

### 3. Some tools for the analysis of sequential strategies based on a Gaussian process prior

# Function approximation with a Gaussian prior

- ▶ Aim: to reconstruct  $f : \mathbb{X} \subseteq \mathbb{R}^d \rightarrow \mathbb{R}$  from a finite set of evaluation results

$$\{(x_1, f(x_1)), \dots, (x_n, f(x_n))\}$$

using a **prior on  $f$** , under the form of a probability measure  $P$  on  $(\Omega, \mathcal{U})$ , with  $\Omega = \mathbb{R}^{\mathbb{X}}$  and  $\mathcal{U}$  the cylindrical  $\sigma$ -algebra on  $\Omega$ .

- ▶  $P$  can be seen as the distribution of the **canonical process  $\xi : \Omega \times \mathbb{X} \rightarrow \mathbb{R}$**  defined as

$$\xi(\omega, \cdot) := \omega, \quad \text{for all } \omega \in \Omega = \mathbb{R}^{\mathbb{X}}$$

→ turns the problem of approximation of  $f$  into a prediction problem for the process  $\xi$

- ▶ Hereafter, we shall assume that  $\mathbb{X} = \mathbb{R}^d$  and that  $\xi$  is a **Gaussian, zero-mean random process with known covariance function  $k : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$**

- ▶ The best predictor of  $\xi(x)$  from observations  $\xi(x_i)$ ,  $i = 1, \dots, n$ , also called the **kriging predictor**, is obtained as the orthogonal projection

$$\hat{\xi}(x; \underline{x}_n) := \sum_{i=1}^n \lambda_i(x; \underline{x}_n) \xi(x_i)$$

of  $\xi(x)$  onto the linear space  $\text{span}\{\xi(x_i), i = 1, \dots, n\}$ .

- ▶ Remark: kriging is a kernel method  
Regression in an RKHS
  - ▶ 1960: splines, (**Schoenberg 1964, Duchon 1976–1979**)
  - ▶ 1980: RBF, (**Micchelli 1986, Powel 1987**)
  - ▶ 1995: SVM, (**Vapnik 1995**)
  - ▶ 1997: SVR, (**Smola 1997**)

#### Prediction

- ▶ 1960: kriging and BLUP (**Matheron, Parzen, Sacks, Ylvisaker**)
  - ▶ 1970: intrinsic kriging (**Matheron, Kimeldorf, Wahba**)
  - ▶ 1997: Gaussian process regression, (**Williams 1997, Neal 1997**)
- ▶ To analyze sequential strategies based on a Gaussian prior, it is useful to understand when and in what sense  $\hat{\xi}(x; \underline{x}_n)$  converges to  $\xi(x)$

## Pointwise consistency in $L^2$ -norm

- ▶ If  $k$  is continuous, the variance of the prediction error, also called **kriging variance** or **power function**

$$\sigma^2(x; \underline{x}_n) := \mathbb{E}[(\xi(x) - \hat{\xi}(x; \underline{x}_n))^2] = k(x, x) - \sum_i \lambda_i(x; \underline{x}_n) k(x, x_i)$$

typically decreases with the **fill-in distance**

$$h_n = \sup_{x \in \mathbb{X}} \min_{1 \leq i \leq n} |x - x_i|$$

- ▶ Assume a continuous covariance of the form  $k(x, y) = \gamma(x - y) \in L^2(\mathbb{R}^d)$ , where the Fourier transform of  $\gamma(h)$  satisfies

$$0 < c_1(1 + \|u\|_2^2)^{-\tau} \leq \tilde{\gamma}(u) \leq c_2(1 + \|u\|_2^2)^{-\tau},$$

with  $\tau > d/2$ , then [see, e.g., Wu et Schaback 93]

$$\sup_{x \in \mathbb{X}} \sigma(x; \underline{x}_n) \leq Ch_n^{\tau - d/2}$$

- ▶ Consistency for sample paths?

## Reminder: the RKHS attached to a Gaussian process

- ▶ Let  $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  be a s.p.d. function
- ▶ There exists a zero-mean Gaussian process  $\xi$  with covariance function  $k$ , s.t.

$$\text{Cov}[\xi(x), \xi(y)] = \mathbb{E}[\xi(x)\xi(y)] = (\xi(x), \xi(y))_{L^2(\Omega, \mathcal{A}, \mathbb{P})} = k(x, y)$$

- ▶ Let  $\mathcal{H}_k$  be the closure of the linear space

$$\tilde{\mathcal{H}}_k = \left\{ X : X \in L^2(\Omega, \mathcal{A}, \mathbb{P}), X = \sum_{i=1}^n \lambda_i \xi(x_i), n \in \mathbb{N}, \lambda_i \in \mathbb{R}, x_i \in \mathbb{R}^d \right\}$$

→  $\mathcal{H}_k$  is the Gaussian Hilbert space spanned by  $\xi$

- ▶ Consider the linear transformation

$$\begin{aligned} \rho : \mathcal{H}_k &\rightarrow \mathbb{R}^{\mathbb{X}} \\ X &\mapsto (\rho X)(\cdot) = (\xi(\cdot), X)_{L^2(\Omega, \mathcal{A}, \mathbb{P})} \end{aligned}$$

- ▶ The linear space  $\mathcal{F}_k = \text{Im } \rho$ , endowed with the inner product

$$(f, g)_{\mathcal{F}_k} = (\rho^{-1}f, \rho^{-1}g)_{L^2(\Omega, \mathcal{A}, \mathbb{P})},$$

is an RKHS.  $\mathcal{F}_k$  is called the **RKHS attached to  $\xi$** . The kernel of  $\mathcal{F}_k$  is  $k$ . Indeed,

i.  $\rho(\xi(x))(\cdot) = (\xi(\cdot), \xi(x))_{L^2(\Omega, \mathcal{A}, \mathbb{P})} = k(\cdot, x) \in \mathcal{F}_k$

ii. for  $f = \rho X$ , with  $X \in \mathcal{H}_k$ ,  $f(x) = (\xi(x), X)_{L^2(\Omega, \mathcal{A}, \mathbb{P})} = (k(x, \cdot), f)_{\mathcal{F}_k}$

- ▶  $\rho$  is a one-to-one isometry from  $\mathcal{H}_k$  onto  $\mathcal{F}_k$  (the **Loève's isometry**)

## Reminder: the RKHS attached to a Gaussian process

- ▶ We say that  $k' \geq k$ , if  $\mathcal{F}_{k'} \supseteq \mathcal{F}_k$
- ▶ Then there exists a linear operator  $L : \mathcal{F}_{k'} \rightarrow \mathcal{F}_{k'}$ , whose range is contained in  $\mathcal{F}_k$ , and such that

$$(f, g)_{\mathcal{F}_{k'}} = (Lf, g)_{\mathcal{F}_k}, \quad \forall f \in \mathcal{F}_{k'} \text{ and } g \in \mathcal{F}_k$$

( $L$  is bounded, symmetric, positive.)

If  $L$  is a nuclear operator (that is,  $\text{tr } L = \sum (Le_n, e_n) < \infty$ ), we write  $k' \gg k$

- ▶ Let  $\xi$  be a zero-mean Gaussian process.  
If  $k' \gg k$ , then there exists a version  $\xi'$  of  $\xi$  such that  $P\{\xi' \in \mathcal{F}_{k'}\} = 1$ .  
If  $k' \not\gg k$ , then  $P\{\xi \in \mathcal{F}_{k'}\} = 0$   
(Hajek 1962, Driscoll 1973, Lukic and Beder 2001)
- ▶ Thus,  $P\{\xi \in \mathcal{F}_k\} = 0$

## Consistency for sample paths

- ▶ Let  $\mathcal{F}_k$  be the RKHS attached to  $\xi$ , and  $\mathcal{F}_k^*$  its dual space
- ▶ Let  $\delta_x \in \mathcal{F}_k^*$  be the evaluation functional at  $x \in \mathbb{R}^d$ , and define

$$\lambda_n^x = \sum_{i=1}^n \lambda_i(x; \underline{x}_n) \delta_{x_i} \in \mathcal{F}_k^*$$

- ▶ Note that  $\xi(x) = \langle \delta_x, \xi(\cdot) \rangle$  and  $\widehat{\xi}(x; \underline{x}_n) = \langle \lambda_n^x, \xi(\cdot) \rangle$
- ▶ Then,

$$\begin{aligned} \|\delta_x - \lambda_n^x\|_{\mathcal{F}_k^*}^2 &= \left\| k(x, \cdot) - \sum_i \lambda_i(x; \underline{x}_n) k(x_i, \cdot) \right\|_{\mathcal{F}_k}^2 \\ &= k(x, x) - \sum_i \lambda_i(x; \underline{x}_n) k(x_i, x) \\ &= \sigma^2(x; \underline{x}_n) \end{aligned}$$

Thus,  $\lambda_n^x \rightarrow \delta_x$  strongly in  $\mathcal{F}_k^*$  iff  $\sigma^2(x; \underline{x}_n) \rightarrow 0$

- ▶ Moreover, since strong convergence in  $\mathcal{F}_k^*$  implies weak convergence in  $\mathcal{F}_k^*$ , we have

$$\lim_{n \rightarrow \infty} \sigma^2(x; \underline{x}_n) = 0 \implies \forall f \in \mathcal{F}_k, \quad \lim_{n \rightarrow \infty} \langle \lambda_n^x, f \rangle = f(x).$$

## Consistency for sample paths

- ▶ However, this result is not satisfying from a Bayesian point of view because  $P(\xi \in \mathcal{F}_k) = 0$ .
- ▶ In fact, we know that for all sequences  $(x_n)_{n \geq 1}$  in  $\mathbb{R}^d$  and all  $x \in \mathbb{R}^d$ , there is a set of functions  $\mathcal{G} \supset \mathcal{F}_k$  such that
  - ▶  $P\{\xi \in \mathcal{G}\} = 1$
  - ▶  $\sigma^2(x; \underline{x}_n) \rightarrow 0 \implies \forall f \in \mathcal{G}, \langle \lambda_n^x, f \rangle \rightarrow f(x)$

(Hence, considering the kriging predictor is relevant from a Bayesian point of view.)

Indeed,  $\widehat{\xi}(x; \underline{x}_n)$  is a martingale sequence, and  $\sup_n E[\widehat{\xi}(x; \underline{x}_n)^2] \uparrow K < \infty$ . Thus,  $(\widehat{\xi}(x; \underline{x}_n))_n$  converges a.s. and in  $L^2$ -norm to a random variable  $\xi_\infty$  (note that  $\xi_\infty = \xi(x)$  if  $\lim_{n \rightarrow \infty} \sigma^2(x; \underline{x}_n) = 0$ ).

- ▶ An important question in the context of optimization is to know whether the sets  $\mathcal{G}$  contain the set  $C(\mathbb{R}^d)$  of all continuous functions.

# Consistency for continuous sample paths

- ▶ A kernel defined on a compact metric space  $\mathbb{X}$  such that the corresponding RKHS is dense in  $C(\mathbb{X})$  for the topology of the uniform convergence has been called a **universal kernel** [I. Steinwart, 2001]
- ▶  $k$  is universal on  $\mathbb{R}^d$  if it is universal on each compact subset of  $\mathbb{R}^d$
- ▶ We have the following result:

Let  $k$  be a universal kernel on  $\mathbb{R}^d$ . Assume that  $\{x_n, n \geq 1\}$  is bounded, and let  $\mathbb{X}_0$  be its (compact) closure in  $\mathbb{R}^d$ .

Then, for all  $x \in \mathbb{X}_0$ , the following assertions are equivalent:

- $\forall f \in C(\mathbb{R}^d), \lim_{n \rightarrow \infty} \langle \lambda_n^x, f \rangle = f(x),$
- the **Lebesgue constant**  $\|\lambda_n^x\|_{TV} = \sum_{i=1}^n |\lambda_i(x; \underline{x}_n)|$  at  $x$  is bounded.

## What can be said about $\|\lambda_n^x\|_{\text{TV}}$ ?

- ▶ Assume a continuous kernel of the form  $k(x, y) = \gamma(x - y) \in L^2(\mathbb{R}^d)$ , where the Fourier transform of  $\gamma(h)$  satisfies

$$0 < c_1(1 + \|u\|_2^2)^{-\tau} \leq \tilde{\gamma}(u) \leq c_2(1 + \|u\|_2^2)^{-\tau},$$

with  $\tau > d/2$ .

- then, it is well-known that  $\mathcal{F}_k$  is the Sobolev space

$$W_2^\tau(\mathbb{R}^d) = \{f \in L^2(\mathbb{R}^d) : \tilde{f}(\cdot)(1 + \|\cdot\|_2^2)^{\tau/2} \in L^2(\mathbb{R}^d)\}$$

- $\mathcal{F}_k$  contains  $C_c^\infty(\mathbb{R}^d)$  and therefore,  $k$  is universal

- ▶ Let  $\mathbb{X} \subset \mathbb{R}^d$  be a compact subset (with an interior cone condition) and define

- ▶ a fill-in distance  $h_n = \sup_{x \in \mathbb{X}} \min_{1 \leq i \leq n} |x - x_i|$

- ▶ a separation distance  $q_n = \frac{1}{2} \min_{i \neq j} |x_i - x_j| \leq h_n$

- ▶ then [Marchi and Schaback, 2008]

$$\|\lambda_n^x\|_{\text{TV}} \leq C\sqrt{n} \left( \frac{h_n}{q_n} \right)^{\tau - d/2}$$

- pointwise consistency for continuous sample path is unknown at present time

## 4. Convergence results

# Convergence results

## Definition

A random process  $\xi$  has the **no-empty-ball** (NEB) property if, for all sequences  $(x_n)_{n \geq 1}$  in  $\mathbb{R}^d$  and all  $x \in \mathbb{R}^d$ , the following assertions are equivalent:

- i)  $x$  is an adherent point of the set  $\{x_n, n \geq 1\}$ ,
- ii)  $\sigma^2(x; \underline{x}_n) \rightarrow 0$  when  $n \rightarrow +\infty$ .

- ▶ If  $\xi$  has the Gaussian covariance written as

$$k(x, y) = s^2 e^{-\alpha \|x-y\|^2}, \quad s > 0, \alpha > 0,$$

then  $\xi$  does not possess the NEB property.

- ▶ If  $\xi$  is second-order stationary and has spectral density  $S$ , with the property that  $S^{-1}$  has at most polynomial growth, then  $\xi$  has the NEB property.  
→ allows consideration of a large class of covariance functions, which includes the class of (non-Gaussian) exponential covariances

$$k(x, y) = s^2 e^{-\alpha \|x-y\|^\beta}, \quad s > 0, \alpha > 0, 0 < \beta < 2,$$

and the class of Matérn covariances popularized by M.L. Stein (1999)

## Proposition

Assume that  $\xi$  is a centered Gaussian process with the NEB property and consider the expected improvement algorithm

$$\begin{cases} X_1 & = x_{\text{init}}, \\ X_{n+1} & = \underset{X_{n+1}}{\operatorname{argmax}} \mathbf{E}_n [ (\widehat{m}_n(\xi) - \xi(X_{n+1}))_+ ], \quad n \geq 1 \end{cases}$$

Then, for all  $x_{\text{init}} \in \mathbb{X}$ , the sequence  $(X_n)_{n \geq 1}$  is P-almost surely dense in the compact search domain  $\mathbb{X}$

- ▶ Convergence rates for Bayesian sequential algorithms are unknown at present time
- ▶ Under some mild assumption, and for  $n$  high enough, we have for the optimization problem:

$$\mathbf{E} [ \widehat{m}_n(\xi) - m(\xi) ] \leq Ch_n^{\kappa/2} \sqrt{\log n}$$

[see, e.g., Grünwalder et al. 2010]

## Concluding remarks

- ▶ (Short and incomplete) overview of the domain of computer experiments
- ▶ The theory of Bayesian sequential algorithms is interesting in the context of expensive-to-evaluate functions, and effective in practical situations where dimension is moderately high
- ▶ Many applications can be found in the literature
- ▶ A great number of methodological and theoretical questions are open