An introduction to optimization under uncertainties

CIMPA, Ulaanbaatar, Mongolia

Didier Lucor

didier.lucor@lisn.upsaclay.fr

July 09th 2021

Course II

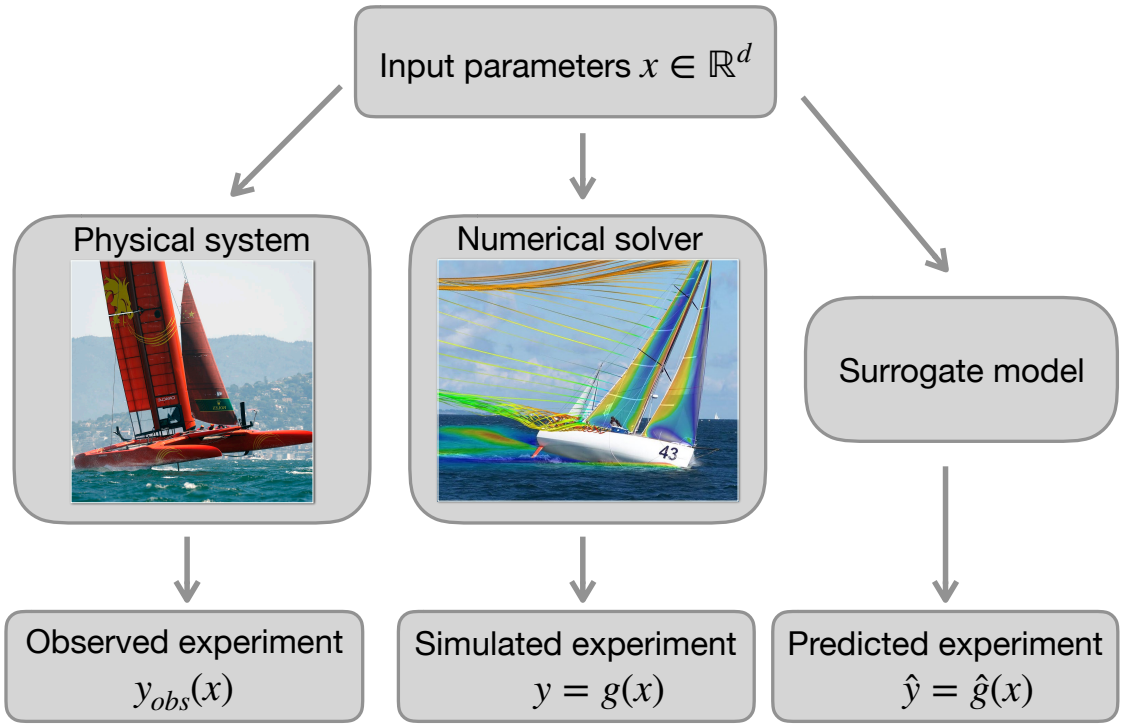
Organization of the class

1. Course I : Concepts, formalism and some classes of resolution methods for simple optimization under uncertainty (1.5h, Thursday)

2. Course II : quantification of uncertainty for more efficient optimization under uncertainties : metamodeling with Gaussian Processes (1.5h, Friday)
   - Surrogate modeling
   - Gaussian vectors and gaussian processes
   - Regression via conditioned gaussian processes
   - Challenge of dimensionality and space-filling designs

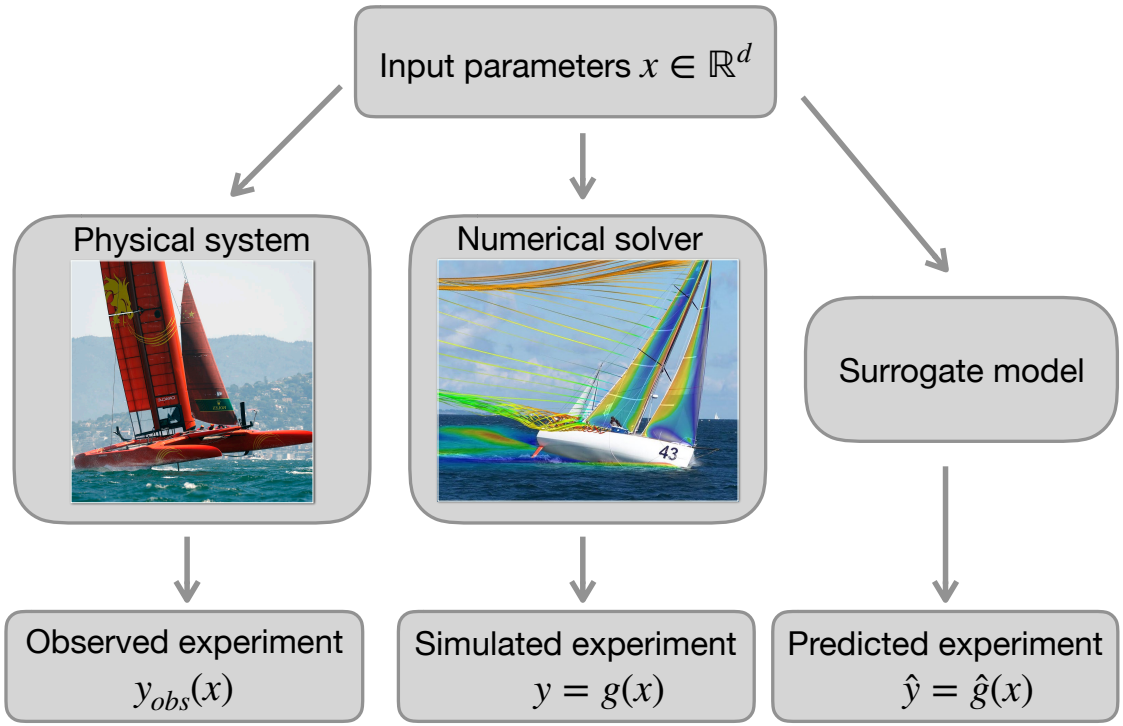
3. Course III : Metamodeling-based Reliability-Based Design Optimization (1.5h, Saturday)

What is a surrogate model ?



Introduction

What is a surrogate model ?



- **Regression** framework with x : input parameters and y is the QoI
- **Surrogate model** (**SM**) $\equiv$ response surface, **metamodel**, simplified model, emulator, . . . ; surrogate modeling, e.g. $\equiv$ linear regression (supervized learning)
- a surrogate model is an **approximation** of our reference model at hand : trading **accuracy** for **speed**

Why do we need a a surrogate model ?

- **Uncertainty** quantifications such as uncertainty propagation, sensitivity analysis (SA), **OUU**, ... need more than a few samples around the nominal solution
- These analyses require **numerous** n **evaluations** (i.e. runs/calls) of the deterministic solver to be useful/accurate

  they may require $n \sim \mathcal{O}(10)$ to determine the response **central moments**, $n \sim \mathcal{O}(10^3 - 10^4)$ for **sensitivity analysis** and $n \sim \mathcal{O}(10^5)$ for the correct estimation of **rare events** $\propto$ **reliability problems**

Why do we need a a surrogate model ?

- We call s a function, associating to each $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ a **scalar** deterministic value $y = s(\boldsymbol{x}) \in \mathbb{R}$; s represents our **black-box model**/solver/code/experiment. **Note that $\boldsymbol{x}$ is not random for now.** We wish to characterize the entire image of the function :

$$\mathcal{S}_s = \{(\boldsymbol{x}, y) \in \mathcal{X} \times \mathbb{R} \mid y = s(\boldsymbol{x})\}\,.$$

- **Assumptions** :

  The number of calls to s is **limited** (e.g. long simulations, slow computers, expensive experimentations)

  We can hold exact $s(\boldsymbol{x}^{(i)})$ or noisy $s(\boldsymbol{x}^{(i)}) + \epsilon^{(i)}$ model observations

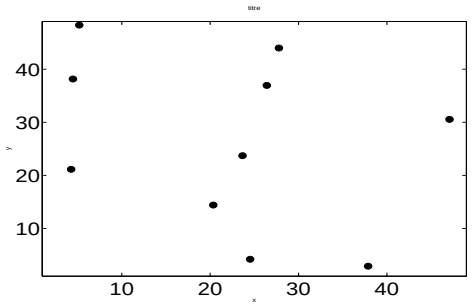
Goal :

maximize the knowledge of the entire system response $\mathcal{S}_s$ (continuous representation) from as little number of observations as possible (discrete representation).

Introduction

main steps of surrogate modeling

1. Design of Experiments (DoE)

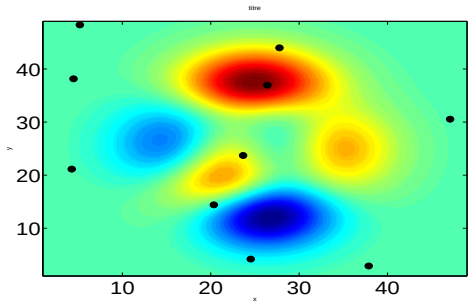


we choose a sample of n points $\left\{\boldsymbol{x}^{(i)},\ 1 \leq i \leq n\right\}$ in $\mathcal{X}$.

2. Simulation

we evaluate n times the model s.

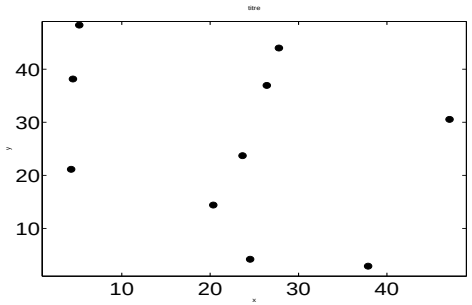
3. Metamodeling



we approximate the response surface "in between" the calculated points obtained from the simulations.

main steps of surrogate modeling

1. Design of Experiments (DoE)

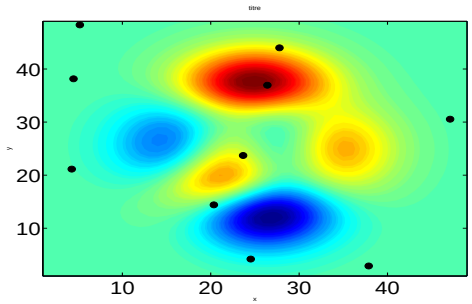


we choose a sample of n points $\left\{\boldsymbol{x}^{(i)},\ 1 \leq i \leq n\right\}$ in $\mathcal{X}$.

2. Simulation

we evaluate n times the model s.

3. Metamodeling



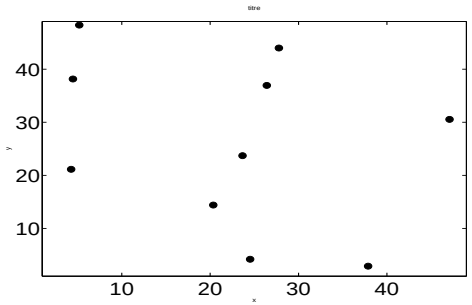
we approximate the response surface "in between" the calculated points obtained from the simulations.

- **Important** : we want our **approximation** of s, called $\hat{s}$, to be capable of predicting well the response for **new parametric value** $\boldsymbol{x}^* \in \mathcal{X}$ (no information available !) :

$$\hat{y} = \hat{s}(\boldsymbol{x}^*) \approx y^* = s(\boldsymbol{x}^*)$$

main steps of surrogate modeling

1. Design of Experiments (DoE)

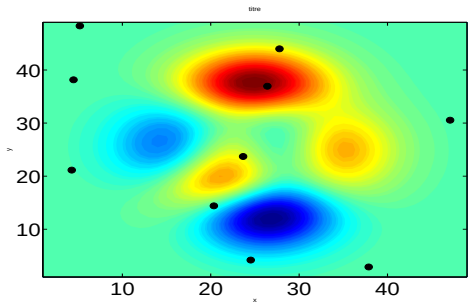


we choose a sample of n points $\left\{\boldsymbol{x}^{(i)},\ 1 \leq i \leq n\right\}$ in $\mathcal{X}$.

2. Simulation

we evaluate n times the model s.

3. Metamodeling



we approximate the response surface "in between" the calculated points obtained from the simulations.

- **Important** : we want our **approximation** of s, called $\hat{s}$, to be capable of predicting well the response for **new parametric value** $\boldsymbol{x}^* \in \mathcal{X}$ (no information available !) :

$$\hat{y} = \hat{s}(\boldsymbol{x}^*) \approx y^* = s(\boldsymbol{x}^*)$$

- We are in the **small data** range : i.e. data are expensive/few (n/d is small).

different types of metamodeling techniques

- (generalized) Linear models (GLM)
- Splines, Radial Basis Functions (RBF)
- Additives models
- Regression trees (CART)
- Support Vector Machines (SVM)
- (Deep) Neural Networks (DNN)
- Kriging or Gaussian Process (GP) models
- Polynomial Chaos expansion (PCE)

Introduction

different types of metamodeling techniques

- (generalized) Linear models (GLM)
- Splines, Radial Basis Functions (RBF)
- Additives models
- Regression trees (CART)
- Support Vector Machines (SVM)
- (Deep) Neural Networks (DNN)
- **Kriging or Gaussian Process (GP) models** **[Krige 1951, Matheron 1963, Santner et al., 2003, Rasmussen & Williams 2006]**
- Polynomial Chaos expansion (PCE)

why GP ?

- **Gaussian distribution is reasonable** for modeling a large variety of random variables
- GP are **simple** to define and simulate
- GP are used in various fields such as **computer experiments**, machine learning, bayesian optimization,...
- They are **fully** characterized by their **mean** and **covariance** functions
- As we will see, Gaussian properties simplify the resolution of problems, and they have been the **most studied theoretically**

Kriging metamodeling :

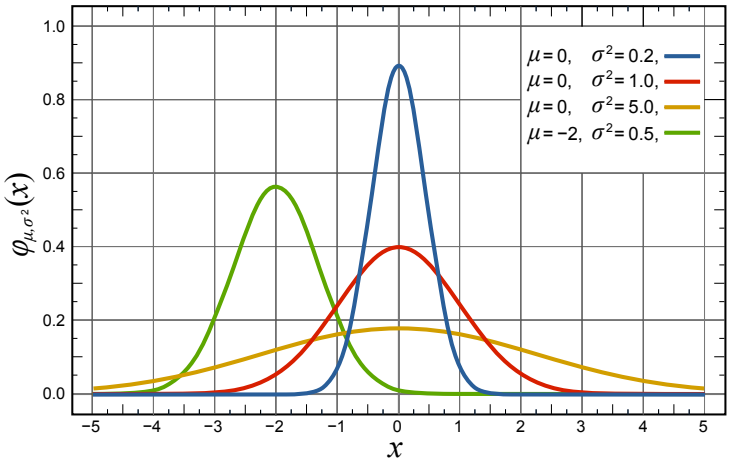
use **stochastic method** (based on GP) to approximate deterministic function !
$\rightarrow$ provides a **stochastic error bound**

Table of contents

Gaussian distribution

- The **gaussian** distribution arises in many different contexts and can be motivated from a variety of different perspectives (maximizes entropy of a single real variable, CLT) and has many important analytical properties.
- a Gaussian random variable Z on $\mathbb{R}$ with **mean** μ_Z and **variance** $\sigma_Z^2 > 0$ has the following distribution :

$$f_{Z\sim\mathcal{N}(\mu_Z,\sigma_Z^2)} = \frac{1}{\sqrt{2\pi}\sigma_Z}\exp\left(-\frac{1}{2\sigma_Z^2}(Z-\mu_Z)^2\right)$$



Gaussian distribution and gaussian vector

- A random vector (**RV**) $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$ is said **gaussian** iif its **joint PDF** $f_{\boldsymbol{Z}}$ writes :

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma_{\boldsymbol{Z}}}} \exp\left\{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{Z}})^T \Sigma_{\boldsymbol{Z}}^{-1} (\boldsymbol{z}-\boldsymbol{\mu}_{\boldsymbol{Z}})\right\},$$

- $\boldsymbol{\mu}_{\boldsymbol{Z}} = \mathbb{E}(\boldsymbol{Z}) = (\mathbb{E}(Z_1), \ldots, \mathbb{E}(Z_n))^T$ its **mean vector**
- $\Sigma_{\boldsymbol{Z}} \equiv C_{\boldsymbol{Z}} = \mathbb{E}\left((\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])^T (\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])\right)$ its **covariance matrix**

$$\Sigma_{\boldsymbol{Z}} = \begin{pmatrix} \sigma_1^2 & \mathrm{cov}(Z_1, Z_2) & \ldots & \mathrm{cov}(Z_1, Z_n) \\ \mathrm{cov}(Z_2, Z_1) & \sigma_2^2 & \ldots & \mathrm{cov}(Z_2, Z_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(Z_n, Z_1) & \mathrm{cov}(Z_n, Z_2) & \ldots & \sigma_n^2 \end{pmatrix}$$

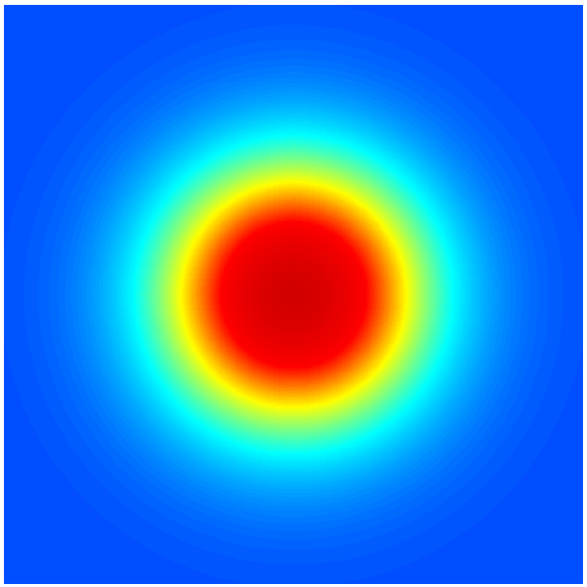
avec $\mathrm{cov}(Z_i, Z_j) = \mathrm{cov}(Z_j, Z_i) = \mathbb{E}\left((Z_i - \mathbb{E}(Z_i))(Z_j - \mathbb{E}(Z_j))\right)$ et $\sigma_i^2 = \mathrm{var}(Z_i)$.

- $\boldsymbol{Z}$ gaussian RV : fully characterized by its mean and covariance (assumed always invertible)

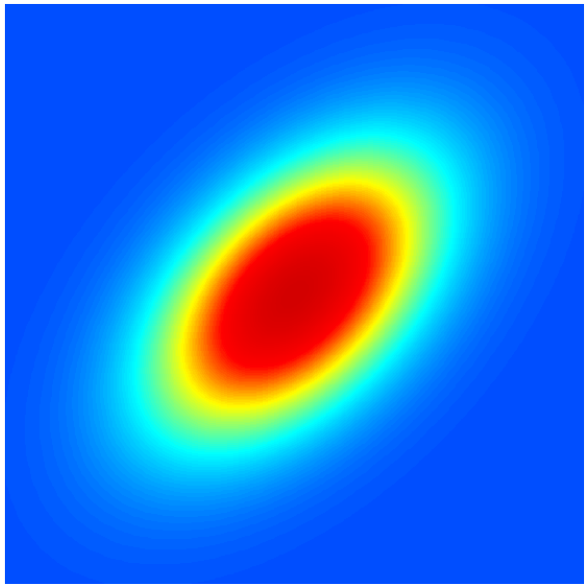
$\rightarrow$ **identifying these 2 quantities is enough !**

Visual 2D representations

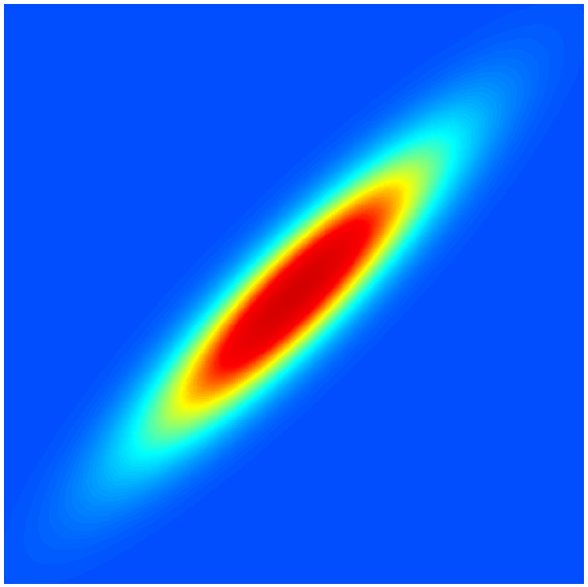
$\mathrm{cov}(Z_1, Z_2) = 0$



$\mathrm{cov}(Z_1, Z_2) = 0.5$



$\mathrm{cov}(Z_1, Z_2) = 0.9$



Gaussian vectors and gaussian processes

Definition

- $\boldsymbol{Z}$: n-dimensional (multivariate) RV is a **gaussian RV**, noted in the following **GRV**, if **any linear combination of its components follows a gaussian distribution** :

$$\forall \boldsymbol{a} \in \mathbb{R}^n \qquad \boldsymbol{a}^T \boldsymbol{Z} = \sum_{i=1}^{d} a_i Z_i \sim \mathcal{N}(\cdot, \cdot)$$

Properties

- a GRV $\Rightarrow$ has Gaussian components (be careful, the reciprocal is not true)
- If a RV components are Gaussian + **independent** $\Rightarrow$ vector is a GRV
- The sum of two independent Gaussian vectors is a Gaussian vector
- GRV components are **independent iff** they are **uncorrelated** (not true for other PDF types !)

Properties

- **stability under linear transformation**

  if $\boldsymbol{Z}$ is a GRV of d-dimension, of mean $\boldsymbol{\mu_Z}$ and covariance $C_{\boldsymbol{Z}}$, then for any matrix M of size (m, d) and any RV : $\boldsymbol{y}$ of size m, $\quad \boldsymbol{X} = M\boldsymbol{Z} + \boldsymbol{y}$ is also a GRV, with :

$$\mathbb{E}(\boldsymbol{X}) = M\boldsymbol{\mu_Z} + \boldsymbol{y}, \quad C_{\boldsymbol{X}} = MC_{\boldsymbol{Z}}M^T.$$

- **generation of iid realizations from GRV**

  to generate random **independent** realizations of any GRV from a standard normal random number generator, we use the following :

$$\boldsymbol{Z} \text{ and } (\boldsymbol{\mu_Z} + R\boldsymbol{\xi}) \text{ follow same joint distribution}$$

  - $\boldsymbol{Z}$ is a GRV, of mean $\boldsymbol{\mu_Z}$ and covariance $C_{\boldsymbol{Z}}$,
  - $\boldsymbol{\xi} \sim \mathcal{N}(\mathbb{0}, \mathbb{1})$,
  - R is a symmetric matrix such that $R^T R = C_{\boldsymbol{Z}}$,

Conditioning properties

- **Gaussian conditioning theorem** Let $\boldsymbol{Z}$ and $\boldsymbol{X}$ be two GRV of size n and m, respectively, with their joint pdf noted

$$\begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{X} \end{pmatrix} \sim \mathcal{N}_{n+m}\left(\begin{pmatrix} \boldsymbol{\mu_Z} \\ \boldsymbol{\mu_X} \end{pmatrix}, \begin{pmatrix} C_{\boldsymbol{Z}} & C_{\boldsymbol{ZX}} \\ C_{\boldsymbol{XZ}} & C_{\boldsymbol{X}} \end{pmatrix} \right)$$

with $\boldsymbol{\mu_Z}$ et $\boldsymbol{\mu_X}$ the mean vectors of size n and m, respectively, the covariance matrix $C_{\boldsymbol{Z}}$ of size $n \times n$, $C_{\boldsymbol{ZX}}$ of size $n \times m$, $C_{\boldsymbol{ZX}} = C_{\boldsymbol{XZ}}^T$ and $C_{\boldsymbol{X}}$ of size $m \times m$.

then **conditionally** on $\boldsymbol{X}$, $\boldsymbol{Z}$ is again a GRV with following moments :

$$(\boldsymbol{Z} \mid \boldsymbol{X} = \boldsymbol{x}) \ \sim \ \mathcal{N}(\boldsymbol{\mu}^{\text{Cond}}(\boldsymbol{x}), [C^{\text{Cond}}(\boldsymbol{x})]),$$

$$\begin{cases} \boldsymbol{\mu}^{\text{Cond}}(\boldsymbol{x}) = \mathbb{E}(\boldsymbol{Z} \mid \boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{\mu_Z} + C_{\boldsymbol{XZ}} C_{\boldsymbol{X}}^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu_X} \right), \\ C^{\text{Cond}}(\boldsymbol{x}) = C_{\boldsymbol{Z}} - C_{\boldsymbol{XZ}} C_{\boldsymbol{X}}^{-1} C_{\boldsymbol{XZ}}^T. \end{cases}$$

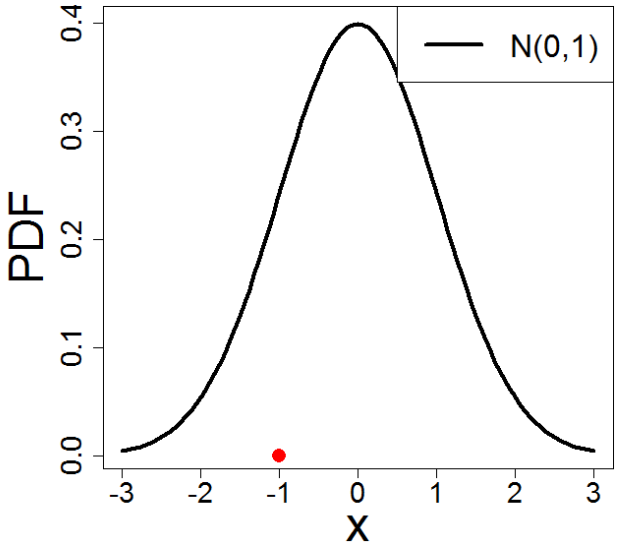
Exercise : we assume $X \sim \mathcal{N}(0,1)$, $Z \sim \mathcal{N}(0,1)$ and are correlated : $\mathbb{E}[XZ] = \rho \leq 1$. Quantify the influence of an observation $X = x$ on f_Z ?

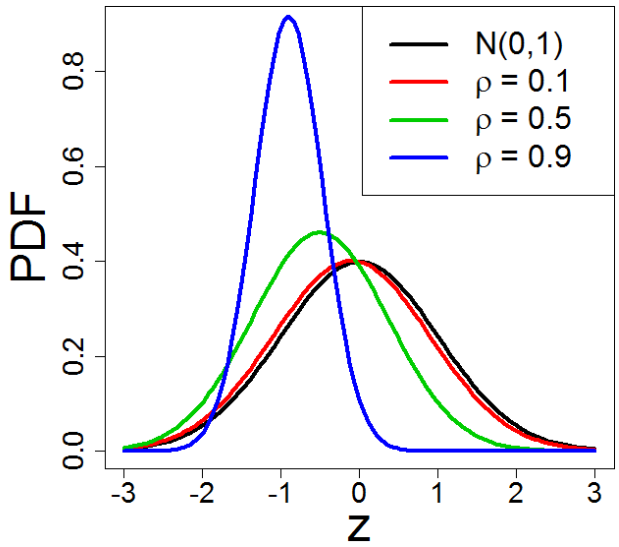
Illustration for different levels of correlation ρ

$X \sim \mathcal{N}(0,1),\ Z \sim \mathcal{N}(0,1),\ -1 \leq \mathbb{E}[XZ] = \rho \leq 1$

$\Rightarrow (Z \mid X = x) \ \sim \ \mathcal{N}(\rho x, 1-\rho^2).$

We condition Y thanks to the knowledge of X, e.g. $X = -1$





Observation :

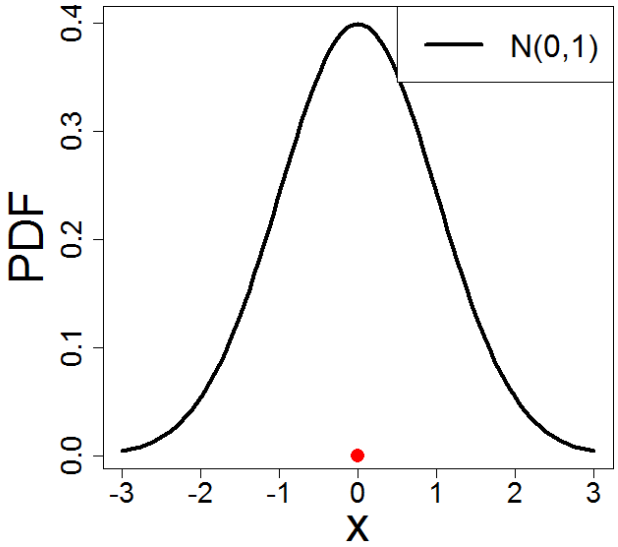
Conditioning changes the mean and reduces the variance !

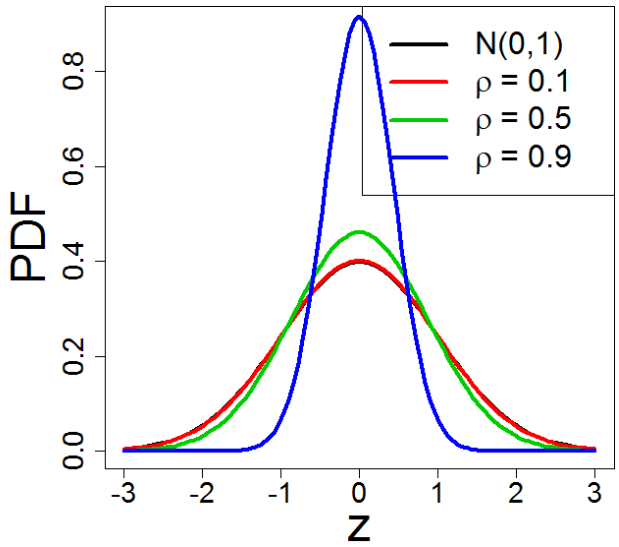
Illustration for different levels of correlation ρ

$X \sim \mathcal{N}(0,1),\ Z \sim \mathcal{N}(0,1),\ -1 \leq \mathbb{E}[XZ] = \rho \leq 1$

$\Rightarrow (Z \mid X = x) \ \sim \ \mathcal{N}(\rho x, 1-\rho^2).$

We condition Z thanks to the knowledge of X, e.g. $X = 0$





Observation :

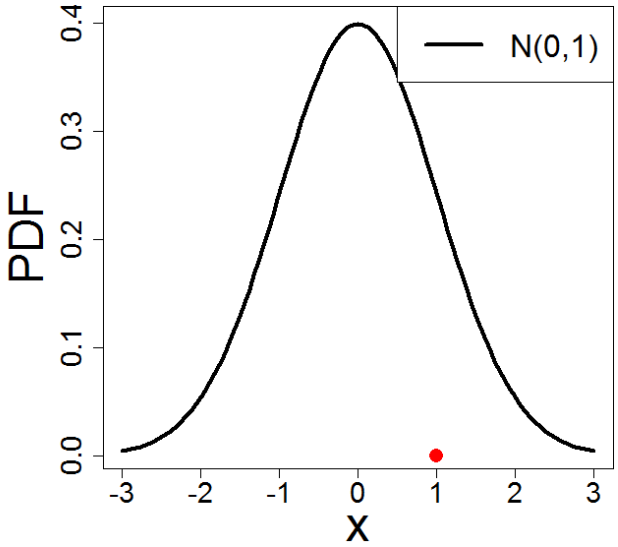
Conditioning changes the mean and reduces the variance !

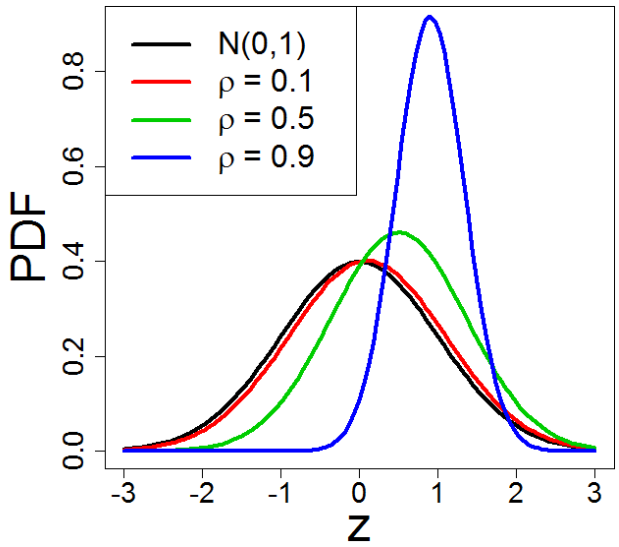
Illustration for different levels of correlation ρ

$X \sim \mathcal{N}(0,1),\ Z \sim \mathcal{N}(0,1),\ -1 \leq \mathbb{E}[XZ] = \rho \leq 1$

$\Rightarrow (Z \mid X = x) \ \sim \ \mathcal{N}(\rho x, 1-\rho^2).$

We condition Z thanks to the knowledge of X, e.g. $X = 1$





Observation :

Conditioning changes the mean and reduces the variance !

Stochastic Processes (SP)

Natural phenomena (wind speed, ground acceleration, wave height, etc) : good **SP** examples. They are **not predictable**, i.e. : their observation never provides the same signal. 2 interpretations :

- a SP on $\mathcal{X}$: a functional $Z : \mathcal{X} \to \mathbb{R}$ / $Z(\boldsymbol{x})$ is a **random variable** for *each* $\boldsymbol{x} \in \mathcal{X}$.
- alternatively a SP is a **function** on $\mathcal{X}$ that is **random**

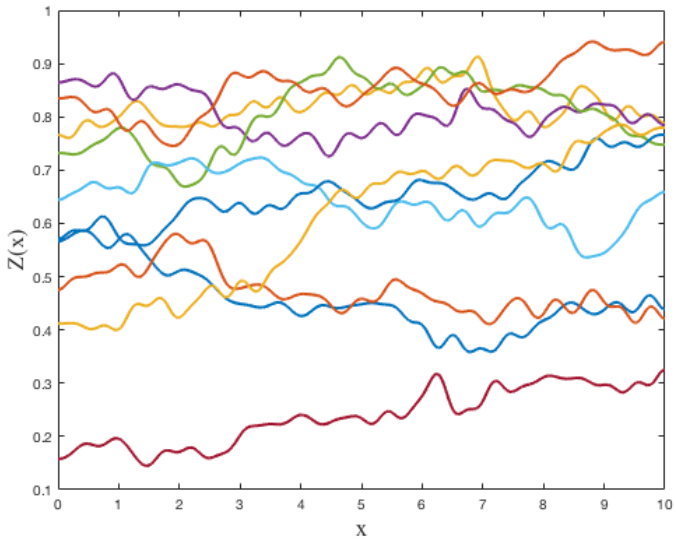
Probability space (Ω_Z, A_Z, P_Z) *indexed* on $\mathcal{X}$

– SP $\equiv$ **parametrized infinite** series of random variables :

$$\{Z_{(\boldsymbol{x})}\}_{\boldsymbol{x}\in\mathcal{X}} \text{ with } Z_{(\boldsymbol{x})} : \omega \mapsto Z(\omega, \boldsymbol{x})$$

Warning ! notation sometimes very much simplified $\Rightarrow Z(\boldsymbol{x})$

– The probability space Ω_Z is the **same** for all $Z(\omega, \boldsymbol{x})$ with $\boldsymbol{x} \in \mathcal{X} \in \mathbb{R}^d$



Stochastic Processes (SP)

Natural phenomena (wind speed, ground acceleration, wave height, etc) : good **SP** examples. They are **not predictable**, i.e. : their observation never provides the same signal. 2 interpretations :

- a SP on $\mathcal{X}$: a functional $Z : \mathcal{X} \to \mathbb{R}$ / $Z(\boldsymbol{x})$ is a **random variable** for *each* $\boldsymbol{x} \in \mathcal{X}$.
- alternatively a SP is a **function** on $\mathcal{X}$ that is **random**

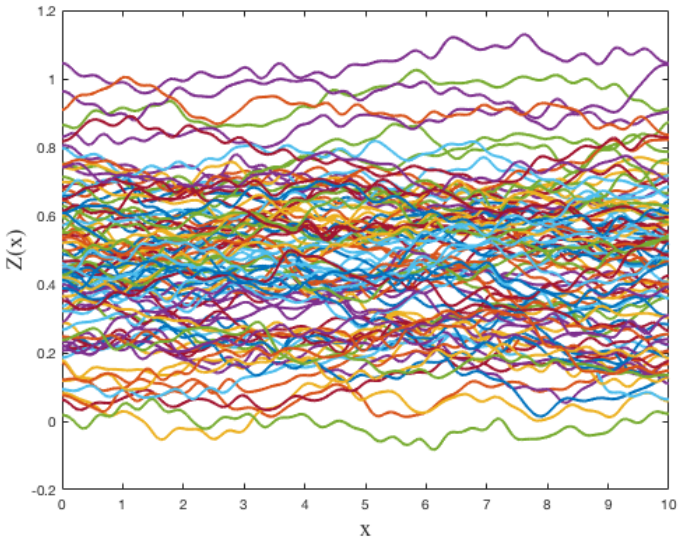
Probability space (Ω_Z, A_Z, P_Z) *indexed* on $\mathcal{X}$

– SP $\equiv$ **parametrized infinite** series of random variables :

$$\{Z_{(\boldsymbol{x})}\}_{\boldsymbol{x}\in\mathcal{X}} \text{ with } Z_{(\boldsymbol{x})} : \omega \mapsto Z(\omega, \boldsymbol{x})$$

Warning ! notation sometimes very much simplified $\Rightarrow Z(\boldsymbol{x})$

– The probability space Ω_Z is the **same** for all $Z(\omega, \boldsymbol{x})$ with $\boldsymbol{x} \in \mathcal{X} \in \mathbb{R}^d$



Stochastic Processes (SP)

Natural phenomena (wind speed, ground acceleration, wave height, etc) : good **SP** examples. They are **not predictable**, i.e. : their observation never provides the same signal. 2 interpretations :

- a SP on $\mathcal{X}$: a functional $Z : \mathcal{X} \to \mathbb{R}$ / $Z(\boldsymbol{x})$ is a **random variable** for *each* $\boldsymbol{x} \in \mathcal{X}$.
- alternatively a SP is a **function** on $\mathcal{X}$ that is **random**

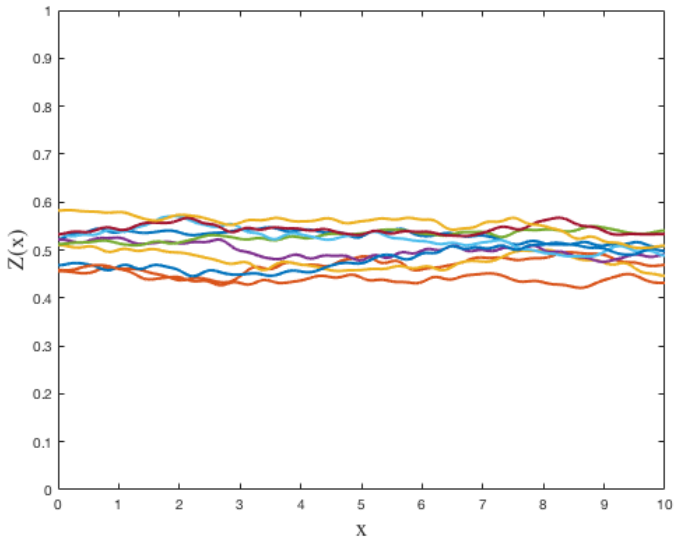
Probability space (Ω_Z, A_Z, P_Z) *indexed* on $\mathcal{X}$

– SP $\equiv$ **parametrized infinite** series of random variables :

$$\{Z_{(\boldsymbol{x})}\}_{\boldsymbol{x}\in\mathcal{X}} \text{ with } Z_{(\boldsymbol{x})} : \omega \mapsto Z(\omega, \boldsymbol{x})$$

Warning ! notation sometimes very much simplified $\Rightarrow Z(\boldsymbol{x})$

– The probability space Ω_Z is the **same** for all $Z(\omega, \boldsymbol{x})$ with $\boldsymbol{x} \in \mathcal{X} \in \mathbb{R}^d$



Stochastic Processes (SP)

Natural phenomena (wind speed, ground acceleration, wave height, etc) : good **SP** examples. They are **not predictable**, i.e. : their observation never provides the same signal. 2 interpretations :

- a SP on $\mathcal{X}$: a functional $Z : \mathcal{X} \to \mathbb{R}$ / $Z(\boldsymbol{x})$ is a **random variable** for *each* $\boldsymbol{x} \in \mathcal{X}$.
- alternatively a SP is a **function** on $\mathcal{X}$ that is **random**

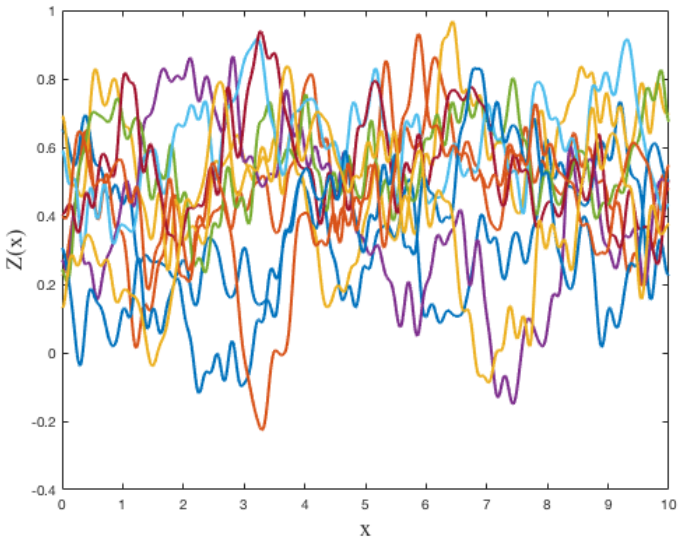
Probability space (Ω_Z, A_Z, P_Z) *indexed* on $\mathcal{X}$

– SP $\equiv$ **parametrized infinite** series of random variables :

$$\{Z_{(\boldsymbol{x})}\}_{\boldsymbol{x}\in\mathcal{X}} \text{ with } Z_{(\boldsymbol{x})} : \omega \mapsto Z(\omega, \boldsymbol{x})$$

Warning ! notation sometimes very much simplified $\Rightarrow Z(\boldsymbol{x})$

– The probability space Ω_Z is the **same** for all $Z(\omega, \boldsymbol{x})$ with $\boldsymbol{x} \in \mathcal{X} \in \mathbb{R}^d$



Stochastic Processes (SP)

Natural phenomena (wind speed, ground acceleration, wave height, etc) : good **SP** examples. They are **not predictable**, i.e. : their observation never provides the same signal. 2 interpretations :

- a SP on $\mathcal{X}$: a functional $Z : \mathcal{X} \to \mathbb{R}$ / $Z(\boldsymbol{x})$ is a **random variable** for *each* $\boldsymbol{x} \in \mathcal{X}$.
- alternatively a SP is a **function** on $\mathcal{X}$ that is **random**

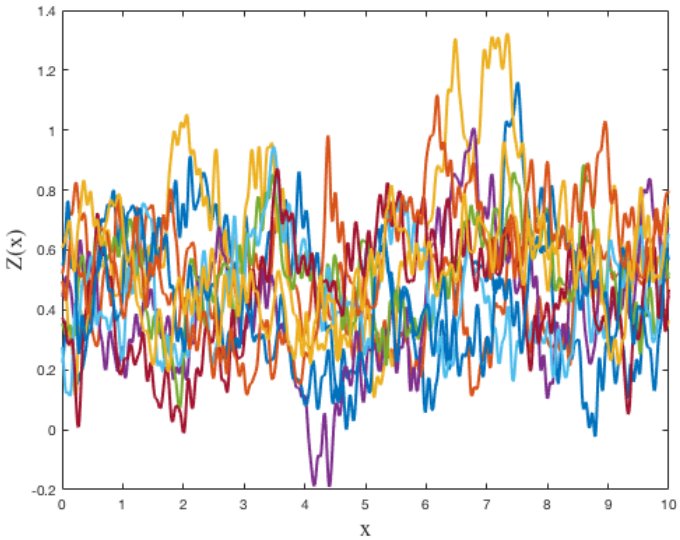
Probability space (Ω_Z, A_Z, P_Z) *indexed* on $\mathcal{X}$

– SP $\equiv$ **parametrized infinite** series of random variables :

$$\{Z_{(\boldsymbol{x})}\}_{\boldsymbol{x}\in\mathcal{X}} \text{ with } Z_{(\boldsymbol{x})} : \omega \mapsto Z(\omega, \boldsymbol{x})$$

Warning ! notation sometimes very much simplified $\Rightarrow Z(\boldsymbol{x})$

– The probability space Ω_Z is the **same** for all $Z(\omega, \boldsymbol{x})$ with $\boldsymbol{x} \in \mathcal{X} \in \mathbb{R}^d$



Gaussian Processes (GP)

Definition :

A SP Z on $\mathbb{R}^d$ is a **Gaussian Process (GP)** when – for all $n \in \mathbb{N}^*$ and – all $\left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\right\} \in \mathcal{X} \times \cdots \times \mathcal{X}$, the random vector $\boldsymbol{z} \equiv \left(Z(\boldsymbol{x}^{(1)}), \ldots, Z(\boldsymbol{x}^{(n)})\right)$ is **Gaussian**.

Gaussian Processes (GP)

Definition :

A SP Z on $\mathbb{R}^d$ is a **Gaussian Process (GP)** when – for all $n \in \mathbb{N}^*$ and – all $\left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\right\} \in \mathcal{X} \times \cdots \times \mathcal{X}$, the random vector $\boldsymbol{z} \equiv \left(Z(\boldsymbol{x}^{(1)}), \ldots, Z(\boldsymbol{x}^{(n)})\right)$ is **Gaussian**.

- a GP is fully characterized by :
  - its **mean** : $\mu(\boldsymbol{x}) = \mathbb{E}\left[Z(\boldsymbol{x})\right]$,
  - and **covariance** : $C(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}\left[\left(Z(\boldsymbol{x}) - \mu(\boldsymbol{x})\right)\left(Z(\boldsymbol{x}') - \mu(\boldsymbol{x}')\right)\right]$ (drives GP amplitude/speed of variation)
  - We sometimes note it as : $Z(\boldsymbol{x}) \sim \mathcal{GP}(\mu(\boldsymbol{x}), C(\boldsymbol{x}, \boldsymbol{x}'))$

Gaussian Processes (GP)

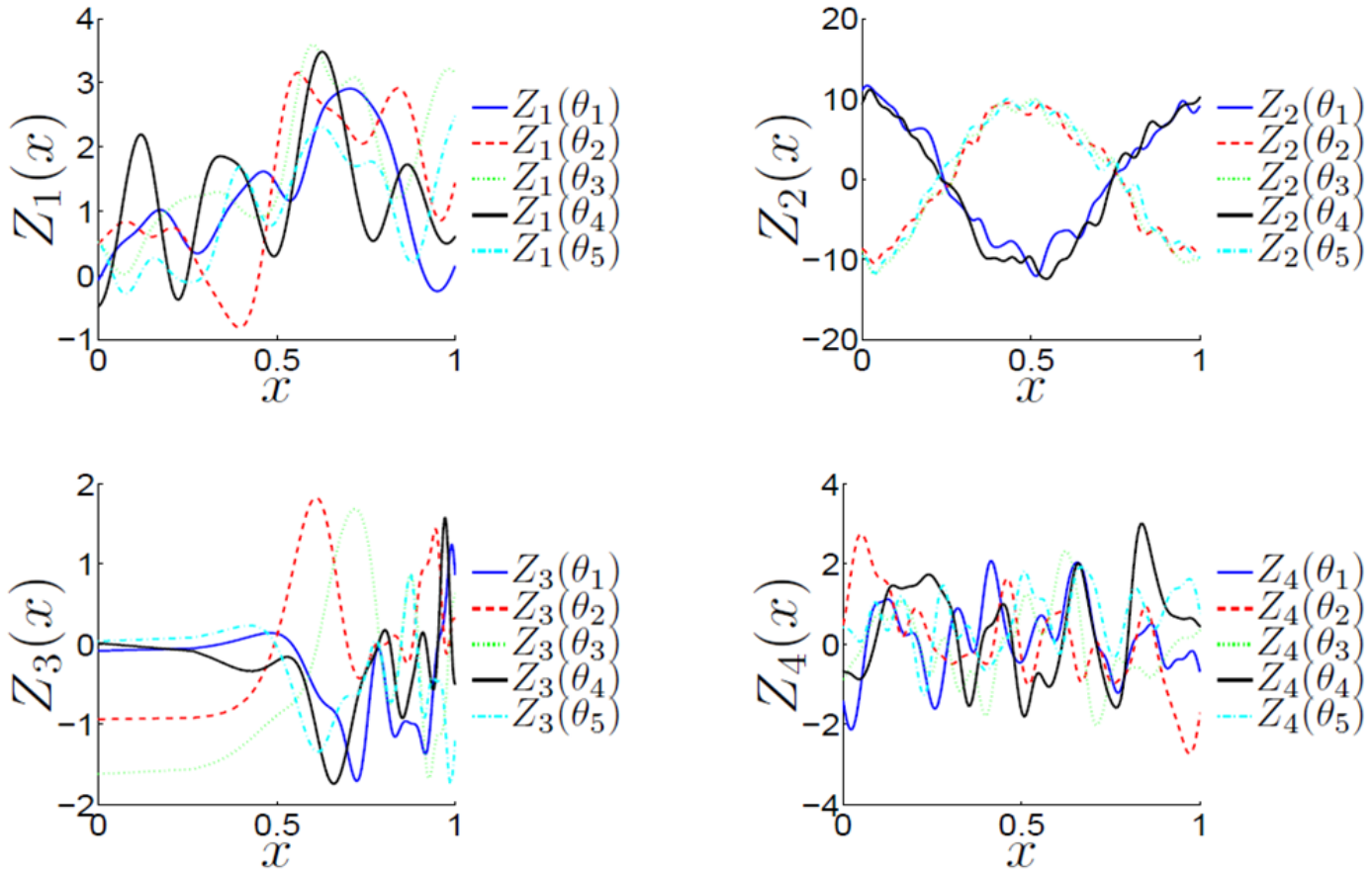
Definition :

A SP Z on $\mathbb{R}^d$ is a **Gaussian Process (GP)** when – for all $n \in \mathbb{N}^*$ and – all $\left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\right\} \in \mathcal{X} \times \cdots \times \mathcal{X}$, the random vector $\boldsymbol{z} \equiv \left(Z(\boldsymbol{x}^{(1)}), \ldots, Z(\boldsymbol{x}^{(n)})\right)$ is **Gaussian**.

- a GP is fully characterized by :
  - its **mean** : $\mu(\boldsymbol{x}) = \mathbb{E}\left[Z(\boldsymbol{x})\right]$,
  - and **covariance** : $C(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}\left[\left(Z(\boldsymbol{x}) - \mu(\boldsymbol{x})\right)\left(Z(\boldsymbol{x}') - \mu(\boldsymbol{x}')\right)\right]$ (drives GP amplitude/speed of variation)
  - We sometimes note it as : $Z(\boldsymbol{x}) \sim \mathcal{GP}(\mu(\boldsymbol{x}), C(\boldsymbol{x}, \boldsymbol{x}'))$
- Roughly speaking a (weakly) **stationary** GP bears a covariance that is stationary, *i.e.* $C(\boldsymbol{x}, \boldsymbol{x}')$ only depends on relative **distances** btw coordinates $|x_i - x_i'|$ (translation invariant)

GP realizations

From the display of 5 of their realizations, can you say that Z_1, Z_2, Z_3 et Z_4 are GP ? stationary ?



Gaussian process conditioning theorem

- given $\boldsymbol{x} \mapsto Z(\boldsymbol{x})$ a GP$(\mu(\boldsymbol{x}), C(\boldsymbol{x}, \boldsymbol{x}'))$, we **condition** the process on the knowledge of n deterministic values $\{Z(\boldsymbol{x}^{(1)}) = z_1, \ldots, Z(\boldsymbol{x}^{(n)}) = z_n\}$ of the physical quantity that it represents

Gaussian process conditioning theorem

- given $\boldsymbol{x} \mapsto Z(\boldsymbol{x})$ a GP$(\mu(\boldsymbol{x}), C(\boldsymbol{x}, \boldsymbol{x}'))$, we **condition** the process on the knowledge of n deterministic values $\{Z(\boldsymbol{x}^{(1)}) = z_1, \ldots, Z(\boldsymbol{x}^{(n)}) = z_n\}$ of the physical quantity that it represents

Notations

$\boldsymbol{Z} = \{Z(\boldsymbol{x}^{(1)}), \ldots, Z(\boldsymbol{x}^{(n)})\}$ observation vector and $\boldsymbol{z} = (z_1, \ldots, z_n)$ observed values

$\boldsymbol{\mu} = (\mu(\boldsymbol{x}^{(1)}), \ldots, \mu(\boldsymbol{x}^{(n)}))$ mean vector of $\boldsymbol{Z}$

R : covariance of $\boldsymbol{Z}$ with $R_{ij} = [C]_{ij} = C(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$

$\boldsymbol{r}(\boldsymbol{x}) = \left(C(\boldsymbol{x}, \boldsymbol{x}^{(1)}), \ldots, C(\boldsymbol{x}, \boldsymbol{x}^{(n)})\right)$: $n \times 1$ covariance vector btw $Z(\boldsymbol{x})$ and $\boldsymbol{Z}$

Gaussian process conditioning theorem

- given $\boldsymbol{x} \mapsto Z(\boldsymbol{x})$ a GP$(\mu(\boldsymbol{x}), C(\boldsymbol{x}, \boldsymbol{x}'))$, we **condition** the process on the knowledge of n deterministic values $\{Z(\boldsymbol{x}^{(1)}) = z_1, \ldots, Z(\boldsymbol{x}^{(n)}) = z_n\}$ of the physical quantity that it represents

Notations

$\boldsymbol{Z} = \{Z(\boldsymbol{x}^{(1)}), \ldots, Z(\boldsymbol{x}^{(n)})\}$ observation vector and $\boldsymbol{z} = (z_1, \ldots, z_n)$ observed values

$\boldsymbol{\mu} = (\mu(\boldsymbol{x}^{(1)}), \ldots, \mu(\boldsymbol{x}^{(n)}))$ mean vector of $\boldsymbol{Z}$

R : covariance of $\boldsymbol{Z}$ with $R_{ij} = [C]_{ij} = C(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$

$\boldsymbol{r}(\boldsymbol{x}) = \left(C(\boldsymbol{x}, \boldsymbol{x}^{(1)}), \ldots, C(\boldsymbol{x}, \boldsymbol{x}^{(n)})\right)$: $n \times 1$ covariance vector btw $Z(\boldsymbol{x})$ and $\boldsymbol{Z}$

- **Conditioning theorem for vectors is generalized to GP**

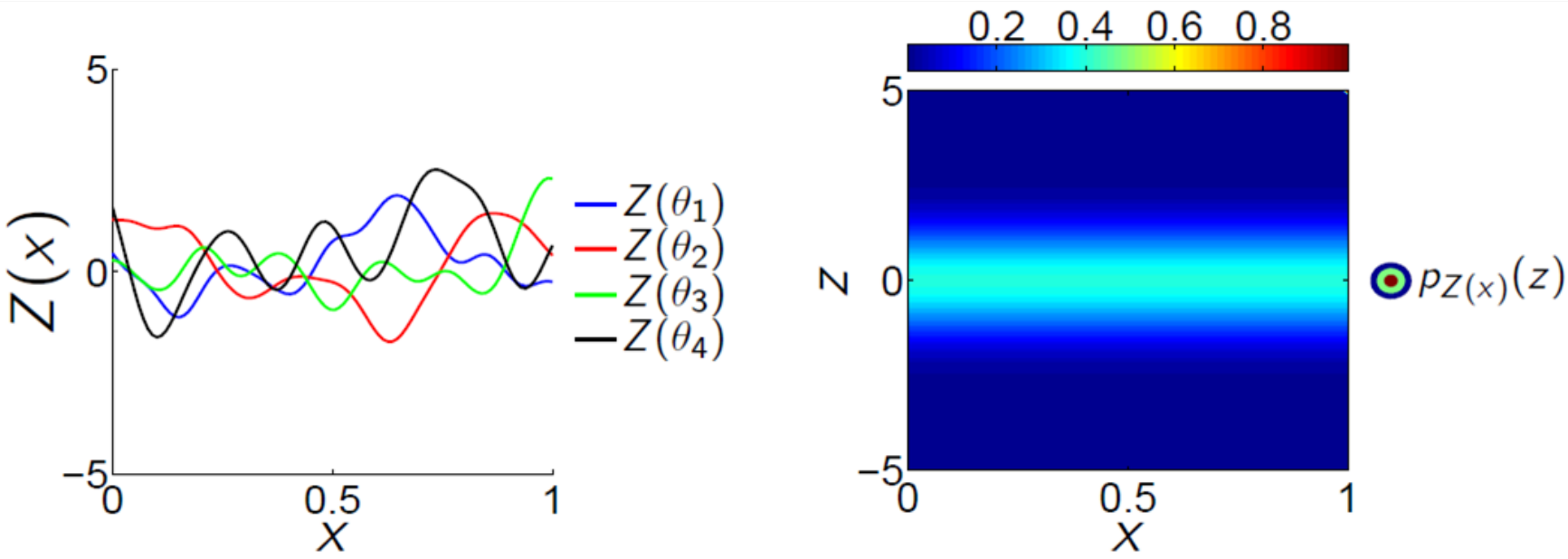
  the **conditioned** process :
  $\boldsymbol{x} \mapsto Z^{\text{cond}}(\boldsymbol{x}) = (Z(\boldsymbol{x}) \mid \{Z(\boldsymbol{x}^{(1)}) = z_1, \ldots, Z(\boldsymbol{x}^{(n)}) = z_n\})$ is a **GP** of **"conditioned"** mean and covariance : $\mu^{\text{cond}}(\boldsymbol{x})$ and $C^{\text{cond}}(\boldsymbol{x}, \boldsymbol{x}')$, such that :

$$\mu^{\text{cond}}(\boldsymbol{x}) \equiv \mathbb{E}(Z(\boldsymbol{x}) \mid \boldsymbol{Z} = \boldsymbol{z}) = \mu(\boldsymbol{x}) + \boldsymbol{r}(\boldsymbol{x})^T R^{-1}(\boldsymbol{z} - \boldsymbol{\mu})$$

$$C^{\text{cond}}(\boldsymbol{x}, \boldsymbol{x}') \equiv \text{cov}(Z(\boldsymbol{x}), Z(\boldsymbol{x}')) \mid \boldsymbol{Z} = \boldsymbol{z}) = C(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{r}(\boldsymbol{x})^T R^{-1} \boldsymbol{r}(\boldsymbol{x}')$$

Visualization of the conditioning

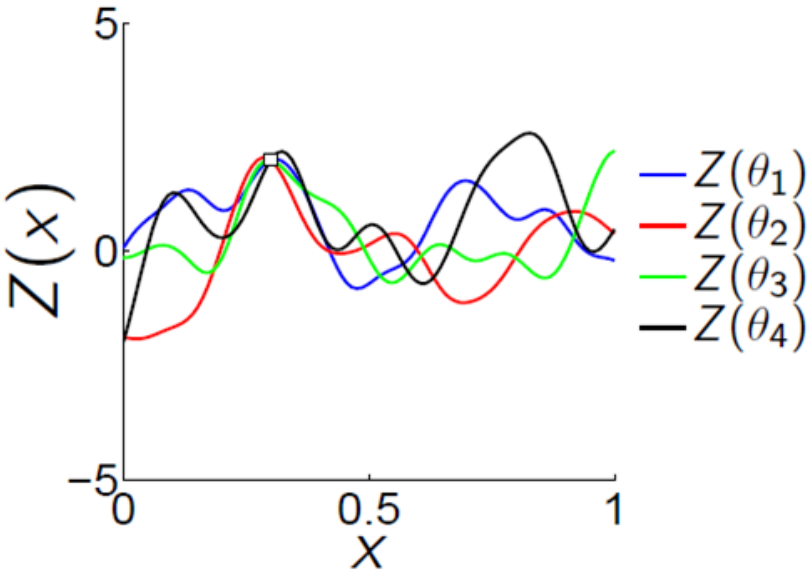
Let $Z(x)$ be a stationary GP of mean $\mu(x) = 0$ and covariance $C(x, x') = \exp\left(-(x - x')^2/100\right)$.

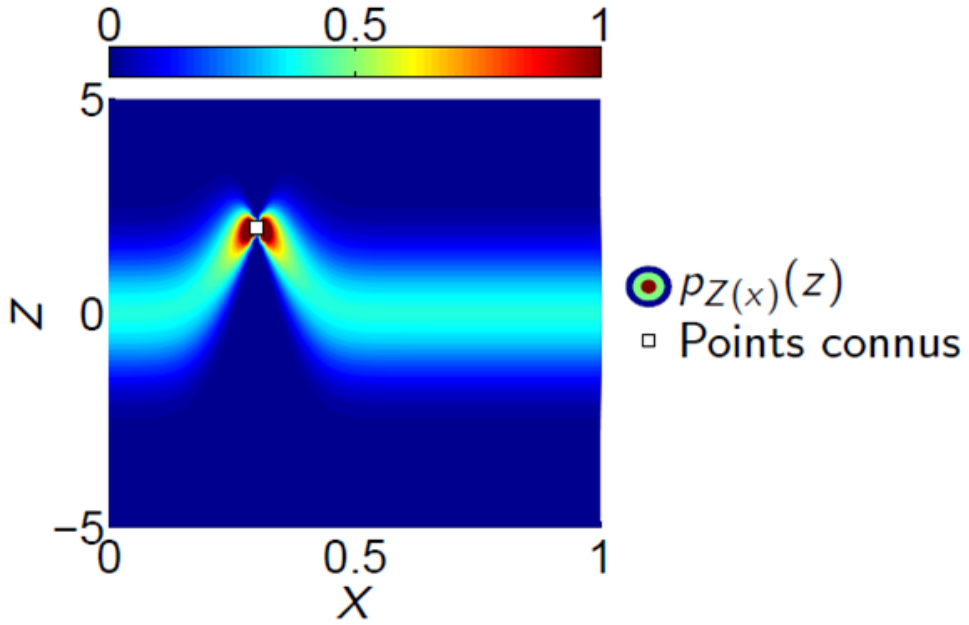


Z is not initially conditioned

Visualization of the conditioning

Let $Z(x)$ be a stationary GP of mean $\mu(x) = 0$ and covariance $C(x, x') = \exp\left(-(x - x')^2/100\right)$.

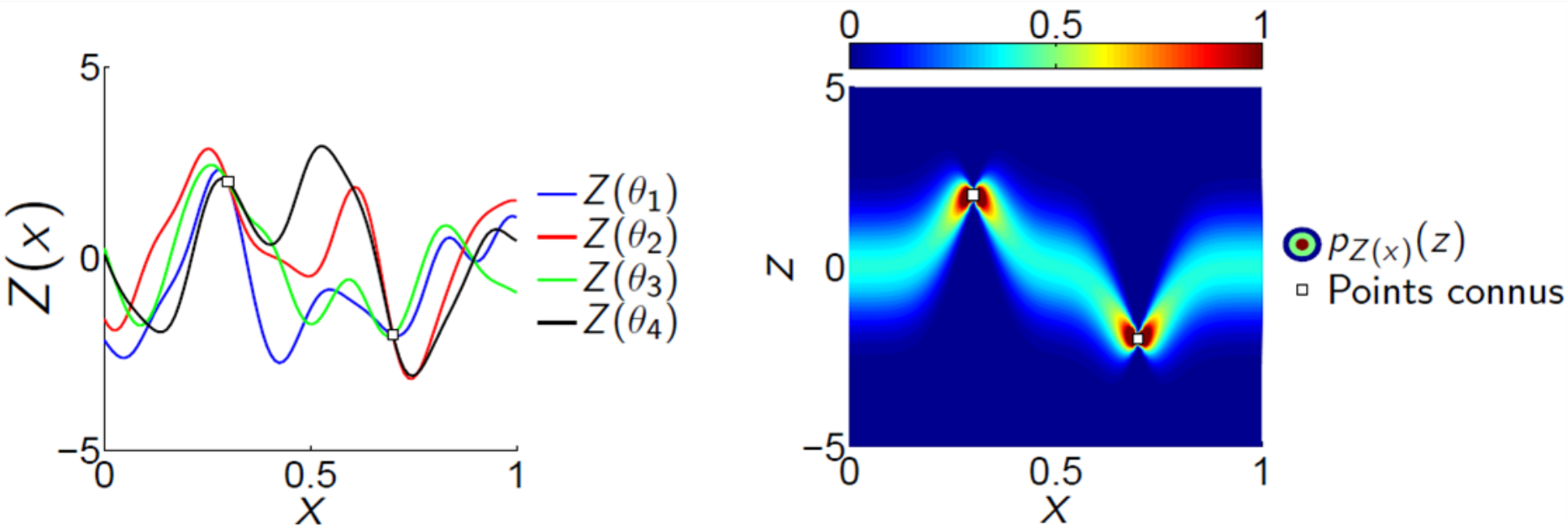




Z conditioned in 1 point

Visualization of the conditioning

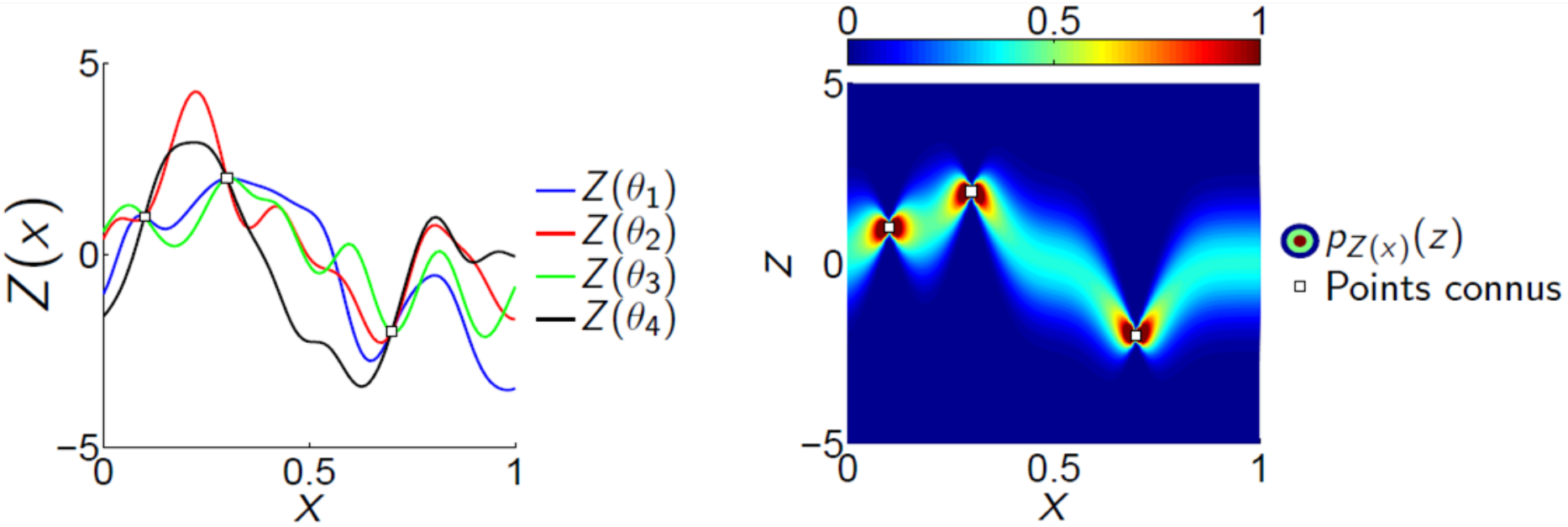
Let $Z(x)$ be a stationary GP of mean $\mu(x) = 0$ and covariance $C(x, x') = \exp\left(-(x - x')^2/100\right)$.



Z conditioned in 2 points

Visualization of the conditioning

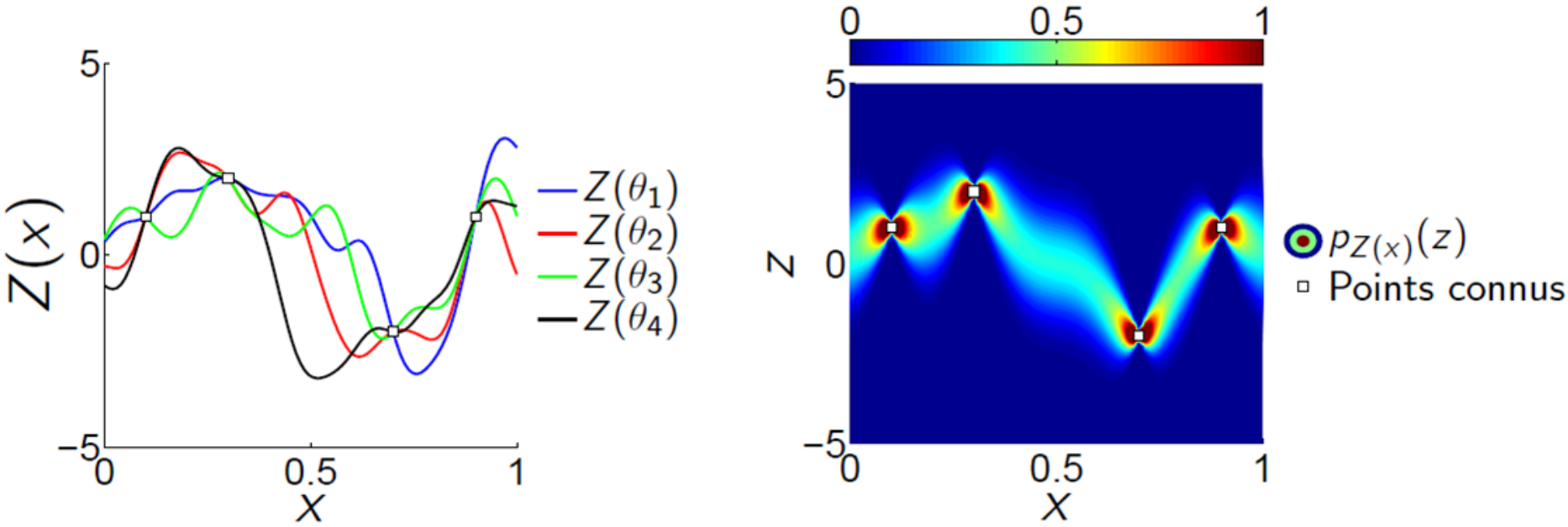
Let $Z(x)$ be a stationary GP of mean $\mu(x) = 0$ and covariance $C(x, x') = \exp\left(-(x - x')^2/100\right)$.



Z conditioned in 3 points

Visualization of the conditioning

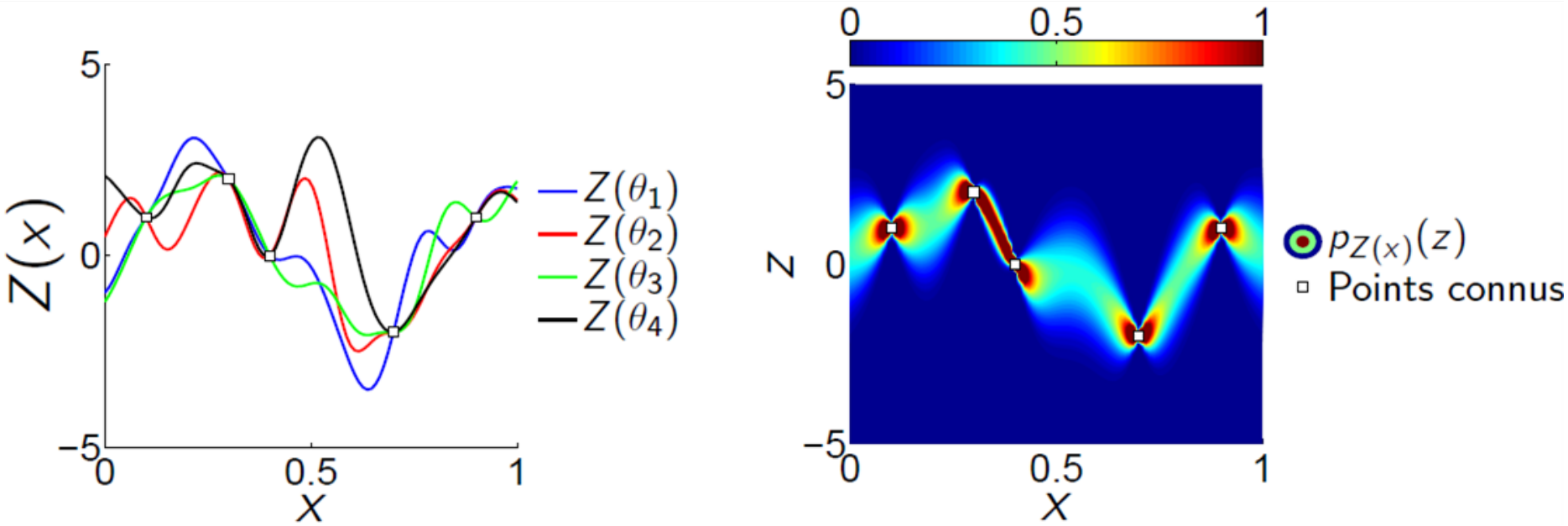
Let $Z(x)$ be a stationary GP of mean $\mu(x) = 0$ and covariance $C(x, x') = \exp\left(-(x - x')^2/100\right)$.



Z conditioned in points

Visualization of the conditioning

Let $Z(x)$ be a stationary GP of mean $\mu(x) = 0$ and covariance $C(x, x') = \exp\left(-(x - x')^2/100\right)$.



Z conditioned in 5 points

Table of contents

Regression via conditioned GP

let us get back to our metamodeling construction :
comparing with a classic linear regression

- linear regression :

$$s(\boldsymbol{x}) \approx \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} + \epsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma).$$

where $\boldsymbol{f}$ are called the regressors

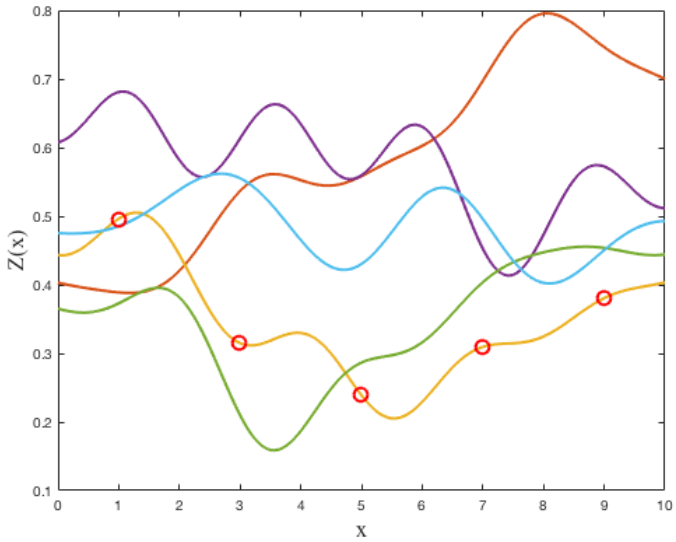
- GP metamodeling :

$$s(\boldsymbol{x}) \approx \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} + \epsilon(\boldsymbol{x}), \quad \varepsilon \sim \mathcal{SCGP},$$

$\mathcal{SCGP} \leftrightarrow$ **"(stationary conditioned) GP"**.

Idea is to add a **stochastic contribution that depends on x !**

In **Bayesian statistics** modeling, we represent a deterministic, unknown number by the realization of a random variable ($\Rightarrow$ enables to incorporate expert knowledge, gives access to Bayes formula...).



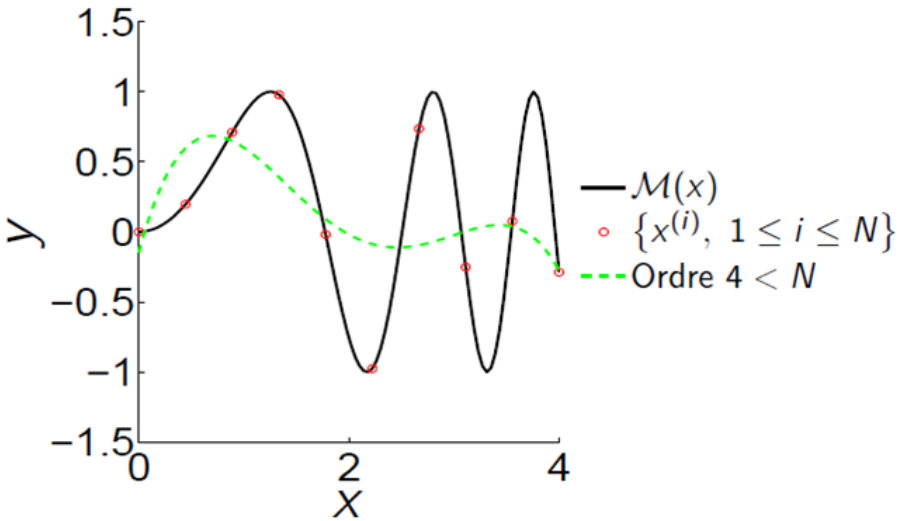
Here, we do the same with functions !

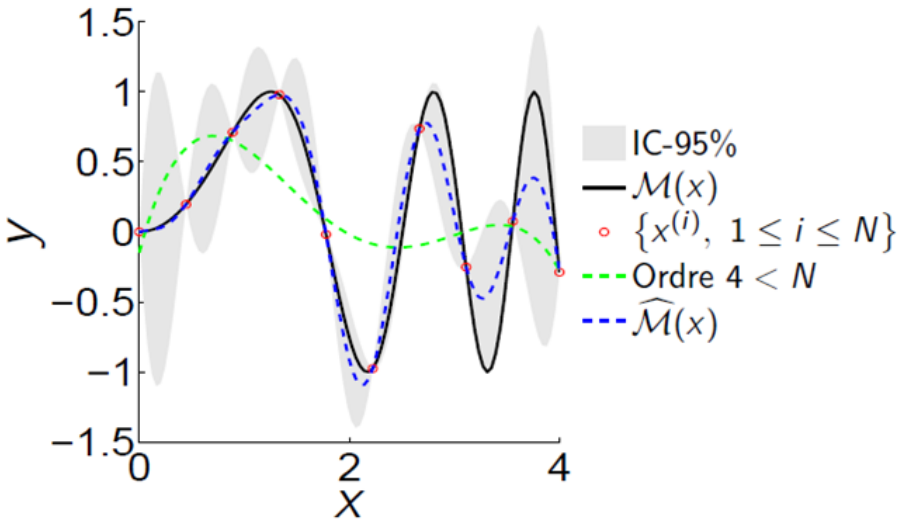
Regression via conditioned GP

Principle

Objective : harness the theory of conditioned GP to improve our prediction,

- **add some confidence** onto the prediction,
- be less dependent on the **choice of the regressors** f.





Example : $d = 1$, $y(x) = \sin(x^2)$, $\mathcal{X} = [0, 4]$.

Regression via conditioned GP

Kriging model :

representing a **deterministic/unknown** function $y = s(\boldsymbol{x} \in \mathcal{X})$, of model s, by **a realization** of a **GP**

- this metamodeling approach assumes that :

$$s(\boldsymbol{x}) \approx \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} + Z(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}$$

  - $\boldsymbol{f}(\boldsymbol{x}) \leftrightarrow$ (**chosen**) deterministic functions are called the **trend**
  - $\boldsymbol{\beta} \leftrightarrow$ (in general unknown) weighting coefficients,
  - $Z(\boldsymbol{x}) \leftrightarrow$ **stationary GP** of **zero** mean and (in general unknown) covariance $C(\boldsymbol{x}, \boldsymbol{x}')$.

Regression via conditioned GP

Kriging model :

representing a **deterministic/unknown** function $y = s(\boldsymbol{x} \in \mathcal{X})$, of model s, by **a realization** of a **GP**

- this metamodeling approach assumes that :

$$s(\boldsymbol{x}) \approx \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} + Z(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}$$

  - $\boldsymbol{f}(\boldsymbol{x}) \leftrightarrow$ (**chosen**) deterministic functions are called the **trend**
  - $\boldsymbol{\beta} \leftrightarrow$ (in general unknown) weighting coefficients,
  - $Z(\boldsymbol{x}) \leftrightarrow$ **stationary GP** of **zero** mean and (in general unknown) covariance $C(\boldsymbol{x}, \boldsymbol{x}')$.

- By **construction**, $\{s(\boldsymbol{x}),\ \boldsymbol{x} \in \mathcal{X}\}$ becomes a GP, of mean $\boldsymbol{x} \mapsto \boldsymbol{f}^T(\boldsymbol{x})$ and covariance $(\boldsymbol{x}, \boldsymbol{x}') \mapsto C(\boldsymbol{x}, \boldsymbol{x}')$.

Kriging conditioning

- again, we **condition** the process on the knowledge of – the covariance $C(\boldsymbol{x}, \boldsymbol{x}')$ and – the n solver values $\{s(\boldsymbol{x}^{(1)}) = y_1, \ldots, s(\boldsymbol{x}^{(n)}) = y_n\}$
- the **conditioned** process : $\hat{s}(\boldsymbol{x}) = (s(\boldsymbol{x}) \mid \{s(\boldsymbol{x}^{(1)}) = y_1, \ldots, s(\boldsymbol{x}^{(n)}) = y_n\})$ is a GP of distribution $\mathcal{N}(\hat{\mu}(\boldsymbol{x}), \hat{\sigma}^2(\boldsymbol{x}))$, such that :

$$\hat{\mu}(\boldsymbol{x}) = \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{r}(\boldsymbol{x})^T R^{-1}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta})$$

$$\hat{\sigma}^2(\boldsymbol{x}) = C(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{r}(\boldsymbol{x})^T R^{-1} \boldsymbol{r}(\boldsymbol{x})$$

Metamodel in practice :

the **conditional mean** $\hat{\mu}(x)$ is taken as the **surrogate model**

Kriging conditioning

- again, we **condition** the process on the knowledge of – the covariance $C(\boldsymbol{x}, \boldsymbol{x}')$ and – the n solver values $\{s(\boldsymbol{x}^{(1)}) = y_1, \ldots, s(\boldsymbol{x}^{(n)}) = y_n\}$
- the **conditioned** process : $\hat{s}(\boldsymbol{x}) = (s(\boldsymbol{x}) \mid \{s(\boldsymbol{x}^{(1)}) = y_1, \ldots, s(\boldsymbol{x}^{(n)}) = y_n\})$ is a GP of distribution $\mathcal{N}(\hat{\mu}(\boldsymbol{x}), \hat{\sigma}^2(\boldsymbol{x}))$, such that :

$$\hat{\mu}(\boldsymbol{x}) = \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{r}(\boldsymbol{x})^T R^{-1}(\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta})$$

$$\hat{\sigma}^2(\boldsymbol{x}) = C(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{r}(\boldsymbol{x})^T R^{-1} \boldsymbol{r}(\boldsymbol{x})$$

Metamodel in practice :

the **conditional mean** $\hat{\mu}(x)$ is taken as the **surrogate model**

Notations

$\boldsymbol{y} = (y_1, \ldots, y_n)$ observed values

$\boldsymbol{F} = (\boldsymbol{f}(\boldsymbol{x}^{(1)}), \ldots, \boldsymbol{f}(\boldsymbol{x}^{(n)}))$ regressors values at x DoE points

$R_{ij} = [C]_{ij} = C(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$

$\boldsymbol{r}(\boldsymbol{x}) = \left(C(\boldsymbol{x}, \boldsymbol{x}^{(1)}), \ldots, C(\boldsymbol{x}, \boldsymbol{x}^{(n)})\right)$: covariance vector btw $s(\boldsymbol{x})$ and $\boldsymbol{s}$

BLUP (Best Linear Unbiased Predictor) or universal kriging

- again, we **condition** the process on the knowledge of – the covariance $C(\boldsymbol{x}, \boldsymbol{x}')$ and – the n solver values $\{s(\boldsymbol{x}^{(1)}) = y_1, \ldots, s(\boldsymbol{x}^{(n)}) = y_n\}$
- if there exists $\boldsymbol{f}(\boldsymbol{x})$, β **(unknown)**, $Z(\boldsymbol{x})$ / for all $\boldsymbol{x} \in \mathcal{X}$, $s(\boldsymbol{x}) = \langle \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{\beta} \rangle + Z(\boldsymbol{x})$, then the **best linear unbiased predictor** of an unobserved quantity $s(\boldsymbol{x})$ is a GP of distribution $\mathcal{N}(\widehat{\mu}_{\mathsf{BLUP}}(\boldsymbol{x}), \widehat{\sigma}^2_{\mathsf{BLUP}}(\boldsymbol{x}))$:

$$\widehat{\mu}_{\mathsf{BLUP}}(\boldsymbol{x}) = \boldsymbol{f}^T(\boldsymbol{x})\widehat{\boldsymbol{\beta}} + \boldsymbol{r}(\boldsymbol{x})^T R^{-1}(\boldsymbol{y} - \boldsymbol{F}\widehat{\boldsymbol{\beta}}),$$

$$\widehat{\sigma}^2_{\mathsf{BLUP}}(\boldsymbol{x}) = C(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{r}(\boldsymbol{x})^T R^{-1} \boldsymbol{r}(\boldsymbol{x}) + \boldsymbol{u}(\boldsymbol{x})^T \left(\boldsymbol{F}^T R^{-1} \boldsymbol{F}\right)^{-1} \boldsymbol{u}(\boldsymbol{x}),$$

Notations

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{F}^T R^{-1} \boldsymbol{F}\right)^{-1} \boldsymbol{F}^T R^{-1} \boldsymbol{y}, \quad \boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{F}^T R^{-1} \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}),$$

with $\widehat{\beta}$ is the **maximum likelihood** approximation of β

BLUP nice properties

- **linear** relative to the observations : there exists $\{a_i(\boldsymbol{x}),\ 1 \leq i \leq n\}$ such that $\hat{s}(\boldsymbol{x}) = \sum_{i=1}^{n} a_i(\boldsymbol{x}) s(\boldsymbol{x}^{(i)})$,
- **unbiased** : $\mathbb{E}\left[\hat{s}(\boldsymbol{x})\right] = \mathbb{E}\left[s(\boldsymbol{x})\right]$,
- **optimal** in the L_2 norm among **unbiased linear predictors** :

$$\hat{s}(\boldsymbol{x}) = \arg \min_{\tilde{s} \in \mathrm{LUP}} \mathbb{E}\left[\left(s(\boldsymbol{x}) - \tilde{s}\right)^2\right],$$

- **interpolant** : $\mathbb{P}\left(\hat{s}(\boldsymbol{x}_i) = y_i\right) = 1$.

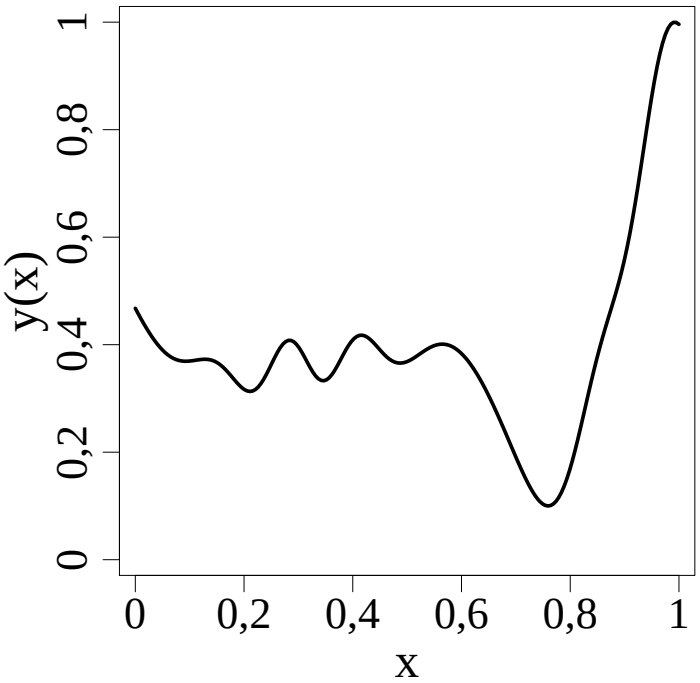
This BLUP estimator is very simple and popular

1D illustration

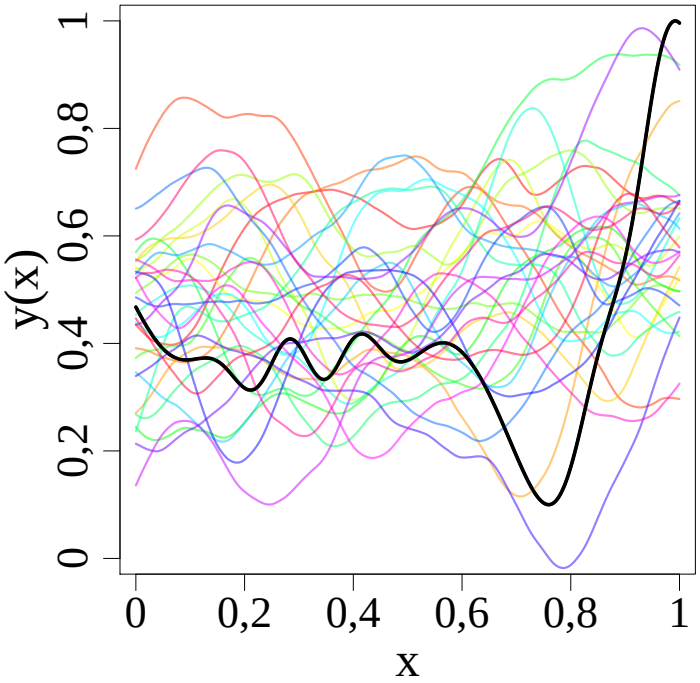
The goal is to approximate the **deterministic/unknown** 1D function.

Hypothesis : the function $y = s(x \in \mathcal{X})$ is **a realization** of a **GP**

Fonction à approcher



Conditionnement

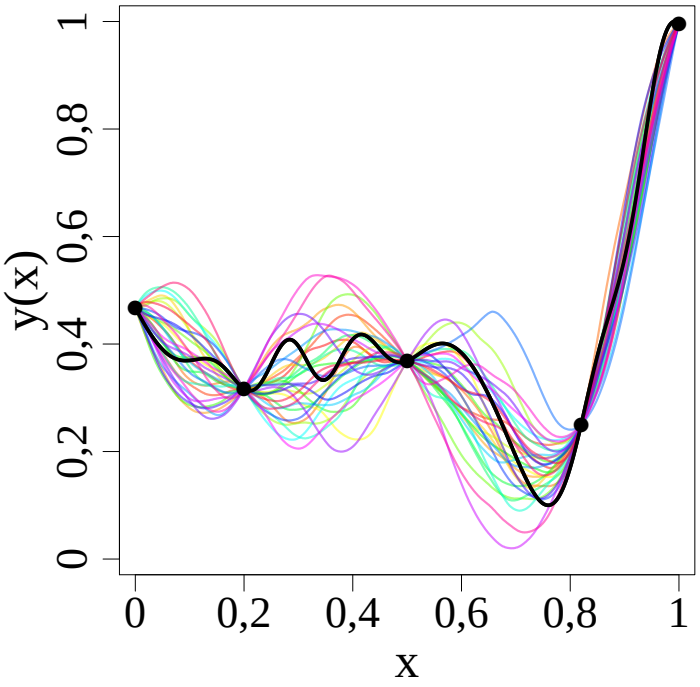


1D illustration

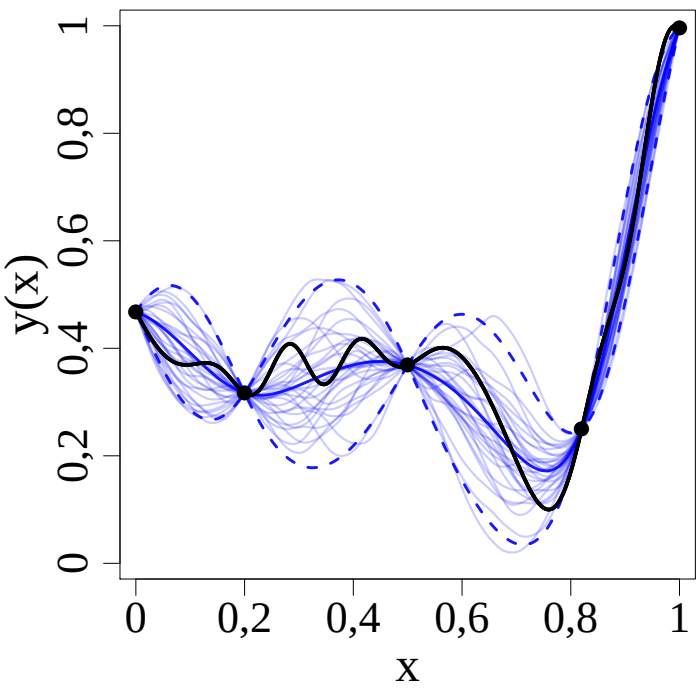
The GP is strongly conditioned by the few observations (black dots).

The surrogate model is obtained as the **mean trajectory of the GP** and its variance allows to build prediction confidence intervals.

Réalisations de PG



Moyenne et variance



Regression via conditioned GP

Importance of the covariance kernel

- $C(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}\left[\left(s(\boldsymbol{x}) - \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta}\right)\left(s(\boldsymbol{x}') - \boldsymbol{f}^T(\boldsymbol{x}')\boldsymbol{\beta}\right)\right]$.
- from knowledge of s at n points in $\mathcal{X}$, how to identify the covariance C, corner stone of the GP metamodeling construction ?
- A covariance function must be **symmetric** et **definite positive**.
- $C(\boldsymbol{x}, \boldsymbol{x}')$ must account for the **regularity** of s.

Roadmap

1. choose a simple **parametric** form (expert judgement).
2. identify the **most-likely** values of the involved **hyperparameters**, thanks to the observations

Standard covariance functions

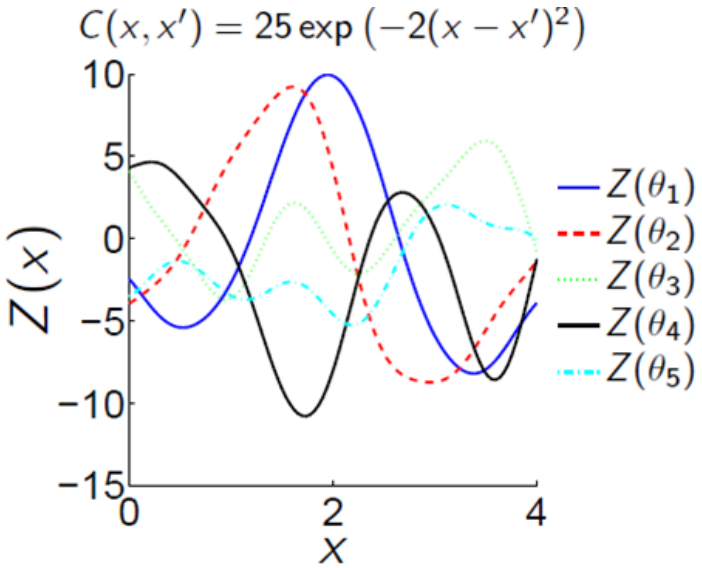
- "nugget" function : $C(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \delta_{\mathbf{0}}(\boldsymbol{x} - \boldsymbol{x}')$,
- linear kernels : $C(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \prod_{i=1}^{d} \max\left(0, 1 - \frac{|x_i - x_i'|}{\ell_i}\right)$,
- exponential kernels : $C(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \exp\left(-\sum_{i=1}^{d} \frac{|x_i - x_i'|}{\ell_i}\right)$,
- gaussian kernels : $C(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \exp\left(-\sum_{i=1}^{d} \frac{(x_i - x_i')^2}{\ell_i^2}\right)$,
- Matern kernels, with Γ Euler function and $\mathcal{B}_\nu^{III}$ Bessel function of the third kind :
  $C(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \prod_{i=1}^{d} \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(2\sqrt{\nu}\frac{|x_i - x_i'|}{\ell_i}\right)^{\nu} \mathcal{B}_\nu^{III}\left(2\sqrt{\nu}\frac{|x_i - x_i'|}{\ell_i}\right)$;
  $\nu = \{1/2, 3/2, 5/2\}$

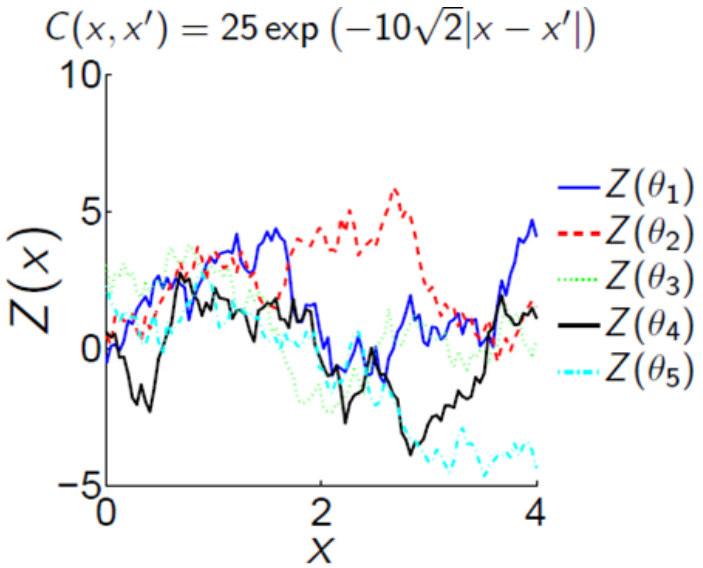
$\rightarrow$ Important hyperparameters : **variance** σ^2, **correlation lengths** ℓ_i and **exponents** ν.

Regression via conditioned GP

Matern covariance $C(\boldsymbol{x}, \boldsymbol{x}') \sim \sigma^2 k\left(\left(\frac{|x_i - x_i'|}{\ell_i}\right)^{\nu}\right)$ - interpreting ν

- exponent of $\frac{|x_i - x_i'|}{\ell_i}$ relates to the differentiability and therefore the regularity of the SP



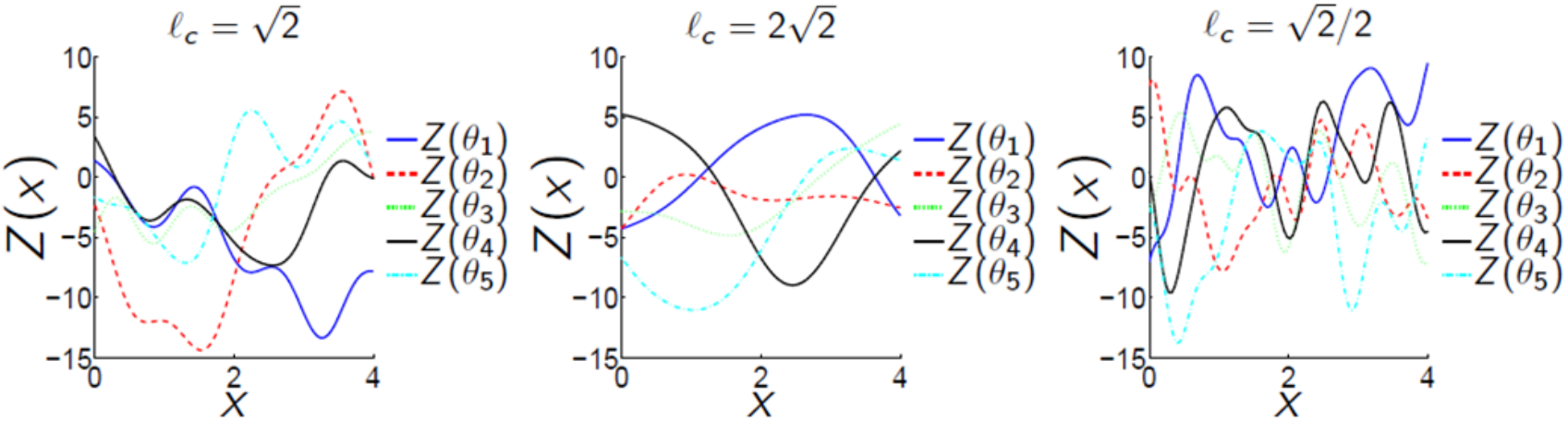


if no *a priori* information about s regularity, then use the Matern function with $\nu = 5/2$

Regression via conditioned GP

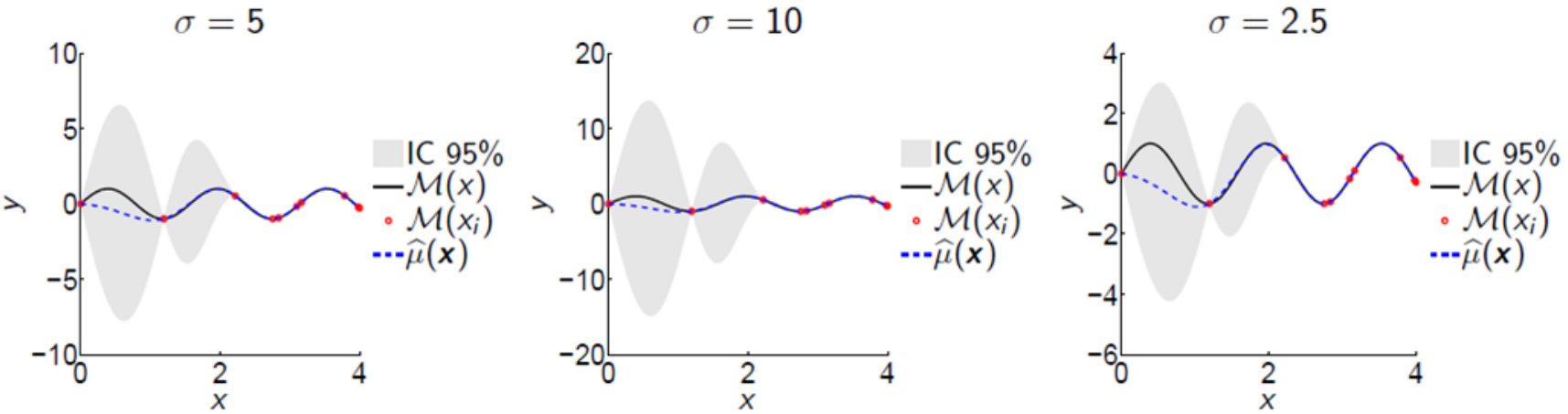
Matern covariance $C(x, x') \sim \sigma^2 k\left(\left(\frac{|x_i - x_i'|}{\ell_i}\right)^{\nu}\right)$ - interpreting ℓ_c

- the correlation length quantities the "radius of influence" across different points
  - if $\ell \to 0$: then the SP becomes "white noise"
  - if $\ell \to +\infty$: then the SP looses its variability and becomes a constant function



Matern covariance $C(\boldsymbol{x}, \boldsymbol{x}') \sim \sigma^2 k\left(\left(\frac{|x_i - x_i'|}{\ell_i}\right)^{\nu}\right)$ - interpreting σ

- The variance describes the amplitude variations of the conditioned GP. It also characterizes the level of confidence of the prediction outside of the observed points.
  - if $\sigma \to 0$: prediction becomes almost deterministic and the surrogate gets closer to the true model
  - if $\sigma \to +\infty$: it means we do not show any confidence to the prediction outside of the observed points



Regression via conditioned GP

Identification of optimal hyperparameters

- if no *a priori* knowledge about (ν, ℓ_j, σ), then we can try to gain some information by exploiting the observations at $\left\{\boldsymbol{x}^{(i)},\ 1 \leq i \leq n\right\}$.
- get **coherent values** of (ν, ℓ_j, σ) with the data $\leftrightarrow$ **maximizing likelihood** $L\left(y_i, \nu, \sigma, \ell_j\right)$ of obtaining data $s(\boldsymbol{x}^{(i)}) = y^{(i)}$ given (ν, σ, ℓ_j).
- we can choose ν *a priori* and not try to estimate it
- in general, we first estimate σ by ML :

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{F}\hat{\beta})^T R^{-1} (\boldsymbol{y} - \boldsymbol{F}\hat{\beta})}{n - p}$$

- estimations of (β, σ) are then plugged into the likelihood in order to estimate ℓ_j... **no analytical solution** (only numerical optimization...)
- another approach for estimation relies on the **cross-validation error**, but still involves an optimization step for the determination of the correlation length

Metamodel validation

- The adequacy of the metamodel can be evaluated on a **separate testing sample** or from a **cross-validation** (CV) procedure
- CV errors : $e_i^2 = \left(s(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{C(-i)}(\boldsymbol{x}^{(i)})\right)^2$

  with $\widehat{\mu}_{(-i)}$ the mean prediction evaluated from $\left\{s(\boldsymbol{x}^{(k)}),\ k \neq i\right\}$.

Metamodel validation

- The adequacy of the metamodel can be evaluated on a **separate testing sample** or from a **cross-validation** (CV) procedure
- CV errors : $e_i^2 = \left(s(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{C(-i)}(\boldsymbol{x}^{(i)})\right)^2$

  with $\widehat{\mu}_{(-i)}$ the mean prediction evaluated from $\left\{s(\boldsymbol{x}^{(k)}),\ k \neq i\right\}$.
- quality of prediction evaluated from $\widehat{\mu}_C(\boldsymbol{x})$ from Q_2 coefficient :

$$Q_2 = 1 - \frac{1}{n}\frac{\sum_{i=1}^{n} e_i^2}{\widehat{\mathrm{var}}(s(\boldsymbol{x}))}$$

- then to have a global error criteria, we introduce ε_{VC}^2 :

$$\varepsilon_{CV}^2(\hat{s}(x)) = \frac{1}{n}\sum_{i=1}^{N}\frac{e_i^2}{\widehat{\sigma}_{(-i)}^2(\boldsymbol{x}^{(i)})}$$

  with $\widehat{\sigma}_{(-i)}^2$, the prediction variance evaluated from $\left\{g(\boldsymbol{x}^{(k)}),\ k \neq i\right\}$.

Regression via conditioned GP

Final remarks

- **Advantages** :
  - easy to build (mostly analytical expressions)
  - **(++) prediction variance provided allowing its precision to be quantified**
  - adding new data points does not disturb the prediction
  - **optimality** among **linear** predictions
  - **moderate dependence on the trend** choice

Regression via conditioned GP

Final remarks

- **Advantages** :
  - easy to build (mostly analytical expressions)
  - **(++) prediction variance provided allowing its precision to be quantified**
  - adding new data points does not disturb the prediction
  - **optimality** among **linear** predictions
  - **moderate dependence on the trend** choice
- **Drawbacks** :
  - modeling rather well suited for **regular/stationary** response surfaces
  - calibrating covariance parameters can be difficult
  - limitation to configurations for which the number of input parameters (the dimension of $\boldsymbol{x}$) is relatively small ($<$10-20).
- **Perspectives** :
  - potential **sequential learning**, **stepwise uncertainty reduction** strategies

Organization of the class

1. Course I : Concepts, formalism and some classes of resolution methods for simple optimization under uncertainty (1.5h, Thursday)

2. Course II : quantification of uncertainty for more efficient optimization under uncertainties : metamodeling with Gaussian Processes (1.5h, Friday)
   - Surrogate modeling
   - Gaussian vectors and gaussian processes
   - Regression via conditioned gaussian processes
   - Challenge of dimensionality and space-filling designs

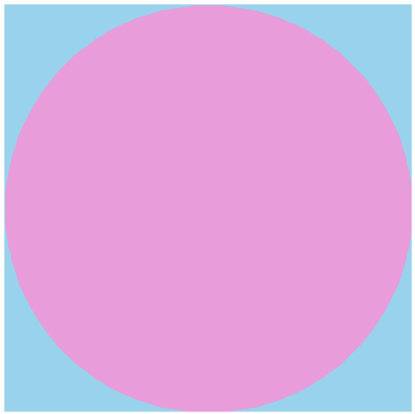
3. Course III : Metamodeling-based Reliability-Based Design Optimization (1.5h, Saturday)

Challenge of dimensionality for metamodeling

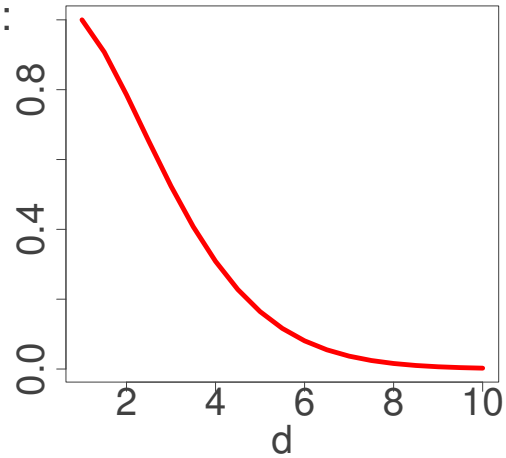
Sphere in d dimension

- VC : cube volume of side $2r$ in d dimension : $(2r)^d$
- VS : sphere volume of radius r in d dimension : $\frac{\left(\sqrt{\pi}r\right)^d}{\Gamma\left(d/2+1\right)}$
- ratio VS/VC :

d	1	2	3	4	5	8	10	16
VS/VC	100%	78.5%	30.1%	52.4%	16.4%	1%	0.25%	4e-6%



Sphere volume / cube volume :



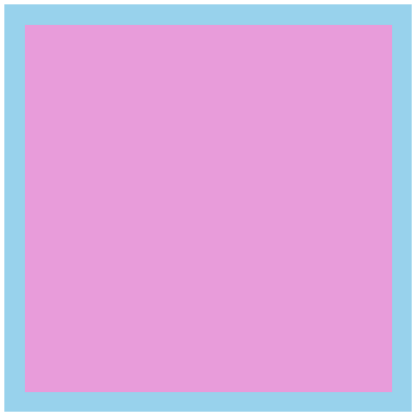
- Hypersphere volume tends to 0 when $d \to \infty$.

Challenge of dimensionality for metamodeling

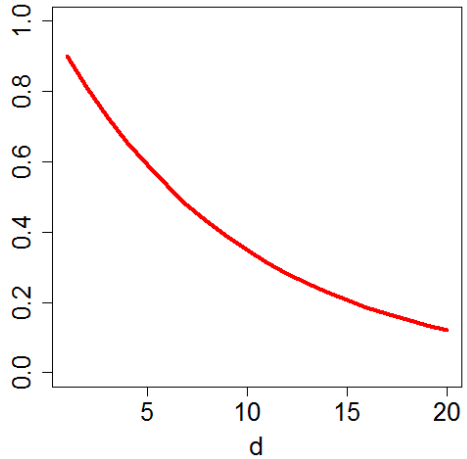
Cube in d dimension

- Cube volume of side c in d dimension : c^d
- Cube volume of side $c - \varepsilon$ in d dimension : $(c - \varepsilon)^d$
- Ratio :

d	1	2	4	8	16	20	30	40
VCe/VC	90%	81%	65.6%	43%	18%	12%	4%	1.5%



Cube volume of side 0.9/ Cube volume of side 1 :

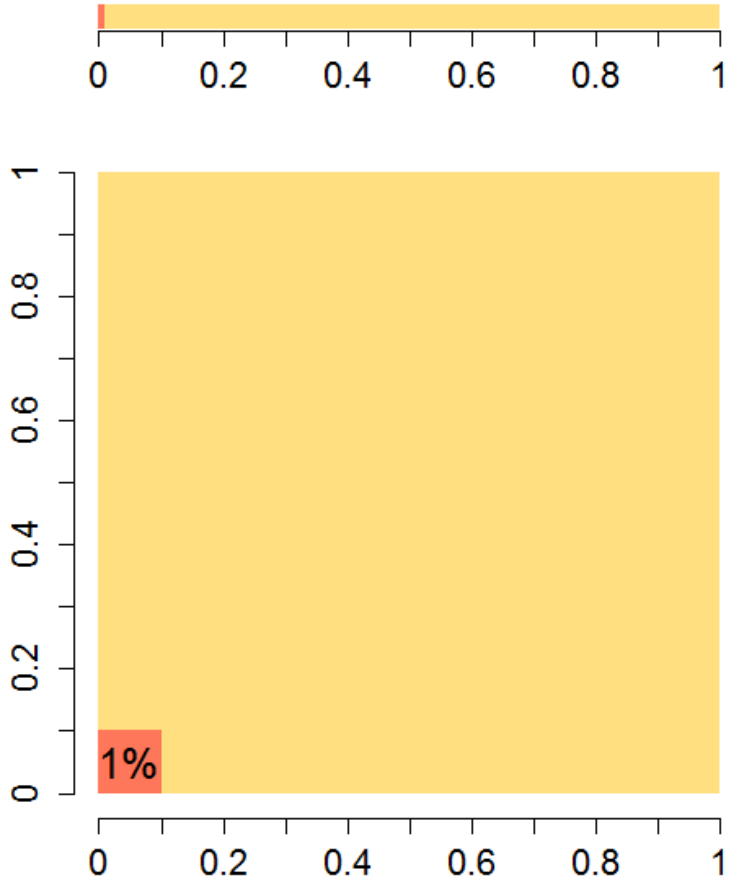


- In high dimension all the volume is contained "in the shell".

  $\rightarrow$ *a posteriori* reducing parametric domain, may disregard all existing computations

Selection of small proportion of data

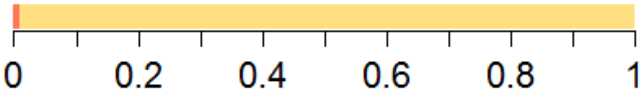
- Selecting only a portion p of data $\leftrightarrow$ considering an hypercube of side $p^{1/d}$.
- Dimension 1 : $p^{1/d} = 0.01$
- Dimension 2 : $p^{1/d} = 0.1$



if we have n points in an hypercube of dimension $d = 4$ and we get rid of 1% of the points (outliers) $\Rightarrow$ same as considering 10% of individual range of each parameter is dismissed.

Selection of small proportion of data

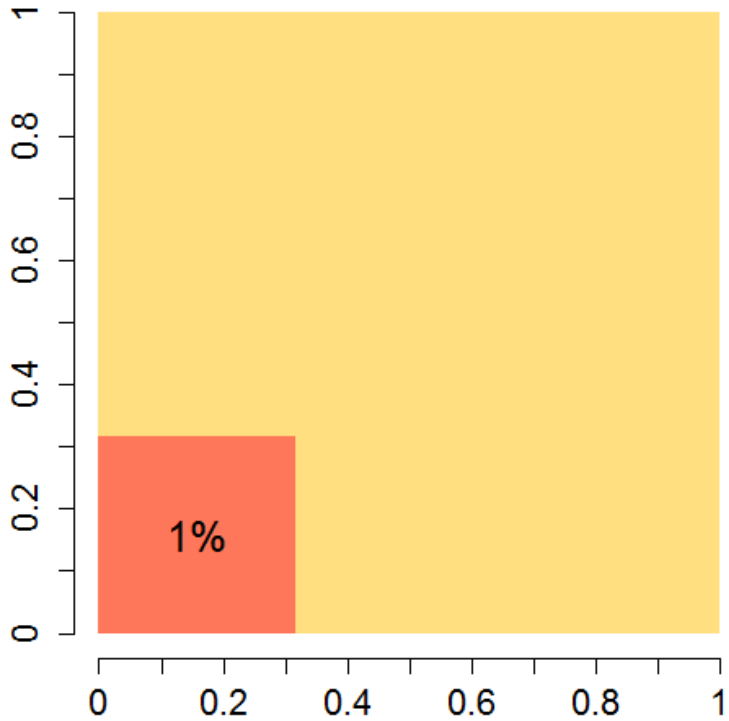
- Selecting only a portion p of data $\leftrightarrow$ considering an hypercube of side $p^{1/d}$.
- Dimension 1 : $p^{1/d} = 0.01$



- Dimension 4 : $p^{1/d} = 0.32$

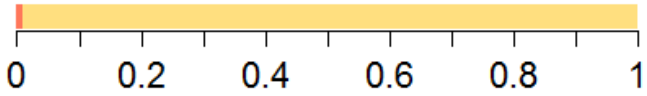
  viz : projecting onto the first two axes

  if we have n points in an hypercube of dimension $d = 4$ and we get rid of 1% of the points (outliers) $\Rightarrow$ same as considering 32% of individual range of each parameter is dismissed.



Selection of small proportion of data

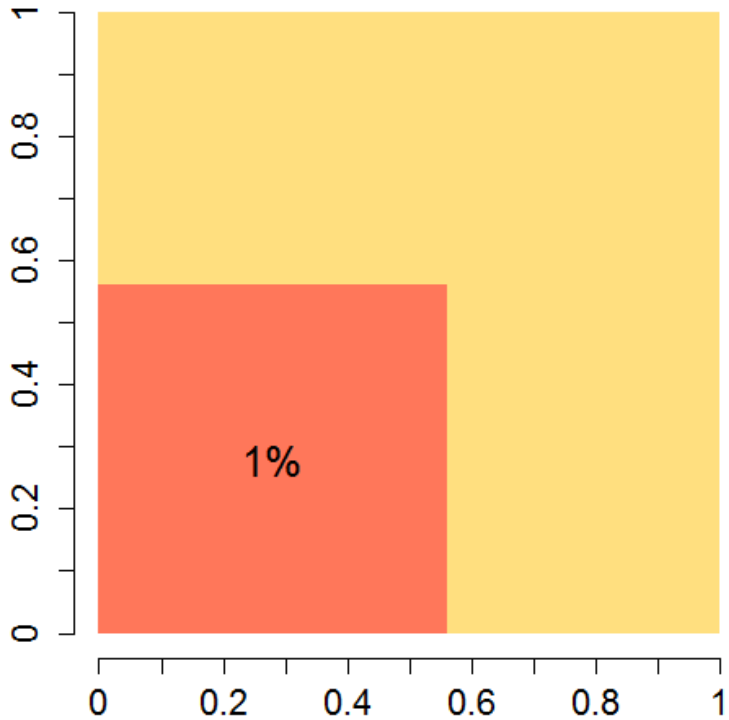
- Selecting only a portion p of data $\leftrightarrow$ considering an hypercube of side $p^{1/d}$.
- Dimension 1 : $p^{1/d} = 0.01$



- Dimension 8 : $p^{1/d} = 0.56$

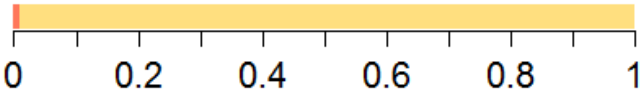
  viz : projecting onto the first two axes

  if we have n points in an hypercube of dimension $d = 8$ and we get rid of 1% of the points (outliers) $\Rightarrow$ same as considering 56% of individual range of each parameter is dismissed.



Selection of small proportion of data

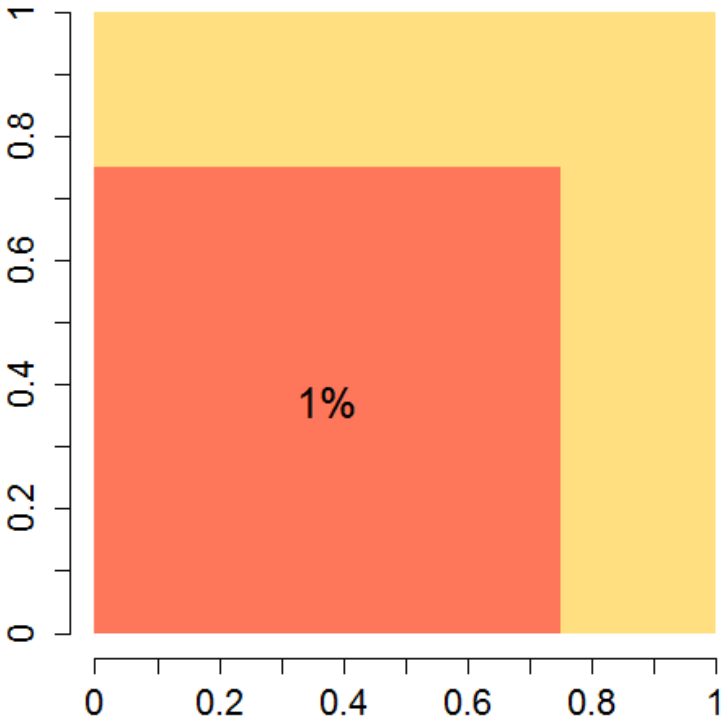
- Selecting only a portion p of data $\leftrightarrow$ considering an hypercube of side $p^{1/d}$.
- Dimension 1 : $p^{1/d} = 0.01$



- Dimension 16 : $p^{1/d} = 0.75$

  viz : projecting onto the first two axes

  if we have n points in an hypercube of dimension $d = 16$ and we get rid of 1% of the points (outliers) $\Rightarrow$ same as considering 75% of individual range of each parameter is dismissed.



Challenge of dimensionality for metamodeling

High-dimensional parametric spaces are "strange" and essentially **empty** ! $\rightarrow$ very hard capturing **local phenomena** with metamodels

Wish list

- try to come up with design of numerical experiments (DoE) that optimize the positioning of our data to build good metamodels, while trying to break free from any *particular* model
- obtain information in any part of the domain. Be able to remove non-influencing parameters and have good projections in subspaces
- **robust** to the **rise in dimension** : in terms of cost, properties, ...
- **sequential capability** (refinements, etc.)

Space filling designs (SFD)

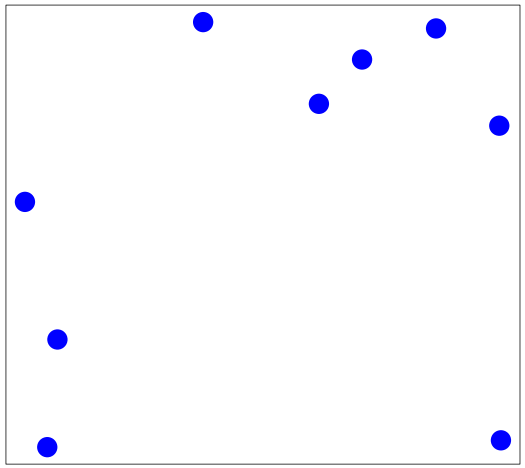
DoE whose points are arranged almost everywhere in the domain of input parameters $\boldsymbol{x} \in \mathbb{R}^d$ in order to explore "as well as possible" the space of possibilities of the system output

$\rightarrow$ domain of optimal designs for computer experiments

Exploratory Design of Experiments

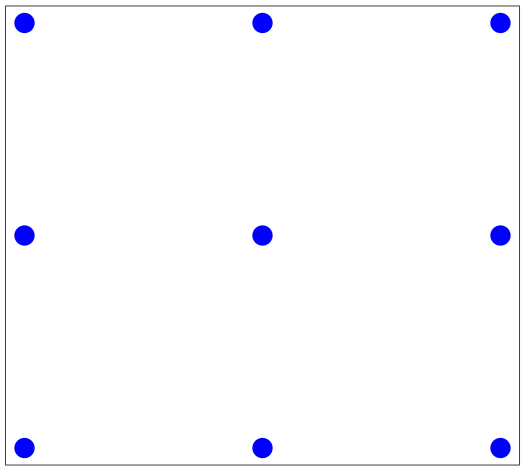
$d = 2$ et $n = 10$

Random DoE (Monte Carlo)



Does not fill up space

Factoriel design with 3 levels



Bad 1D projection

Optimal exploration of an hypercube

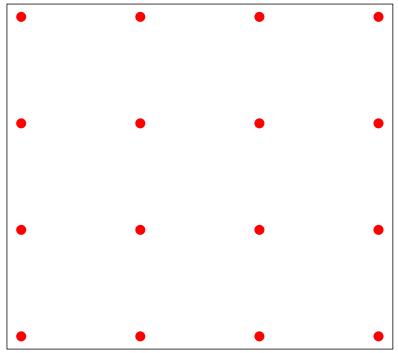
- the precision of the exploration (and therefore the numerical cost, ie. number of simulations n) strongly depends on the dimension d of the space.

  regular grid with p levels (e.g. quadrature) : $n = p^d$ simulations

  $d = 2$, $p = 3$: n=9

  $d = 10$, $p = 3$: n=59049

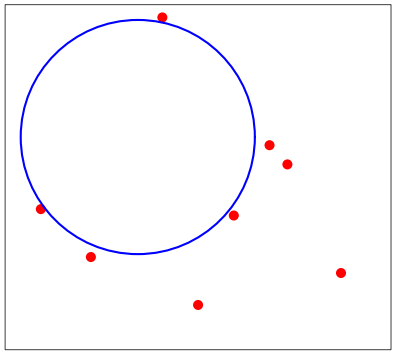
  $\longrightarrow$ **Curse of dimensionality !**

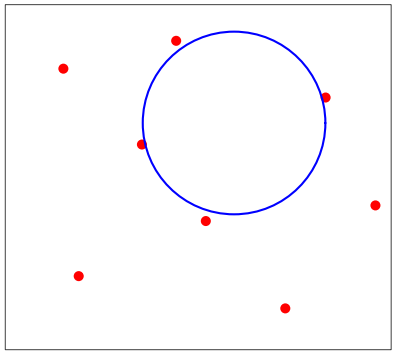


- In order to minimize n, we need samples with good input space cover

  purely random sample does not meet this requirement : example ($d = 2$, $n = 8$)

Monte Carlo





Optimized design (Space Filling Design)

Space-filling designs (1)

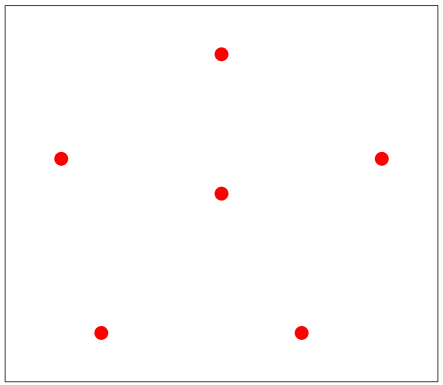
Geometrical criteria

1. **Minimax D_{Minimax}**

   - **minimize the maximum distance** btw any domain point and the closest point of the design

$$\min_{D} \max_{x} d(x, D) = \max d(x, D_{\text{Minimax}}) \text{ with } d(x, D) = \min_{x^{(0)} \in D} d(x, x^{(0)})$$

   - no point in $[0, 1]^d$ is too far from a point of the D_{Minimax} DoE
   - very good DoE but too expensive to build for $d > 3$ (optimization problem with $d \times n$ parameters).



Space-filling designs (1)

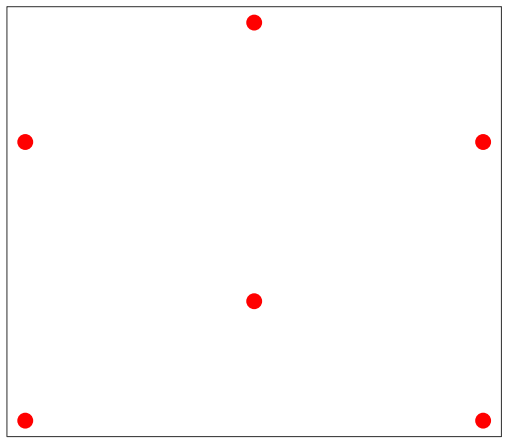
Geometrical criteria

2. Maximin D_{Maximin}

- **maximise the minimum distance** btw DoE points :

$$\max_{D} \min_{x^{(1)},x^{(2)} \in D} d(x^{(1)}, x^{(2)}) = \min_{x^{(1)},x^{(2)} \in D_{\text{Maximin}}} d(x^{(1)}, x^{(2)}) \quad (L^2\text{norme})$$

- Tendency to place points close to the domain boundaries.



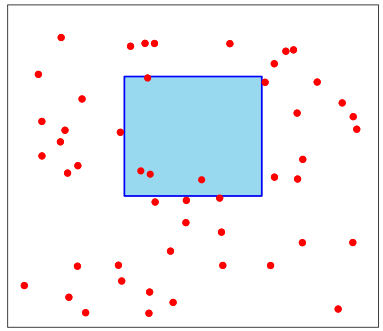
Space-filling designs (2)

Discrepancy measure

- Discrepancy : statistical criteria measuring the maximum deviation the curent sample points distribution and a uniform distribution.
- geometrical interpretation : comparaison btw subdomains volume and number of points in subdomains

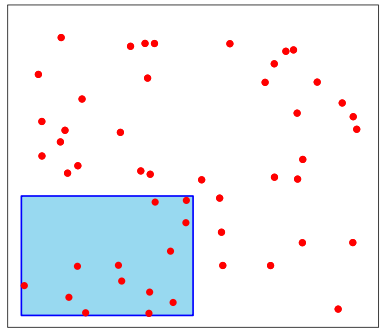
$$Q(t) \subset \mathcal{X} = [0,1[^d,\ Q(t) = [0,t_1[\times[0,t_2[\times\cdots\times[0,t_d[$$

$$\text{Discrepancy(DoE)} = \sup_{Q(t)\in[0,1[^d} \left| \frac{n_{Q(t)}}{n} - \prod_{i=1}^{d} t_i \right|$$



- Low discrepancy : "uniform" distribution of the points in the design space
- In practice : we choose a discrepancy with an L^2 norm

  $\longrightarrow$ Analytical form (discrepancy is centered for instance)

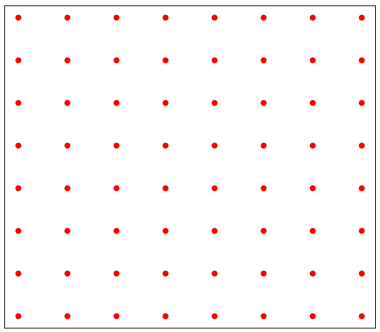


Space-filling designs (2)

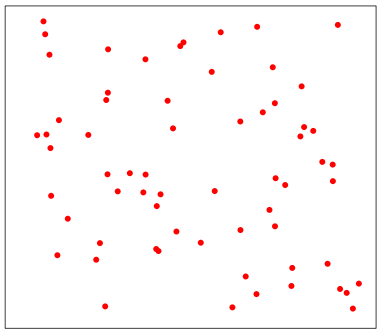
Low discrepancy sequences, i.e. Quasi Monte-Carlo (QMC)

- there exists many sequences with low discrepancy that can be used for space designs with good filling properties : Halton, Sobol, Van der Corput sequences ... and fast to build
- same framework as Monte Carlo techniques but faster convergence $\mathcal{O}((\ln n)^s/n)$ when $d \leq 10$

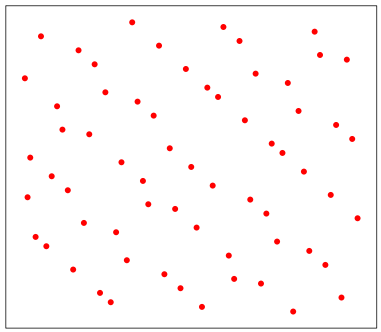
Regular grid



Monte Carlo



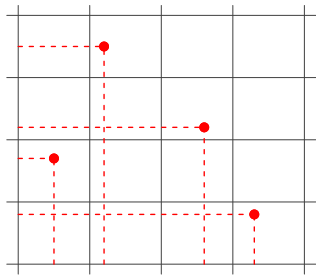
Sobol' sequence



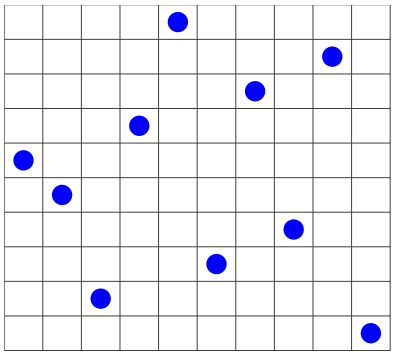
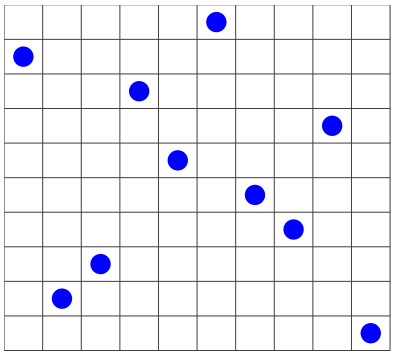
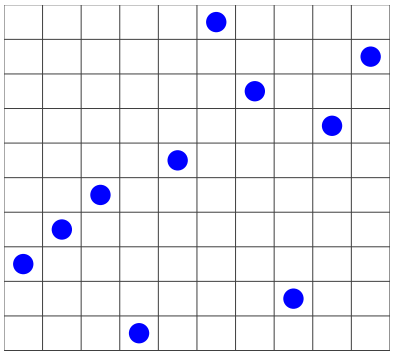
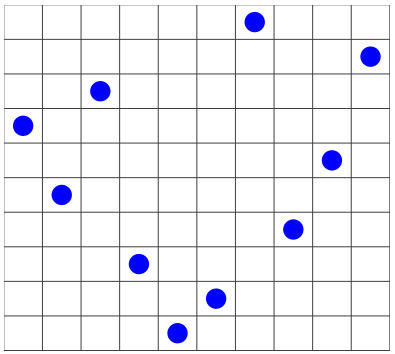
Space-filling designs (3)

Latin Hypercube Sampling (LHS)

- Properties : uniform projections onto the marginals
- Principle of an $\mathrm{LHS}(n, d)$ (d input parameters, n points) :
  - each dimension is divided into n intervals
  - random drawing of a point into each stratum



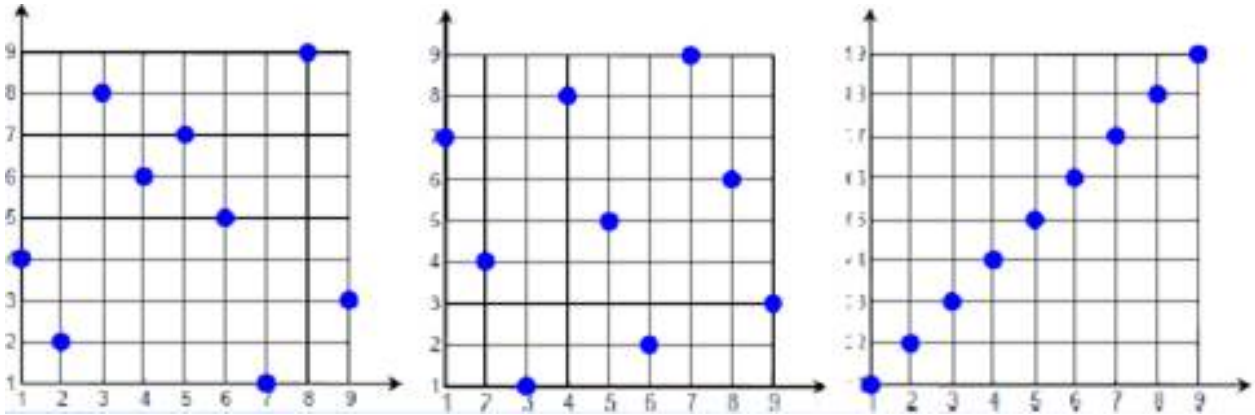
- each of the strata levels is occupied a single time for each parameter

Space-filling designs (3)

Avantages/drawbacks

- very easy to build : each of the design columns is a permutation of $\{1, 2, \ldots, n\}$ (strata choice)
- no difficulties for large n and d
- no duplicates when projecting onto subspaces
- (–) **BUT** does not necessarily fills up the space !



→ possibility of improving LHS thanks to optimization of different criteria (space filling criterion introduced before)

To recap on space filling designs

Goal

Sample a large-dimensional space in an "optimal" way, it ie get as much information as possible about the behavior of the output $Y = g(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$.

- Exploratory plans are good candidates for filling the space well. They are optimized either : :
  - on a criterion of distances between the points (minimax, maximin)
  - on a criterion of uniform distribution of points (discrepancy)
- The property of uniform projections on the margins can be obtained *via* Latin hypercube sampling (LHS)
- It is possible to couple the 2 properties by building optimized LHS.
- There are other types of designs for filling up the input parameter space (Voronoi centroidal tessellations, maximum entropy DoE, Strauss design space, etc.)

Short conclusion about Course II

- many choices for **metamodels** for OUU, but **Kriging** is appealing, for low to moderate number of model parameters, and **regular** and relatively **stationary** design functions
- its mathematical foundation relying on **Gaussian Processes** is very strong, GP provide a **Bayesian prior** over unknown functions with the benefit of uncertainty quantification
- its **prediction of variance**, essentially a **measure of error**, allows its precision to be quantified
- considering adequate **space-filling Design of Experiments** is beneficial to building more accurate surrogate models in high parametric dimension