Consistency of stepwise uncertainty reduction strategies for Gaussian processes

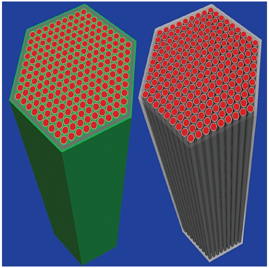
François Bachoc

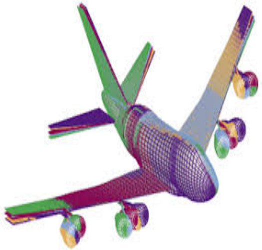
Institut de Mathématiques de Toulouse

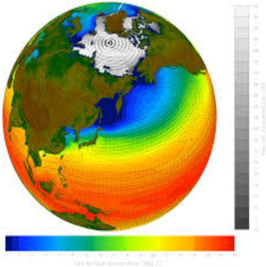
Joint work with **Julien Bect** (Centrale-Supélec) and **David Ginsbourger** (IDIAP Martigny and University of Bern)

Séminaire CERMICS - April 2017

Computer models have become essential in science and industry !







For clear reasons : cost reduction, possibility to explore hazardous or extreme scenarios...

Computer models as expensive functions

A computer model can be seen as a deterministic function

$$\begin{aligned} f : \mathbb{X} \subset \mathbb{R}^d &\to \mathbb{R} \\ x &\mapsto f(x) \end{aligned}$$

- x : tunable simulation parameter (e.g. geometry)
- $f(x)$: scalar quantity of interest (e.g. energetic efficiency)

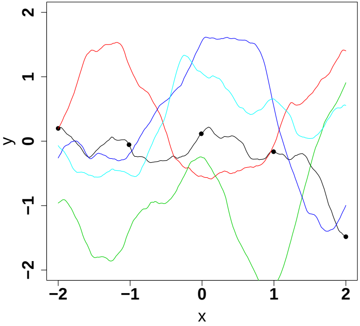
The function f is usually

- continuous (at least)
- non-linear
- only available through evaluations $x \mapsto f(x)$

$\Longrightarrow$ black box model

Gaussian processes (Kriging model)

Modeling the **black box function** as a **single realization** of a Gaussian process $\xi(x)$ on the domain $\mathbb{X} \subset \mathbb{R}^d$



Usefulness

Predicting the continuous realization function, from a finite number of **observation points**

Gaussian processes

Definition

A stochastic process $\xi : \mathbb{X} \to \mathbb{R}$ is Gaussian if for any $x_1, ..., x_n \in \mathbb{X}$, the vector $(\xi(x_1), ..., \xi(x_n))$ is a Gaussian process

Mean and covariance functions

The distribution of a Gaussian process is characterized by

- Its mean function : $x \mapsto m(x) = \mathbb{E}(\xi(x))$. Can be any function $\mathbb{X} \to \mathbb{R}$
- Its covariance function $(x_1, x_2) \mapsto k(x_1, x_2) = Cov(\xi(x_1), \xi(x_2))$

The covariance function

- The function $k : \mathbb{X}^2 \to \mathbb{R}$, defined by $k_1(x_1, x_2) = cov(\xi(x_1), \xi(x_2))$

In most classical cases :

- Stationarity : $k(x_1, x_2) = k(x_1 - x_2)$
- Continuity : $k(x)$ is continuous $' \Rightarrow'$ Gaussian process realizations are continuous
- Decrease : $k(x)$ decreases with $||x||$ and $\lim_{||x|| \to +\infty} k(x) = 0$

The covariance function

$$k : (x_1, x_2) \to k(x_1, x_2) = cov(\xi(x_1), \xi(x_2))$$

k must me symmetric non-negative definite

$$\forall n \in \mathbb{N}, \forall x_1, ..., x_n \in \mathbb{R}^d, \forall \lambda_1, ..., \lambda_n \in \mathbb{R} : \quad \sum_{i,j=1}^{n} \lambda_i \lambda_j k(x_i, x_j) \geq 0$$

$\Longrightarrow$ the covariance matrix $[k(x_i, x_j)]_{i,j=1,\ldots,n}$ must be non-negative definite
$\Longrightarrow$ Many possibilities on $\mathbb{R}^d$

Often, we require the covariance function to be positive definite :

$$\text{if } (x_1, ..., x_n) \text{ are 2-by-2 distinct and } (\lambda_1, ..., \lambda_n) \neq (0, ..., 0) : \quad \sum_{i,j=1}^{n} \lambda_i \lambda_j k(x_i, x_j) > 0$$

$\Longrightarrow$ the covariance matrix $[k(x_i, x_j)]_{i,j=1,\ldots,n}$ must be positive definite
$\Longrightarrow$ No $\xi(x)$ can be expressed as a linear combination of $\xi(x_1), ..., \xi(x_n)$ when $x_1 \neq x, ..., x_n \neq x$
$\Longrightarrow \approx$ the realizations of ξ are sufficiently complex

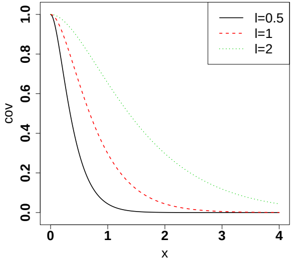
Example of the Matérn $\frac{3}{2}$ covariance function on $\mathbb{R}$

The Matérn $\frac{3}{2}$ covariance function, for a Gaussian process on $\mathbb{R}$ is parameterized by

- A variance parameter $\sigma^2 > 0$
- A correlation length parameter $\ell > 0$

It is defined as

$$k_{\sigma^2,\ell}(x_1,x_2) = \sigma^2 \left(1 + \sqrt{6}\frac{|x_1 - x_2|}{\ell}\right) e^{-\sqrt{6}\frac{|x_1 - x_2|}{\ell}}$$



Interpretation

- Stationarity, continuity, decrease
- σ^2 corresponds to the order of magnitude of the functions that are realizations of the Gaussian process
- ℓ corresponds to the speed of variation of the functions that are realizations of the Gaussian process

$\Rightarrow$ Natural generalization on $\mathbb{R}^d$

Choice of the mean and covariance functions

- In practice the mean and covariance functions are estimated from the observations $\xi(x_1), ..., \xi(x_n)$
- Typical estimation techniques are maximum likelihood (Mardia 83, Zhang 04) and cross validation (Bachoc 13)
- In the rest of the talk, we assume that the mean function m and the covariance function k are known

Conditional distribution

Gaussian process ξ observed at $x_1, ..., x_n$

Notation

- $\mathbf{y} = (\xi(x_1), ..., \xi(x_n))^t$
- $\mathbf{R}$ is the $n \times n$ matrix $[k(x_i, x_j)]$
- $\mathbf{r}(x) = (k(x, x_1), ..., k(x, x_n))^t$
- $\mathbf{m} = (m(x_1), ..., m(x_n))^t$

Conditional mean

The conditional mean is $m_n(x) := \mathbb{E}(\xi(x)|\xi(x_1), ..., \xi(x_n)) = m(x) + \mathbf{r}(x)^t \mathbf{R}^{-1}(\mathbf{y} - \mathbf{m})$.

Conditional variance

The conditional variance is
$k_n(x, x) = var(\xi(x)|\xi(x_1), ..., \xi(x_n)) = \mathbb{E}\left[(\xi(x) - m_n(x))^2\right] = k(x, x) - \mathbf{r}(x)^t \mathbf{R}^{-1} \mathbf{r}(x)$.

Conditional distribution

Conditionally to $\xi(x_1), ..., \xi(x_n)$, ξ is a Gaussian process with (conditional) mean function m_n and (conditional) covariance function $(x_1, x_2) \to k_n(x_1, x_2) = k(x_1, x_2) - \mathbf{r}(x_1)^t \mathbf{R}^{-1} \mathbf{r}(x_2)$

Illustration of conditional mean and variance

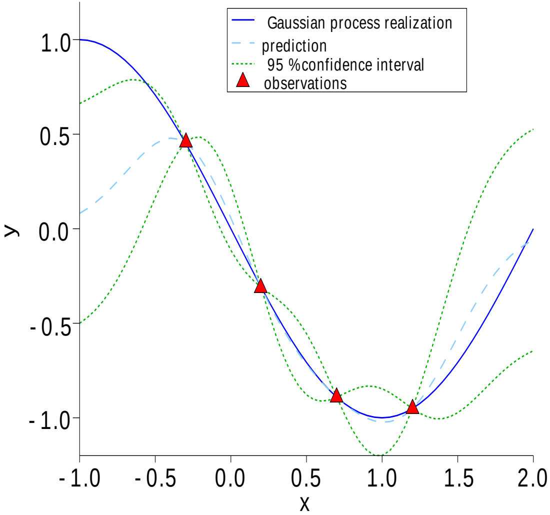
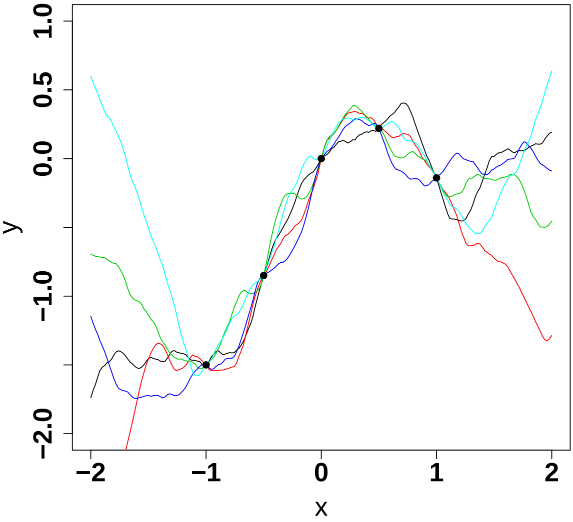


Illustration of the conditional distribution



Application to computer experiments

Gaussian process model for computer experiments

Basic idea : representing the code function $\mathbb{X} \subset \mathbb{R}^d \to \mathbb{R}$ by a realization of a Gaussian process

- Bayesian framework on a fixed function

What we obtain

- Metamodel of the code : the Gaussian process conditional mean function approximates the code function, and its evaluation cost is negligible
- Error indicator with the conditional variance
- Full conditional Gaussian process $\Rightarrow$ possible goal-oriented iterative strategies for optimization, failure domain estimation, probability estimation, code calibration...

$\Longrightarrow$ In the rest of the talk we focus on these iterative strategies

Sequential designs

We consider a Gaussian process ξ on a compact $\mathbb{X} \subset \mathbb{R}^d$ with continuous mean function m, continuous covariance function k and continuous sample paths

Motivation

- When we observe $\xi(x_1), ..., \xi(x_n)$, the mean and covariance functions become m_n and k_n
- $\Longrightarrow$ We want to choose $x_1, ..., x_n$ so that m_n and k_n become maximally informative (e.g. $k_n(x, x)$ small, or $k_n(x, x)$ small when $m_n(x)$ is large)

Sequential design

It is more efficient to select x_{i+1} after $\xi(x_1), ..., \xi(x_i)$ are observed

The observation points $x_1, ..., x_n$ become random observation points $X_1, ..., X_n$

Sequential designs : definitions

Definition

A sequence $(X_n)_{n \geq 1}$ of random points in $\mathbb{X}$ will be said to form a (non-randomized) sequential design if, for all $n \geq 1$, X_n is $\mathcal{F}_{n-1}$-measurable, where

$$\mathcal{F}_k = \sigma(X_1, \xi(x_1), ..., X_k, \xi(x_k))$$

Gaussian measures

- A Gaussian measure ν is a measure on $\mathcal{C}(\mathbb{X})$ corresponding to a Gaussian process with continuous sample paths (see e.g. Bogachev 98).
- ν is characterized by the mean function m_ν and the covariance function k_ν
- We let $\mathcal{GP}(m_\nu, k_\nu)$ denote the Gaussian measure ν

The conditioning mapping

We let $\mathrm{Cond}_{x_1,z_1,\ldots,x_n,z_n}(\nu)$ be the Gaussian measure $\mathcal{GP}(m_{\nu,n}, k_{\nu,n})$ where

$$m_{\nu,n}(x) = m_\nu(x) + \mathbf{r}(x)^t \mathbf{R}^{-1}(\mathbf{z} - \mathbf{m})$$

and

$$k_n(x_1, x_2) = k_\nu(x_1, x_2) - \mathbf{r}(x_1)^t \mathbf{R}^{-1} \mathbf{r}(x_2)$$

with

- $\mathbf{z} = (z_1, ..., z_n)^t$
- $\mathbf{R}$ is the $n \times n$ matrix $[k_\nu(x_i, x_j)]$
- $\mathbf{r}(x) = (k_\nu(x, x_1), ..., k_\nu(x, x_n))^t$
- $\mathbf{m} = (m_\nu(x_1), ..., m_\nu(x_n))^t$

A convenient result

For any sequential design of experiment (X_i), the conditional distribution of ξ (with Gaussian measure ν) given $X_1, \xi(X_1), ..., X_n, \xi(X_n)$ is $\mathrm{Cond}_{X_1,\xi(X_1),\ldots,X_n,\xi(X_n)}(\nu)$

$\Longrightarrow$ conditioning 'as if' $X_1, ..., X_n$ were deterministic

Let $\nu = \mathcal{GP}(m_\nu, k_\nu)$ be a Gaussian measure and let ξ_ν be a Gaussian process with measure ν

Uncertainty functional

It is a function $\mathcal{H} : \nu \mapsto \mathcal{H}(\nu) \in [0, \infty)$

- Expected global improvement (EGO) (Mockus 78, Jones et al 98)

$$\mathcal{H}(\nu) = \mathbb{E}(\max_{u \in \mathbb{X}} \xi_\nu(u)) - \max_{u \in \mathbb{X}; k_\nu(u,u)=0} \mathbb{E}(\xi_\nu(u))$$

- Knowledge gradient (Frazier et al 08, 09)

$$\mathcal{H}(\nu) = \mathbb{E}(\max_{u \in \mathbb{X}} \xi_\nu(u)) - \max_{u \in \mathbb{X}} \mathbb{E}(\xi_\nu(u))$$

- Integrated Bernoulli variance (Bect at al 12, Chevalier et al 14)

$$\mathcal{H}(\nu) = \int_{\mathbb{X}} p_\nu(u)(1 - p_\nu(u)) du$$

  with $p_\nu(u) = \mathbb{P}(\xi_\nu(u) \leq T)$ for fixed $T \in \mathbb{R}$

- Variance of excursion volume (Bect at al 12, Chevalier et al 14)

$$\mathcal{H}(\nu) = \mathrm{Var}\left(\int_{\mathbb{X}} \mathbf{1}_{\xi_\nu(u) \leq T} du \right)$$

Let

$$\mathcal{J}_x(\nu) = \mathbb{E}\left(\mathcal{H}(\mathrm{Cond}_{x,\xi_\nu(x)}(\nu))\right)$$

$\mathcal{J}_x(\nu)$ is the expected uncertainty after observing $\xi(x)$

Stepwise Uncertainty Reduction (SUR)

The sequential design (X_i) follows a SUR strategy when

$$X_{i+1} \in \underset{x \in \mathbb{X}}{\operatorname{argmin}}\, \mathcal{J}_x(\mathrm{Cond}_{X_1,\xi(X_1),\ldots,X_i,\xi(X_i)}(\nu_0))$$

with ν_0 the distribution of the Gaussian process ξ

For the examples

Let $\mathbb{E}_n$, Cov_n and $\mathbb{P}_n$ denote conditional mean, covariance and probability for the distribution of ξ given $\mathcal{F}_n$

- Expected global improvement

$$X_{n+1} \in \underset{x \in \mathbb{X}}{\operatorname{argmax}} \, \mathbb{E}_n \left(\left(\xi(x) - \max_{u \in \mathbb{X}; k_{n+1,x}(u,u)=0} \right)^+ \right)$$

with $k_{n+1,x}(u,v) = \text{Cov}_n(\xi(u), \xi(v) | \xi(x))$

- Knowledge gradient

$$X_{i+1} \in \underset{x \in \mathbb{X}}{\operatorname{argmax}} \, \mathbb{E} \left(\max_{u \in \mathbb{X}} \mathbb{E}_n(\xi(u) | \xi(x)) \right)$$

- Integrated Bernoulli variance

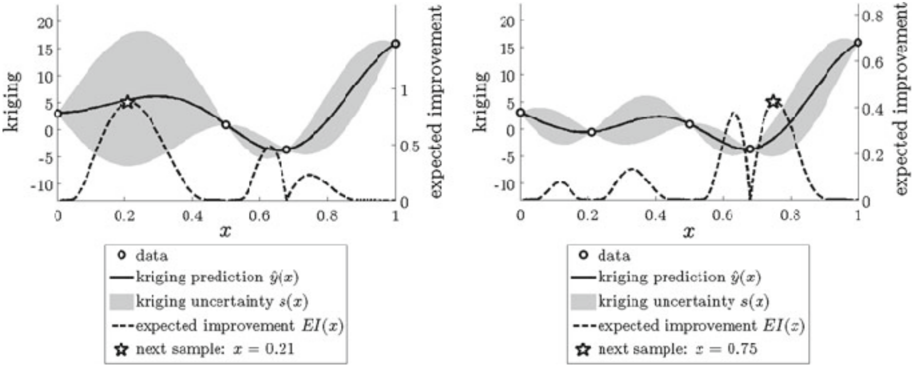
$$X_{n+1} \in \underset{x \in \mathbb{X}}{\operatorname{argmin}} \, \mathbb{E} \left(\int_{\mathbb{X}} p_{n+1,x}(u)(1 - p_{n+1,x}(u)) du \right)$$

with $p_{n+1,x}(u) = \mathbb{P}_n(\xi \leq T | \xi(x))$

- Variance of excursion volume

$$X_{n+1} \in \underset{x \in \mathbb{X}}{\operatorname{argmin}} \, \mathbb{E} \left(\text{Var}_n \left(\int_{\mathbb{X}} \mathbf{1}_{\xi(u) \leq T} du \middle| \xi(x) \right) \right)$$

(for minimization)



(Figure borrowed from Viana et al 13, Journal of Global Optimization)

Some applications

- Expected global improvement is the most used SUR strategy
  - optimal design (car industry...)
  - optimal fitting of parametric model (chemistry...)
- Integrated Bernoulli variance and Variance of excursion volume are used in failure domain estimation
  - nuclear engineering...
- Knowledge gradient can be used when Expected global improvement is used
  - drug discovery...

Goal

The rest of the talk is based on joint work with Julien Bect and David Ginsbourger

Preliminary version

Bect, Bachoc and Ginsbourger ; A supermartingale approach to Gaussian process based sequential design of experiments, Arxiv 1608.01118v1

A final version is in preparation

We want to provide general conditions ensuring that

$$\mathcal{H}\left(\mathrm{Cond}_{X_1,\xi(X_1),\ldots,X_n,\xi(X_n)}(\nu_0)\right) \to^{a.s.}_{n\to\infty} 0$$

with ν_0 the distribution of the Gaussian process ξ
$\Longrightarrow$ Uncertainty going to zero

Existing consistency results for sequential strategies

- Srinivas et al 12 provide rates of convergence for the sequential strategy GP-UCB (optimization)
- Bull 11 provide rates of convergence for expected improvement. Here the function f to optimize is deterministic and belongs to the RKHS of k
  However in general $P(\xi \in RKHS(k)) = 0 \implies$ problematic from a Bayesian point of view
- Bect et Vazquez 10 prove the consistency of Expected Global Improvement. They work with covariance functions which are not too smooth and not degenerate (we will improve this point here)

Convergence

For any sequential design of experiments (X_i), a.s. as $n \to \infty$

- The conditional mean function m_n converges to a random continuous function $m_\infty : \mathbb{X} \to \mathbb{R}$
- The conditional covariance function k_n converges to a random continuous function $k_\infty : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$
- The above convergences are uniform on $\mathbb{X}$ and $\mathbb{X} \times \mathbb{X}$

Proof : the conditional variance is decreasing + martingale arguments

Limit conditionning

Let $\mathcal{F}_\infty$ be the sigma-algebra generated by $\cup_{n \geq 1} \mathcal{F}_n$. Then conditionally to $\mathcal{F}_\infty$, ξ is a Gaussian process with mean function m_∞ and covariance function k_∞

An 'Ad Hoc' convergence for Gaussian measures

Definition

Let (ν_n) denote a sequence of Gaussian measures. We will say that (ν_n) is an *almost surely convergent sequence of conditional distributions* if

i) there exists a random Gaussian measure ν_∞ such a.s., as $n \to \infty$, m_{ν_n} and k_{ν_n} converge to m_{ν_∞} and k_{ν_∞} uniformly on $\mathbb{X}$ and $\mathbb{X} \times \mathbb{X}$;

ii) there exists a Gaussian process ξ such that, for all $n \in \mathbb{N} \cup \{+\infty\}$, $\nu_n = \mathrm{P}\big(\xi \in \cdot \mid \widetilde{\mathcal{F}}_n\big)$ for some σ-algebra $\widetilde{\mathcal{F}}_n \subset \mathcal{F}$.

Two Examples

- For any sequential design, the conditional distribution $P_n^\xi = \mathrm{P}(\xi \in .|\mathcal{F}_n)$ converges almost surely to $P_\infty^\xi = \mathrm{P}(\xi \in .|\mathcal{F}_\infty)$
- Let $x_\infty \in \mathbb{X}$ such that $k(x_\infty, x_\infty) > 0$. Let (x_k) be a sequence in $\mathbb{X}$ such that $x_k \to x_\infty$. For each $k \in \mathbb{N} \cup \{+\infty\}$, let $\nu_k = \mathrm{Cond}_{x_k,\, \xi(x_k)}\big(\mathrm{P}_0^\xi\big)$. Then (ν_k) is an almost surely convergent sequence of conditional distributions with limit ν_∞.

Supermartingale property

Definition

The functional $\mathcal{H}$ is said to have the *supermartingale property* if, for any sequential design X_1, X_2, ..., the sequence $\left(\mathcal{H}(\mathrm{P}_n^\xi)\right)$ is an $(\mathcal{F}_n)$-supermartingale.

The supermartingale propery holds for the four examples.

Expected global improvement

with $P^\xi_{n+1,\xi(x)} = \text{Cond}_{x,\xi(x)}(P^\xi_n)$

$$\begin{aligned}
\mathcal{H}(P_n^\xi) - \mathrm{E}_n[\mathcal{H}(P^\xi_{n+1,\xi(X_{n+1})})] &= \mathrm{E}_n(\max_{u\in\mathbb{X}} \xi(u)) - \mathrm{E}_n\left(\mathrm{E}_n(\max_{u\in\mathbb{X}} \xi(u)|\xi(X_{n+1}))\right) \\
&\quad - \max_{k_n(u,u)=0} \mathrm{E}_n(\xi(u)) + \mathrm{E}_n\left(\max_{k_n(u,u|\xi(X_{n+1}))=0} \mathrm{E}_n(\xi(u)|\xi(X_{n+1}))\right) \\
&\geq \mathrm{E}_n\left(\max_{k_n(u,u)=0} \mathrm{E}_n(\xi(u)|\xi(X_{n+1}))\right) - \max_{k_n(u,u)=0} \mathrm{E}_n(\xi(u)) \\
&= \max_{k_n(u,u)=0} \xi(u) - \max_{k_n(u,u)=0} \xi(u) \\
&= 0
\end{aligned}$$

from law of total variance and since $k_n(u,u|\xi(x)) = \text{Var}_n(\xi(u)|\xi(u)) \leq k_n(u,u)$

Integrated Bernoulli variance

Let $p_{n+1,x,z}(u) = \mathrm{E}_n(\mathbf{1}_{\xi(u)\leq T}|\xi(x) = z)$

$$
\begin{aligned}
\mathcal{H}(P^{\xi}_{n+1}) &= \mathbb{E}_n\left(\int_{\mathbb{X}} p_{n+1,X_{n+1},\xi(X_{n+1})}(u)(1 - p_{n+1,X_{n+1},\xi(X_{n+1})}(u))du\right) \\
&= \int_{\mathbb{X}} \mathbb{E}\left(\mathrm{var}_n(\mathbf{1}_{\xi_u\leq T}|\xi(X_{n+1}))\right) du \\
&\leq \int_{\mathbb{X}} \mathrm{var}_n(\mathbf{1}_{\xi_u\leq T})du \\
&= \mathcal{H}(P^{\xi}_n)
\end{aligned}
$$

The convergence result

Let

$$\mathcal{G}(\nu) = \sup_{x \in \mathbb{X}} \left(\mathcal{H}(\nu) - \mathrm{E}(\mathcal{H}(\mathrm{Cond}_{x,\xi_{\nu(x)}}(\nu))) \right)$$

(maximum expected uncertainty reduction)

Theorem

Let $\mathcal{H}$ denote an uncertainty functional with the supermartingale property.
Let (X_n) denote a SUR sequential design for $\mathcal{H}$

$$X_{n+1} \in \operatorname*{argmin}_{x \in \mathbb{X}} \mathrm{E}(\mathcal{H}(\mathrm{Cond}_{x,\xi(x)}(P_n^{\xi})))$$

Then $\mathcal{G}(\mathrm{P}_n^{\xi}) \to 0$ almost surely. If, moreover,

i) $\mathcal{H}(\mathrm{P}_n^{\xi}) \to \mathcal{H}(\mathrm{P}_\infty^{\xi})$ almost surely,

ii) $\mathcal{G}(\mathrm{P}_n^{\xi}) \to \mathcal{G}(\mathrm{P}_\infty^{\xi})$ almost surely ;

iii) $\mathcal{G}(\nu) = 0 \Longrightarrow \mathcal{H}(\nu) = 0$;

then $\mathcal{H}(\mathrm{P}_n^{\xi}) \to 0$ almost surely.

Assumptions i) and ii) are continuity assumptions
Assumption iii) is essential, it means

no possible uncertainty reduction with one more observation $\Longrightarrow$ no uncertainty

Summary of additional results

- We prove that the general results apply to the four examples
- We introduce the notion of regular loss function, where $\mathcal{H}$ is an average loss when estimating a quantity of interest (e.g. maximum of ξ, $\{u \in \mathbb{X} : \xi(u) \leq T\}$,...).
- We provide a specific convergence result for regular loss function, with easier to check assumptions

Conclusion

Summary

- Gaussian process provide a Bayesian framework on deterministic function (e.g. computer models)
- The probabilistic framework enables to define expected uncertainties and Stepwise Uncertainty Reduction (SUR) strategies
- We prove convergence of SUR strategies
- Remark : Our proof does not rely on showing that (X_i) is almost surely dense in $\mathbb{X}$. We allow for degenerate or very smooth covariance functions. Sometimes we do not need $\sup_{u \in \mathbb{X}} k_n(u, u) \to 0$

Two open questions

- When the covariance function is estimated (frequentist or Bayesian)
- Rate of convergence

Thank you for your attention !