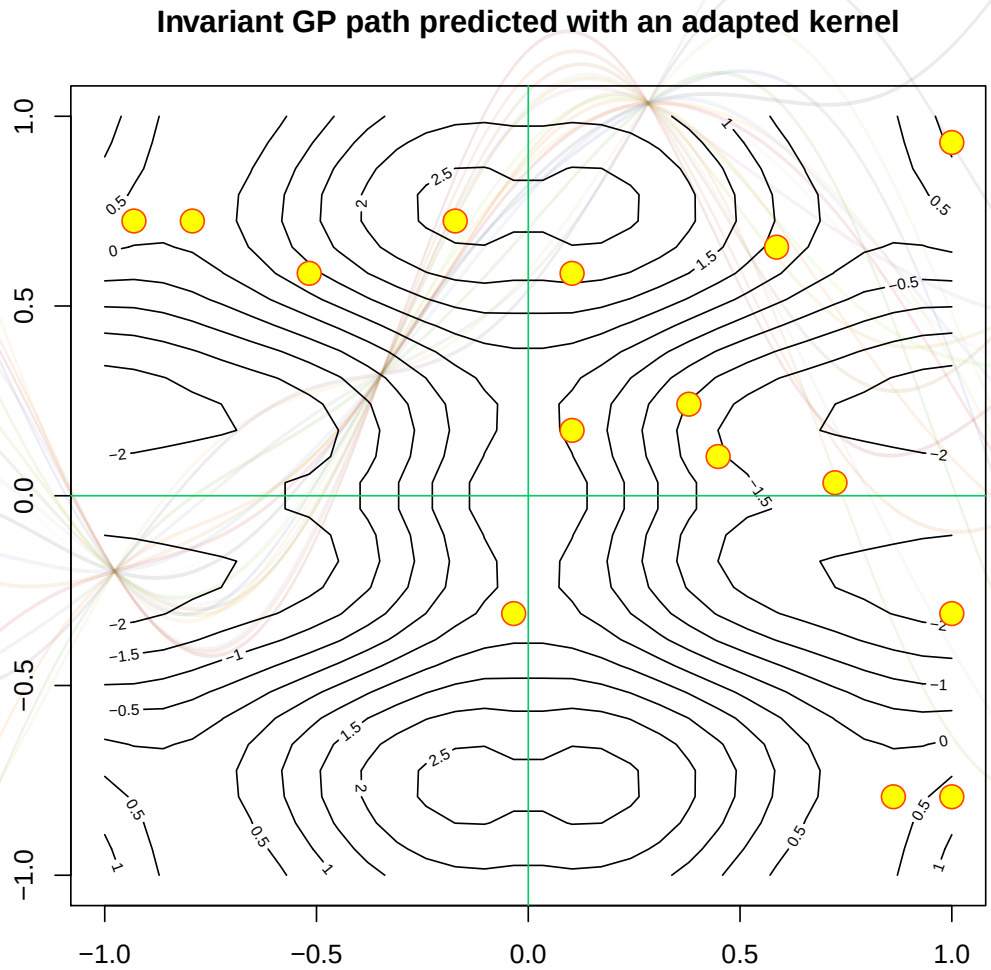

Making new from old : Choose a (non-symmetric) GP $Y$, and define

$$Z_{x_1,x_2} = \sum_{g \in G} Y_{g.\mathbf{x}} = Y_{x_1,x_2} + Y_{-x_1,x_2} + Y_{x_1,-x_2} + Y_{-x_1,-x_2}$$

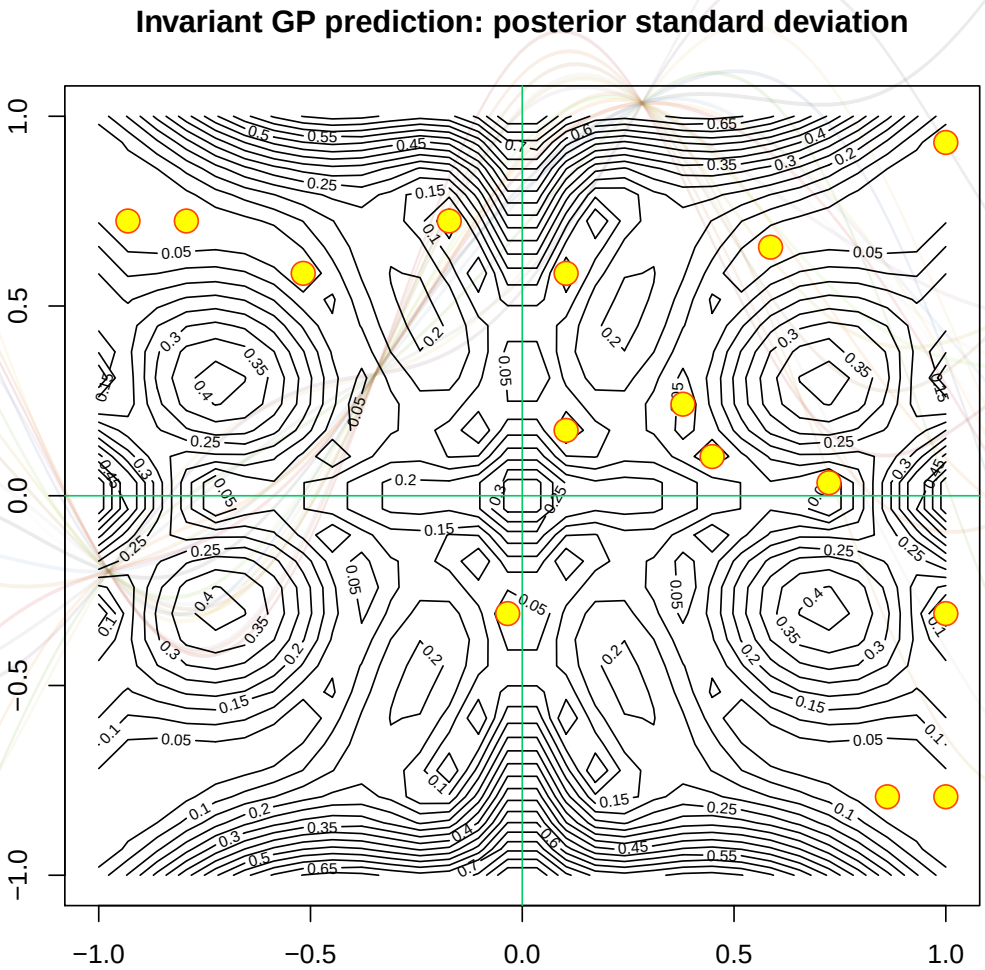
# Illustration : Symmetric GPs



# Illustration : Symmetric GPs



# Illustration : Symmetric GPs

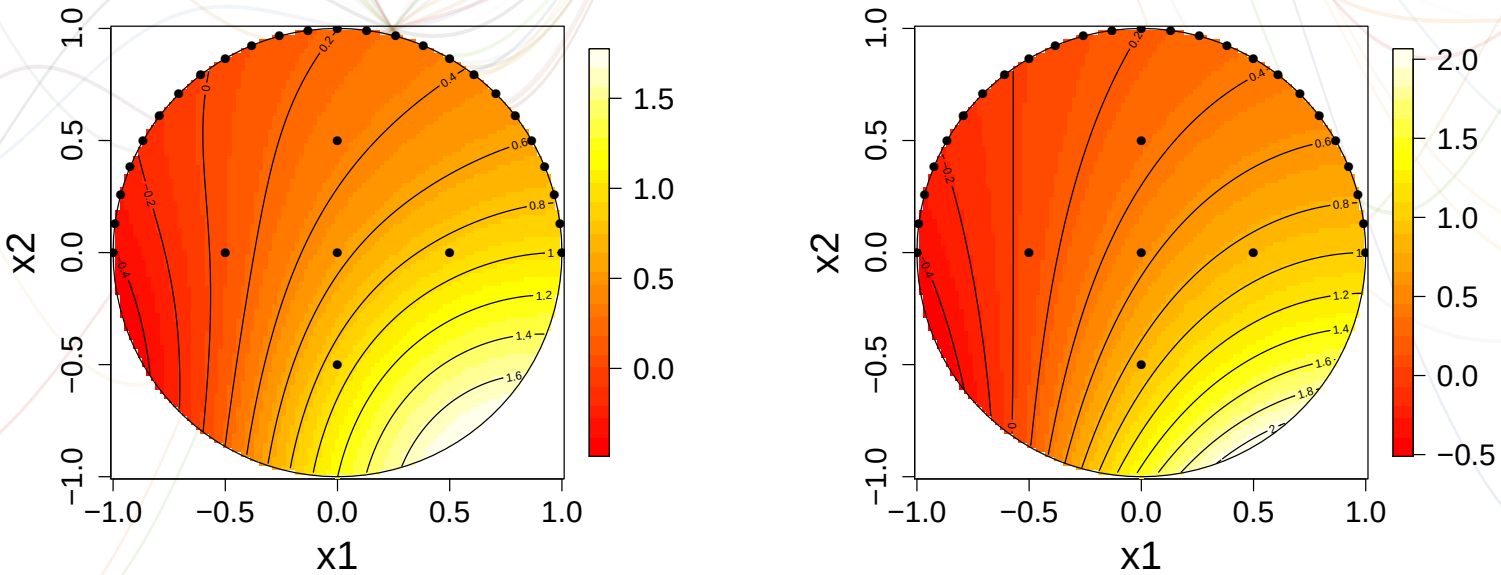


# Illustration : Maximum of a harmonic function [Ginsbourger et al., 2016]

We compare two GPs for predicting the maximum of a 2D harmonic function

Gaussian kernel $$k(\mathbf{x},\mathbf{x}') = \sigma^2 \exp\left(-\frac{(x_1-x_1')^2}{\ell_1^2} - \frac{(x_2-x_2')^2}{\ell_2^2}\right)$$

Harmonic kernel $$k(\mathbf{x},\mathbf{x}') = \sigma^2 \exp\left(\frac{x_1x_1'+x_2x_2'}{\ell^2}\right)\cos\left(\frac{x_2x_1'-x_1x_2'}{\ell^2}\right)$$



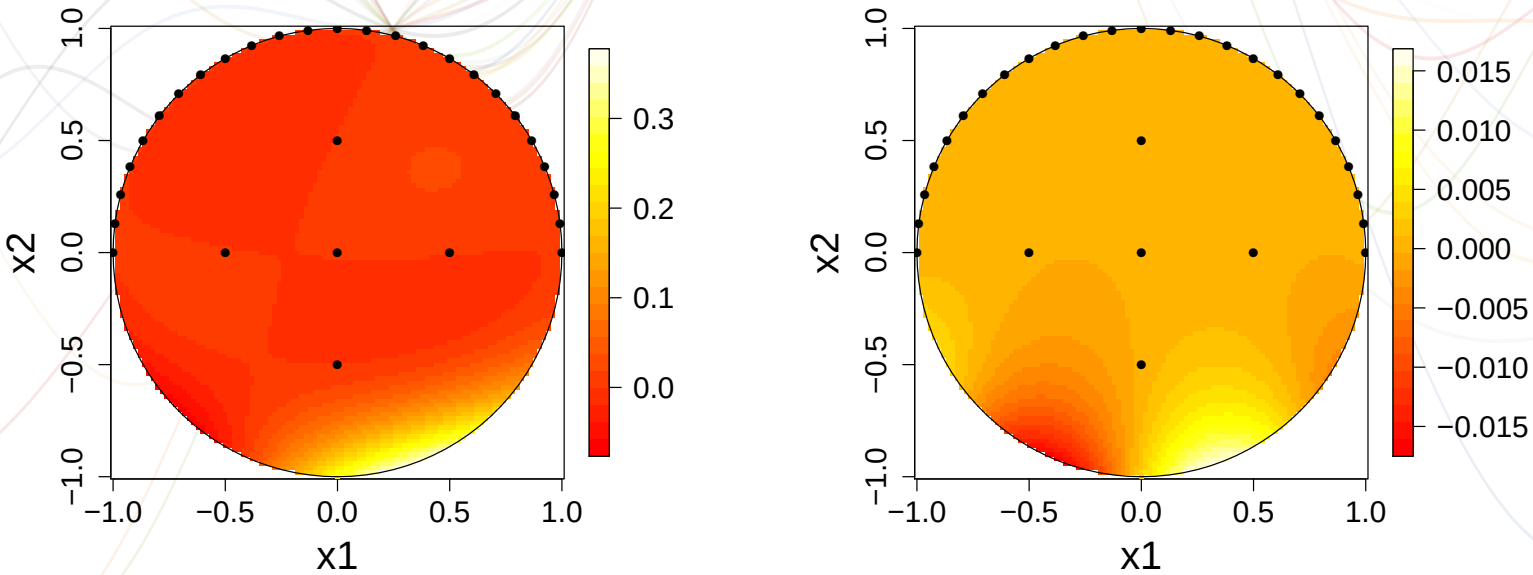

Prediction mean – Left : Gaussian kernel ; Right : Harmonic kernel.

# Illustration : Maximum of a harmonic function [Ginsbourger et al., 2016]

We compare two GPs for predicting the maximum of a 2D harmonic function

Gaussian kernel $$k(\mathbf{x},\mathbf{x}') = \sigma^2 \exp\left(-\frac{(x_1-x_1')^2}{\ell_1^2} - \frac{(x_2-x_2')^2}{\ell_2^2}\right)$$

Harmonic kernel $$k(\mathbf{x},\mathbf{x}') = \sigma^2 \exp\left(\frac{x_1x_1'+x_2x_2'}{\ell^2}\right)\cos\left(\frac{x_2x_1'-x_1x_2'}{\ell^2}\right)$$



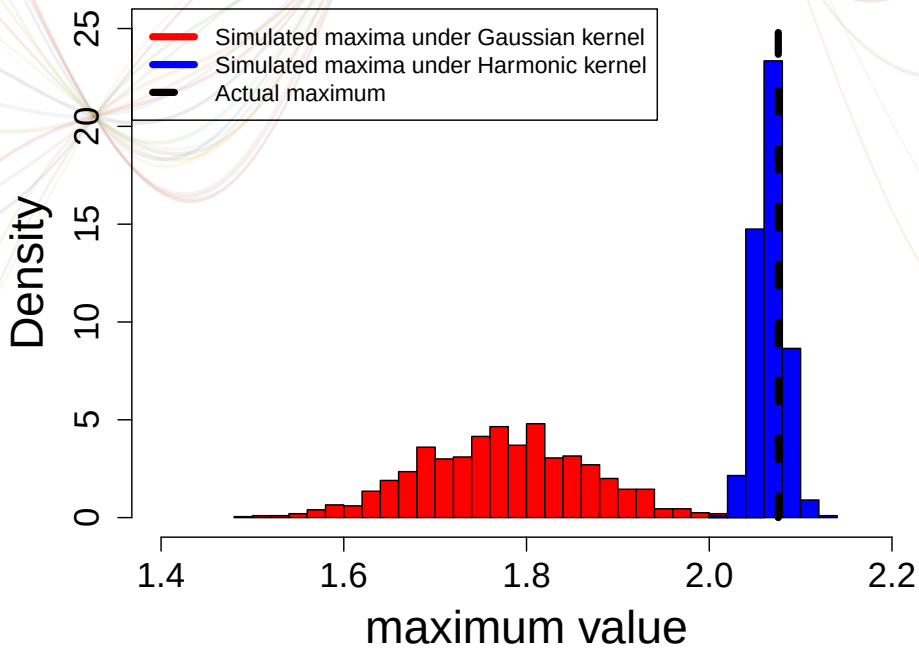

Prediction st. dev. – Left : Gaussian kernel ; Right : Harmonic kernel.

# Illustration : Maximum of a harmonic function [Ginsbourger et al., 2016]

We compare two GPs for predicting the maximum of a 2D harmonic function

Gaussian kernel $$k(\mathbf{x},\mathbf{x}') = \sigma^2 \exp\left(-\frac{(x_1-x_1')^2}{\ell_1^2} - \frac{(x_2-x_2')^2}{\ell_2^2}\right)$$

Harmonic kernel $$k(\mathbf{x},\mathbf{x}') = \sigma^2 \exp\left(\frac{x_1x_1'+x_2x_2'}{\ell^2}\right)\cos\left(\frac{x_2x_1'-x_1x_2'}{\ell^2}\right)$$




Conditional simulations of the maximum under the two GPs.

# GP under linear inequalities : Impact on uncertainty quantification

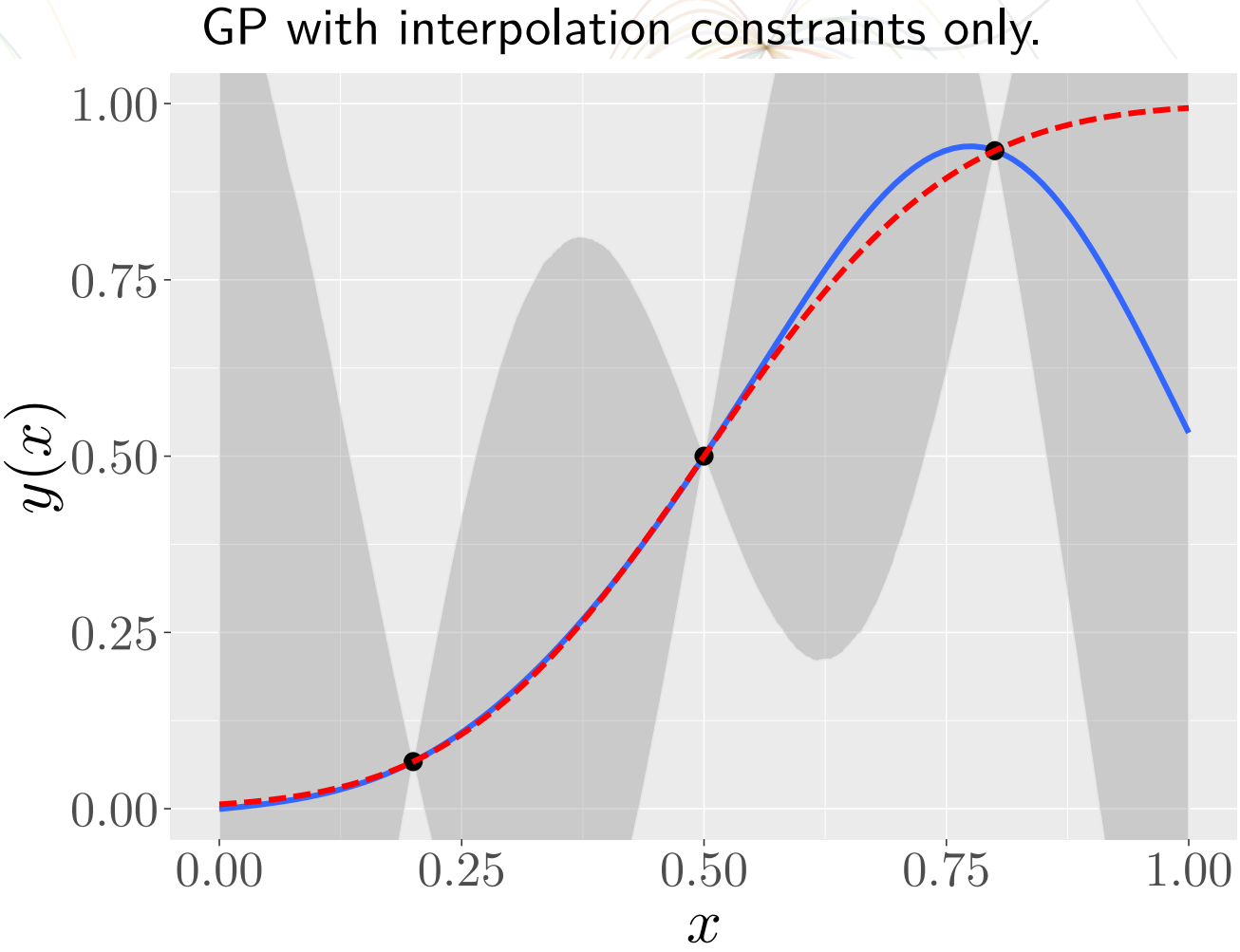



Illustration on a toy example (cdf of a Normal distribution)

# GP under linear inequalities : Impact on uncertainty quantification

GP with boundedness + monotonicity additional constraints.

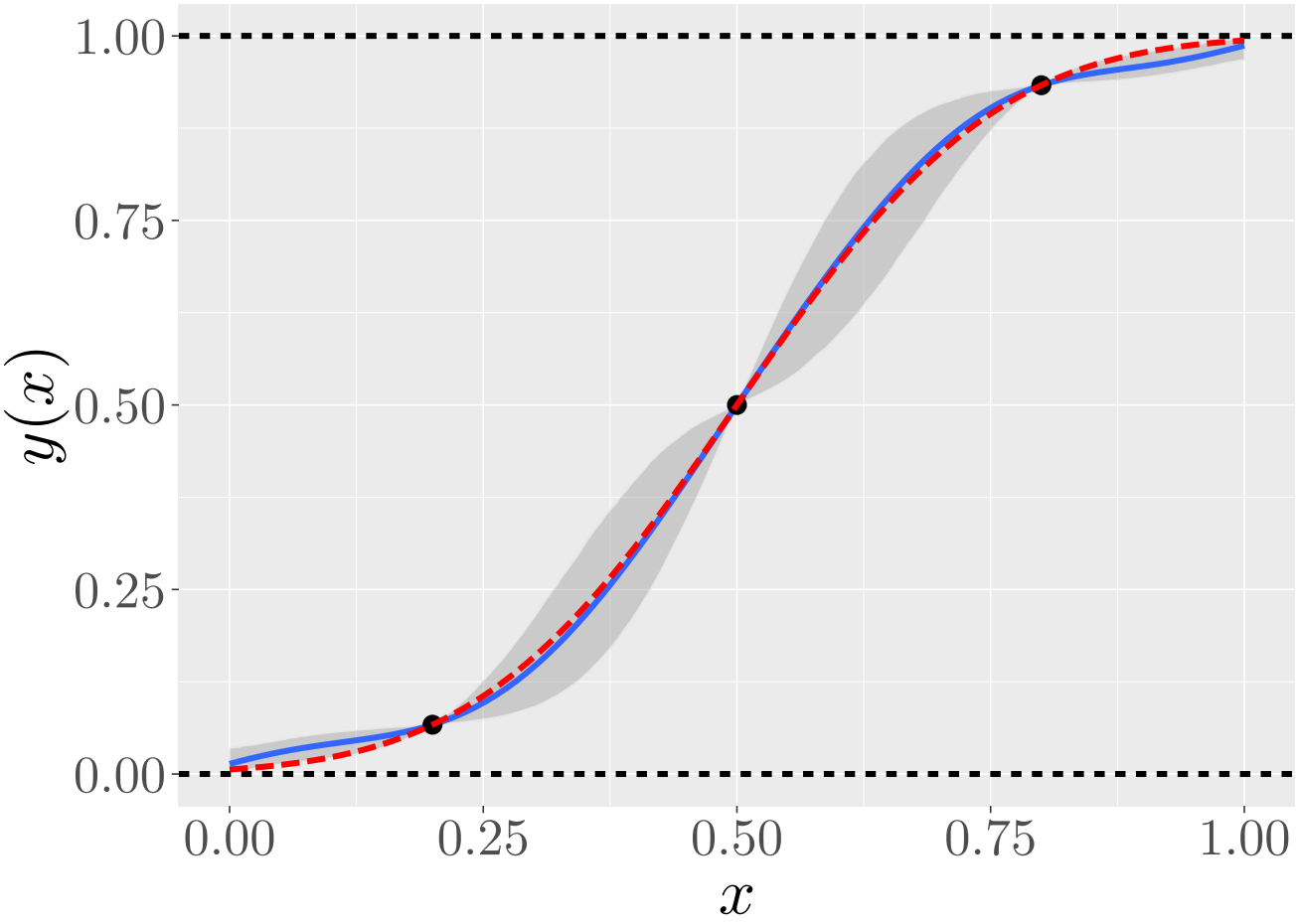



Illustration on a toy example (cdf of a Normal distribution)

# GP and linear inequalities : Some theory

## A finite elements ($P_1$) model for 1D GPs

**[Maatouk and Bay, 2017, López-Lopera et al., 2018]**

Each sample path of a GP $Y$ is approximated by an **affine function**

$$Y_m(x) = \sum_{j=1}^{m} \xi_j \phi_j(x)$$

where $\phi_j$ are "hat" functions and $\boldsymbol{\xi}$ is a Gaussian vector extracted from $Y$

- Key point : Boundedness, monotonicity (and others) for an affine function can be checked only at knots $\rightarrow$ finite number of conditions only

# GP and linear inequalities : Some theory

## A finite elements ($P_1$) model for 1D GPs

**[Maatouk and Bay, 2017, López-Lopera et al., 2018]**

Each sample path of a GP $Y$ is approximated by an **affine function**

$$Y_m(x) = \sum_{j=1}^{m} \xi_j \phi_j(x)$$

where $\phi_j$ are "hat" functions and $\boldsymbol{\xi}$ is a Gaussian vector extracted from $Y$

- Key point : Boundedness, monotonicity (and others) for an affine function can be checked only at knots $\rightarrow$ finite number of conditions only

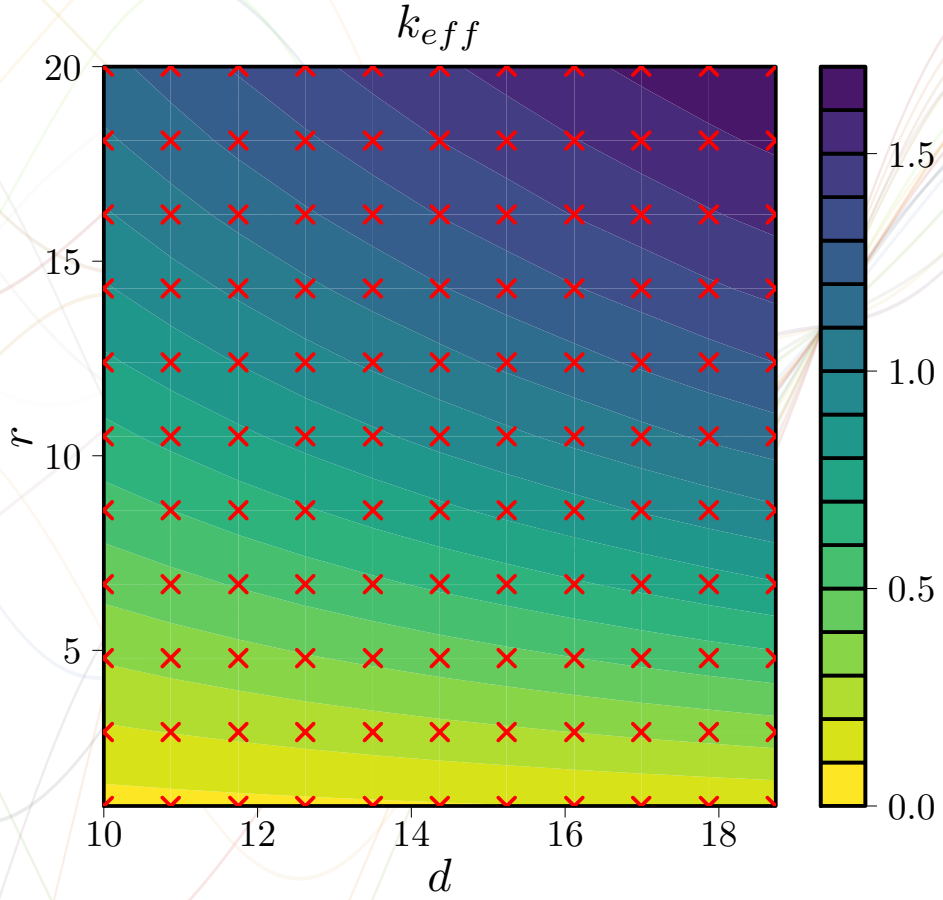
## Correspondence with spline under inequality [Bay et al., 2016]

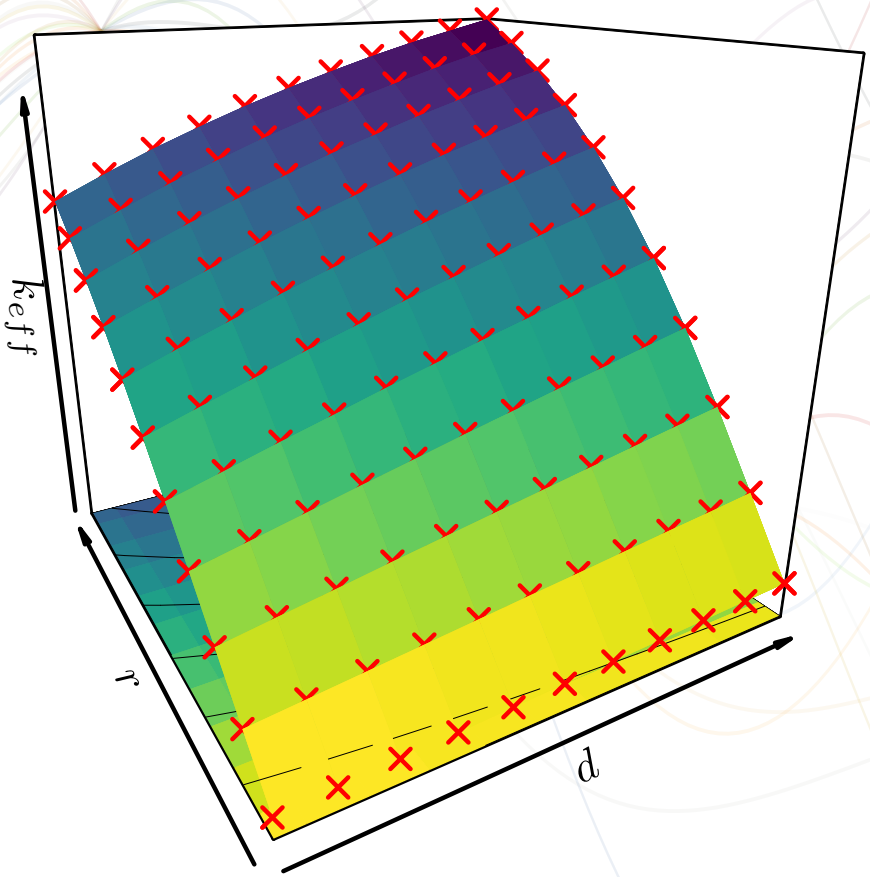
Asymptotically, computing the mode a posteriori of $Y_m$ is equivalent to

$$\min_{f \in \mathcal{H} \cap \mathcal{C}} \|f\| \qquad \text{s.t.} \quad f(x_i) = y_i, \quad (i = 1, \ldots, n)$$

where $\mathcal{C}$ is a convex set of inequalities, and $\mathcal{H}$ is the RKHS associated to $Y$.

# Example of application in 2D

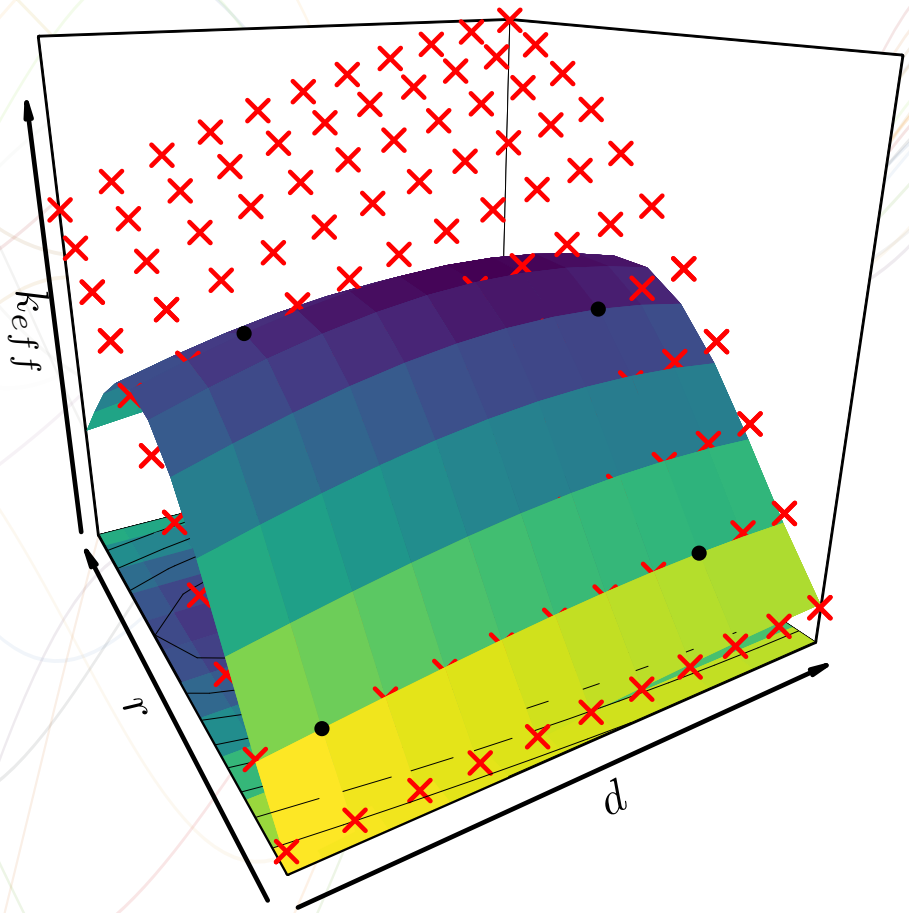


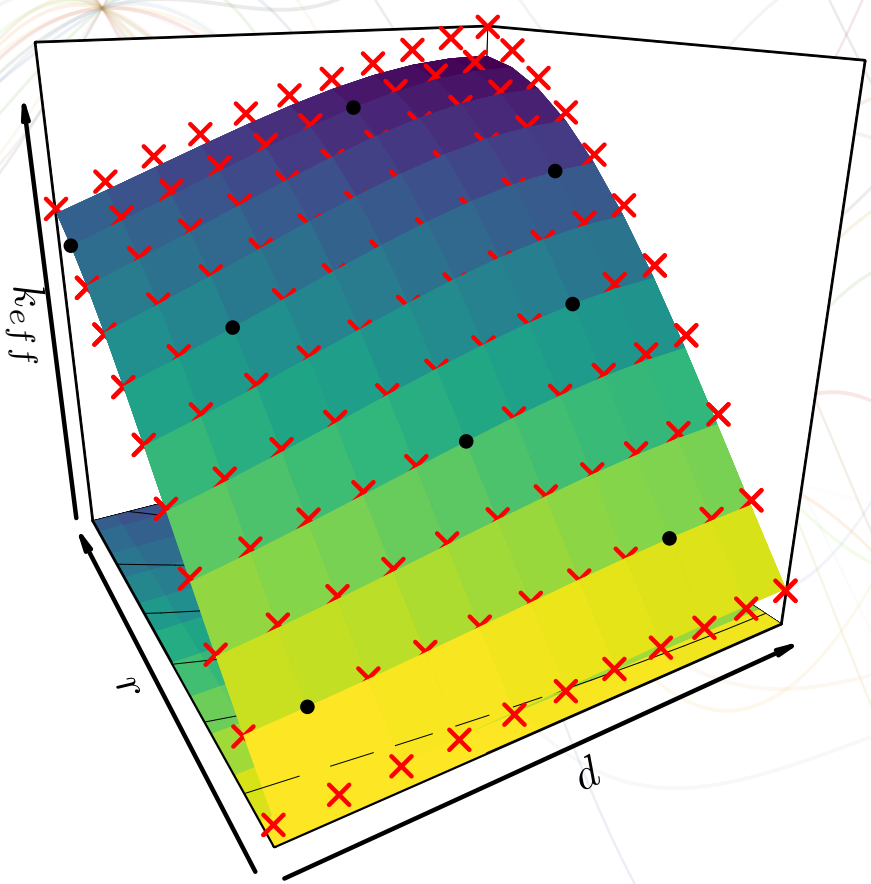




Nuclear criticality safety assessments : IRSN's dataset.

Extra information : $k_{\text{eff}}$ is positive and non-decreasing.

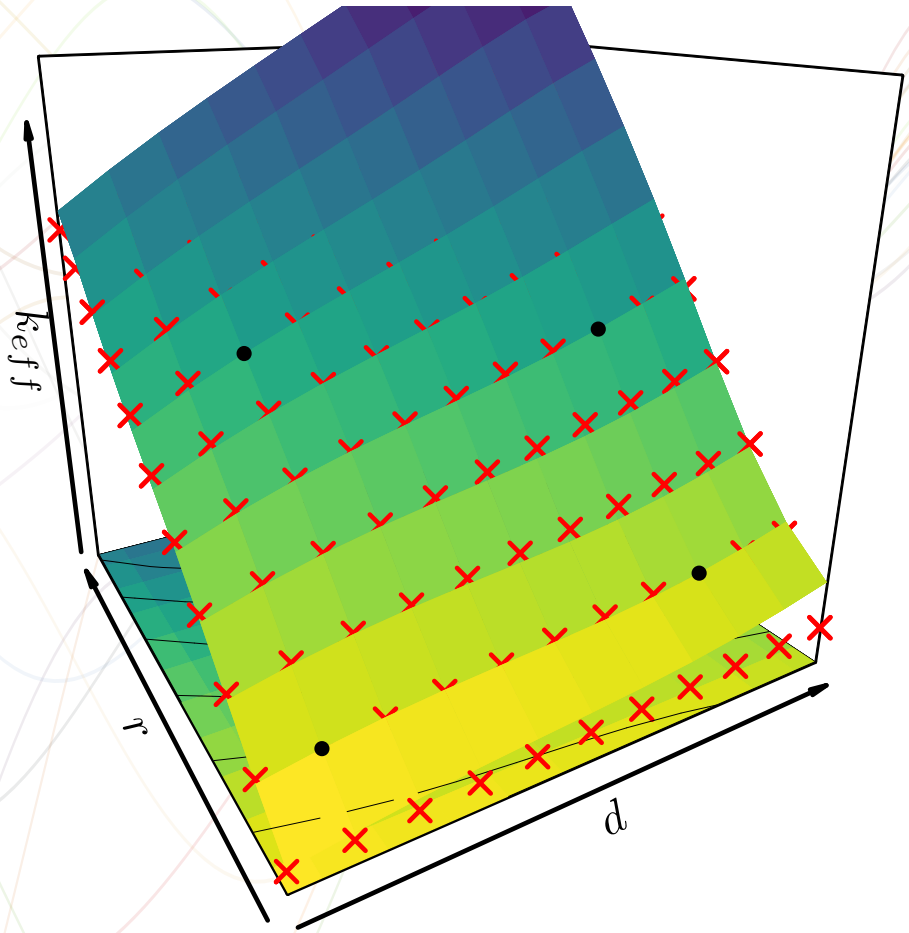
# Example of application in 2D

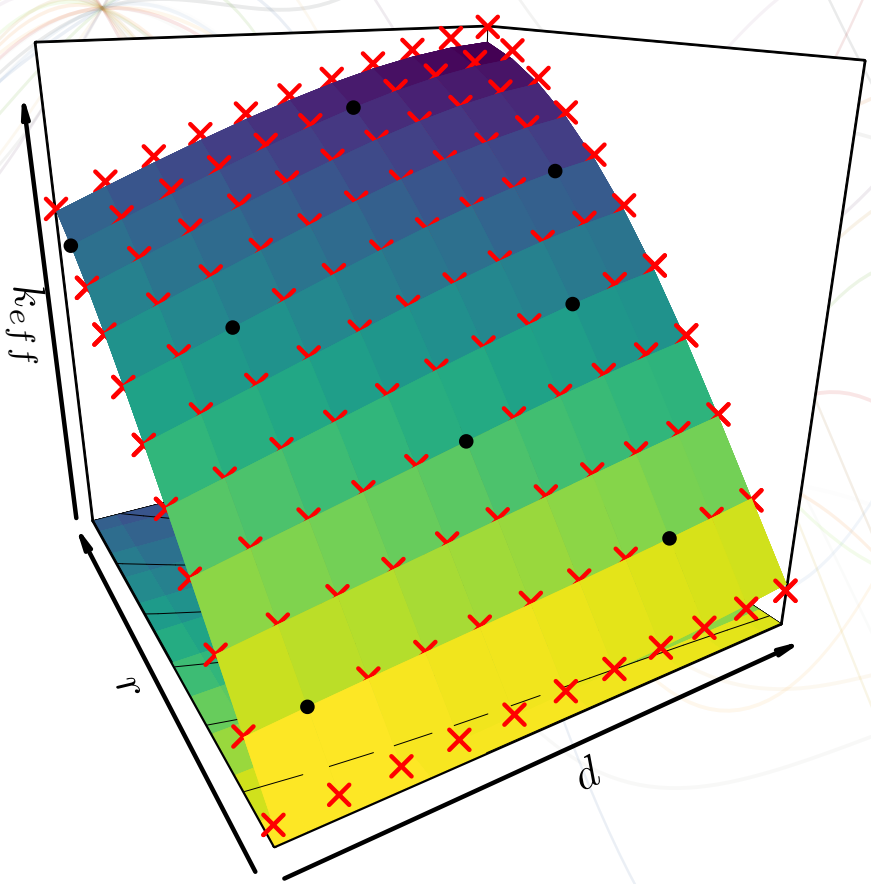



Unconstrained model + MLE.

# Example of application in 2D

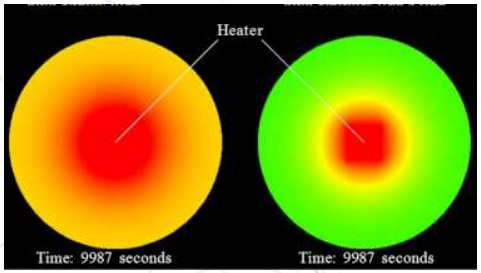


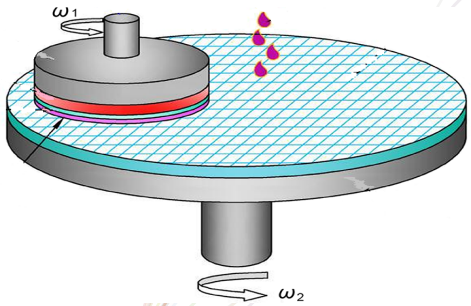

Constrained model + constrained MLE.

# Outline

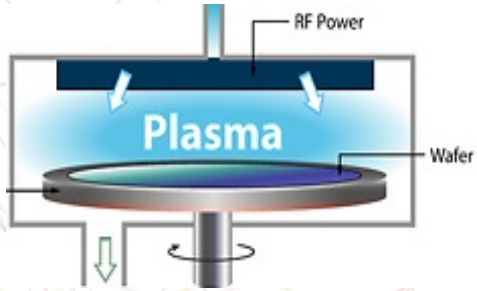
# A motivating industrial context



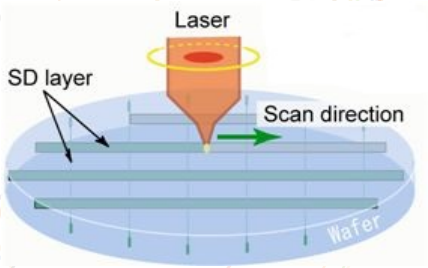

Heating



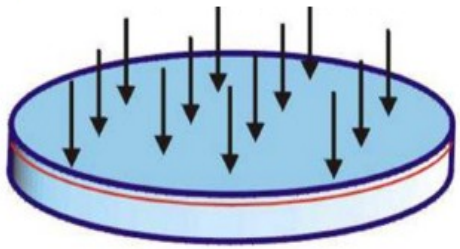

Polishing



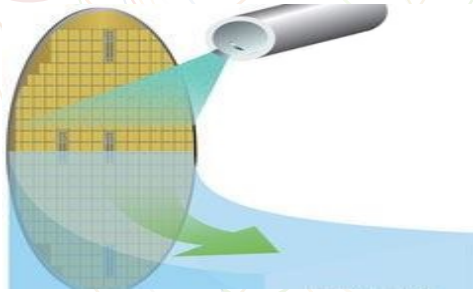

Vapor deposition




Laser processing



Doping



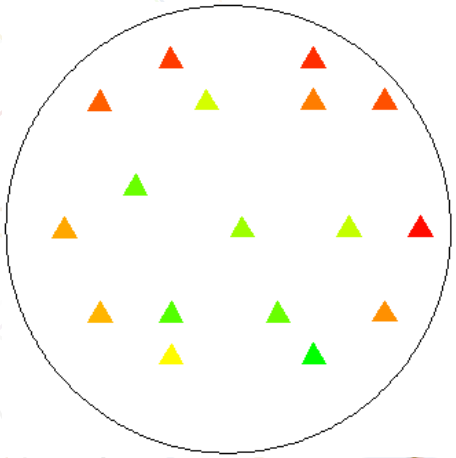
Cleaning

## Surface quality control in microelectronics

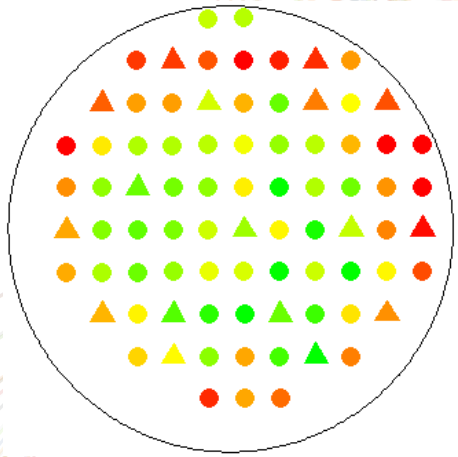
The infinite dimensional problem is made feasible via two steps :

1. *Construction of a spatial interpolation model*
2. Standard quality control of the model parameters

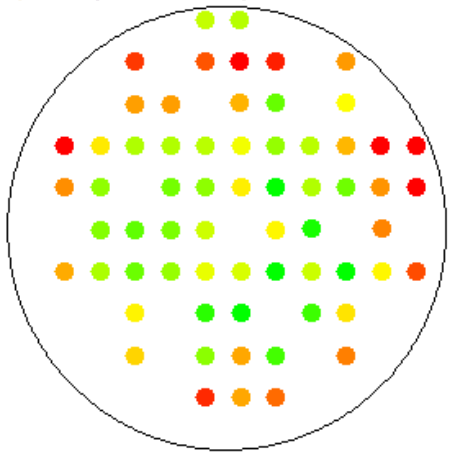
# Zoom on a wafer



DoE



DoE + test set



test set

## Interpolation problem on the unit disk

- The design of experiments (DoE) has 17 fixed (not controllable) points
- Here, presence of strong (but not pure) radial effects

# How to build a GP model on the unit disk ?

## Cartesian GP

Use the restriction of a GP defined on the unit square

$\rightarrow$ "Neighbours" are defined through horizontal and vertical distances

## Polar GP [Padonou and Roustant, 2016]

Combine a GP for the radius (on $[0, 1]$) and a GP for the angle (on the circle)

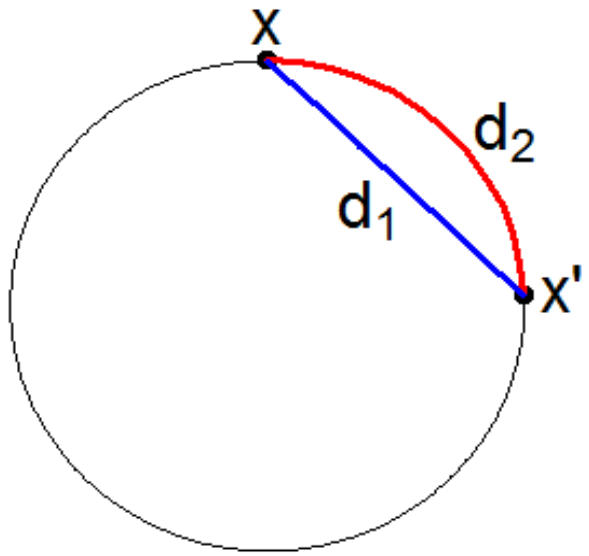
$$Z(x, y) = R(r) * A(\theta)$$

$\rightarrow$ "Neighbours" are defined through radial and angular distances

# How to build a GP model on the unit circle $\mathbb{S}$ ?

## Distances on $\mathbb{S}$

- Chordal distance $d_1$
- Geodesic distance $d_2$

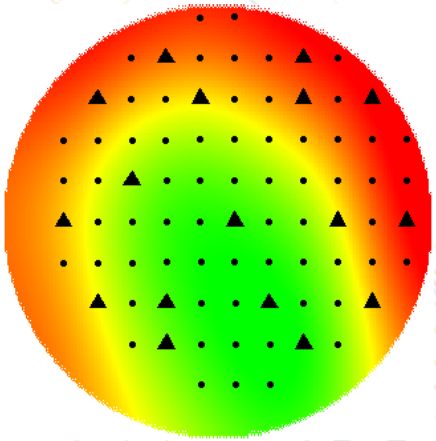



## Stationary (isotropic) kernels on $\mathbb{S} \times \mathbb{S}$

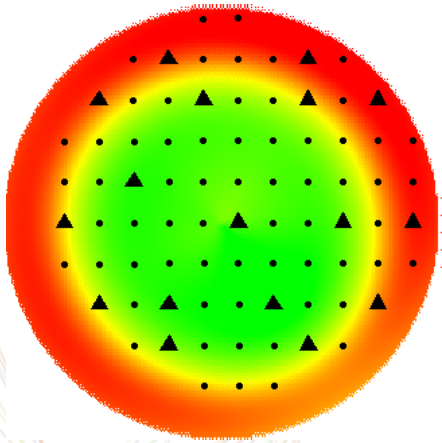
*Examples of construction*

- For $d_1$, just consider the restriction of a 2D isotropic GP to $\mathbb{S}$
- For $d_2$, plug-in $d_2$ into a 1D compactly supported stationary kernel $k(r)$, with support included in $[0, \pi[$

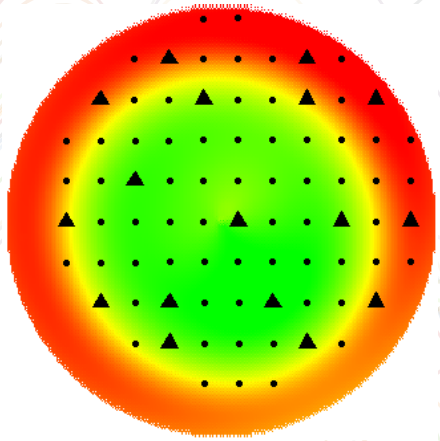
# Results (prediction accuracy)



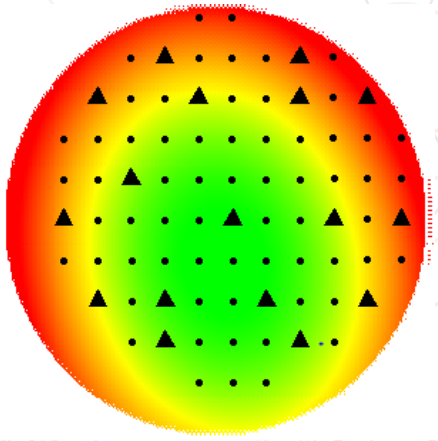
Cartesian GP



Polar GP (chord.)



Polar GP (geo.)



Zernike regression

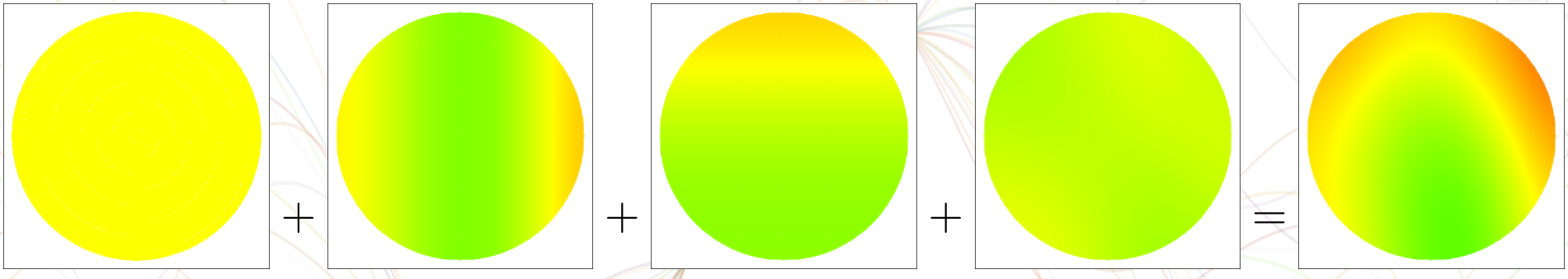
| GP type | Cartesian | | Polar (chordal) | | Polar (geodesic) | |
|---|---|---|---|---|---|---|
| Kernel type | $k_{\text{prod}}$ | $k_{\text{add}}$ | $k_{\text{prod}}$ | $k_{\text{add}}$ | $k_{\text{prod}}$ | $k_{\text{add}}$ |
| RMSE | 0.75 * | 0.77 | 0.69 | 0.60 * | 0.68 | 0.61 * |

Polar GPs better capture the radial pattern

$\Leftarrow$ Parameter estimation detects a strong angular correlation

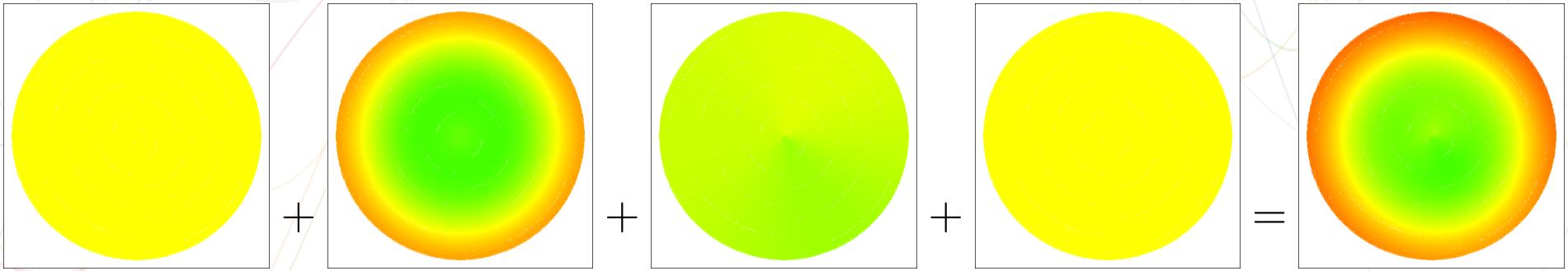
# Results (Hoeffding-Sobol [ANOVA] decomposition)

## Sobol decomposition of the Kriging mean, Cartesian GP, uniform measure over $[-1,1]^2$



| $\mu(x,y)$ | $m_1(x)$ | $m_2(y)$ | $m_{12}(x,y)$ | $m(x,y)$ |
|---|---|---|---|---|
| $-$ | $S = 46$ | $S = 52$ | $S = 2$ | Kriging mean |

## Sobol decomposition of the Kriging mean, polar GP, uniform measure over $\mathcal{D}$

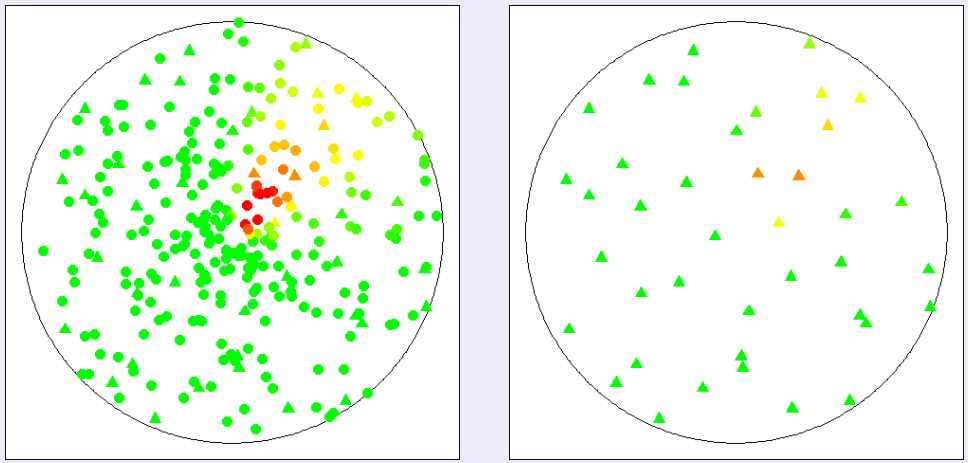


| $\mu(\rho,\theta)$ | $m_1(\rho)$ | $m_2(\theta)$ | $m_{12}(\rho,\theta)$ | $m(\rho,\theta)$ |
|---|---|---|---|---|
| $-$ | $S = 85$ | $S = 8$ | $S = 7$ | Kriging mean |

# Application in environments

## A model of atmospheric dispersion of a greenhouse gas

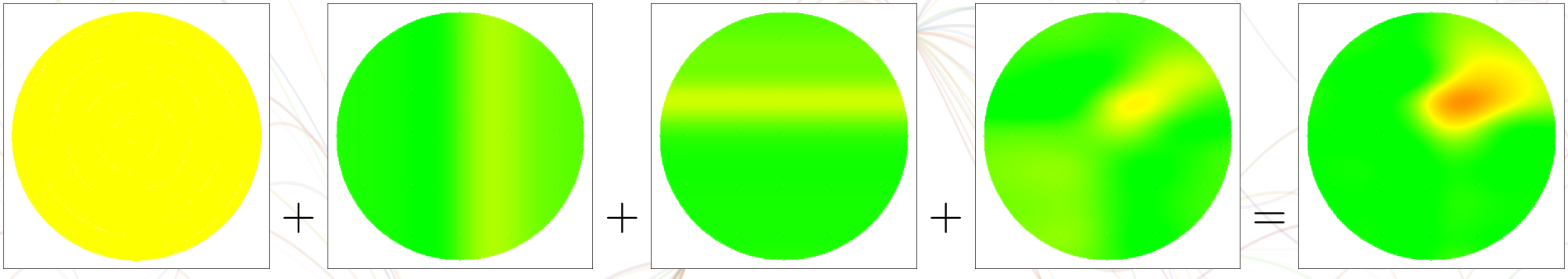
- Complex phenomenon simulated by a computer code [Batton-Hubert et al., 2013]
- Inputs : wind speed (in $[0, v_{\max}]$), wind direction (in $[0, 2\pi]$, directional input)



DoE + test set | DoE

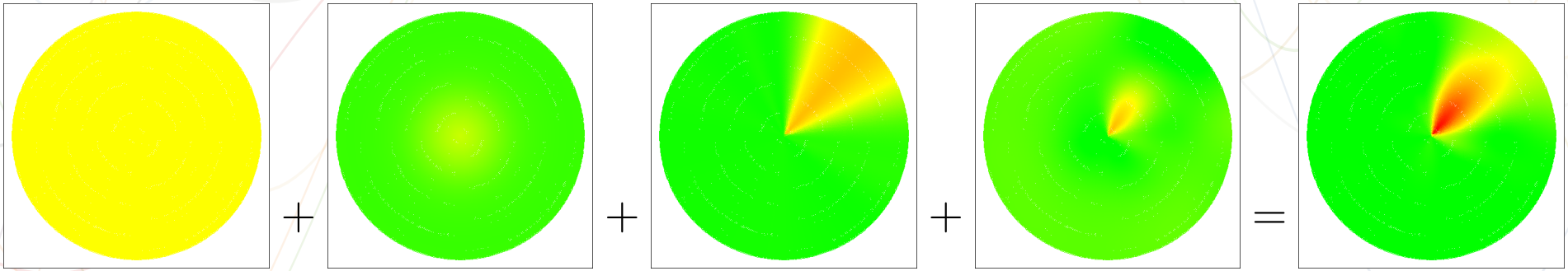
# Results (ANOVA decomposition) [Padonou and Roustant, 2017]

## Sobol decomposition of the Kriging mean, Cartesian GP, uniform measure over $[-1,1]^2$



| $\mu(x,y)$ | $m_1(x)$ | $m_2(y)$ | $m_{12}(x,y)$ | $m(x,y)$ |
|---|---|---|---|---|
| – | $S = 26$ | $S = 29$ | $S = 45$ | Kriging mean |

## Sobol decomposition of the Kriging mean, polar GP, uniform measure over $\mathcal{D}$



| $\mu(\rho,\theta)$ | $m_1(\rho)$ | $m_2(\theta)$ | $m_{12}(\rho,\theta)$ | $m(\rho,\theta)$ |
|---|---|---|---|---|
| – | $S = 6$ | $S = 80$ | $S = 14$ | Kriging mean |

# More about GPs on $d$-dimensional sphere $\mathbb{S}^d$

## Some entries in the literature

- A characterization of isotropic kernels, due to [Schoenberg, 1942]
  - $d = 1$ : $k(\theta) = \sum_{n=0}^{\infty} b_n \cos(n\theta)$ $\quad (b_n \geq 0, \quad \sum_n b_n < +\infty)$
  - $d \geq 2$ : invokes Gegenbauer polynomials.
- A comprehensive review and study by [Gneiting, 2013], among which :
  - Extension of Lévy's theorem : plugging the geodesic distance into a compactly supported 1-dimensional kernel works up to $d \leq 3$
  - Kernels defined with completely monotone functions
- Time-varying GPs on a sphere : see e.g. [Porcu et al., 2016]

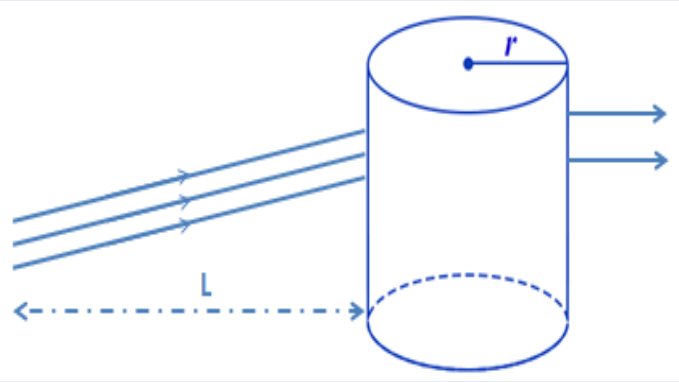
## See also...

- Detecting periodicities in signals with RKHS [Durrande et al., 2016]
- Geodesic GPs to reconstruct free-form surfaces [del Castillo et al., 2015]

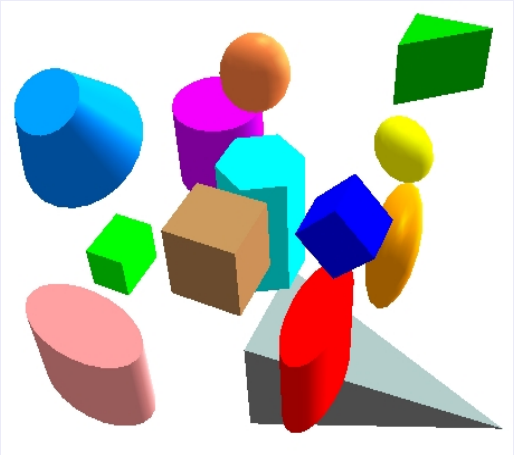
# GP for categorical inputs : groups of levels & trees

## A guiding case study : Nuclear activity quantification

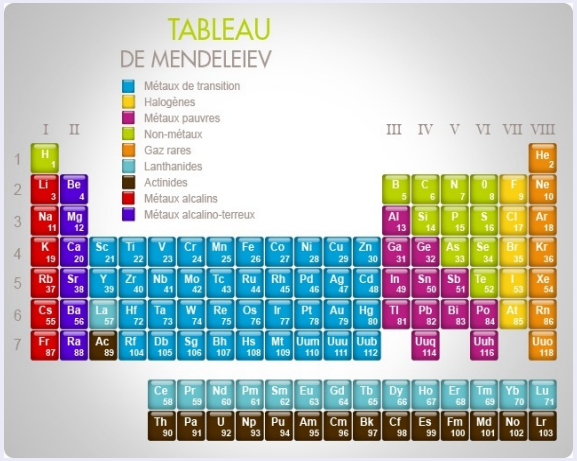
1. Computation using Monte Carlo
2. 4 continuous inputs : $L$, density, mean width, lateral surface
3. 3 categorical inputs : energy, form, chemical element.



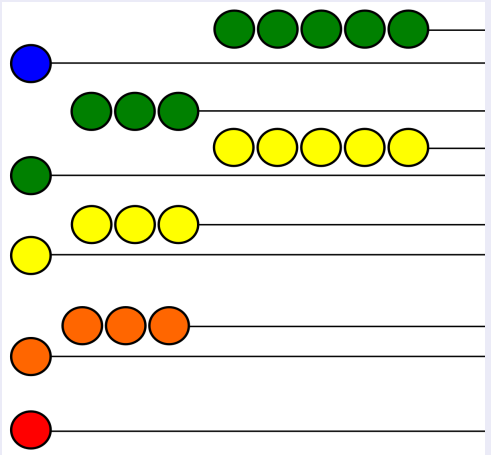

## Specific problem : large number of levels



Form (s, c, p)




$Z \in \{1, \ldots, 94\}$

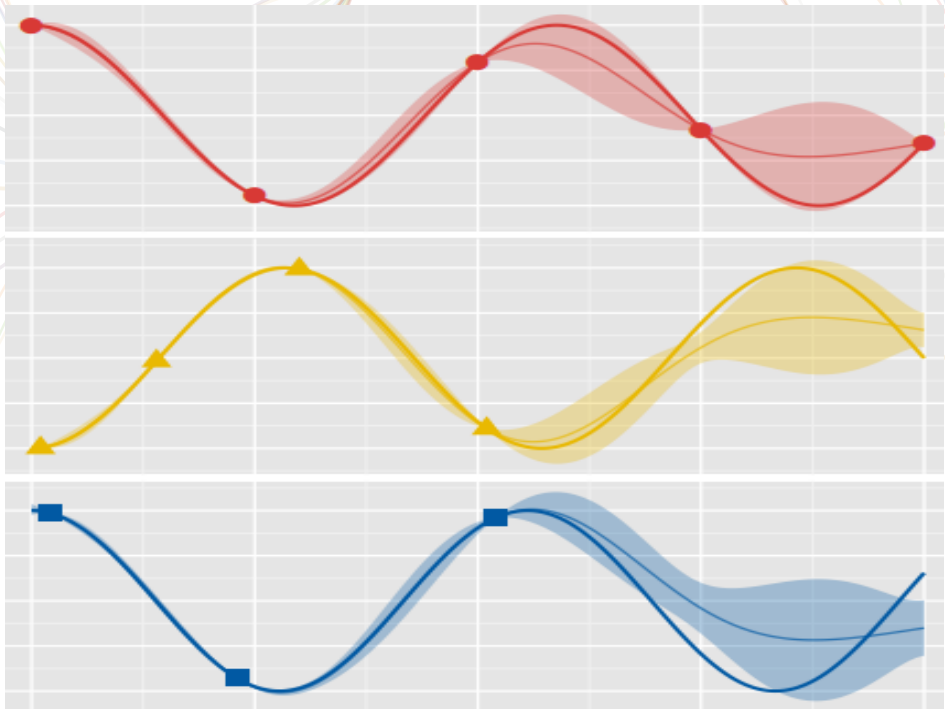


$E \in \{E_1, E_2, E_3, E_4, E_5, E_6\}$

## GP interpretation when no distance is available

A GP for $(x, u) \in [0,1] \times \{"red", "yellow", "blue"\}$ can be defined with :

- a kernel on $[0,1]$, i.e. a **covariance function**
- a kernel on $\{"red", "yellow", "blue"\}$ , i.e. a **covariance matrix**
- a valid operation between them, such as *, $+$, ...



Example : $\mathrm{Cov}(Y(x, "blue"), Y(x', "red")) = k(x, x') \times 0.8$

# GP for categorical inputs : group kernels for partitioned levels

A group kernel is a block covariance matrix

## Block covariance matrix

$$\mathbf{T} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{B}_{1,2} & \cdots & \mathbf{B}_{1,G} \\ \mathbf{B}_{2,1} & \mathbf{W}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{B}_{G-1,G} \\ \mathbf{B}_{G,1} & \cdots & \mathbf{B}_{G,G-1} & \mathbf{W}_G \end{pmatrix}$$

with constant between-group blocks

## Hierarchical (group/level) process

$$\eta_{g/\ell} = \mu_g + \lambda_{g/\ell}$$

with :

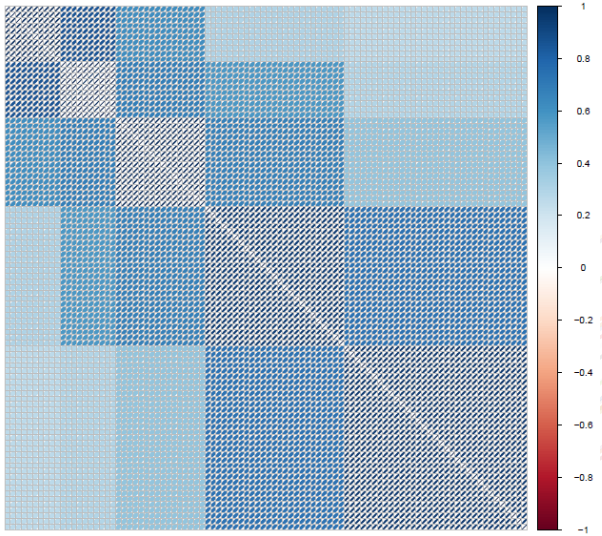
- Gaussian ind. priors for $\boldsymbol{\mu}, \boldsymbol{\lambda}_{g/.}$
- Centering cond. : $\sum_\ell \lambda_{g/\ell} = 0$

## Main results [Roustant et al., 2018]

- Connection with hierarchical GPs : $\mathbf{T} = \mathrm{cov}(\boldsymbol{\eta})$.
- Characterization & parameterization of valid group kernels

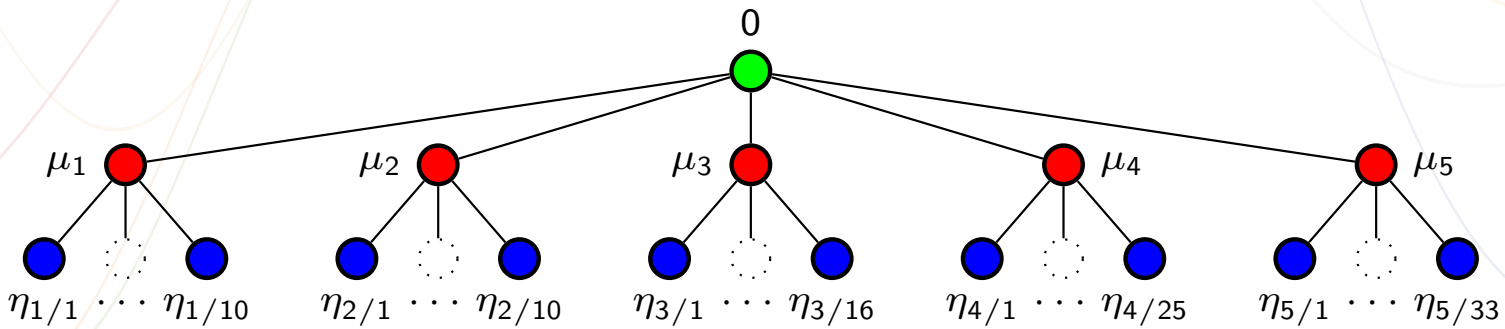
# Result on the case study

For the categorical input 'chemical element', 5 groups are identified by experts
$\rightarrow$ Parsimonious parametrization with only 20 parameters (instead of $94 \times 95/2$)




- 10 between-group covariances
- 5 within-group covariances
- 5 within-group variances

With a stratified LHS design of size $3 \times 94$, the $Q^2$ of the whole model is $> 0.95$

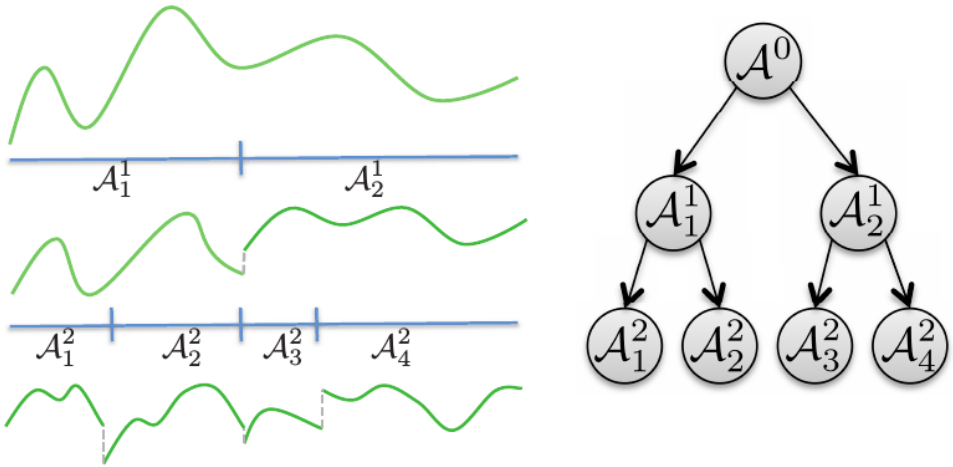


# More on hierarchical GPs

- Wavelet kernels [Amato et al., 2006]
- Treed Gaussian processes [Gramacy, 2007]
- Lattice Kriging [Nychka et al., 2015]
- Multiresolution GPs [Fox and Dunson, 2012]
- Hierarchical GPs [Park and Choi, 2010]
- ...

*Remark : In these models, the children ("details in subareas") are independent conditionaly on the mother ("trend").*

*This was not the case before since children sum to 0 (cond. on the mother).*



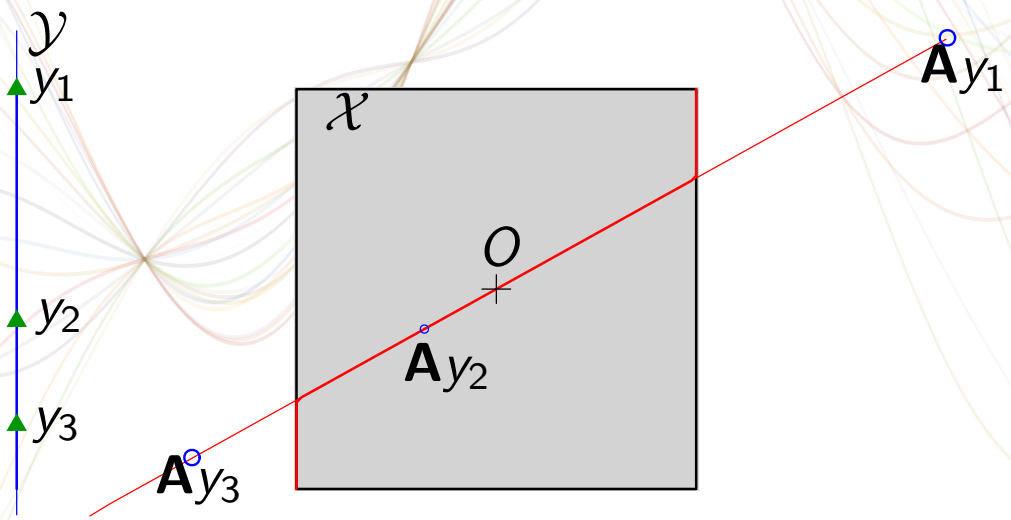

Source : [Fox and Dunson, 2012], Figure 2.

# Outline

# Random EMbedding Bayesian Optimization (REMBO, [Wang et al., 2013] )

**Hypothesis** $(\mathcal{H})$ : Only $d_e \ll D$ of the $D$ variables of $f$ are influential

**Principle** : Embed $\mathcal{Y} \subset \mathbb{R}^d$ to $\mathcal{X}$ ($d \ll D$) with a random matrix of normals.

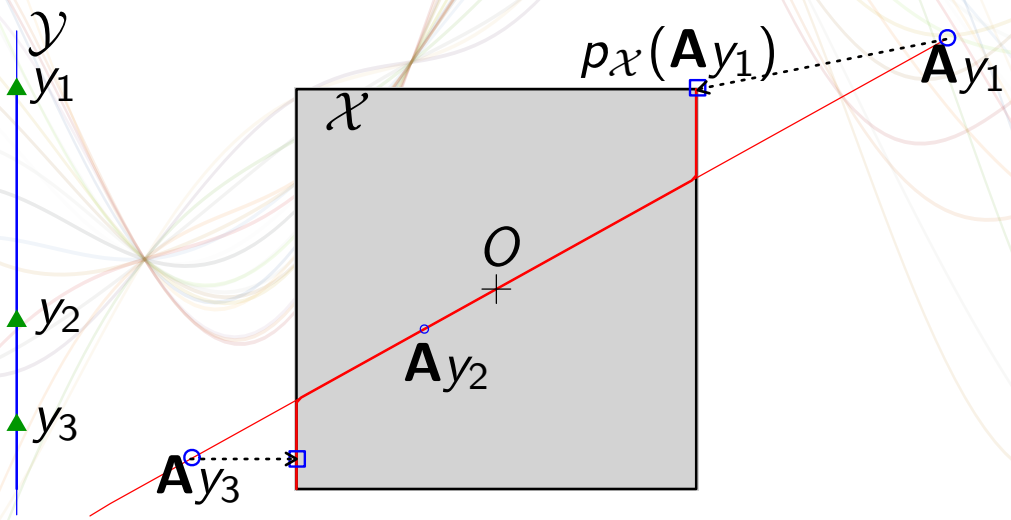



A convex projection on $\mathcal{X}$, $p_{\mathcal{X}}$, is applied to points mapped outside of $\mathcal{X}$.

# Random EMbedding Bayesian Optimization (REMBO, [Wang et al., 2013] )

**Hypothesis** $(\mathcal{H})$ : Only $d_e \ll D$ of the $D$ variables of $f$ are influential

**Principle** : Embed $\mathcal{Y} \subset \mathbb{R}^d$ to $\mathcal{X}$ ($d \ll D$) with a random matrix of normals.



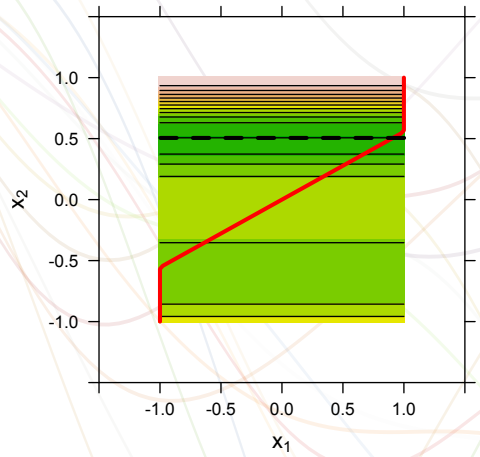

A convex projection on $\mathcal{X}$, $p_{\mathcal{X}}$, is applied to points mapped outside of $\mathcal{X}$.

# REMBO principle

**Interest** : Optimization of $f$ is carried out on

$$g : \begin{array}{ccc} \mathcal{Y} \subset \mathbb{R}^d & \to & \mathbb{R} \\ \mathbf{y} & \mapsto & f(p_{\mathcal{X}}(\mathbf{A}\mathbf{y})) \end{array}$$

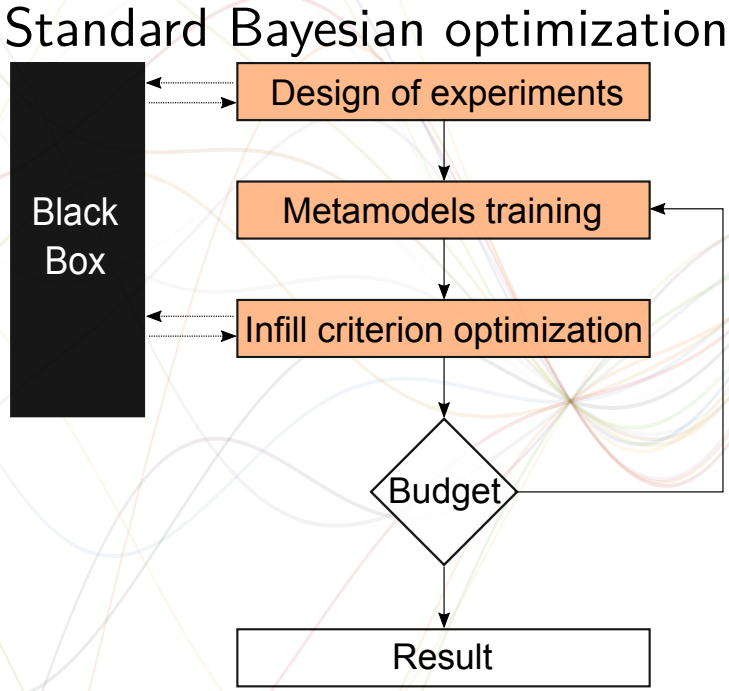
$\to$ Much smaller search space




**Theorem**

If $\mathcal{Y} \supset \mathcal{B}(0, d/\varepsilon)$ and under $(\mathcal{H})$, finding $\mathbf{y}^* \in \mathcal{Y}$ such that $g(\mathbf{y}^*) = \min f$ has a solution with probability at least $(1 - \varepsilon)$.

# REMBO procedure for a costly function $f$

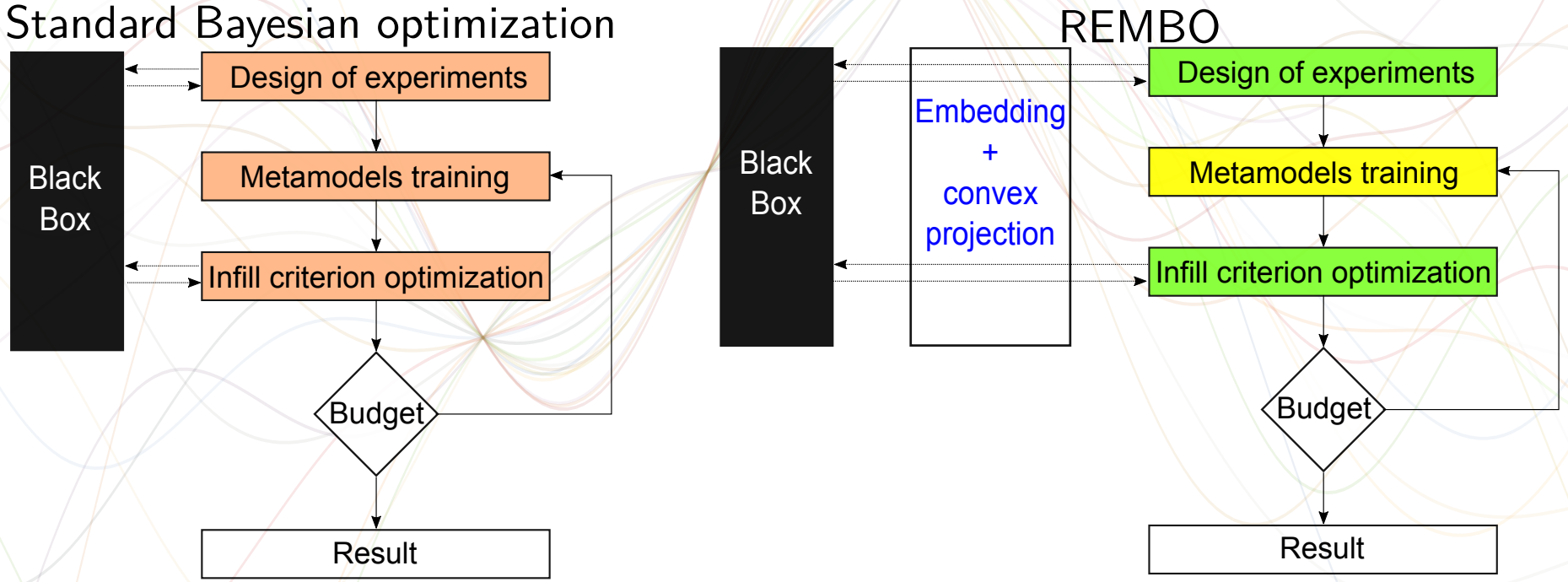



Orange : dimension $D$

Yellow : dimension $d$ or $D$

Green : dimension $d$

# REMBO procedure for a costly function $f$




Orange : dimension $D$

Yellow : dimension $d$ or $D$

Green : dimension $d$

# Choice of the covariance kernel for sequential optimization

**Main issue** : Incorporating high-dimensional information on $f$ via a kernel

[Wang et al., 2013] proposed two covariance kernels :

- kernel defined between projections in $\mathcal{X}$,
  e.g. $k_{\mathcal{X}}(\mathbf{y}, \mathbf{y'}) = exp(-\frac{\|p_{\mathcal{X}}(\mathbf{Ay}) - p_{\mathcal{X}}(\mathbf{Ay'})\|_D^2}{2l^2})$
  - kriging in dimension $D$

# Choice of the covariance kernel for sequential optimization

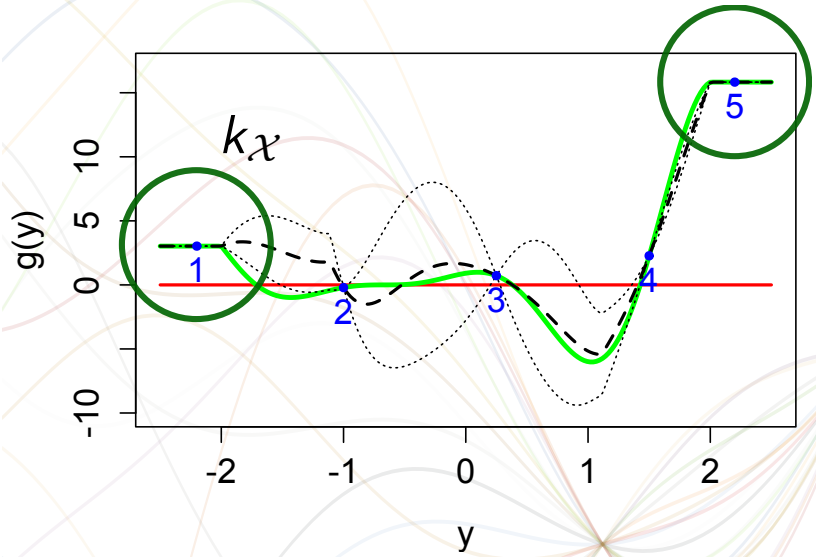
**Main issue** : Incorporating high-dimensional information on $f$ via a kernel

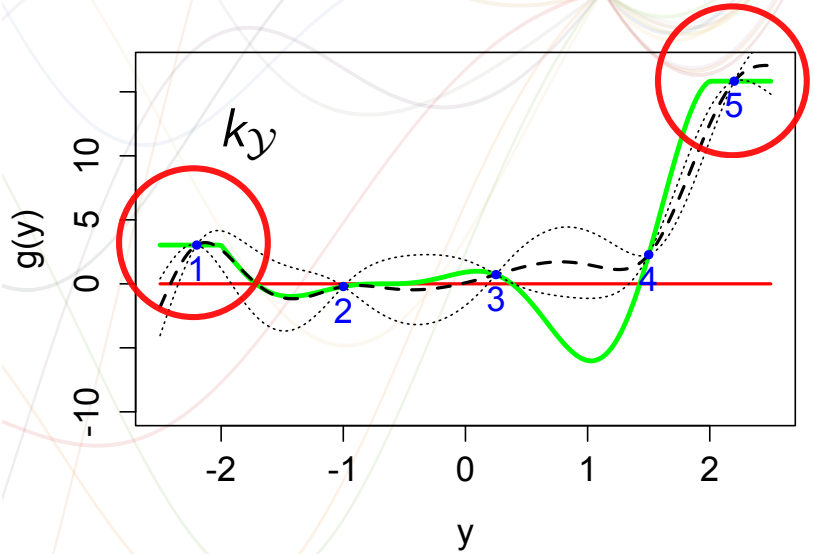
[Wang et al., 2013] proposed two covariance kernels :

- kernel defined between projections in $\mathcal{X}$,
  e.g. $k_{\mathcal{X}}(\mathbf{y}, \mathbf{y'}) = exp(-\frac{\|p_{\mathcal{X}}(\mathbf{Ay}) - p_{\mathcal{X}}(\mathbf{Ay'})\|_D^2}{2l^2})$
  - kriging in dimension $D$

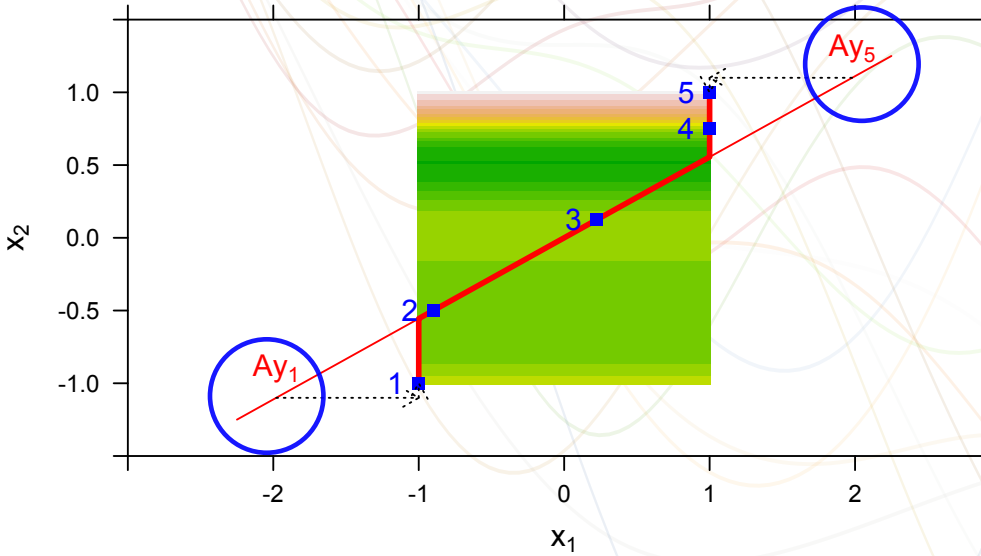
- kernel defined between points of $\mathcal{Y}$,
  e.g. $k_{\mathcal{Y}}(\mathbf{y}, \mathbf{y'}) = exp(-\frac{\|\mathbf{y} - \mathbf{y'}\|_d^2}{2l^2})$
  - kriging in dimension $d$
  - suffers from non-injectivity of the mapping

# Illustration : GP predictions with $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$



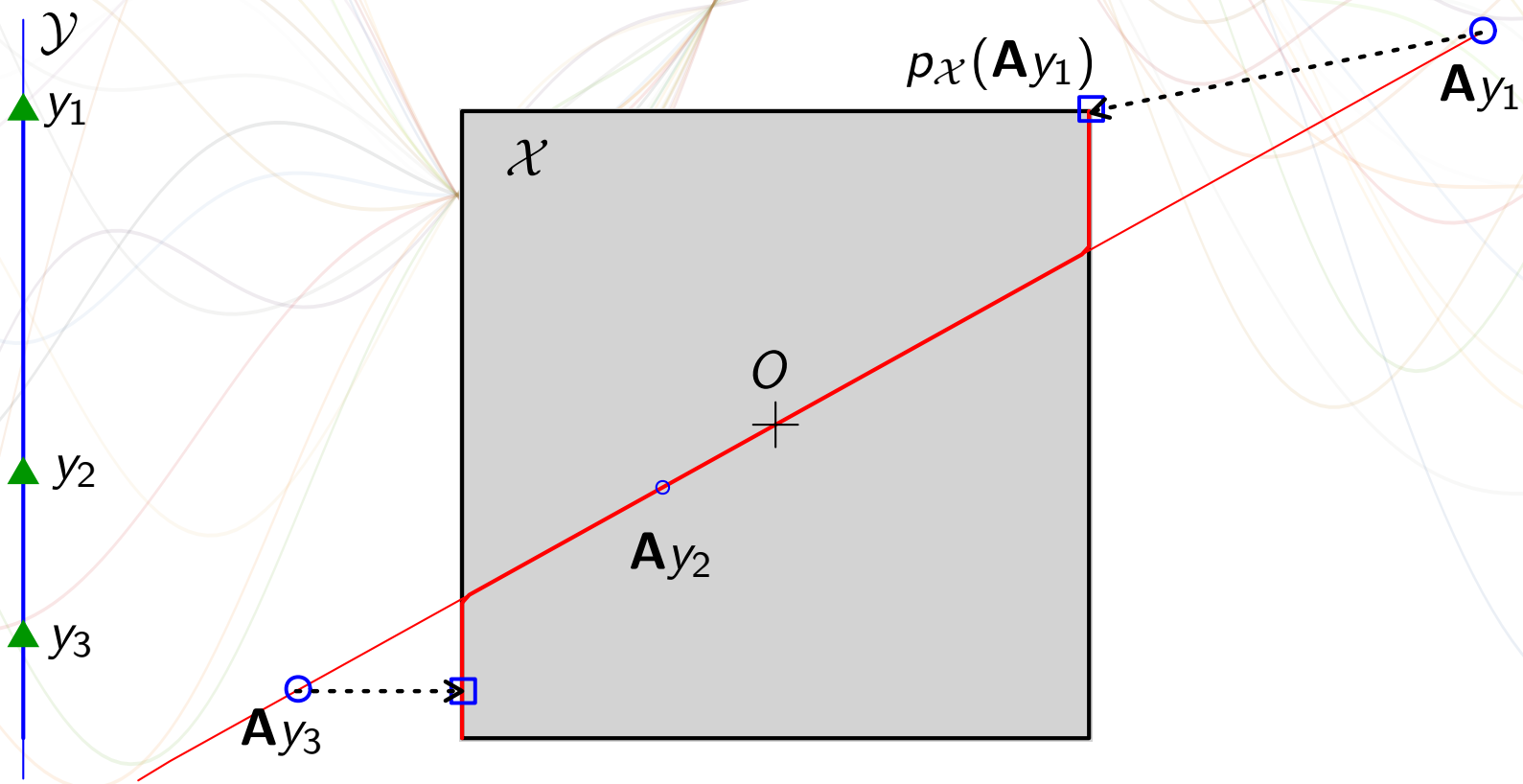

$$\mathcal{Y} = [-2.5, 2.5],\ \mathcal{X} = [-1, 1]^2$$




$$\mathbf{A} = [0.9, 0.5]^T$$

# Towards a new kernel : new mapping [Binois et al., 2015]

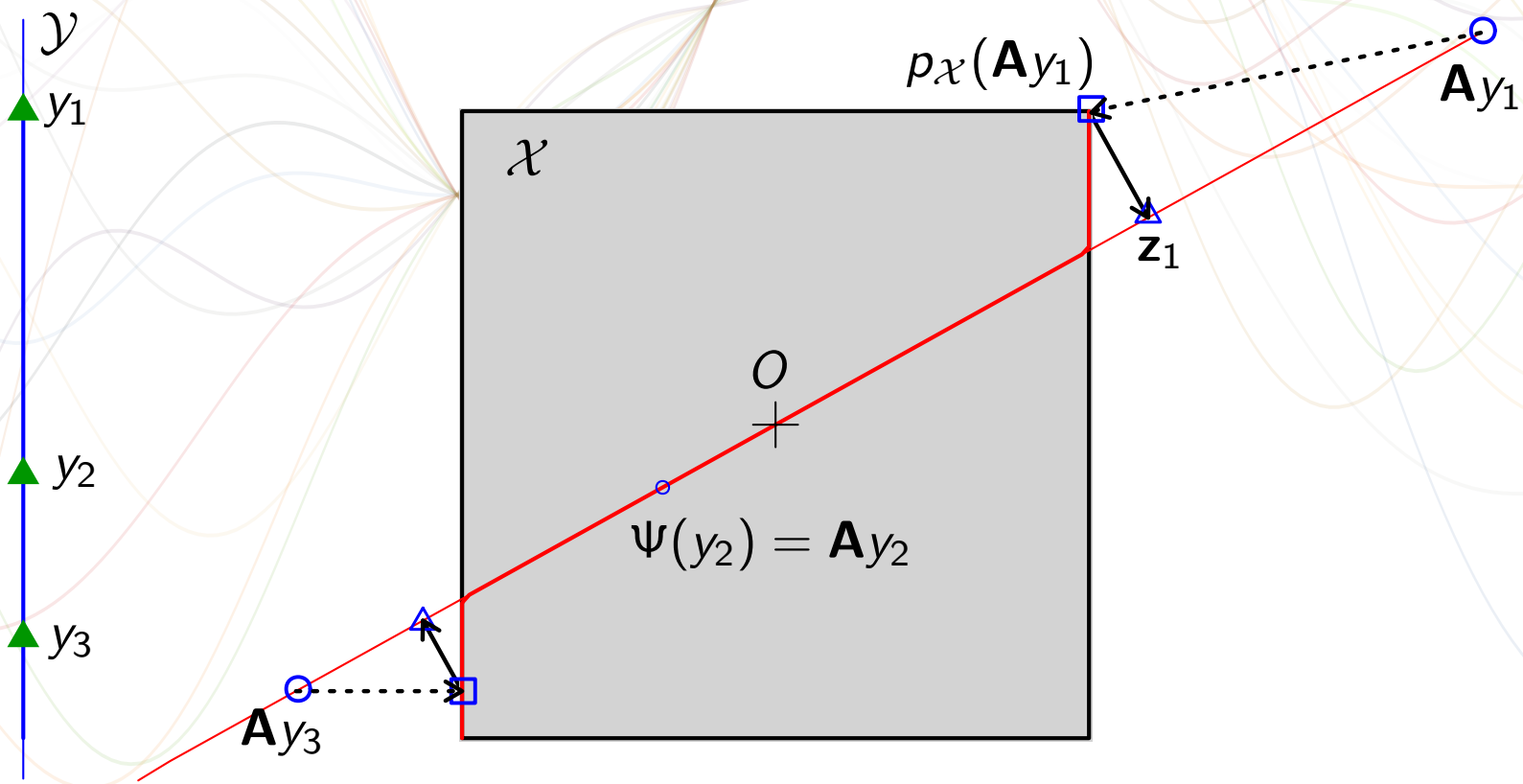
**Aim** : accounting for real distances within a low-dimensional kernel



# Towards a new kernel : new mapping [Binois et al., 2015]

**Aim** : accounting for real distances within a low-dimensional kernel

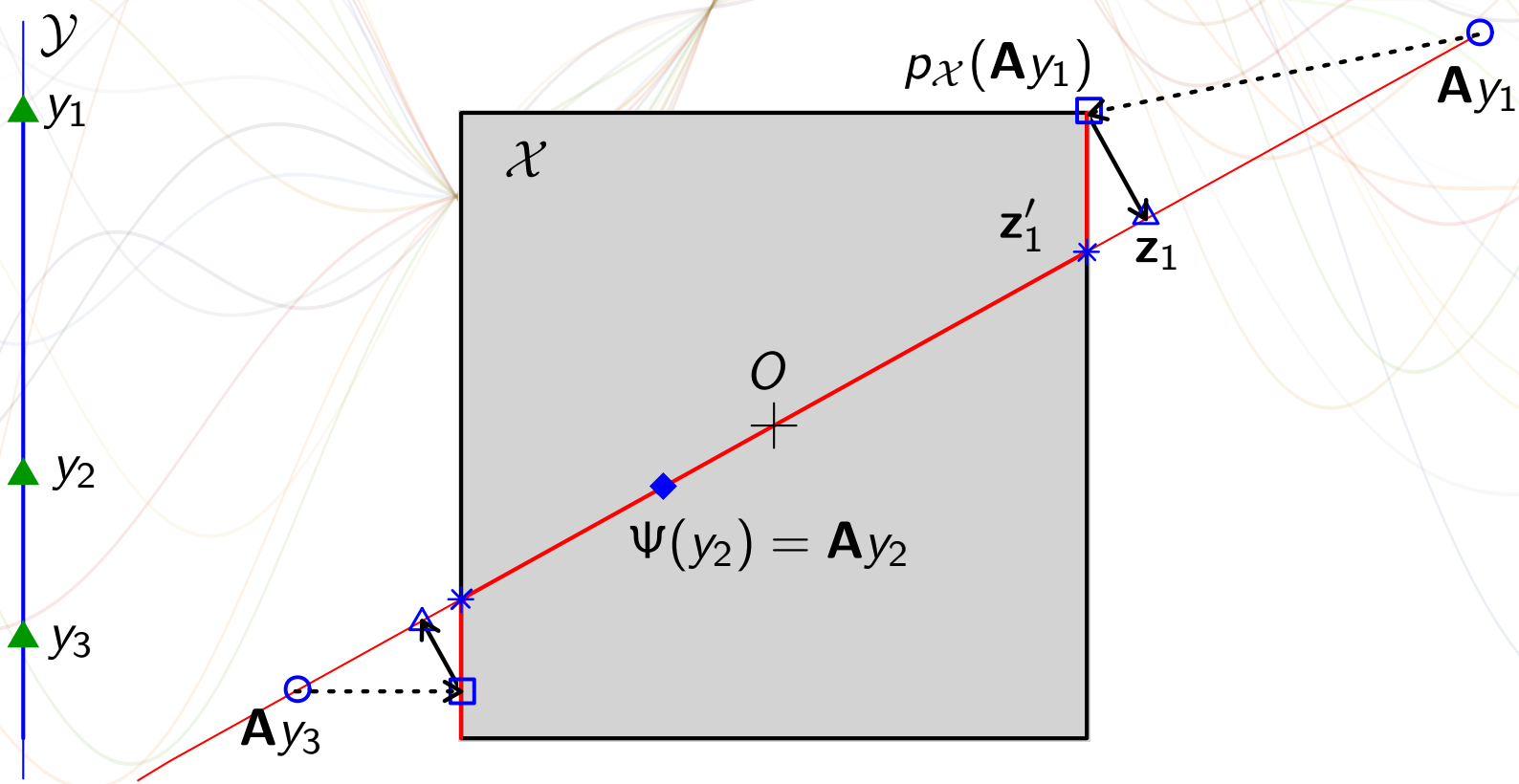
1) Projection in a low dimensional space : orthogonal projection onto Ran(**A**)



# Towards a new kernel : new mapping [Binois et al., 2015]

**Aim** : accounting for real distances within a low-dimensional kernel

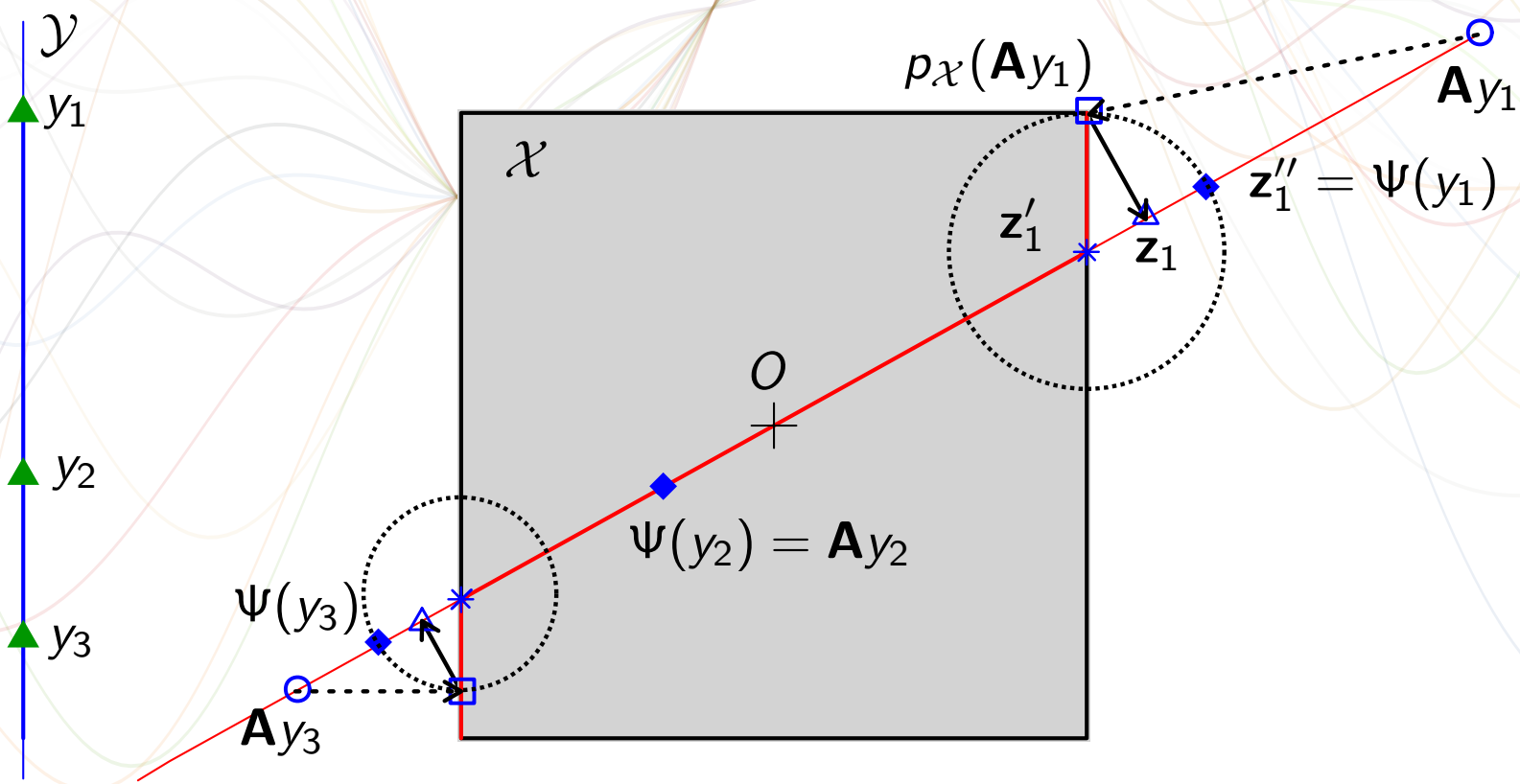
1) Projection in a low dimensional space : orthogonal projection onto Ran($\mathbf{A}$)
2) Reproducing high-dimensional distances on $p_{\mathcal{X}}(\mathbf{A}\mathbb{R}^d)$, using a pivot point



# Towards a new kernel : new mapping [Binois et al., 2015]

**Aim** : accounting for real distances within a low-dimensional kernel

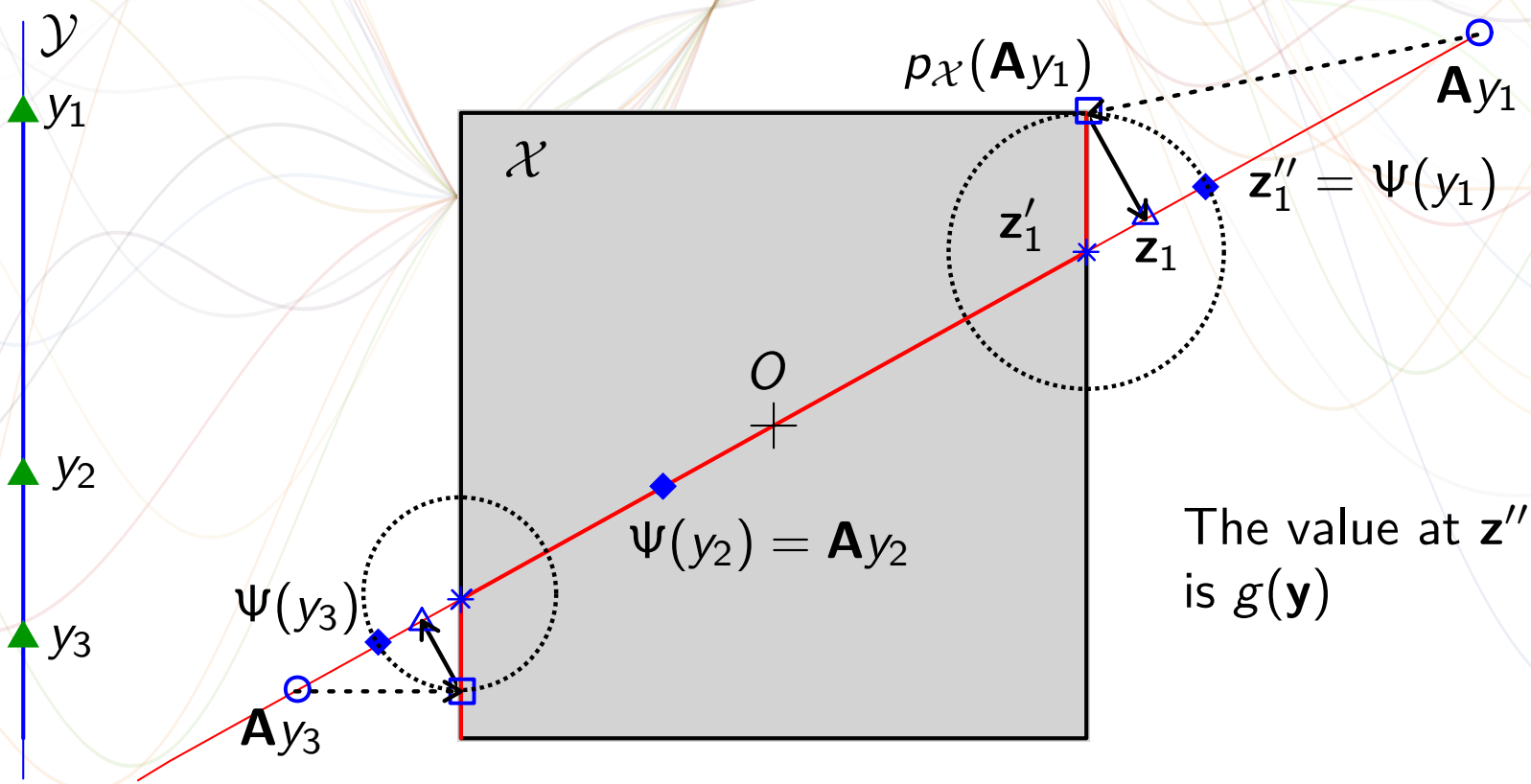
1) Projection in a low dimensional space : orthogonal projection onto Ran(**A**)
2) Reproducing high-dimensional distances on $p_{\mathcal{X}}(\mathbf{A}\mathbb{R}^d)$, using a pivot point



# Towards a new kernel : new mapping [Binois et al., 2015]

**Aim** : accounting for real distances within a low-dimensional kernel

1) Projection in a low dimensional space : orthogonal projection onto Ran($\mathbf{A}$)
2) Reproducing high-dimensional distances on $p_{\mathcal{X}}(\mathbf{A}\mathbb{R}^d)$, using a pivot point

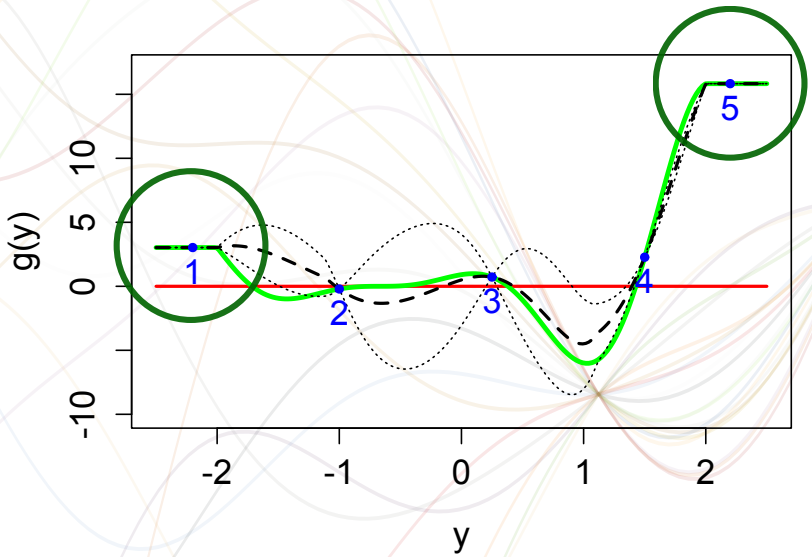



The value at $\mathbf{z}''=\Psi(\mathbf{y})$ is $g(\mathbf{y})$

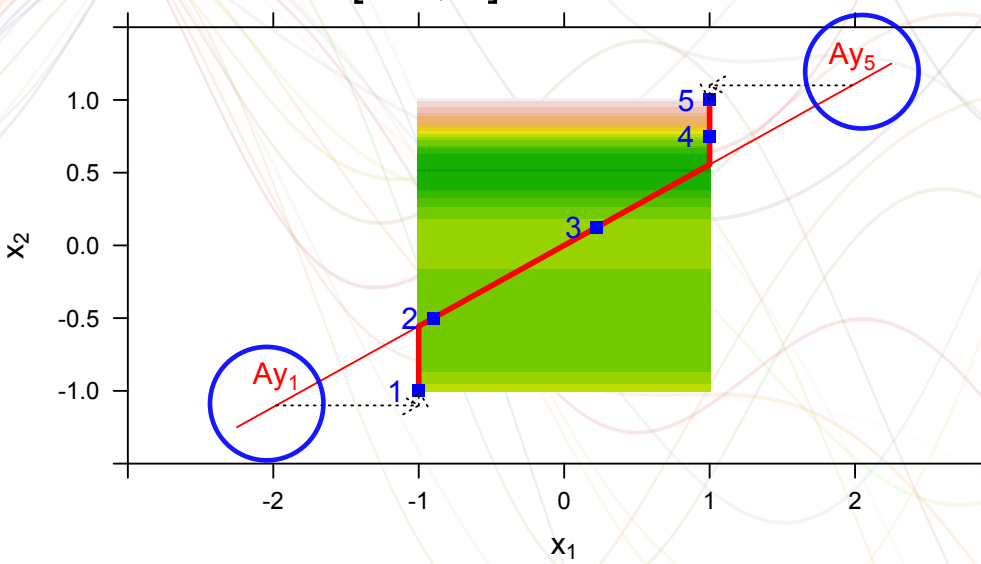
# Illustration : GP predictions with $k_\Psi$

$$\mathbf{A} = [0.9, 0.5]^T$$

$\mathcal{Y} = [-2.5, 2.5]$




$\mathcal{X} = [-1, 1]^2$

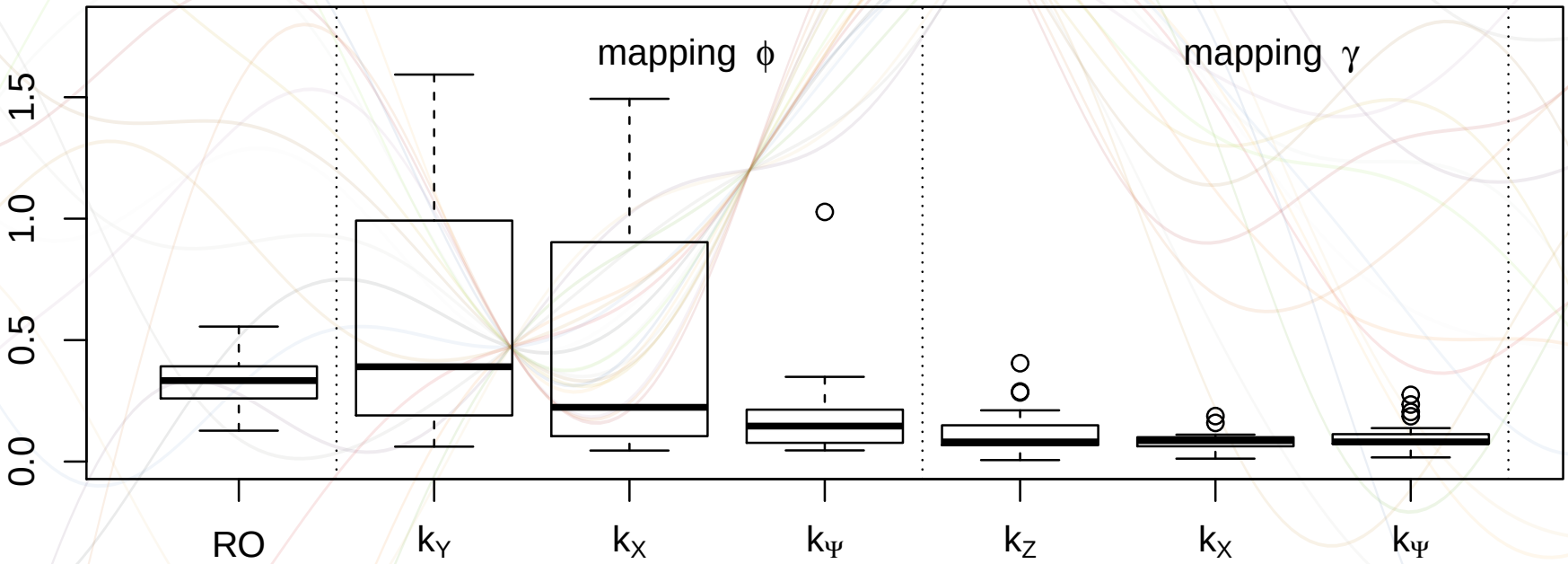



Warped kernel,

e.g. $k_\Psi(\mathbf{y}, \mathbf{y'}) = exp(-\frac{\|\Psi(\mathbf{y}) - \Psi(\mathbf{y}')\|_d^2}{2l^2})$

# An example of optimization performance in 50 dim. [Binois et al., 2018]

Hartman 6 function, $d = 6$, $D = 50$, 250 evaluations over 25 runs, optimality gap :




RO : Random optimization
$\phi(y)$ : embedding from $\mathcal{Y} = [-\sqrt{d}, \sqrt{d}]^d$ to $\mathcal{X}$
$\gamma(z)$ : embedding from $\mathcal{B} \subseteq Ran(\mathbf{A})$ to $\mathcal{X}$.

# Outline

# Some conclusions

- We illustrated the versatility of GPs on small data (metamodeling)
  - It was an overview, not a review
    $\rightarrow$ There are new GP methods every day ! (Ex : Create your own one)

# Some conclusions

- We illustrated the versatility of GPs on small data (metamodeling)
  - It was an overview, not a review
    $\rightarrow$ There are new GP methods every day ! (Ex : Create your own one)
- There is a florishing literature on GPs for big data :
  - neural networks for GPs : deep GPs
  - scalable GPs with variational inference
  - ...

# Some conclusions

- We illustrated the versatility of GPs on small data (metamodeling)
  - It was an overview, not a review
    $\rightarrow$ There are new GP methods every day ! (Ex : Create your own one)
- There is a florishing literature on GPs for big data :
  - neural networks for GPs : deep GPs
  - scalable GPs with variational inference
  - ...
- What statisticians & mathematicians can add more
  - A better theoretical understanding (e.g. convergence results on EGO)
  - Connections with math. theory (e.g. here : spline and inequalities)
  - Different point of views (e.g. identifiability "vs" prediction only)
  - ...

## Acknowledgements


- Some works are coming from the Chair in Applied Mathematics OQUAIDO, gathering partners in technological research (BRGM, CEA, IFPEN, IRSN, Safran, Storengy) and academia (CNRS, Ecole Centrale de Lyon, Mines Saint-Etienne, University of Grenoble, University of Nice, University of Toulouse) around advanced methods for Computer Experiments.
- Other ones are coming from the former ReDICE consortium
- Many thanks to several slides co-writers : M. Binois, Y. Deville, N. Durrande, D. Ginsbourger, R. Le Riche, A. Lopez Lopéra, E. Padonou.
- All collaborators involved in the presented works !


# Références I


Amato, U., Antoniadis, A., and Pensky, M. (2006).
Wavelet kernel penalized estimation for non-equispaced design regression.
*Statistics and Computing*, 16(1) :37–55.

Aronszajn, N. (1950).
Theory of reproducing kernels.
*Transactions of the American mathematical society*, 68(3) :337–404.

Batton-Hubert, M., Binois, M., and Padonou, E. (2013).
Inverse modeling to estimate methane surface emission with optimization and reduced models : application of waste landfill plants.
In *13th Annual Conference of the European Network for Business and Industrial*, Ankara, Turkey.

Bay, X., Grammont, L., and Maatouk, H. (2016).
Generalization of the Kimeldorf-Wahba correspondence for constrained interpolation.
*Electronic journal of statistics* , 10(1) :1580–1595.

Bect, J., Bachoc, F., and Ginsbourger, D. (2017).
A supermartingale approach to Gaussian process based sequential design of experiments.
working paper or preprint.

# Références II


Berlinet, A. and Thomas-Agnan, C. (2011).

*Reproducing kernel Hilbert spaces in probability and statistics*.

Springer Science & Business Media.

Binois, M., Ginsbourger, D., and Roustant, O. (2015).

A warped kernel improving robustness in Bayesian optimization via random embeddings.

In Dhaenens, C., Jourdan, L., and Marmion, M., editors, *Learning and Intelligent Optimization. LION 2015. Lecture Notes in Computer Science*, volume 8994. Springer, Cham.

Binois, M., Ginsbourger, D., and Roustant, O. (2018).

On the choice of the low-dimensional domain for global optimization via random embeddings.

Technical report, https ://hal.inria.fr/hal-01508196.

Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V., and Richet, Y. (2014).

Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set.

*Technometrics*, 56(4) :455–465.

# Références III


del Castillo, E., Colosimo, B. M., and Tajbakhsh, S. D. (2015).
Geodesic Gaussian processes for the parametric reconstruction of a free-form surface.
*Technometrics*, 57(1) :87–99.

Durrande, N., Hensman, J., Rattray, M., and Lawrence, N. (2016).
Detecting periodicities with Gaussian processes.
*PeerJ Computer Science*, 2 :e50.

Fang, K., Li, R., and Sudjianto, A. (2006).
*Design and Modeling for Computer Experiments*.
Chapman & Hall/CRC.

Fox, E. and Dunson, D. B. (2012).
Multiresolution Gaussian processes.
In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 737–745. Curran Associates, Inc.

Ginsbourger, D., Roustant, O., and Durrande, N. (2016).
On degeneracy and invariances of random fields paths with applications in Gaussian process modelling.
*Journal of Statistical Planning and Inference*, 170 :117 – 128.

# Références IV


Gneiting, T. (2013).
Strictly and non-strictly positive definite functions on spheres.
*Bernoulli*, 19(4) :1327–1349.

Gramacy, R. B. (2007).
An R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian process models.
*Journal of Statistical Software*, 19(9) :1–46.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998).
Efficient global optimization of expensive black-box functions.
*Journal of Global Optimization*, 13(4) :455–492.

Kimeldorf, G. and Wahba, G. (1971).
Some results on Tchebycheffian spline functions.
*Journal of mathematical analysis and applications*, 33(1) :82–95.

Krige, D. G. (1951).
A statistical approach to some basic mine valuation problems on the witwatersrand.
*Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52(6) :119–139.

# Références V


López-Lopera, A. F., Bachoc, F., Durrande, N., and Roustant, O. (2018).
Finite-dimensional Gaussian approximation with linear inequality constraints.
Technical report, https ://arxiv.org/abs/1710.07453.

Maatouk, H. and Bay, X. (2017).
Gaussian process emulators for computer experiments with inequality constraints.
*Mathematical Geosciences*, 49(5) :557–582.

Matheron, G. (1963).
Principles of geostatistics.
*Economic Geology*, 58 :1246–1266.

Močkus, J. (1975).
On Bayesian methods for seeking the extremum.
In Marchuk, G. I., editor, *Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974*, pages 400–404, Berlin, Heidelberg. Springer Berlin Heidelberg.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015).
A multiresolution Gaussian process model for the analysis of large spatial datasets.
*Journal of Computational and Graphical Statistics*, 24(2) :579–599.

# Références VI


Padonou, E. and Roustant, O. (2016).

Polar Gaussian processes and experimental designs in circular domains.

*SIAM/ASA Journal on Uncertainty Quantification*, 4(1) :1014–1033.

Padonou, E. and Roustant, O. (2017).

Analyse de sensibilité en domaine circulaire.

In *SFdS, 49èmes journées de Statistique*.

Park, S. and Choi, S. (2010).

Hierarchical Gaussian process regression.

In Sugiyama, M. and Yang, Q., editors, *Proceedings of 2nd Asian Conference on Machine Learning*, volume 13 of *Proceedings of Machine Learning Research*, pages 95–110.

Porcu, E., Bevilacqua, M., and Genton, M. (2016).

Spatio-temporal covariance and cross-covariance functions of the great circle distance on a sphere.

*JASA*, 111(514) :888–898.

Rasmussen, C. E. and Williams, C. K. (2006).

*Gaussian processes for machine learning*.

the MIT Press.

# Références VII


Roustant, O., Padonou, E., Deville, Y., Clément, A., Perrin, G., Giorla, J., and Wynn, H. (2018).

Group kernels for Gaussian process metamodels with categorical inputs.

*ArXiv e-prints*.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989).

Design and analysis of computer experiments.

*Statistical Science*, 4(4) :409–435.

Santner, T. J., Williams, B. J., and Notz, W. (2003).

*The Design and Analysis of Computer Experiments*.

Springer-Verlag, New York.

Schoenberg, I. (1942).

Positive definite functions on spheres.

*Duke Math. J.*, 9 :96–108.

Vazquez, E. and Bect, J. (2010).

Convergence properties of the expected improvement algorithm with fixed mean and covariance functions.

*Journal of Statistical Planning and Inference*, 140(11) :3088 – 3095.

# Références VIII

Wang, Z., Zoghi, M., Hutter, F., Matheson, D., and de Freitas, N. (2013).
Bayesian optimization in high dimensions via random embeddings.
In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*.