Training on Gaussian processes

Chair PILearnWater

O. Roustant

INSA Toulouse, 20-22 January 2025

Objectives

At the end of the training, the participants should:

- Know basics on Gaussian processes (GP), kernels, RKHS with application to the approximation of functions
- Put in action these notions to customize GPs subject to linear information (e.g. derivatives, linear PDEs)

Program

We will follow the textbook Metamodeling with Gaussian processes, available online.

- Session 1: Generalities on random processes and GPs (Chapters 2 & 7)

Demo Simulation of (conditional) GPs

- Session 2: Kernels (Chapter 3)

Demo *Application of GPs to predict flooding maps*

- Session 3: RKHS and Gaussian process regression (Chapters 3 & 4)
- Session 4: Physics-informed GPs (Chapters 2, 3 & 4 + extra)

Outline

1. **Gaussian processes (GP)**
   - Random processes
   - Gaussian processes
   - Demo 1: simulation of (conditional) GPs
2. Kernels and RKHS
3. GP regression
4. Physics-informed GP

Random processes

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which all the (real-valued) random variables will be defined. We denote by $L^2(\mathbb{P})$ the Hilbert space of square integrable random variables (defined on Ω).

For a given set $\mathbb{X}$, a *random process* (RP) is a family of random variables $Y(x) : \Omega \to \mathbb{R}$, indexed by $x \in \mathbb{X}$.

We will denote $Y := (Y(x))_{x \in \mathbb{X}}$.

Trajectory, realization or sample path

Let Y be a RP. For a fixed $w \in \Omega$, a *trajectory or realization or sample path* of Y is the function $x \mapsto Y(x)(w)$.

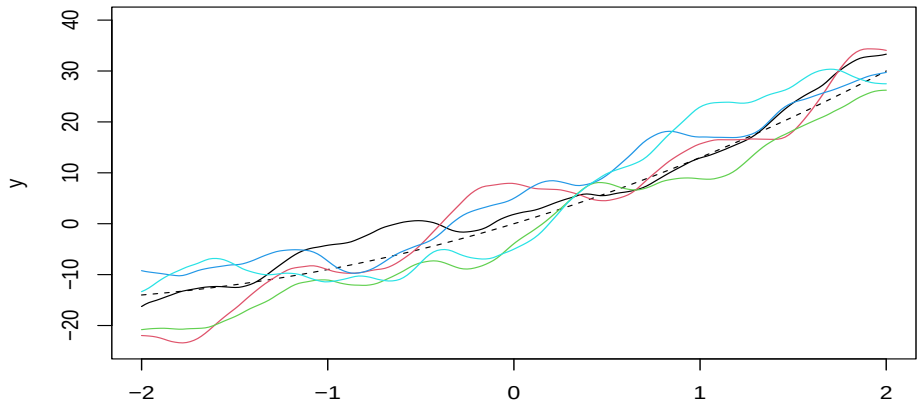


Figure: Five sample paths of a Gaussian process on $\mathbb{X} = [-2, 2]$

Second-order random process, mean, kernel.

Y is a *second-order RP* when all the r.v. $Y(x)$ belong to $L^2(\mathbb{P})$.

By Cauchy-Schwartz inequality, this implies that first moments (expectation) as well as second moments (covariances) are well-defined.

- The *mean* of Y is the function $x \in \mathbb{X} \mapsto \mathbb{E}(Y(x))$.
- The *covariance function* or *kernel* of Y is the function $(x, x') \in \mathbb{X} \times \mathbb{X} \mapsto \mathbb{C}\mathrm{ov}(Y(x), Y(x'))$.

Similarly, the *variance* of Y denotes the function $x \in \mathbb{X} \mapsto k(x, x) = \mathbb{V}\mathrm{ar}(Y(x))$.

Stationarity

- Y is *strongly stationary* if for all locations $x_1, \ldots, x_n \in \mathbb{X}$, the law of $(Y(x_1 + t), \ldots, Y(x_n + t))$ does not depend on the translation t.

Stationarity

- Y is *strongly stationary* if for all locations $x_1, \ldots, x_n \in \mathbb{X}$, the law of $(Y(x_1 + t), \ldots, Y(x_n + t))$ does not depend on the translation t.
- Y is *weakly stationary* if for all locations $x_1, \ldots, x_n \in \mathbb{X}$, the first two moments of the law of $(Y(x_1 + t), \ldots, Y(x_n + t))$ do not depend on t. Equivalently:

$$\mathbb{E}(Y(x)) = m, \qquad k(x, x') = c(x - x')$$

with $m = \mathbb{E}(Y(x_0))$ (for some $x_0 \in \mathbb{X}$) and $c(h) = k(0, h) = k(0, -h)$.

Stationarity

- Y is *strongly stationary* if for all locations $x_1, \ldots, x_n \in \mathbb{X}$, the law of $(Y(x_1 + t), \ldots, Y(x_n + t))$ does not depend on the translation t.
- Y is *weakly stationary* if for all locations $x_1, \ldots, x_n \in \mathbb{X}$, the first two moments of the law of $(Y(x_1 + t), \ldots, Y(x_n + t))$ do not depend on t. Equivalently:

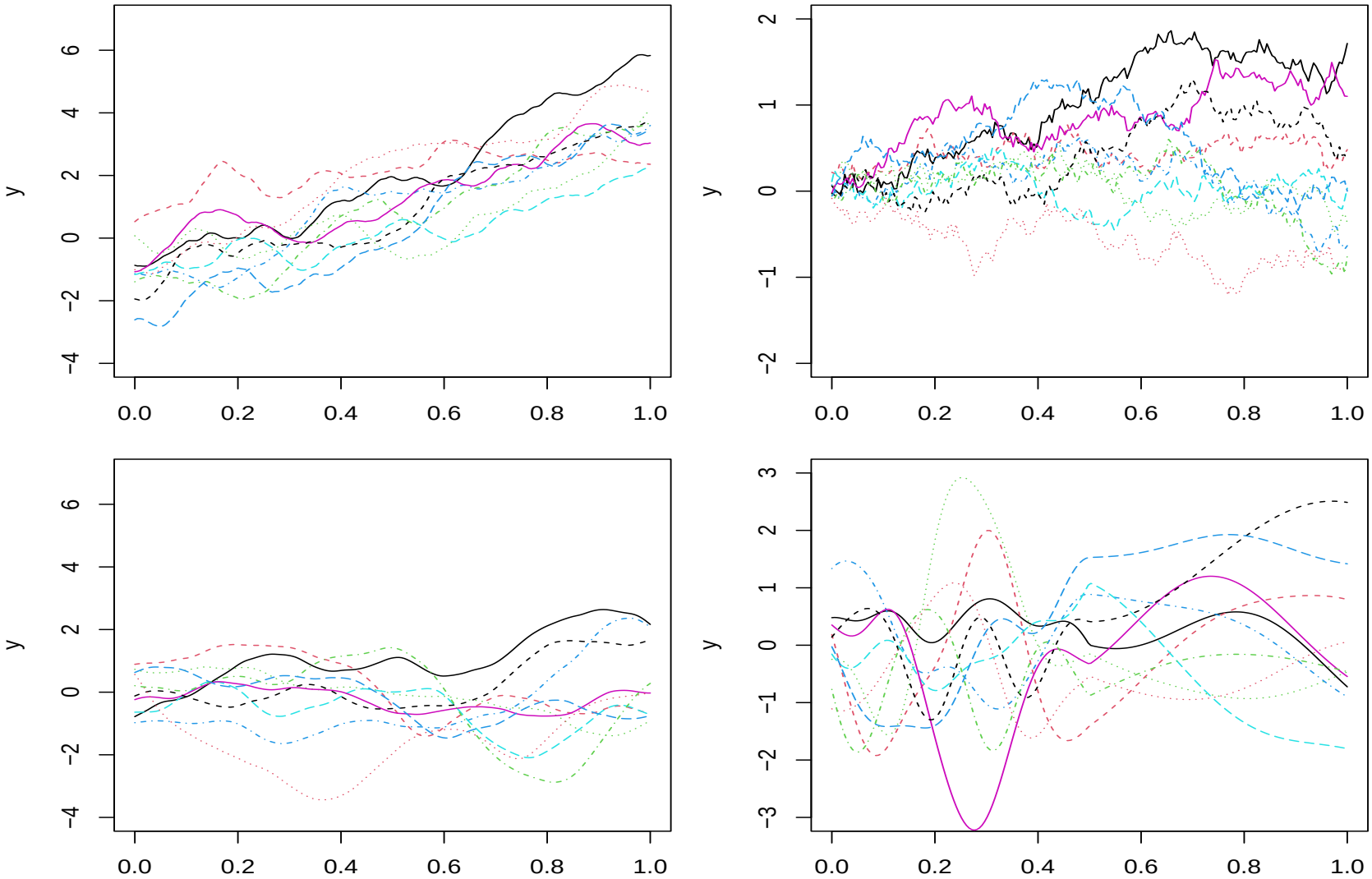
$$\mathbb{E}(Y(x)) = m, \qquad k(x, x') = c(x - x')$$

with $m = \mathbb{E}(Y(x_0))$ (for some $x_0 \in \mathbb{X}$) and $c(h) = k(0, h) = k(0, -h)$.

Obviously, strong stationarity implies weak stationarity.

Exercise

Only one graph corresponds to a stationary process. Which one?
For the others, why the corresponding random process is non stationary?



Gaussian processes

A random process Y defined on $\mathbb{X}$ is a *Gaussian process* (GP) if for all locations $x_1, \ldots, x_n \in \mathbb{X}$, the random vector $(Y(x_1), \ldots, Y(x_n))$ is a Gaussian vector.

The law of such random vectors is fully characterized by the mean m and the kernel k of Y. We will denote $Y \sim GP(m, k)$.

Reminder on Gaussian vectors

$X := (X_1, \ldots, X_d)^\top$ is a Gaussian vector iff it is the affine transformation of independent standard Normal random variables: $\exists \mu \in \mathbb{R}^d, A \in \mathbb{R}^{d \times m}$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)^\top$ where $\varepsilon_1, \ldots, \varepsilon_m$ are i.i.d. $\mathcal{N}(0,1)$, such that

$$X = \mu + A\varepsilon$$

Reminder on Gaussian vectors

$X := (X_1, \ldots, X_d)^\top$ is a Gaussian vector iff it is the affine transformation of independent standard Normal random variables: $\exists \mu \in \mathbb{R}^d, A \in \mathbb{R}^{d \times m}$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)^\top$ where $\varepsilon_1, \ldots, \varepsilon_m$ are i.i.d. $\mathcal{N}(0,1)$, such that

$$X = \mu + A\varepsilon$$

The mean of X is equal to μ, and its covariance matrix is

$$\mathbb{C}\mathrm{ov}(X) := \mathbb{E}\left[(X-\mu)(X-\mu)^\top\right] = AA^\top$$

Reminder on Gaussian vectors

$X := (X_1, \ldots, X_d)^\top$ is a Gaussian vector iff it is the affine transformation of independent standard Normal random variables: $\exists \mu \in \mathbb{R}^d, A \in \mathbb{R}^{d \times m}$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)^\top$ where $\varepsilon_1, \ldots, \varepsilon_m$ are i.i.d. $\mathcal{N}(0,1)$, such that

$$X = \mu + A\varepsilon$$

The mean of X is equal to μ, and its covariance matrix is

$$\mathbb{C}\mathrm{ov}(X) := \mathbb{E}\left[(X-\mu)(X-\mu)^\top\right] = AA^\top$$

If $\Gamma := \mathbb{C}\mathrm{ov}(X)$ is invertible, X is called non degenerated.
We denote $X \sim \mathcal{N}(\mu, \Gamma)$.

Density function of the multivariate normal distribution

If $X \sim \mathcal{N}(\mu, \Gamma)$ is a non degenerated Gaussian vector in $\mathbb{R}^d$, then X admits the density function

$$f_X(x) = \frac{1}{(2\pi)^{d/2}|\Gamma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Gamma^{-1}(x-\mu)\right)$$

where $|\Gamma| = \det(\Gamma)$.

Density function of the multivariate normal distribution

If $X \sim \mathcal{N}(\mu, \Gamma)$ is a non degenerated Gaussian vector in $\mathbb{R}^d$, then X admits the density function

$$f_X(x) = \frac{1}{(2\pi)^{d/2}|\Gamma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Gamma^{-1}(x-\mu)\right)$$

where $|\Gamma| = \det(\Gamma)$.

This comes directly from the definition, using the theorem of change of variables.

The level sets of the density function (the sets of $x \in \mathbb{R}^d$ such that $f_X(x) = y$, for a given y) are ellipsoids centered at μ, whose axis are given by the eigenvectors of Γ.

The linear combination property

$X := (X_1, \ldots, X_d)^\top$ is a Gaussian vector iff all linear combination of its components follow a (one-dimensional) Normal distribution:

$$\forall t_1, \ldots, t_d \in \mathbb{R}, \quad t_1 X_1 + \cdots + t_d X_d \quad \text{follows a Normal distribution}$$

The linear combination property

$X := (X_1, \ldots, X_d)^\top$ is a Gaussian vector iff all linear combination of its components follow a (one-dimensional) Normal distribution:

$$\forall t_1, \ldots, t_d \in \mathbb{R}, \quad t_1 X_1 + \cdots + t_d X_d \quad \text{follows a Normal distribution}$$

This can be proved with the characteristic function of a probability measure.

This is a practical way to show that X is a Gaussian vector. Mind that it is *not sufficient* that $X_1, \ldots, X_d$ are normally distributed.

Stability by linear mapping

A linear mapping of a Gaussian vector is a Gaussian vector. More precisely, if $X \sim \mathcal{N}(\mu, \Gamma)$ is Gaussian vector on $\mathbb{R}^d$, and $L : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a $d \times d'$ matrix, then LX is a Gaussian vector on $\mathbb{R}^{d'}$ with $LX \sim \mathcal{N}(L\mu, L\Gamma L^\top)$

Stability by linear mapping

A linear mapping of a Gaussian vector is a Gaussian vector. More precisely, if $X \sim \mathcal{N}(\mu, \Gamma)$ is Gaussian vector on $\mathbb{R}^d$, and $L : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a $d \times d'$ matrix, then LX is a Gaussian vector on $\mathbb{R}^{d'}$ with $LX \sim \mathcal{N}(L\mu, L\Gamma L^\top)$

The result is obvious from the definition: if $X = \mu + A\varepsilon$, then $LX = L\mu + LA\varepsilon$.

Non-correlation and independence

If $X = (X_1, \ldots, X_d)$ is a Gaussian vector, then for all (i, j), the random variables X_i and X_j are independent if and only if $\mathbb{C}\mathrm{ov}(X_i, X_j) = 0$.

Non-correlation and independence

If $X = (X_1, \ldots, X_d)$ is a Gaussian vector, then for all (i, j), the random variables X_i and X_j are independent if and only if $\mathbb{C}\mathrm{ov}(X_i, X_j) = 0$.

This is because the probability distribution of X only depends on the mean and the covariances of its components.

The property is FALSE if X is not a Gaussian vector (even if $X_1, \ldots, X_n$ are normally distributed!).

Conditioning of Gaussian vectors

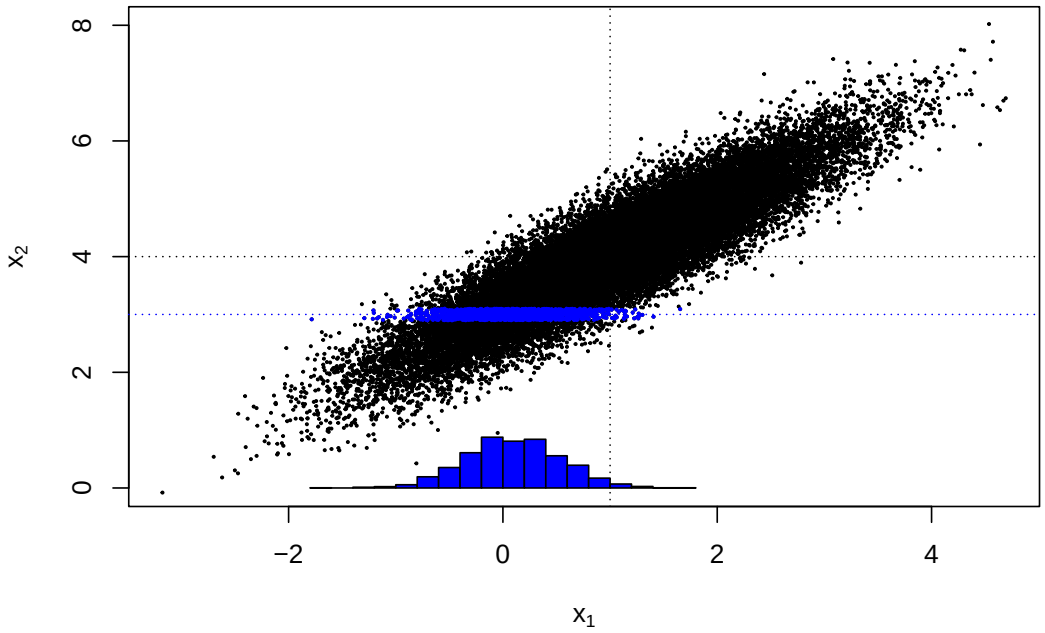


Figure: Illustration of the conditioning of Gaussian vectors on a simulated sample.

Let $U = (V, W) \sim \mathcal{N}(\mu, \Gamma)$ be a Gaussian vector on $\mathbb{R}^d$, where V, W are subvectors of dimension d_V, d_W respectively. Write $\mu = (\mu_V, \mu_W)^\top$ with $\mu_V = \mathbb{E}(V), \mu_W = \mathbb{E}(W)$ and

$$\Gamma = \begin{bmatrix} \Gamma_V & \Gamma_{V,W} \\ \Gamma_{W,V} & \Gamma_W \end{bmatrix}$$

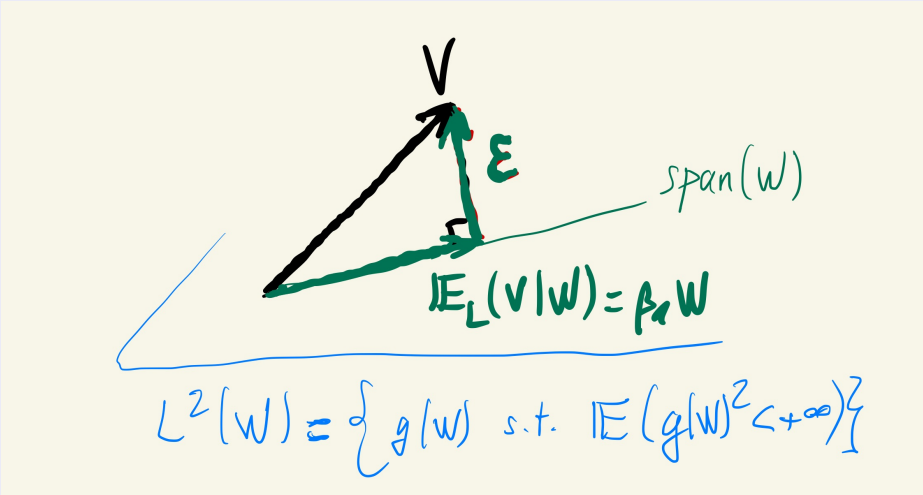
where $\Gamma_V = \mathbb{C}\mathrm{ov}(V), \Gamma_W = \mathbb{C}\mathrm{ov}(W)$, and $\Gamma_{V,W} = \mathbb{C}\mathrm{ov}(V, W) := \mathbb{E}[(V - \mu_V)(W - \mu_W)^\top]$ (similar definition for $\Gamma_{W,V}$).

Then $V|W = w$ is a Gaussian vector on $\mathbb{R}^{d_W}$ with mean and covariance matrix

$$\begin{aligned} \mathbb{E}(V|W = w) &= \mu_V + \Gamma_{V,W}\Gamma_W^{-1}(w - \mu_W) && \text{linear with respect to } w \\ \mathbb{C}\mathrm{ov}(V|W = w) &= \Gamma_V - \Gamma_{V,W}\Gamma_W^{-1}\Gamma_{W,V} && \text{does not depend on } w \end{aligned}$$

Exercise: Proof for centered random variables i.e. $d_V = 1, d_W = 1$

Let us write $\mathbb{E}_L(V|W) = \beta_1 W$ (we admit that there is no constant term) and define $\varepsilon = V - \mathbb{E}_L(V|W)$.



1. Let us show that linear / non-linear regressions of V on W coincide.
   - Prove that (ε, W) is a Gaussian vector.
   - Deduce that ε and W are independent.
   - Deduce that $\mathbb{E}(V|W) = \mathbb{E}_L(V|W)$
2. Prove that $V|W = w \overset{law}{=} \varepsilon + \beta_1 w \overset{law}{=} \mathcal{N}(\beta_1 w, \mathbb{V}\mathrm{ar}(\varepsilon))$.
3. Show that $\beta_1 = \Gamma_{V,W}\Gamma_W^{-1}$ and $\mathbb{V}\mathrm{ar}(\varepsilon) = \|\varepsilon\|^2 = \Gamma_V - \Gamma_{V,W}\Gamma_W^{-1}\Gamma_{W,V}$.

Linear and non-linear regression

Let $Y, X_1, \ldots, X_d$ be random variables in $L^2(\mathbb{P})$, and $X = (X_1, \ldots, X_d)$. Define:

- $\mathbb{E}(Y|X_1, \ldots, X_d)$, **the *non-linear regression* of Y on $X_1, \ldots X_d$**, as the best approximation of Y by functions of $X_1, \ldots, X_d$ in the L^2 sense.

  It is the orthogonal projection of Y onto $L^2(X_1, \ldots, X_d)$, the Hilbert space of square integrable random variables: $\mathbb{E}(Y|X_1, \ldots, X_d) = h(X)$ where $h(X)$ is such that $\mathbb{E}([Y - h(X)]^2)$ is minimal.

Linear and non-linear regression

Let $Y, X_1, \ldots, X_d$ be random variables in $L^2(\mathbb{P})$, and $X = (X_1, \ldots, X_d)$. Define:

- $\mathbb{E}(Y|X_1, \ldots, X_d)$, **the *non-linear regression* of Y on $X_1, \ldots X_d$**, as the best approximation of Y by functions of $X_1, \ldots, X_d$ in the L^2 sense.

  It is the orthogonal projection of Y onto $L^2(X_1, \ldots, X_d)$, the Hilbert space of square integrable random variables: $\mathbb{E}(Y|X_1, \ldots, X_d) = h(X)$ where $h(X)$ is such that $\mathbb{E}([Y - h(X)]^2)$ is minimal.

- $\mathbb{E}_L(Y|X_1, \ldots, X_d)$, **the *linear regression* of Y on $X_1, \ldots X_d$**, as the best approximation of Y by *linear combinations* of $1, X_1, \ldots, X_d$ in the L^2 sense.

  It is the orthogonal projection of Y onto the vector space spanned by $1, X_1, \ldots, X_d$:
  $\mathbb{E}_L(Y|X_1, \ldots, X_d) = \beta_0 + \beta^\top X$ where β_0, β are such that $\mathbb{E}([Y - (\beta_0 + \beta^\top X)]^2)$ is minimal.

Linear and non-linear regression

Let $Y, X_1, \ldots, X_d$ be random variables in $L^2(\mathbb{P})$, and $X = (X_1, \ldots, X_d)$. Define:

- $\mathbb{E}(Y|X_1, \ldots, X_d)$, **the *non-linear regression* of Y on $X_1, \ldots X_d$**, as the best approximation of Y by functions of $X_1, \ldots, X_d$ in the L^2 sense.

  It is the orthogonal projection of Y onto $L^2(X_1, \ldots, X_d)$, the Hilbert space of square integrable random variables: $\mathbb{E}(Y|X_1, \ldots, X_d) = h(X)$ where $h(X)$ is such that $\mathbb{E}([Y - h(X)]^2)$ is minimal.

- $\mathbb{E}_L(Y|X_1, \ldots, X_d)$, **the *linear regression* of Y on $X_1, \ldots X_d$**, as the best approximation of Y by *linear combinations* of $1, X_1, \ldots, X_d$ in the L^2 sense.

  It is the orthogonal projection of Y onto the vector space spanned by $1, X_1, \ldots, X_d$:
  $\mathbb{E}_L(Y|X_1, \ldots, X_d) = \beta_0 + \beta^\top X$ where β_0, β are such that $\mathbb{E}([Y - (\beta_0 + \beta^\top X)]^2)$ is minimal.

If $(Y, X_1, \ldots, X_d)$ is a Gaussian vector, then

$$\mathbb{E}(Y|X_1, \ldots, X_d) = \mathbb{E}_L(Y|X_1, \ldots, X_d)$$

Some properties of GPs inherited from Gaussian vectors

For GPs, properties of Gaussian vectors directly imply:

- **strong stationarity $\Leftrightarrow$ weak stationarity.**
  Thus, we simply speak of stationary GP.

Some properties of GPs inherited from Gaussian vectors

For GPs, properties of Gaussian vectors directly imply:

- **strong stationarity $\Leftrightarrow$ weak stationarity.**
  Thus, we simply speak of stationary GP.
- **independence $\Leftrightarrow$ non correlation.**
  If Y is a GP, $Y(x)$ and $Y(x')$ are independent iff $\mathbb{C}\mathrm{ov}(Y(x), Y(x')) = 0$.

Some properties of GPs inherited from Gaussian vectors

For GPs, properties of Gaussian vectors directly imply:

- **strong stationarity $\Leftrightarrow$ weak stationarity.**
  Thus, we simply speak of stationary GP.
- **independence $\Leftrightarrow$ non correlation.**
  If Y is a GP, $Y(x)$ and $Y(x')$ are independent iff $\mathbb{C}\mathrm{ov}(Y(x), Y(x')) = 0$.
- **a GP is stable by linear mapping.**
  Formally, if Y is a GP, and L is a linear mapping operating on the sample paths of Y, then LY is a GP.

Some properties of GPs inherited from Gaussian vectors

For GPs, properties of Gaussian vectors directly imply:

- **strong stationarity $\Leftrightarrow$ weak stationarity.**
  Thus, we simply speak of stationary GP.
- **independence $\Leftrightarrow$ non correlation.**
  If Y is a GP, $Y(x)$ and $Y(x')$ are independent iff $\mathbb{C}\mathrm{ov}(Y(x), Y(x')) = 0$.
- **a GP is stable by linear mapping.**
  Formally, if Y is a GP, and L is a linear mapping operating on the sample paths of Y, then LY is a GP.
- **a GP Y conditioned on $Y(x_i) = y_i, (i = 1, \ldots, n)$ is still a GP.**
  This is the basis of Gaussian process regression, developed in Chapter 4. By stability of GPs under linearity, this property is true for linear equality constraints (and not only interpolation ones).

Gaussian processes and linear operations

If $Y \sim GP(0, k)$ and L is a linear function acting on the sample paths. of Y, then $LY \sim GP(0, k_L)$ with $k_L(s,t) = L_s L_t k(s,t)$. Here L_s (resp. L_t) means that we apply L on the function $s \mapsto k(s,t)$ (resp. $t \mapsto k(s,t)$).

Gaussian processes and linear operations

If $Y \sim GP(0, k)$ and L is a linear function acting on the sample paths. of Y, then $LY \sim GP(0, k_L)$ with $k_L(s,t) = L_s L_t k(s,t)$. Here L_s (resp. L_t) means that we apply L on the function $s \mapsto k(s,t)$ (resp. $t \mapsto k(s,t)$).

The expression of k_L comes, formally, from the bilinearity of covariance:

$$\mathbb{C}\mathrm{ov}(LY(s), LY(t)) = L_t(\mathbb{C}\mathrm{ov}(LY(s), Y(t)) = L_s L_t \mathbb{C}\mathrm{ov}(Y(s), Y(t))$$

Gaussian processes and linear operations

If $Y \sim GP(0, k)$ and L is a linear function acting on the sample paths. of Y, then $LY \sim GP(0, k_L)$ with $k_L(s,t) = L_s L_t k(s,t)$. Here L_s (resp. L_t) means that we apply L on the function $s \mapsto k(s,t)$ (resp. $t \mapsto k(s,t)$).

The expression of k_L comes, formally, from the bilinearity of covariance:

$$\mathbb{C}\mathrm{ov}(LY(s), LY(t)) = L_t(\mathbb{C}\mathrm{ov}(LY(s), Y(t)) = L_s L_t \mathbb{C}\mathrm{ov}(Y(s), Y(t))$$

Example if L is the derivative. If Y is a 1D centered GP with kernel k (smooth enough), then when it exists, $(Y'(x))_{x\in\mathbb{X}}$ is a centered GP, with kernel:

$$k_{Y'}(s,t) = \mathbb{C}\mathrm{ov}\left(\frac{\partial Y(s)}{\partial s}, \frac{\partial Y(t)}{\partial t}\right) = \frac{\partial}{\partial t}\mathbb{C}\mathrm{ov}\left(\frac{\partial Y(s)}{\partial s}, Y(t)\right) = \frac{\partial}{\partial s}\frac{\partial}{\partial t}\mathbb{C}\mathrm{ov}(Y(s), Y(t)) = \frac{\partial^2 k}{\partial s \partial t}(s,t)$$

Simulation of a GP

A simulation of $Y \sim GP(m, k)$ is possible on a set of discrete locations $X = \{x_1, \ldots, x_n\}$.

Simulation of a GP

A simulation of $Y \sim GP(m, k)$ is possible on a set of discrete locations $X = \{x_1, \ldots, x_n\}$. Indeed, then the law of $(Y(x_1), \ldots, Y(x_n))^\top$ is $\mathcal{N}(m(X), k(X, X))$ where

- $m(X)$ is the vector of size n whose component i is equal to $m(x_i)$
- $k(X, X)$ is the matrix of size n whose coefficient (i, j) is equal to $k(x_i, x_j)$

Obtaining a realization of Y at X, it is thus equivalent to simulating from $\mathcal{N}(m(X), k(X, X))$.

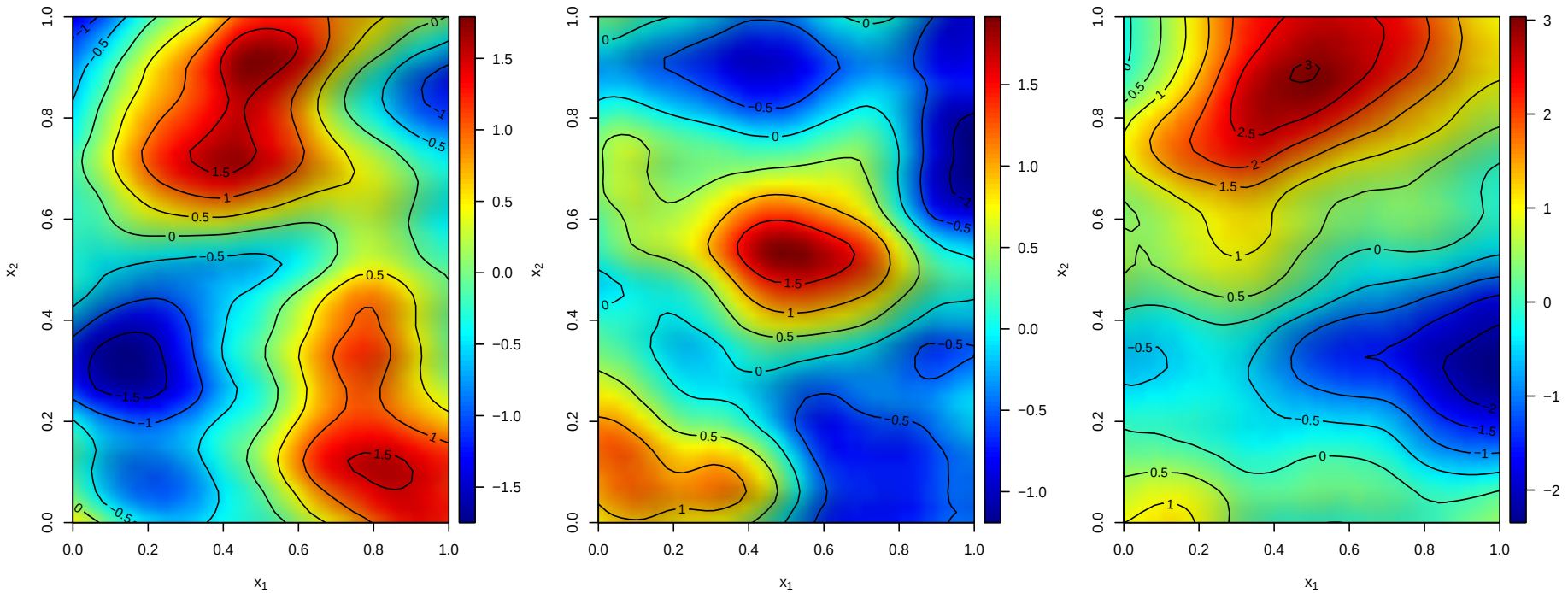


Figure: Simulations of a Gaussian process on $\mathbb{X} = [0,1]^2$ (obtained by simulating a Gaussian vector on a fine grid of $\mathbb{X}$), with kernel $k(x, x'; \ell) = k_1(x_1, x_1'; 2\ell)k_2(x_2, x_2'; \ell)$, where k_1 is a one-dimensional Matérn 5/2 kernel (see Chap. 3) and ℓ is a parameter.

Demo 1: simulation of GPs and conditional GPs

See the R code

Transition: find a GP with odd sample paths! (exercise 2.1)

Outline

1 Gaussian processes (GP)

2 **Kernels and RKHS**
- Kernels
- Demo 2: prediction of costal flooding maps
- RKHS

3 GP regression

4 Physics-informed GP

Positive semi-definite functions

Let $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a symmetric function, i.e. $\forall x, x' \in \mathbb{X},\ k(x, x') = k(x', x)$.

k is *positive semi-definite* (psd) if for all finite set $X = \{x_1, \ldots, x_n\}$, the matrix $k(X, X) = (k(x_i, x_j))_{1 \leq i,j \leq n}$ is (symmetric) psd.

Equivalently,

$$k \text{ is psd} \quad \Leftrightarrow \quad \begin{array}{c} \forall n \geq 1, \forall x_1, \ldots, x_n \in \mathbb{X}, \forall \alpha_1, \ldots, \alpha_n \in \mathbb{R}, \\ \sum_{1 \leq i,j \leq n} \alpha_i \alpha_j k(x_i, x_j) \geq 0 \end{array}$$

Kernels, covariance functions and psd functions

Covariance functions and psd functions coincide

- if k is the covariance of a (second-order) random process, then k is psd.
- if k is psd, there exists a second-order centered RP with covariance k. Moreover, there is a unique (in law) centered GP with kernel k.

Thus, the word *kernel* will denote either a covariance function or a psd function.

Exercise 3.1. Prove the first point.

All kernels are scalar products in a feature space

If k is a kernel, there exists a Hilbert space $(\mathcal{H}, \langle ., . \rangle)$, called *feature space*, and a function $\Phi : \mathbb{X} \to \mathcal{H}$, often called *kernel embedding*, such that

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle \quad \text{for all} \quad x, x' \in \mathbb{X}$$

For instance, if Z is a centered RP with kernel k, we can take

$$\mathcal{H} = L^2(\mathbb{P}), \qquad \Phi(x) = Z(x).$$

General operations on kernels (valid for all domain $\mathbb{X}$)

- If k is a kernel, then $\sigma^2 k$ is a kernel for all $\sigma \in \mathbb{R}$.
- (Stability by sum) If k_1, k_2 are two kernels on $\mathbb{X}^2$, then $k_1 + k_2$ is a kernel. Similarly if k_1, k_2 are kernels on $(\mathbb{X}_1)^2, (\mathbb{X}_2)^2$ respectively, the tensor sum[†] $k_1 \oplus k_2$ is a kernel on $(\mathbb{X}_1 \times \mathbb{X}_2)^2$.
- (Stability by product) If k_1, k_2 are two kernels on $\mathbb{X}^2$, then $k_1 k_2$ is a kernel. Similarly, if k_1, k_2 are kernels on $(\mathbb{X}_1)^2, (\mathbb{X}_2)^2$ respectively, the tensor product[†] $k_1 \otimes k_2$ is a kernel on $(\mathbb{X}_1 \times \mathbb{X}_2)^2$.
- (*warping* or *embedding*) If k is a kernel on $\mathbb{X}^2$, and $f : \mathbb{U} \to \mathbb{X}$ is a function, then $k_f(u, u') := k(f(u), f(u'))$ is a kernel on $\mathbb{U}^2$.

[†] The tensor sum is $k_{\mathbf{1}} \oplus k_{\mathbf{2}} \left(\begin{pmatrix} x_{\mathbf{1}} \\ x_{\mathbf{2}} \end{pmatrix}, \begin{pmatrix} x'_{\mathbf{1}} \\ x'_{\mathbf{2}} \end{pmatrix} \right) = k_{\mathbf{1}}(x_{\mathbf{1}}, x'_{\mathbf{1}}) + k_{\mathbf{2}}(x_{\mathbf{2}}, x'_{\mathbf{2}})$. Similar def. for the product.

Exercise 3.2. (Proof of operations on kernels)

Let Y, Y_1, Y_2 be centered GPs associated to k, k_1, k_2, with Y_1, Y_2 independent.

Find a centered random process Z corresponding to the target kernel, built in function of Y, Y_1, Y_2. Is Z a GP? Summarize your findings in the table below.

kernel	associated RP	is this RP a GP?
$\sigma^2 k$		
$k_1 + k_2$		
$k_1 k_2$		
k_f		

Mercer representation of continuous kernels on compact domains

Assume that $\mathbb{X}$ is a *compact* metric space, and k is *continuous* on $\mathbb{X} \times \mathbb{X}$. Let ν be a finite measure on $\mathbb{X}$ and $L^2(\mathbb{X}, \nu) = \{f : \mathbb{X} \to \mathbb{R},\ \int_T f(x)^2 d\nu(x) < +\infty\}$

Then there exists a Hilbert basis of eigenfunctions $(\phi_n)_{n \geq 0}$ of $L^2(\mathbb{X}, \nu)$ and eigenvalues $(\lambda_n)_{n \geq 0}$, with $\lambda_0 \geq \lambda_1 \geq ... \geq \lambda_n \geq \cdots \geq 0$ such that

$$k(x, x') = \sum_{n \geq 0} \lambda_n \phi_n(x) \phi_n(x')$$

where the convergence is uniform on $\mathbb{X} \times \mathbb{X}$. The $\phi_n' s$ are eigenfunctions of the Hilbert-Schmidt operator defined on $L^2(\mathbb{X}, \nu)$ by

$$Tf(x) = \int_T k(x, x') f(x') d\nu(x')$$

and thus verify $T\phi_n(x) = \lambda_n \phi_n(x)$ for all $n \geq 0$.

This extends the fact that psd matrices admit a spectral decomposition.

Radial kernels on Hilbert spaces

Let $\mathbb{X} = \mathcal{H}$ be a Hilbert space, with norm $\|.\|$

The function

$$k(x, x') = F(\|x - x'\|)$$

is a kernel iff there exists a Borel measure[†] on $\mathbb{R}^+$ s.t. $F(t) = \int_{\mathbb{R}} e^{-ut^2} d\nu(u)$.
It is a (strict) pd function iif we add the condition: $\nu \neq \delta_0$

N.B. If we write $k(x, x') = G(\|x - x'\|^2)$ *then* $G(t) = F(\sqrt{t})$ *is the Laplace transform of* ν.

Examples:

Kernel name	**Expression of** $F(t)$
Squared exponential	$\exp(-t^2)$
Power-exponential	$\exp(-t^s)$ with $s \in (0, 2]$
Multiquadric	$(c^2 + t^2)^{-\beta}$ with $\beta > 0, c > 0$

[†] ν is a Borel measure if for all compact C we have $\nu(C) < +\infty$

Kernels of stationary GPs on $\mathbb{R}^d$ (Bochner's theorem).

The kernel of a real-valued stationary GP on $\mathbb{R}^d$ is the Fourier transform of a finite measure

$$k(x, x') = \int_{\mathbb{R}^d} \cos(2\pi\langle x - x', t\rangle) d\mu(t)$$

where $\langle ., . \rangle$ is the usual scalar product on $\mathbb{R}^d$.
The finite measure μ is called *spectral measure*.

Exercise 3.3. Proof that all expressions of these forms are valid kernels.
Hint: Using that for all u in $\mathbb{R}$, $\cos(u) = Re\left(e^{iu}\right)$, prove that the function $k_t : (x, x') \mapsto \cos(2\pi\langle x - x', t\rangle)$ is psd.

Kernel name	Kernel form	Spectral measure
cosine	$\cos(2\pi h)$	Dirac δ_1
sinc	$\frac{\sin(\pi h)}{\pi h}$	Uniform
Squared exponential	$k(h) = \exp\left(-\frac{1}{2}\frac{h^2}{\ell^2}\right)$	Gaussian
Exponential	$\exp\left(-\frac{\lvert h\rvert}{\ell}\right)$	Student $t_{1/2}$
Matérn 3/2	$\left(1+\sqrt{3}\frac{\lvert h\rvert}{\ell}\right)\exp\left(-\sqrt{3}\frac{\lvert h\rvert}{\ell}\right)$	Student $t_{3/2}$
Matérn 5/2	$\left(1+\sqrt{5}\frac{\lvert h\rvert}{\ell}+\frac{5}{3}\frac{h^2}{\ell^2}\right)\exp\left(-\sqrt{5}\frac{\lvert h\rvert}{\ell}\right)$	Student $t_{5/2}$

Table: Examples of kernels of 1-dim. stationary GPs on $\mathbb{X} = \mathbb{R}$. Here $h = x - x'$.

Demo 2: prediction of costal flooding maps

Illustration and discussion

RKHS (definition)

Reproducing kernel Hilbert spaces (RKHS) are Hilbert spaces of functions $f : \mathbb{X} \to \mathbb{R}$ such that for all $x \in \mathbb{X}$ the evaluation $f \mapsto f(x)$ is continuous.

This automatically gives a kernel. Indeed, by Riesz theorem, for all $x \in \mathbb{X}$, there exists a unique $k_x \in \mathcal{H}$ s.t. for all $f \in \mathcal{H}$,

$$f(x) = \langle f, k_x \rangle_{\mathcal{H}} \qquad (*)$$

Define $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ by $k(x, x') = k_x(x')$. Choosing $f = k_{x'}$ in (*) gives

$$\langle k(x, .), k(x', .) \rangle_{\mathcal{H}} = k(x, x') \qquad (**)$$

which shows in particular that k is a kernel.
(*), (**) are called reproducing properties and k the reproducing kernel of $\mathcal{H}$.

RKHS (equivalence with psd functions)

We have just seen that if $\mathcal{H}$ is a RKHS, then it provides a kernel k.

Moore-Aronszajn theorem

Conversely, if k is a kernel, one can construct a unique RKHS $\mathcal{H}_k$ with kernel k. This RKHS is given by

$$\mathcal{H}_k = \overline{\text{span}(k(x,.), x \in \mathbb{X})}$$

with inner product defined on the $k(x,.)$'s by $\langle k(x,.), k(y,.)\rangle := k(x,y)$, and extended to $\mathcal{H}_k$ by linearity and continuity.

RKHS (Translation between RKHS and random processes)

Let Z be a (second-order) RP with kernel k.
Let $\bar{\mathcal{L}}(Z) = \overline{\text{span}(Z(t),\ t \in \mathbb{X})}$,

Loève isometry

$\bar{\mathcal{L}}(Z)$ is isometric to $\mathcal{H}_k$ through the map defined on the $k(x,.)$'s by

$$\phi : \begin{array}{c} \mathcal{H}_k \to \bar{\mathcal{L}}(Z) \\ k(x,.) \mapsto Z(x) \end{array} \tag{1}$$

and extended by linearity and continuity.

RKHS (examples)

- Finite-dimensional Hilbert spaces (i.e. Euclidean spaces)
- Infinite dimensional Hilbert spaces with minimal regularity
  - If $\mathbb{X}$ is a bounded interval, $L^2(\mathbb{X})$ is NOT a RKHS
  - When $X = \mathbb{R}^d$, the Sobolev space $H^m(\mathbb{X})$ is a RKHS iff $m > d/2$ $\rightarrow H^1(\mathbb{R}), H^2(\mathbb{R}), \ldots$ are RKHS
  - Matérn kernels correspond to Sobolev spaces with specific norms

Summary: the 3 faces of a kernel

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

where $\mathcal{H}$ is a reproducing kernel Hilbert Space with dot product:

$$\langle k(x, .), k(x', .) \rangle = k(x, x')$$

Outline

1 Gaussian processes (GP)

2 Kernels and RKHS

3 **GP regression**
- The Gaussian process approach
- The geostatistical approach: Kriging
- The functional approach: approximation in RKHS

4 Physics-informed GP

A functional interpolation problem

Given a set of observations $(x_i, y_i), i = 1, \ldots, n$ with $x_i \in \mathbb{X}$ and $y_i \in \mathbb{R}$, we search for an interpolator i.e. a function $f : \mathbb{X} \to \mathbb{R}$ such that

$$f(x_i) = y_i, \qquad \text{for all } i = 1, \ldots, n$$

This is an ill-posed problem, so we need more assumptions.

GP regression

Here we view f as one realization of a GP Y with mean m and kernel k

Then, interpolating is equivalent to condition Y on $Y(x_i) = y_i$ $(i = 1, \ldots, n)$

In a Gaussian setting, the conditional process is known explicitely:
$Y|\{Y(x_i) = y_i, i = 1, \ldots, n\}$ is a GP with mean m_c and kernel k_c, with:

$$\begin{aligned} m_c(x) &= m(x) + k(x, X)k(X, X)^{-1}(y - m(X)) \\ k_c(x, x') &= k(x, x') - k(x, X)k(X, X)^{-1}k(X, x') \end{aligned}$$

Notations:

- $y = (y_{\mathbf{1}}, \ldots, y_n)^\top$
- $m(X) = (m(x_{\mathbf{1}}), \ldots, m(x_n))^\top$
- $k(x, X) = (k(x_{\mathbf{1}}, x), \ldots, k(x_n, x))$, $k(X, x) = k(x, X)^\top$
- $k(X, X) = (k(x_i, x_j))_{\mathbf{1} \leq i,j \leq n}$.

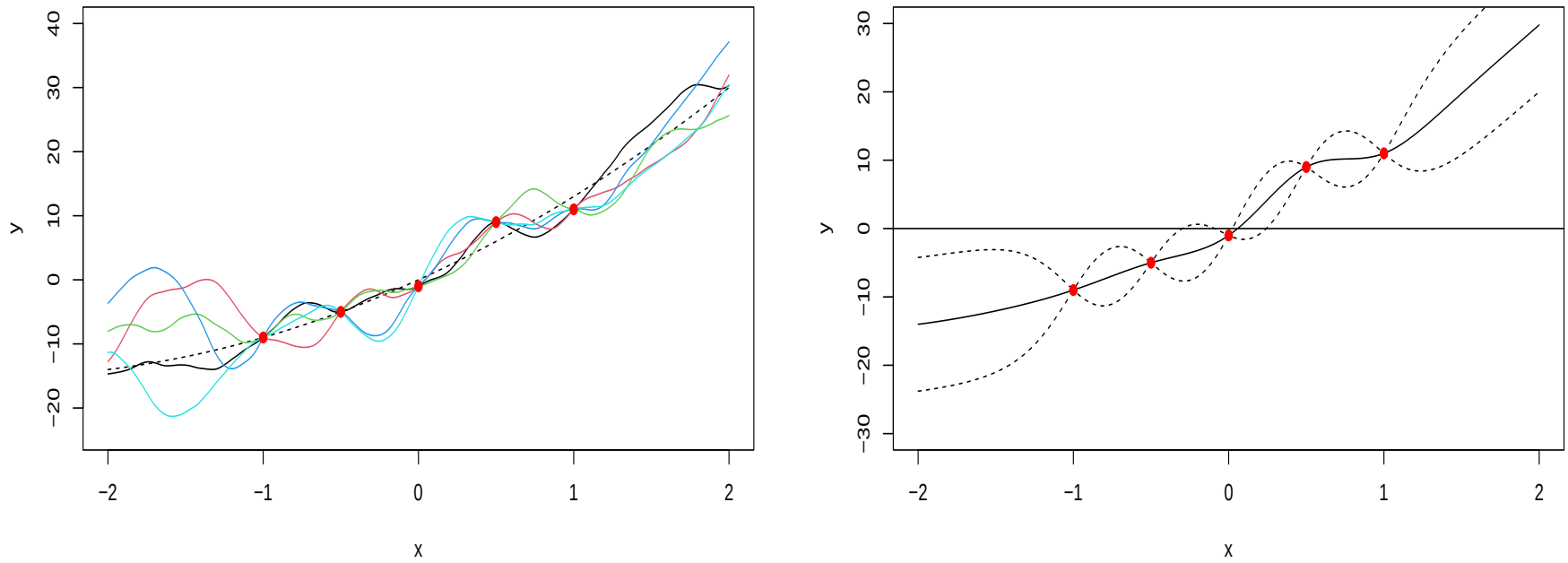


Figure: Example of GP conditional on $\{Y(x_i) = y_i, i = 1, \ldots, n\}$ for $n = 5$ points. Left: simulations. Right: mean $m_c(x)$ and 95% prediction intervals $\left[m_c(x) \pm 1.96\sqrt{k_c(x,x)}\right]$.

From geostatistics to GP regression models

Geostatistical approach for spatial interpolation, Simple Kriging

Let Y be a centered random process (or with known mean).
In geostatistics, the prediction of $Y(x)$ knowing $Y(x_1), \ldots, Y(x_n)$ is computed by the Best Linear Unbiased Predictor (BLUP). It means, to find $w_1, \ldots, w_n$ s.t.

$$\hat{Y}(x) := w_0 + w_1 Y(x_1) + \cdots + w_n Y(x_n)$$

minimizes $\mathrm{MSE} := \mathbb{E}([Y(x) - \hat{Y}(x)]^2)$ under $\mathbb{E}(\hat{Y}(x)) = \mathbb{E}(Y(x))$.

Link between Simple Kriging and Gaussian process interpolation

- If Y is Gaussian, the conditional expectation coincides with the orthogonal projection onto a linear space
  $\rightarrow$ BLUP = conditional expectation and min(MSE) = conditional variance
- If Y is not Gaussian, the two approaches are different in general
  $\rightarrow$ Advantage of BLUP: closed-form expressions
  $\rightarrow$ Drawback: we do not know the conditional law

Time line

- 1951: Spatial interpolation in geosciences (Krige, 1951)
  $\rightarrow$ "Kriging"
- 1963: Foundations of geostatistics (Matheron, 1963)
- 1989: Computer experiments, metamodelling (Sacks, Welch, Mitchell and Wynn, 1989)
  $\rightarrow$ Application to dimensions ≥ 4

Gaussian processes, splines and RKHS

Correspondence between interpolation spline and GP conditional mean

The interpolation spline is defined by the functional problem

$$(*) \qquad \min_{h \in \mathcal{H}} \|h\|_{\mathcal{H}} \qquad s.t. \quad h(x_i) = y_i,\ i = 1, \ldots, n$$

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

$$\begin{aligned} h^\star(x) &= \mathbb{E}\left[Y(x) \,|\, Y(x_i) = y_i,\ i = 1, \ldots, n\right] \\ &= k(x, X) k(X, X)^{-1} y \end{aligned}$$

where $Y \sim GP(0, k)$, and $\|h\|^2_{\mathcal{H}} = y^\top k(X, X)^{-1} y$.

$\rightarrow$ In this sense, GP regression is generalizing interpolation splines.

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

Gaussian processes, splines and RKHS

Link with the approximation error and conditional variance

For all $x \in \mathbb{X}$ and all $h \in \mathcal{H}$ verifying the interpolation constraints $h(X) = y$:

$$|h(x) - h^\star(x)| \leq \|h\|_{\mathcal{H}} \times k_c(x,x)^{1/2}$$

N.B., The 'x' part of the upper bound, $k_c(x,x)^{1/2}$, is called *power function*.

Gaussian processes, splines and RKHS

Correspondence between approximation spline and GP conditional mean for noisy observations

The approximation spline is defined by the regularized regression problem

$$(*) \qquad \min_{h \in \mathcal{H}} \quad \sum_{i=1}^{n} (h(x_i) - y_i)^2 + \lambda \|h\|_{\mathcal{H}}^2$$

If $\mathcal{H}$ is the RKHS of kernel k, and if $k(X,X) + \lambda I_n$ is invertible, then (*) has a unique solution in the finite dimensional space spanned by the $k(x_i, .)$:

$$\begin{aligned} h_{\text{opt}}(x) &= \mathbb{E}\left[Y(x) \,|\, Y(x_i) + \varepsilon_i = y_i,\ i = 1, \ldots, n\right] \\ &= k(x,X)(k(X,X) + \lambda I_n)^{-1} y \end{aligned}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent $\mathcal{N}(0, \lambda)$, and indep. of the GP Y.

Outline

1 Gaussian processes (GP)

2 Kernels and RKHS

3 GP regression

4 **Physics-informed GP**
- Linear PDEs
- Non-linear PDEs

Various information... same framework

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

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

G: Finite group of symmetries.

Various information... same framework

Answer : The information can be reformulated with a linear operator L

Centered	$\int Y(x)\,\mu(dx) = 0$	$L(f) = \int f(x)\,\mu(dx)$
Harmonic	$\frac{\partial^2 Y(x)}{\partial x_1^2} + \frac{\partial^2 Y(x)}{\partial x_2^2} = 0$	

Various information... same framework

Answer : The information can be reformulated with a linear operator L

Centered	$\int Y(x)\,\mu(dx) = 0$	$L(f) = \int f(x)\,\mu(dx)$
Harmonic	$\frac{\partial^2 Y(x)}{\partial x_1^2} + \frac{\partial^2 Y(x)}{\partial x_2^2} = 0$	$L(f) = \frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2}$
Symmetric	$Y(g.x) = Y(x), \forall g \in G$	

Various information... same framework

Answer : The information can be reformulated with a linear operator L

Centered	$\int Y(x)\,\mu(dx) = 0$	$L(f) = \int f(x)\,\mu(dx)$
Harmonic	$\frac{\partial^2 Y(x)}{\partial x_1^2} + \frac{\partial^2 Y(x)}{\partial x_2^2} = 0$	$L(f) = \frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2}$
Symmetric	$Y(g.x) = Y(x), \forall g \in G$	$L(f)(x) = f(x) - \frac{1}{\lvert G\rvert}\sum_{g\in G} f(g.x)$
Additive	$Y(x) = \sum_{j=1}^{d} Y^j(x_j)$	

Various information... same framework

Answer : The information can be reformulated with a linear operator L

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

Two results for GPs constrained by linear equalities (Ginsbourger et al., 2016)

- $Y : GP(m, k)$
- L: linear operator on Y such that $L(m) = 0$ (+ integrability condition)

Argumentwise property for kernels

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

$$\forall x : L(Y)(x) = 0 \quad \Leftrightarrow \quad \forall x' : L(k(x', .)) = 0$$

Proof idea: $\mathbb{C}\mathrm{ov}(L(Y)(x), Y(x')) = L(\mathbb{C}\mathrm{ov}(Y(x), Y(x'))$

Two results for GPs constrained by linear equalities (Ginsbourger et al., 2016)

- $Y : GP(m, k)$
- L: linear operator on Y such that $L(m) = 0$ (+ integrability condition)

Argumentwise property for kernels

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

$$\forall x : L(Y)(x) = 0 \quad \Leftrightarrow \quad \forall x' : L(k(x', .)) = 0$$

Proof idea: $\mathbb{C}\mathrm{ov}(L(Y)(x), Y(x')) = L(\mathbb{C}\mathrm{ov}(Y(x), Y(x'))$

Inheritance to conditional distributions

Let Y_c the GP Y conditional on $Y(x_1), \ldots, Y(x_n)$. Then

$$\forall x : L(Y)(x) = 0 \quad \Rightarrow \quad \forall x : L(Y_c)(x) = 0$$

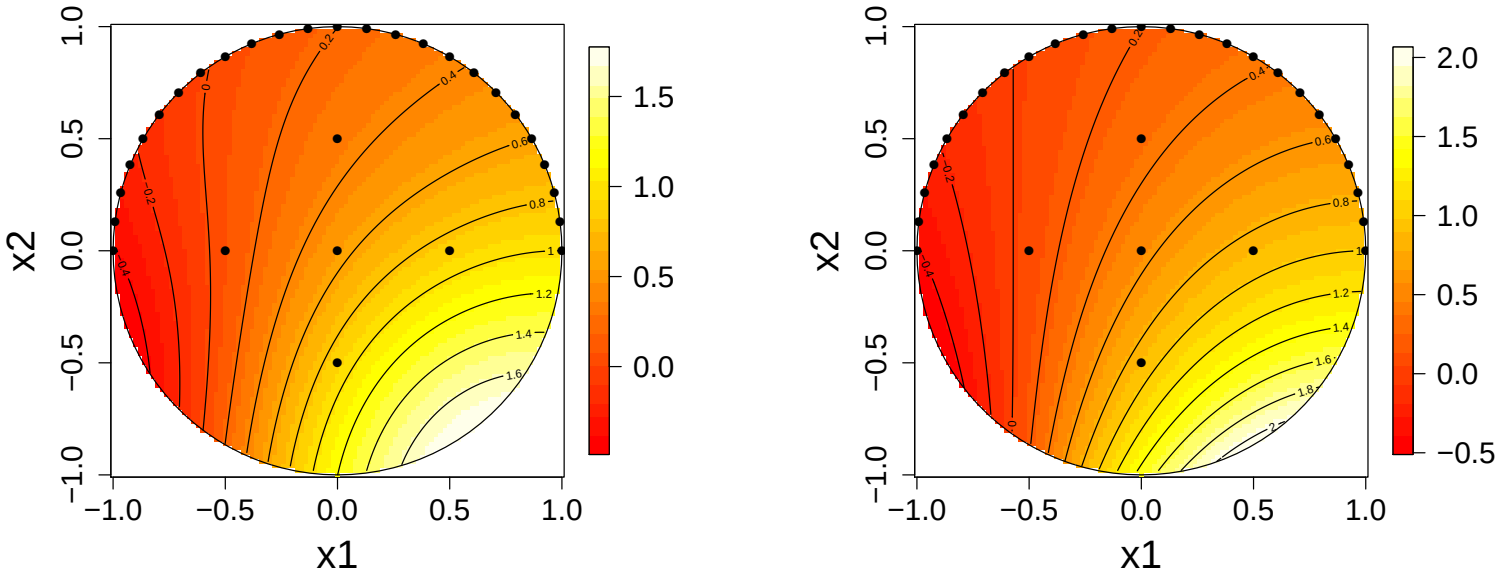
Proof idea: The cond. mean and cond. cov. are linear functions of $k(., x')$.

Illustration: Maximum of a harmonic function

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

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

Harmonic kernel $$k(x, x') = \sigma^2 \exp\left(\frac{x_1 x_1' + x_2 x_2'}{\ell^2}\right) \cos\left(\frac{x_2 x_1' - x_1 x_2'}{\ell^2}\right)$$



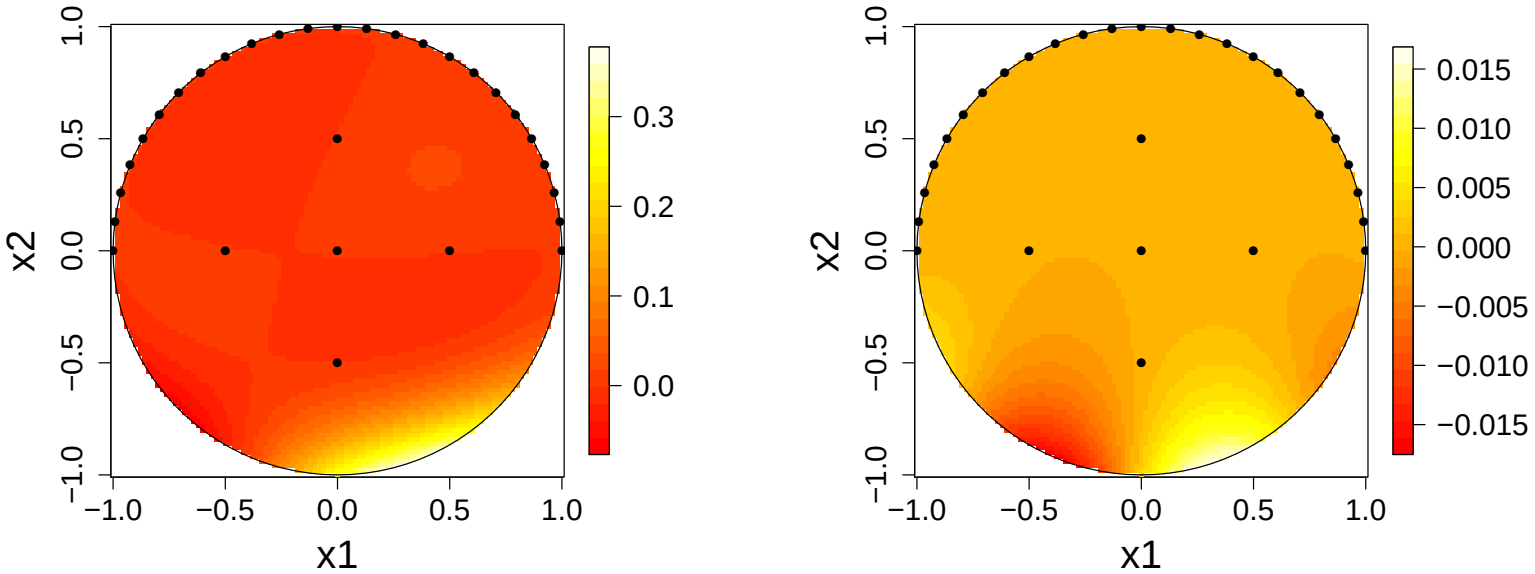
Prediction mean – Left: Gaussian kernel; Right: Harmonic kernel.

Illustration: Maximum of a harmonic function

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

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

Harmonic kernel $$k(x, x') = \sigma^2 \exp\left(\frac{x_1 x_1' + x_2 x_2'}{\ell^2}\right) \cos\left(\frac{x_2 x_1' - x_1 x_2'}{\ell^2}\right)$$



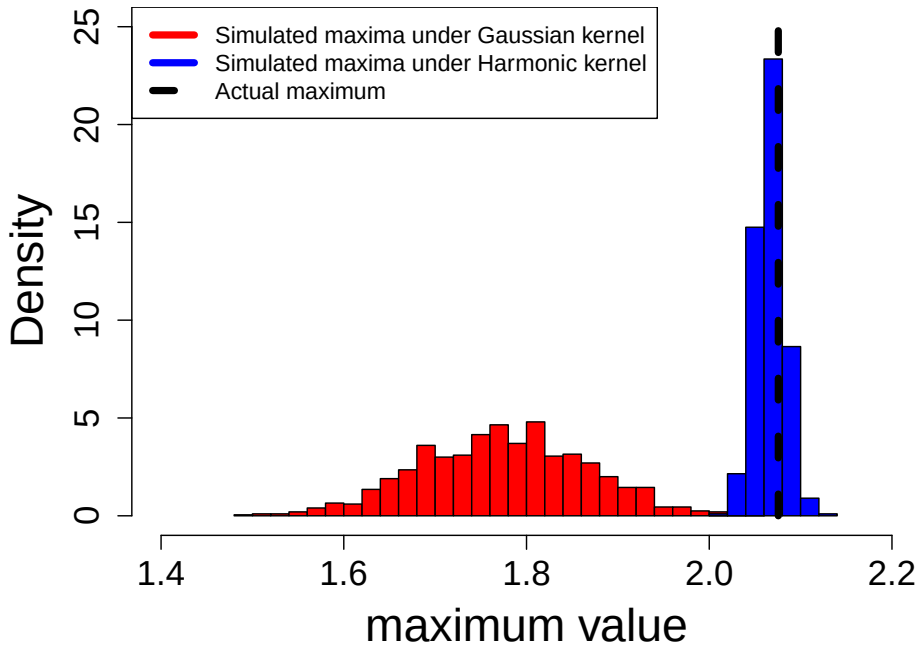
Prediction st. dev. – Left: Gaussian kernel; Right: Harmonic kernel.

Illustration: Maximum of a harmonic function

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

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

Harmonic kernel $$k(x, x') = \sigma^2 \exp\left(\frac{x_1 x_1' + x_2 x_2'}{\ell^2}\right) \cos\left(\frac{x_2 x_1' - x_1 x_2'}{\ell^2}\right)$$



Conditional simulations of the maximum under the two GPs.

An example of physics-informed GP

Exercise 2.3. Let us consider the heat equation

$$\frac{\partial u}{\partial t}(x,t) - \alpha \frac{\partial^2 u}{\partial x^2}(x,t) = 0$$

with initial condition $u(x,0) = \phi(x)$, with $\alpha > 0$. By applying the Fourier transform to the equation, the solution on $\mathbb{R}^2$ is (under suitable conditions):

$$u(x,t) = \int_{\mathbb{R}} S(x-y,t)\phi(y)dy. \tag{2}$$

where $S(x,t) = (4\pi\alpha t)^{-1/2} \exp\left(-\frac{x^2}{4\alpha t}\right)$.
We now consider that ϕ is unknown, and that $(\phi(x))_{x\in\mathbb{R}}$ is a GP$(0, k_\phi)$.

- What can you say of the mapping $L : \phi \mapsto \int_{\mathbb{R}} S(\bullet - y, \bullet)\phi(y)dy$?
- Explain why $(u(x,t))_{(x,t)\in\mathbb{R}^2}$ should be a GP on $\mathbb{R}^2$
- Compute formally its mean and its kernel in function of k_ϕ.

An example of physics-informed GP

For some k_ϕ, the kernel k_u is given explicitly.
For instance, when $k_\phi(y, y') = \exp\left(-\frac{1}{2}\frac{(y-y')^2}{\theta^2}\right)$ is the square exponential kernel, then k_u has the form :

$$k_u\left(\begin{pmatrix} x \\ t \end{pmatrix}, \begin{pmatrix} x' \\ t' \end{pmatrix}\right) = \frac{\sigma_u^2}{\sqrt{2\pi}\sqrt{\theta^2 + 2\alpha(t+t')}} \exp\left(-\frac{1}{2}\frac{(x-x')^2}{\theta^2 + 2\alpha(t+t')}\right).$$

(see https://www.mdpi.com/1099-4300/22/2/152).

More on physics-informed GP for linear PDEs

A nice reference is the PhD of Iain Henderson (and related references) containing a recent state-of-the-art, and developments suitable to general PDEs where solutions are not defined pointwise.

Solving non-linear PDEs with GPs

The idea is to use the functional view of GPs, i.e. RKHS.

The next slides are based on Chen, Hosseini, Owhadi, Stuart, 2021

Solving non-linear PDEs with GPs

As an example, we focus on the PDE

$$\begin{aligned} -\Delta u(x) + (u(x))^3 &= f(x) \qquad \forall x \in \Omega \\ u(x) &= g(x) \qquad \forall x \in \partial\Omega \end{aligned}$$

where Ω is an open set of $\mathbb{R}^d$ and $f : \Omega \to \mathbb{R}$, $g : \partial\Omega \to \mathbb{R}$ are continuous functions.

The aim is to predict $u(x)$ knowing values of f on Ω and values of g on $\partial\Omega$.

Solving non-linear PDEs with GPs

Let $\mathcal{H}$ be a RKHS with kernel k.

Let us consider the minimization problem

$$\min_{u\in\mathcal{H}} \|u\|_{\mathcal{H}} \quad s.t. \quad \begin{cases} -\Delta u(x_m) + (u(x_m))^3 = f(x_m) & m = 1,\dots,M_\Omega \\ u(x_m) = g(x_m) & m = M_\Omega+1,\dots,M \end{cases}$$

where $x_1,\dots,x_M$ are fixed points (collocation points).

As the PDE is non-linear, we cannot use the representer theorem. The trick is to recast this problem as the bilevel optimization problem:

$$\min_{z=(z_m^{(1)},z_m^{(2)})_{m=1,\dots,M}} \left(\min_{\substack{u\in\mathcal{H} \\ u(x_m)=z_m^{(1)} \\ \Delta u(x_m)=z_m^{(2)} \\ m=1,\dots,M}} \|u\|_{\mathcal{H}} \right) \quad s.t. \quad \begin{cases} -z_m^{(2)} + (z_m^{(1)})^3 = f(x_m) & m = 1,\dots,M_\Omega \\ z_m^{(1)} = g(x_m) & m = m_\Omega+1,\dots,M \end{cases}$$

Solving non-linear PDEs with GPs

Inner optimization

$$\min_{u\in\mathcal{H}} \|u\|_{\mathcal{H}} \quad s.t. \quad \begin{cases} u(x_m) = z_m^{(1)} & m = 1,\dots,M \\ \Delta u(x_m) = z_m^{(2)} & m = 1,\dots,M \end{cases}$$

where $z_M = (z_m^{(1)}, z_m^{(2)})_{m=1,\dots,M}$ is a vector of real numbers.

As Δ is a linear operator, we can use a version of the representer theorem. Let $Y \sim GP(0,k)$. The solution is unique and given by

$$u_z^\star(x) = \mathbb{E}\left(Y(x)|\{Y(x_m) = z_m^{(1)}, \Delta Y(x_m) = z_m^{(2)}, m = 1,\dots,M\}\right)$$

and $\|u_z^\star\|_{\mathcal{H}}^2 = z_M^\top K(X,X)^{-1} z_M$ where $K(X,X)$ is the covariance matrix of the vector $(Y(x_1),\dots,Y(x_M),\Delta Y(x_1),\dots,\Delta Y(x_M))$.

Exercise. Take $d = 1, M = 1$. Express $u_z^\star$ and $K(X,X)$ in function of k.

Solving non-linear PDEs with GPs

Outer minimization problem

Step 2. Minimize $\|u_z^\star\|_{\mathcal{H}}^2$ w.r.t $z \in \mathbb{R}^M$ subject to non-linear constraints inherited from the PDE

$$\min_{z\in\mathbb{R}^M} \|u_z^\star\|_{\mathcal{H}}^2 \quad s.t. \quad \begin{cases} -z_m^{(2)} + (z_m^{(1)})^3 = f(x_m) & m = 1,\dots,M_\Omega \\ z_m^{(1)} = g(x_m) & m = m_\Omega + 1,\dots,M \end{cases}$$

This is a finite-dimensional optimization problem, solvable with ad-hoc numerical routines.

The authors prove the convergence of the solution of this problem to the solution of the PDE when the number of collocation points M goes to infinity.