**Rencontres GDR MascotNum**

décembre 2008 – IHP

# Autour des méthodes de régression dans les RKHS

**J. Bect[1], B. Iooss[2], L. Pronzato[3], E. Vazquez[1]**

1. SUPELEC, Gif-sur-Yvette

2. CEA, Cadarache

3. CNRS I3S, Sophia Antipolis

# Régression dans les RKHS

**Trois sujets** présentés :

1. Théorie de la pratique du krigeage
2. Planification d'expériences pour codes numériques
3. Optimisation globale bayésienne

Encadrement possible par

- Julien Bect et Emmanuel Vazquez : enseignants - chercheurs à SUPELEC
- Bertrand Iooss : CEA
- Luc Pronzato : DR au CNRS

# Contexte et applications visées

- Conception de produit/système via simulations numériques, avec pour objectifs
  - d'optimiser les performances d'un produit (problème d'optimisation globale)
  - d'estimer la probabilité de défaillance d'un produit
  - d'optimiser les performances en garantissant une probabilité de défaillance donnée (optimisation sous contrainte)

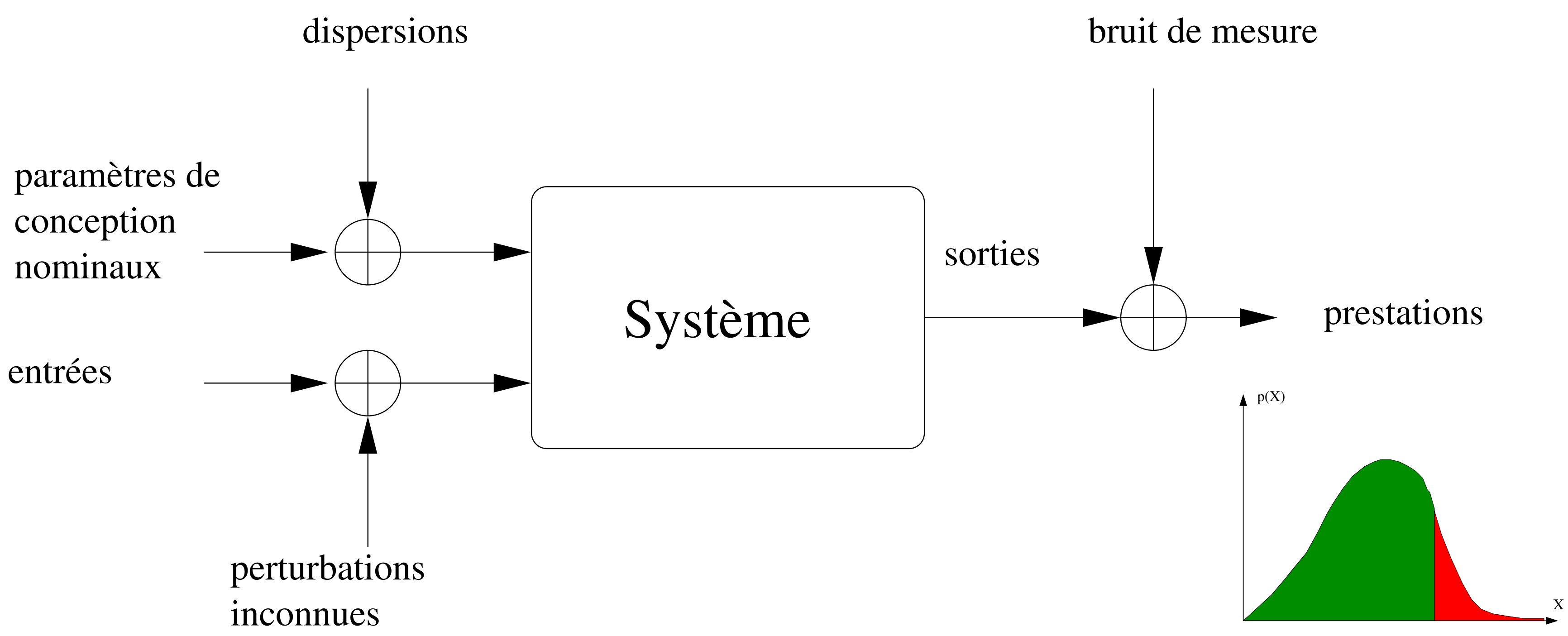
# Exemple

Conception d'un conduit d'admission

(thèse CIFRE Renault de J. Villemonteix)

- optimisation de la forme du conduit pour réduire les émissions polluantes
- prendre en compte les dispersions de fabrication
- 1 simulation = plusieurs heures

# Formalisation



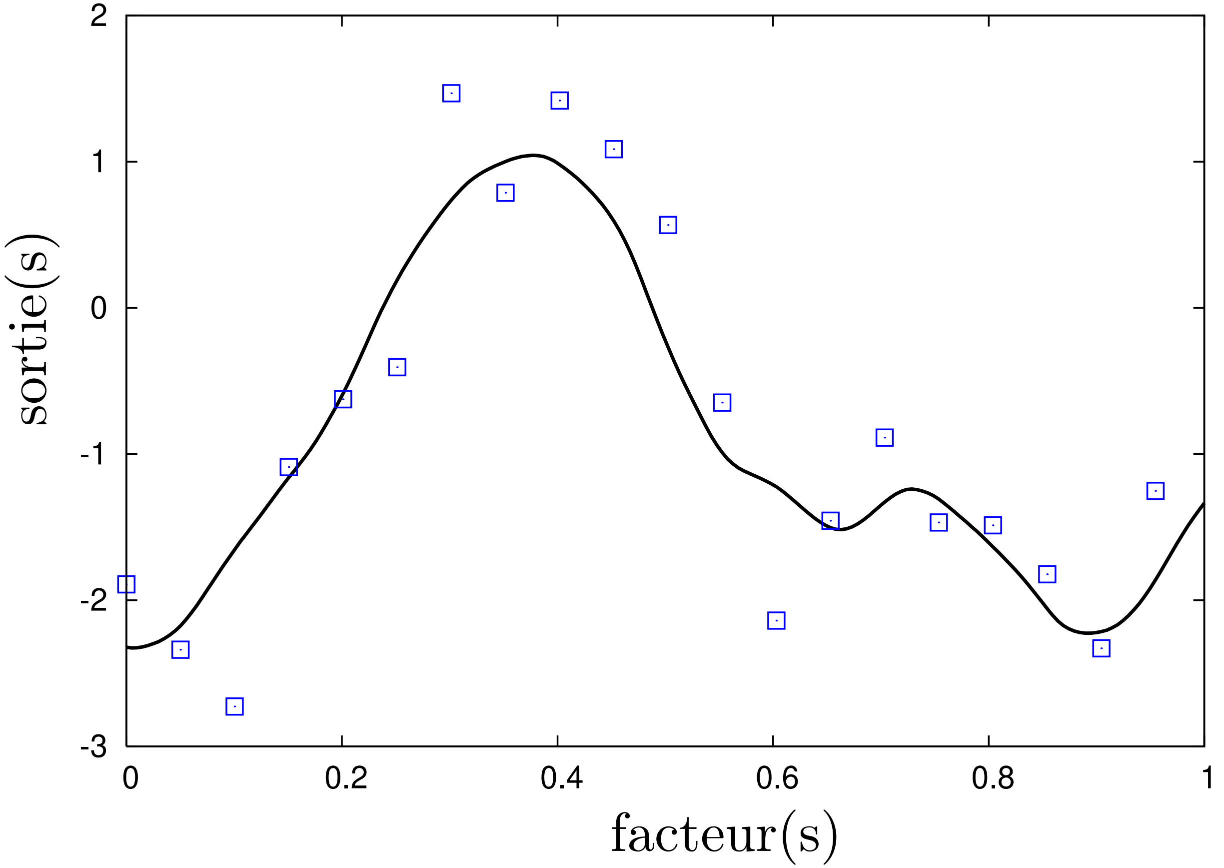
- La simulation numérique d'un système est vue comme une fonction

$$f : \mathbb{X} \subseteq \mathbb{R}^d \longrightarrow \mathbb{R}^q\,,$$

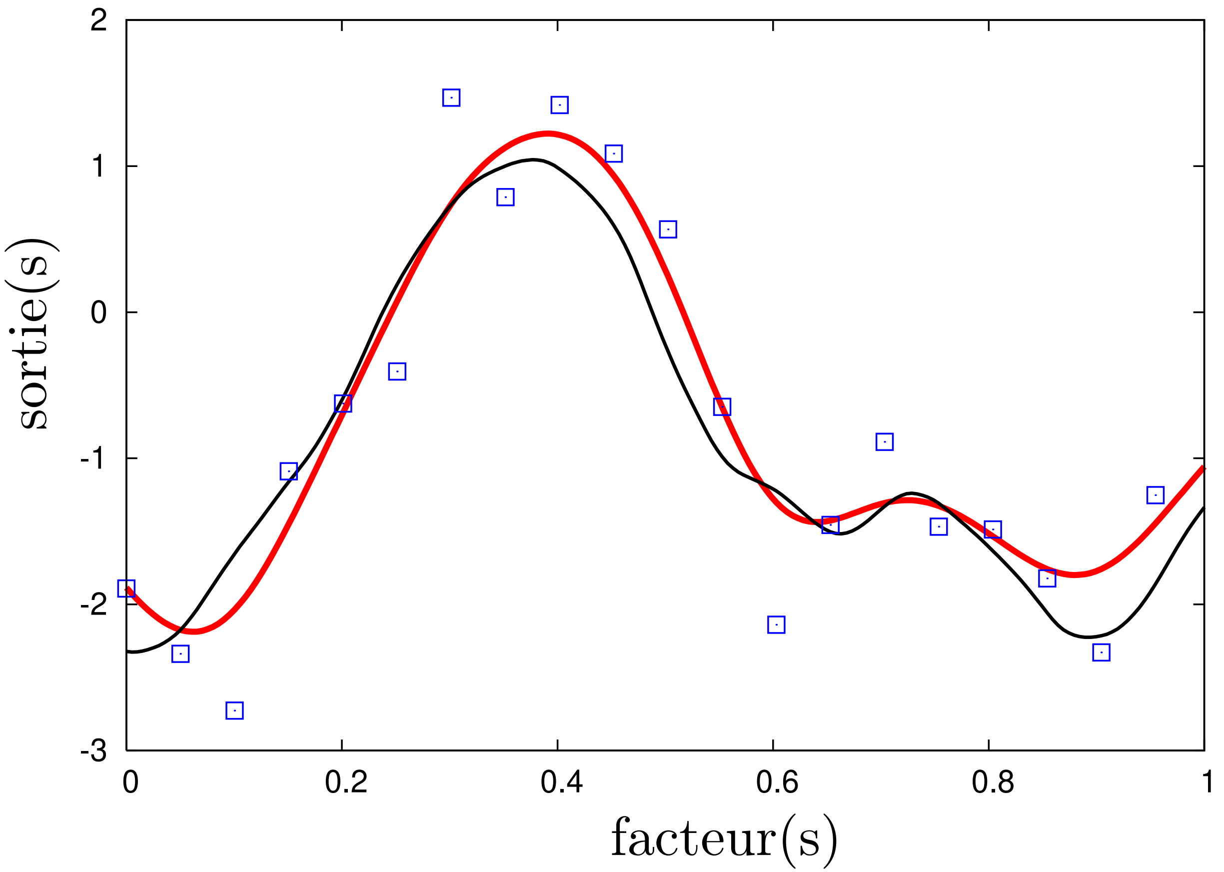
  avec $d$ et $q$ des entiers petits (disons $1 \leq d \leq 50$ et $1 \leq q \leq 4$)
- Difficulté spécifique liée à la nature des simulations numériques : une évaluation de $f$ souvent coûteuse (p. ex. plusieurs minutes/heures) $\longrightarrow$ budget d'évaluations très réduit en pratique
- Objectif : établir une relation facile à calculer entre les sorties du système et plusieurs facteurs (entrées) à partir de simulations déjà effectuées

  ➜ construire une approximation $\widehat{f}$ de $f$
- Utiliser $\widehat{f}$ au lieu de $f$ pour concevoir le système

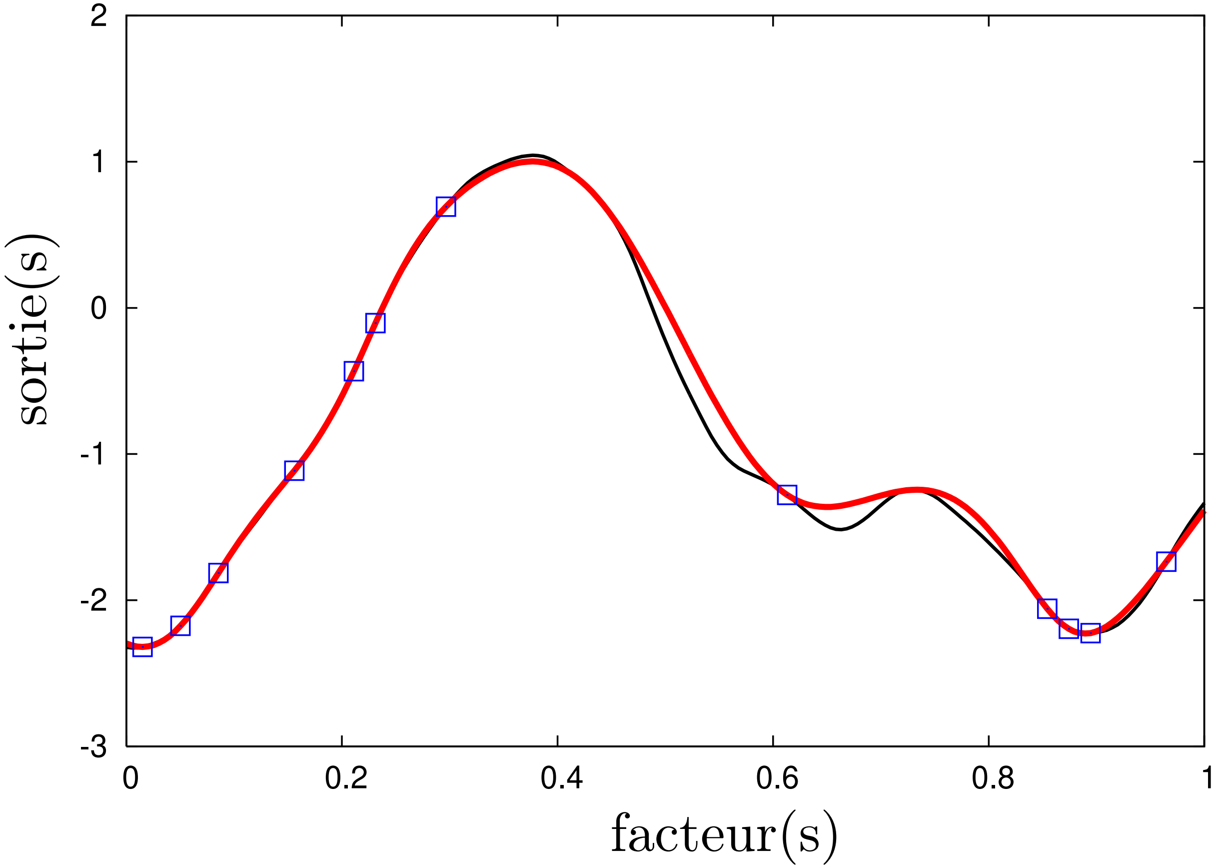
# Données



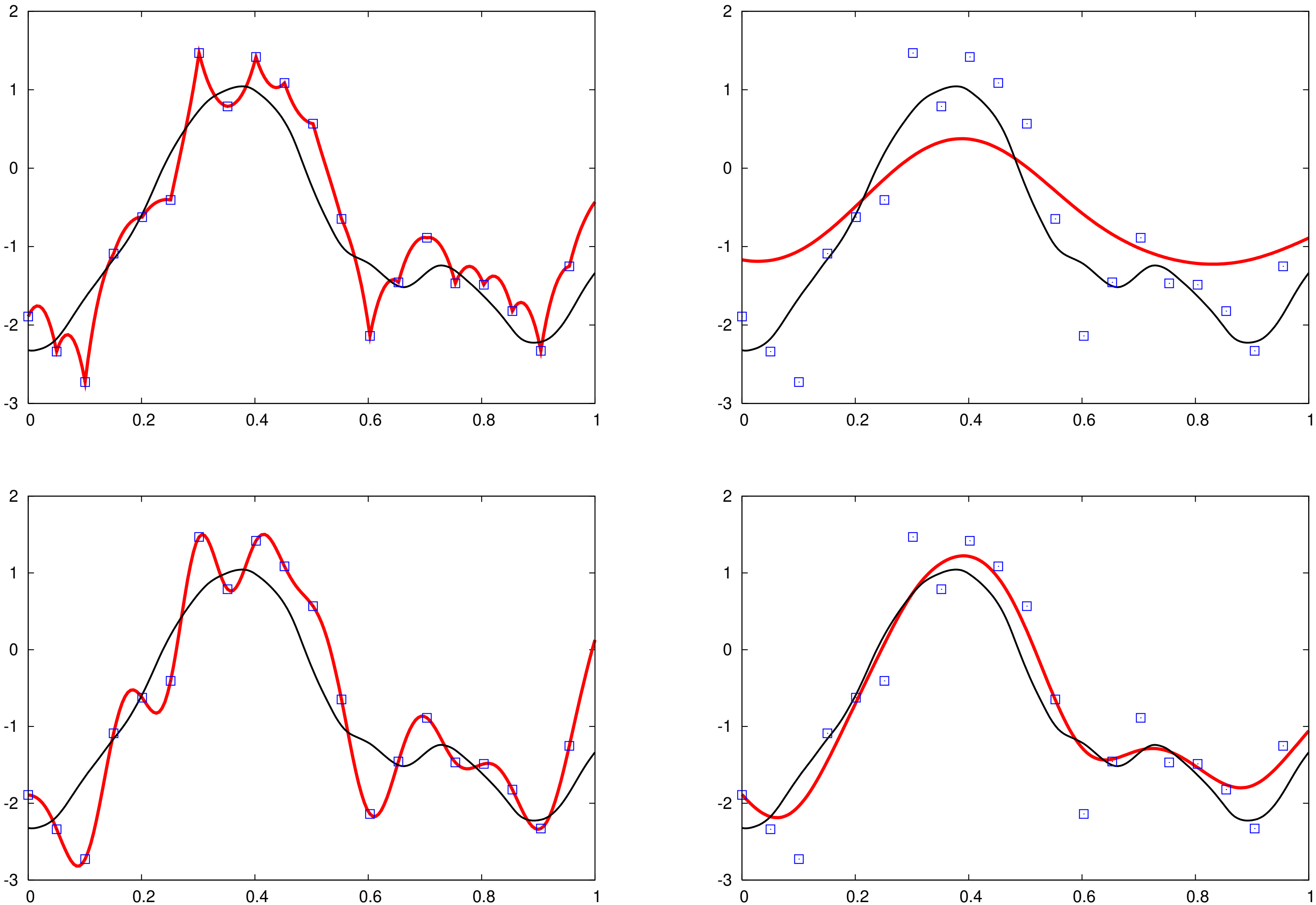
# Approximation



# Interpolation



Problème : choisir une bonne approximation


2
1
0
-1
-2
-3
0
0.2
0.4
0.6
0.8
1

Essentiel :

- ◦ choix de l'espace d'approximation
- ◦ contrôle de la régularité

➠ classe des méthodes à noyaux, *i.e.* approximation par des fonctions dans des espaces hilbertiens à noyaux reproduisant (RKHS)

➠ approche probabiliste, krigeage

# Espace hilbertien à noyau reproduisant

- $\mathcal{F}$ espace hilbertien de fonctions $\mathbb{X} \subseteq \mathbb{R}^d \to \mathbb{R}$, avec produit scalaire noté $(.,.)_{\mathcal{F}}$
- $k(x,y) : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ noyau reproduisant de $\mathcal{F}$, si pour tout $f \in \mathcal{F}$, tout $x \in \mathbb{X}$,

$$f(x) = (k(x,.), f)_{\mathcal{F}}\,. \tag{1}$$

➠ $\mathcal{F}$ appelé espace hilbertien à noyau reproduisant ou RKHS (**Aronszajn, 1950**)

Régression régularisée dans un RKHS

Approximation (**Tikhonov et Arsenin 1977**)

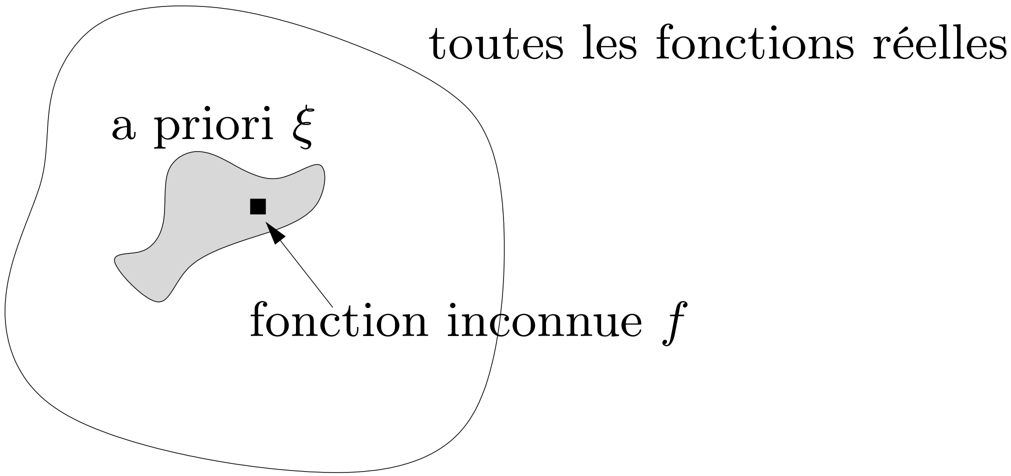
$$\text{minimiser} \quad \underbrace{\|f\|^2_{\mathcal{F}}}_{\text{régularité}} + C \underbrace{\sum_i l(f(x_i) - f^{\text{obs}}_{x_i})}_{\text{adéquation aux données}}$$

Interpolation

$$\begin{aligned} &\text{minimiser} && \|f\|^2_{\mathcal{F}}\,, \quad f \in \mathcal{F} \\ &\text{sous contraintes} && f(x_i) = f^{\text{obs}}_{x_i}\,, \quad \forall i = 1, \ldots, n\,. \end{aligned}$$

## Approche probabiliste de l'approximation

- Point de vue bayésien $\leftrightarrow$ choix d'un a priori sous la forme d'une mesure de probabilité sur $\mathbb{R}^{\mathbb{X}}$, ce qui revient à choisir la loi d'un processus aléatoire $\xi : \mathbb{X} \to \mathbb{R}$.



Si $\xi$ est un processus gaussien,

- il est très facile d'obtenir les **expressions analytiques**
  - de la moyenne a posteriori, ou krigeage : $\widehat{\xi}(x; \underline{x}_n) := \mathrm{E}\left[\xi(x) \mid \underline{\xi}_n\right]$, avec $\underline{\xi}_n = \{\xi(x_1), \ldots, \xi(x_n)\}$
  - la variance de l'erreur de prédiction, ou variance de krigeage : $\widehat{\sigma}_n^2(x\,;\,\underline{x}_n) := \mathrm{var}\left[\widehat{\xi}(x; \underline{x}_n) - \xi(x)\right]$
- il est possible de **simuler des trajectoires** de $\xi$ conditionnellement aux évaluations précédentes
  - permet par exemple l'estimation de la densité a posteriori du maximum $M = \max_x \xi(x)$
  - ...
- en pratique, on suppose que $\xi$ **est un processus gaussien**...

# Équivalence des méthodes

➜ Méthodes à noyaux

❍ 1960 : splines, (**Schoenberg 1964, Duchon 1976–1979**)

❍ 1980 : RBF, (**Micchelli 1986, Powel 1987**)

❍ 1995 : SVM, (**Vapnik 1995**)

❍ 1997 : SVR, (**Smola 1997**)

❍ 1999 : SVR semi-paramétrique (**Smola 1999**)

➜ Krigeage – prédiction linéaire

❍ 1950 : prédiction pour la recherche minière (**Krige 1951**)

❍ 1960 : krigeage, géostatistique (**Matheron 1963**) – École des Mines

❍ 1970 : krigeage intrinsèque (**Matheron 1971**)

❍ 1997 : prédiction par « processus gaussiens », (**Williams 1997, Neal 1997**)

# Sujet 1. Théorie de la pratique du krigeage

❑ Étapes de la modélisation d'un système :

- ➠ Choix des facteurs, inclusion d'informations a priori ...
- ➠ Choix d'un noyau paramétré
- ➠ Estimation des paramètres du noyau
- ➠ Régression

Il reste encore beaucoup de travail théorique sur chacune des étapes !

**Exemples** :

◦ Choix des facteurs : approximation en dimension élevée ?
  – analyse de sensibilité
  – réduction de la dimension : sélection de variables, projections...

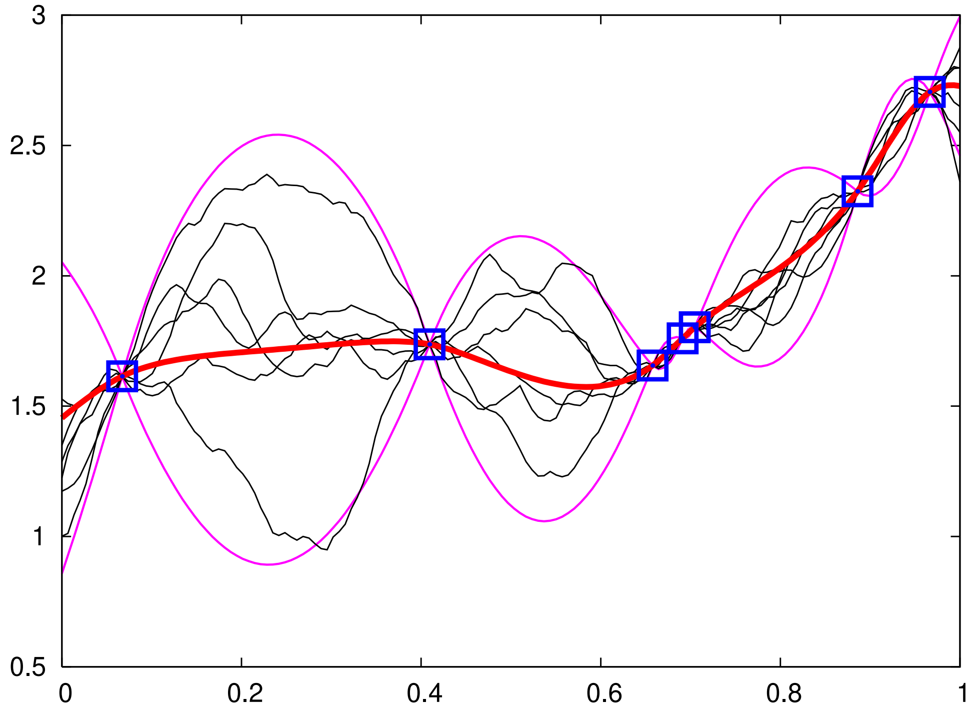
◦ Choix d'un noyau ↔ usuellement, structures classiques (exponentielle, Matérn) + estimation de paramètres
  – structures anisotropes, non stationnaires
  – technique d'estimation des paramètres ?
  – krigeage bayésien ?

◦ Régression : gérer efficacement les grands nombres de données

**Objectif** : intégrer une réflexion théorique sur ces questions dans un logiciel libre proposé à la communauté scientifique et industrielle

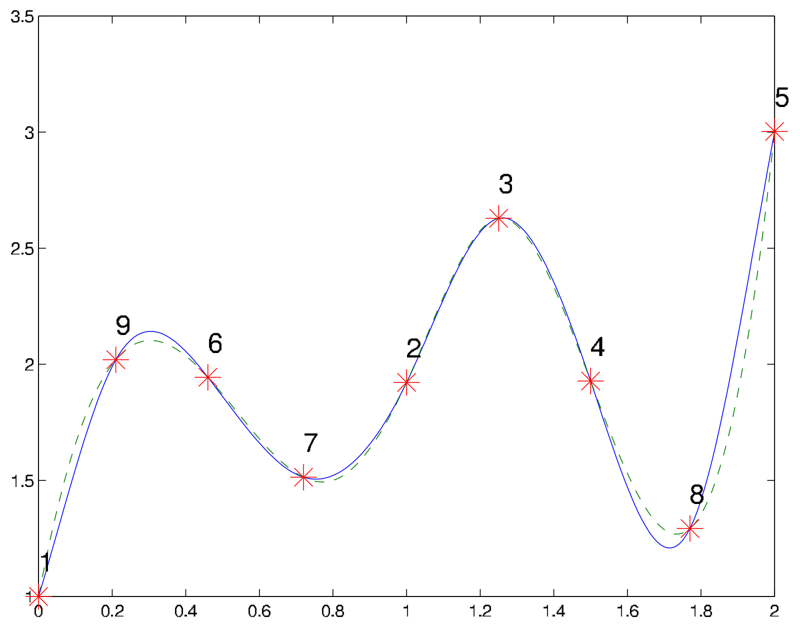
# Sujet 2. Planification d'expériences

Question : quelles simulations effectuer pour rendre précis le modèle peu coûteux construit à partir des résultats de simulation ?



**Ex :** $X_1$ à $X_5$ fixés puis $X_{n+1}$ au point de variance de krigeage maximale

- - - $f(x)$, — $\eta_n(x)$



# II) Un objectif peut en cacher un autre ...

- Modèle paramétrique : prédire $\Leftrightarrow$ estimer $\theta$
- Modèle non paramétrique (krigeage) : on a supposé $\theta$ (vecteur des paramètres de la covariance) connu ...

Estimation de $\beta, \sigma^2$ et $\theta$ par maximum de vraisemblance (processus gaussien) :
$\hat{\beta}^n(\theta) = (\mathbf{1}^\top \mathbf{C}_n^{-1}(\theta)\mathbf{y}_n)/(\mathbf{1}^\top \mathbf{C}_n^{-1}(\theta)\mathbf{1})$
$\hat{\sigma}_n^2(\theta) = \frac{1}{n}\,[\mathbf{y}_n - \hat{\beta}^n(\theta)\mathbf{1}]^\top \mathbf{C}_n^{-1}(\theta)[\mathbf{y}_n - \hat{\beta}^n(\theta)\mathbf{1}]$
et $\hat{\theta}^n_{MV} = \arg\min_\theta \det[\hat{\sigma}_n^2(\theta)\mathbf{C}_n(\theta)]$
$\longrightarrow$ le plan d'expériences a une influence sur la précision de l'estimation de $\theta$

Posons $\alpha = (\sigma^2, \theta)$, $\nu = (\beta, \alpha) = (\beta, \sigma^2, \theta)$, $\mathbf{C}_{n,\alpha} = \sigma^2 \mathbf{C}_n(\theta)$

• Si $\mathbb{E}\{Z(x,\omega)Z(u,\omega)\} \searrow 0$ pour $\|x-u\| \to \infty$ et

$\boxed{\text{les } X_i \text{ sont dans un domaine de taille croissante}}$

$\hat{\nu}^n_{MV}$ est consistant et $\mathbf{M}_n^{1/2}(\nu)(\nu^n_{MV} - \nu) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathbf{I})$ avec

$$\mathbf{M}_n(\nu) = \mathbb{E}\{[\partial l_n(\nu)/\partial \nu]\,[\partial l_n(\nu)/\partial \nu]^\top\} = \begin{pmatrix} m_n(\beta) & \mathbf{0}^\top \\ \mathbf{0} & \tilde{\mathbf{M}}_n(\alpha) \end{pmatrix}$$

$$m_n(\beta) = \mathbf{1}^\top \mathbf{C}_{n,\alpha}^{-1} \mathbf{1}\,,\ [\tilde{\mathbf{M}}_n(\alpha)]_{i,j} = \frac{1}{2}\mathrm{trace}\left[\mathbf{C}_{n,\alpha}^{-1} \frac{\partial \mathbf{C}_{n,\alpha}}{\partial \theta_i} \mathbf{C}_{n,\alpha}^{-1} \frac{\partial \mathbf{C}_{n,\alpha}}{\partial \theta_j}\right]$$

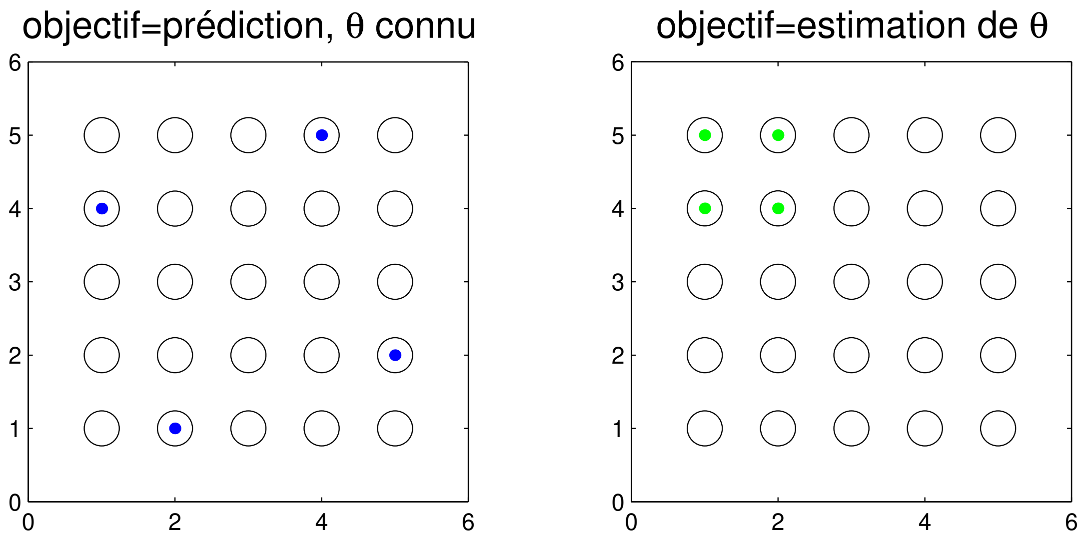
avec $l_n(\nu)$ la log-vraisemblance

Si on s'intéresse à la précision sur $\alpha$

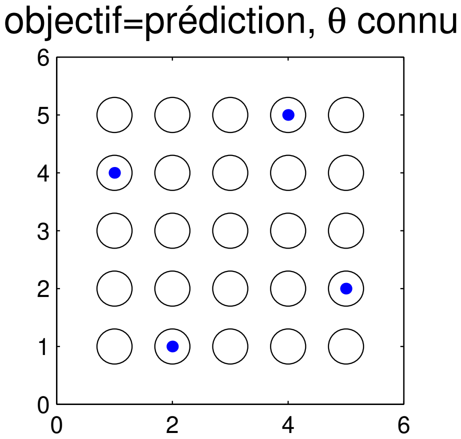
$\to$ maximiser $\det \tilde{\mathbf{M}}_n(\alpha)$

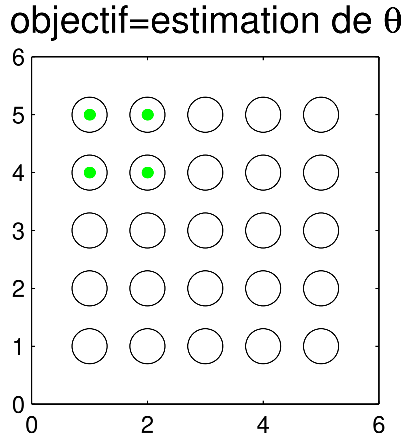
**Ex** [Zimmerman, 2006] : $\mathbb{E}\{Z(x,\omega)Z(u,\omega)\} = \sigma^2\theta^{\|x-u\|}$, $\theta = 0.3$ ($\sigma^2$ connu), $\mathcal{X}$ grille régulière $5 \times 5$

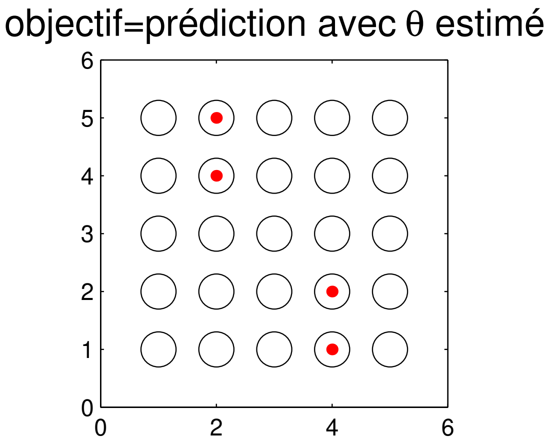
- $X_1, \ldots, X_4$ minimisent $\max_{x\in\mathcal{X}} \rho_4(x)$ ($\rightarrow$ prédiction)
- $X_1, \ldots, X_4$ maximisent $\det \tilde{\mathbf{M}}_4(\theta)$ ($\rightarrow$ estimation de $\theta$)



Suite de l'exemple de [Zimmerman, 2006] : $X_1, \ldots, X_4$ minimisent $\max_{x \in \mathcal{X}} \rho_n(x) + \mathbf{trace}\left[\tilde{\mathbf{M}}_n^{-1}(\theta) \frac{\partial \mathbf{u}^\top(x)}{\partial \theta} \mathbf{C}_{n,\theta} \frac{\partial \mathbf{u}(x)}{\partial \theta^\top}\right]$

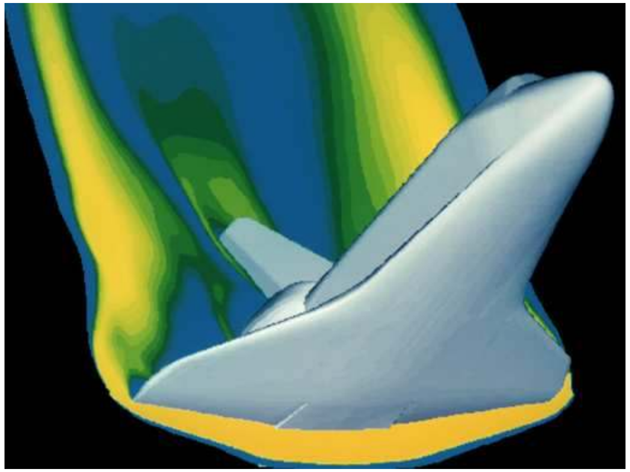


# Sujet 3. Optimisation bayésienne

Optimisation d'un système avec simulations numériques

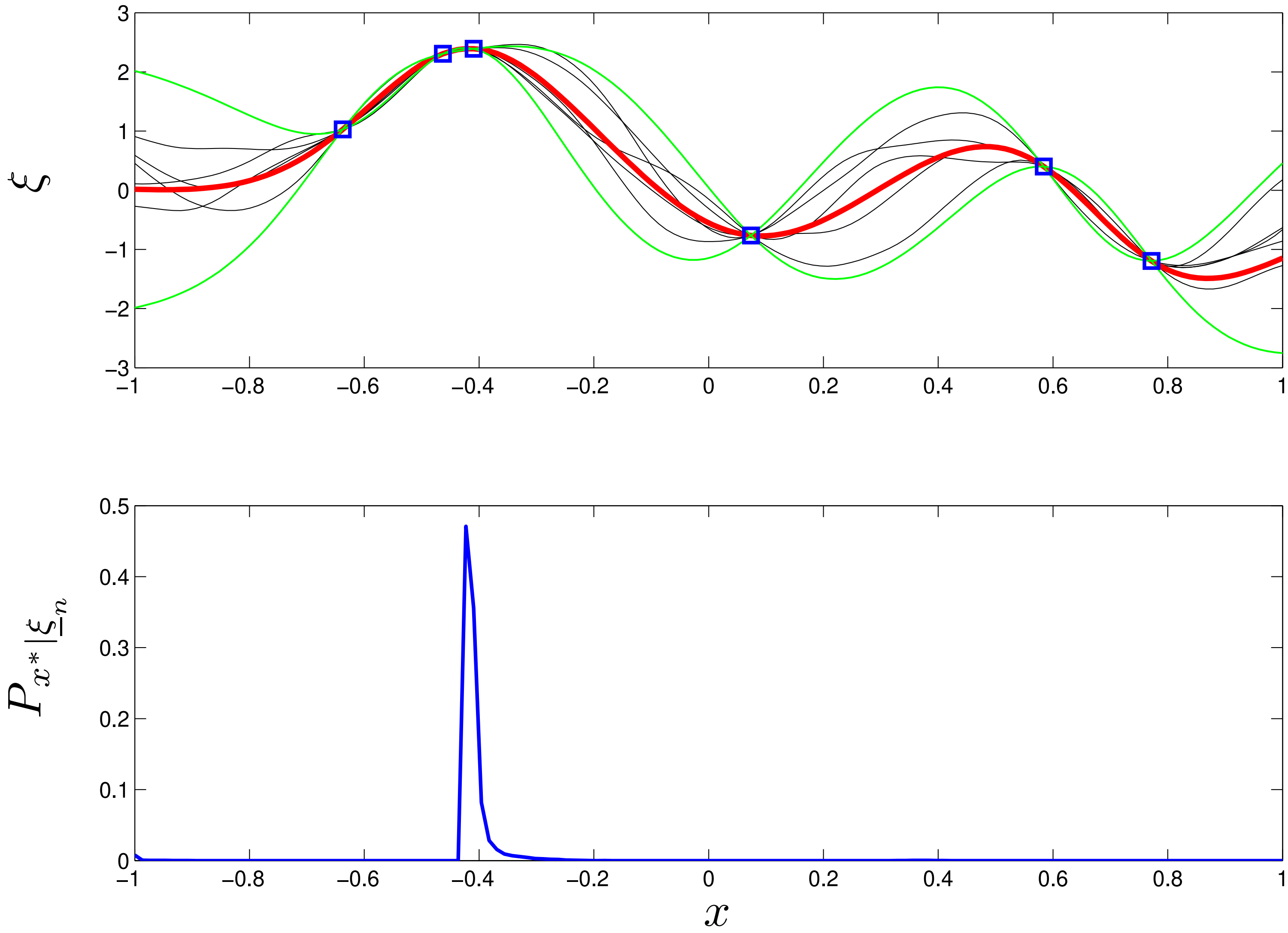


- Soit $f : \mathbb{X} \to \mathbb{R}$, avec $\mathbb{X} \subset \mathbb{R}^d$ borné, une fonction à maximiser
- Soit $x^*$ un maximiseur global :

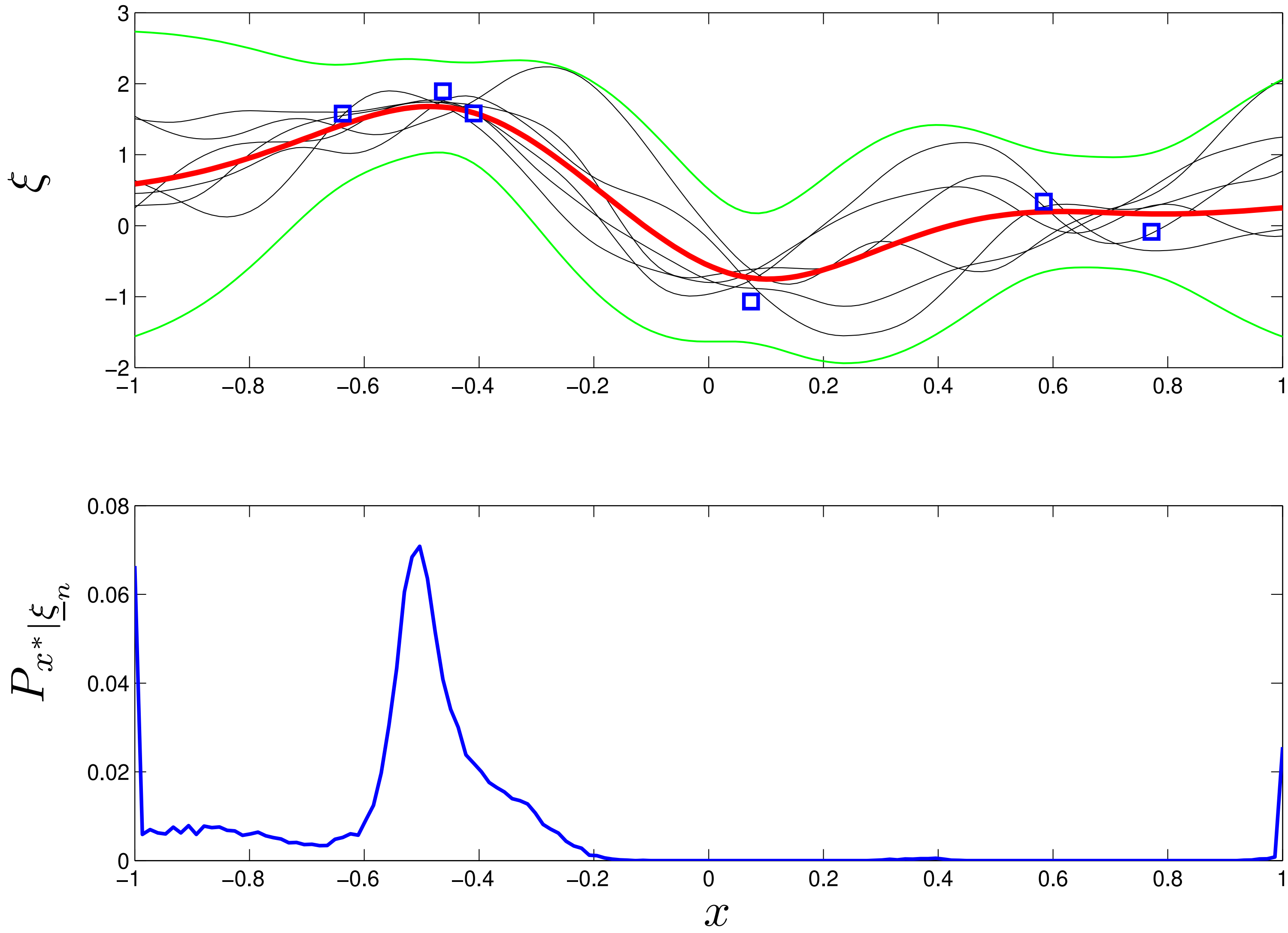
$$x^* = \arg\max_{x \in \mathbb{X}} f(x)$$

- Intérêt du modèle probabiliste : distribution a posteriori du maximiseur $x^*$ ou du maximum

## Exemple : distribution a posteriori de $x^*$ dans le cas d'évaluations sans bruit



## ... dans le cas d'évaluations bruitées



Comment construire une suite $(x_n)_{n\in\mathbb{N}}$ de manière à ce que $x^*$ soit une valeur d'adhérence ?

# L'algorithme EI

- Expected improvement $\leftrightarrow$ excursion moyenne de $\xi$ au dessus du maximum des évaluations précédentes :

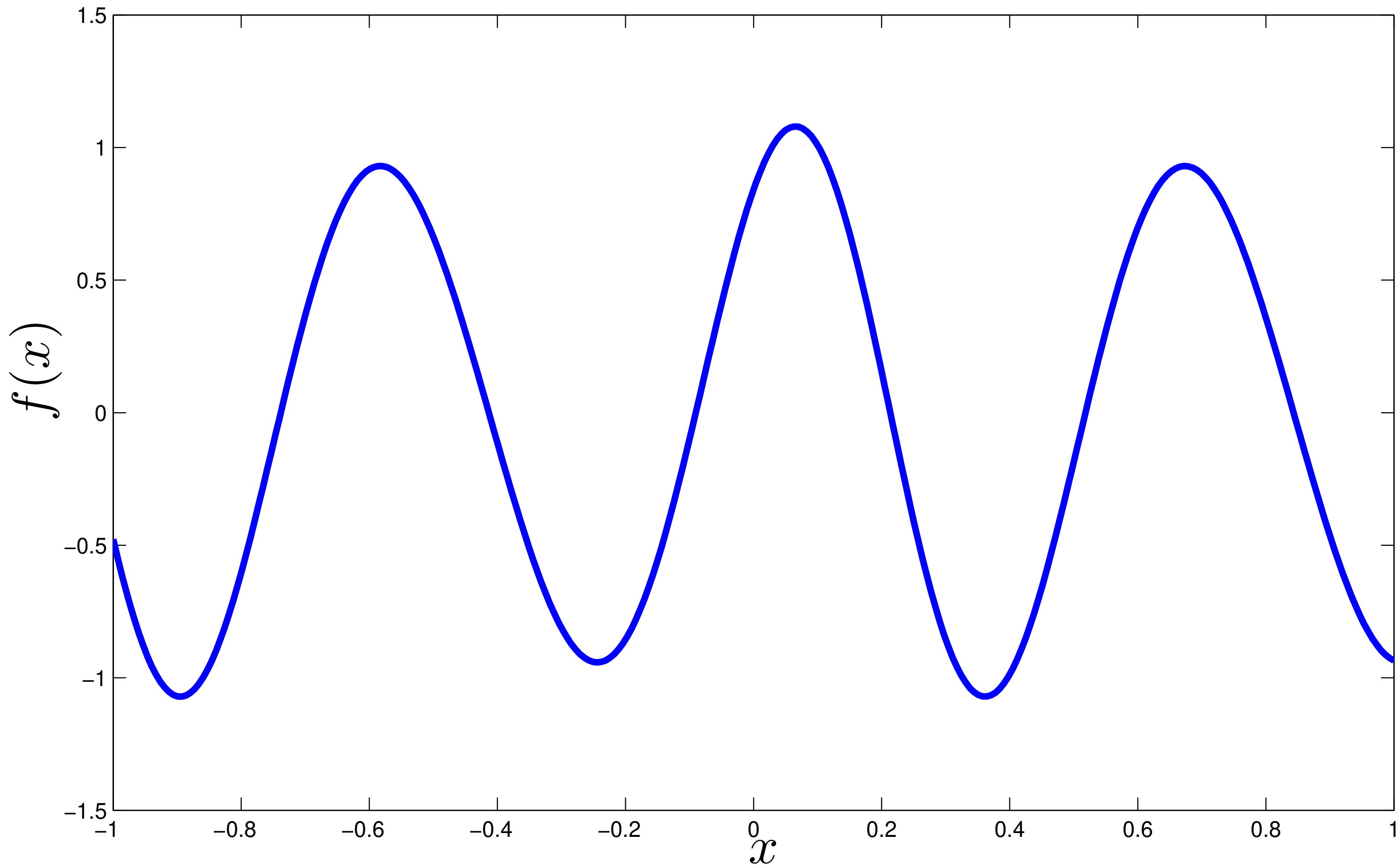
$$\rho_n(x) := \mathrm{E}\left[\ \max(0, \xi(x) - M_n)\ \middle|\ \underline{\xi}_n\ \right]$$

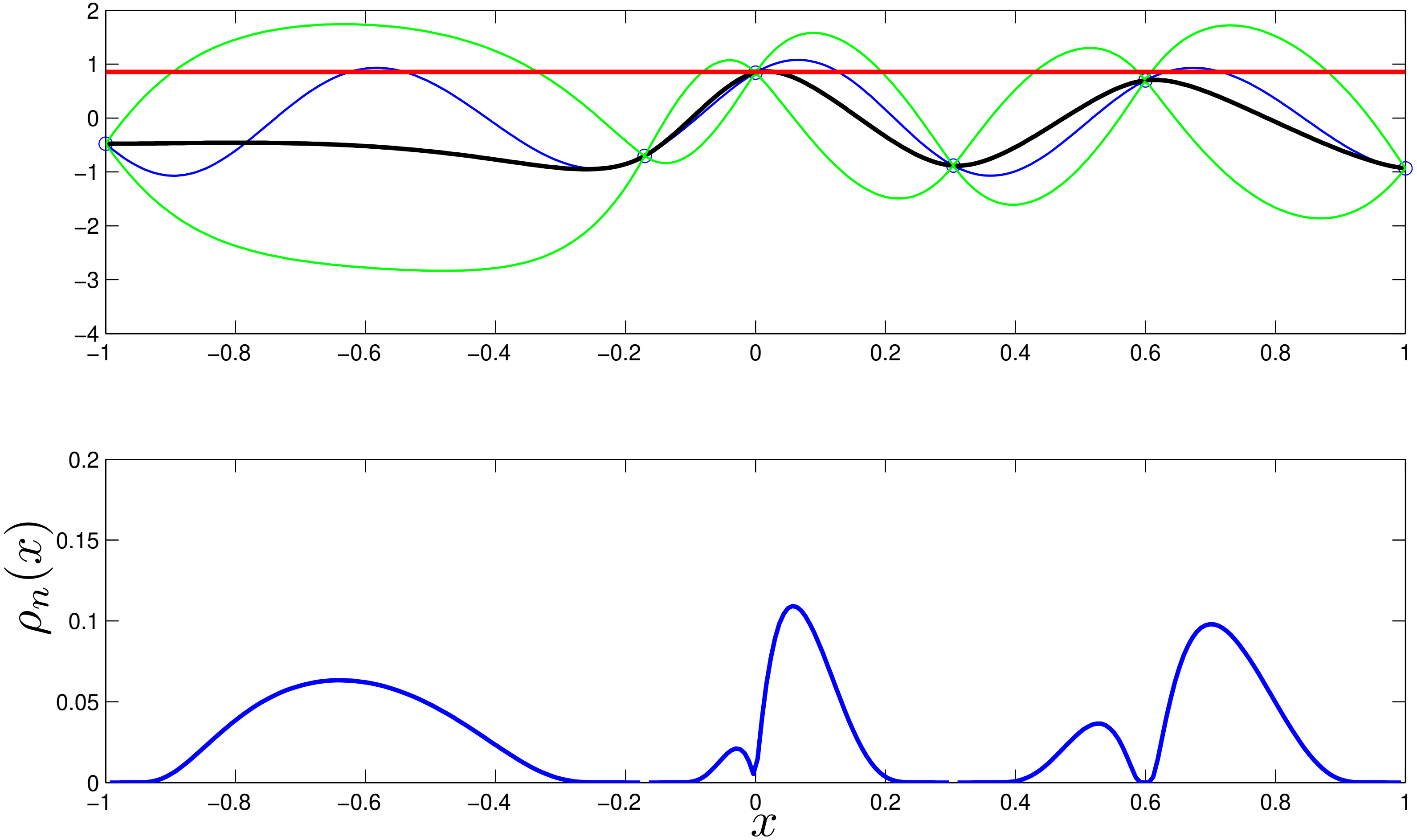
  avec $M_n = \max\{\xi(x_1), \ldots \xi(x_n)\}$.
- $\rho_n(x)$ peut être calculé analytiquement
- l'algorithme EI :

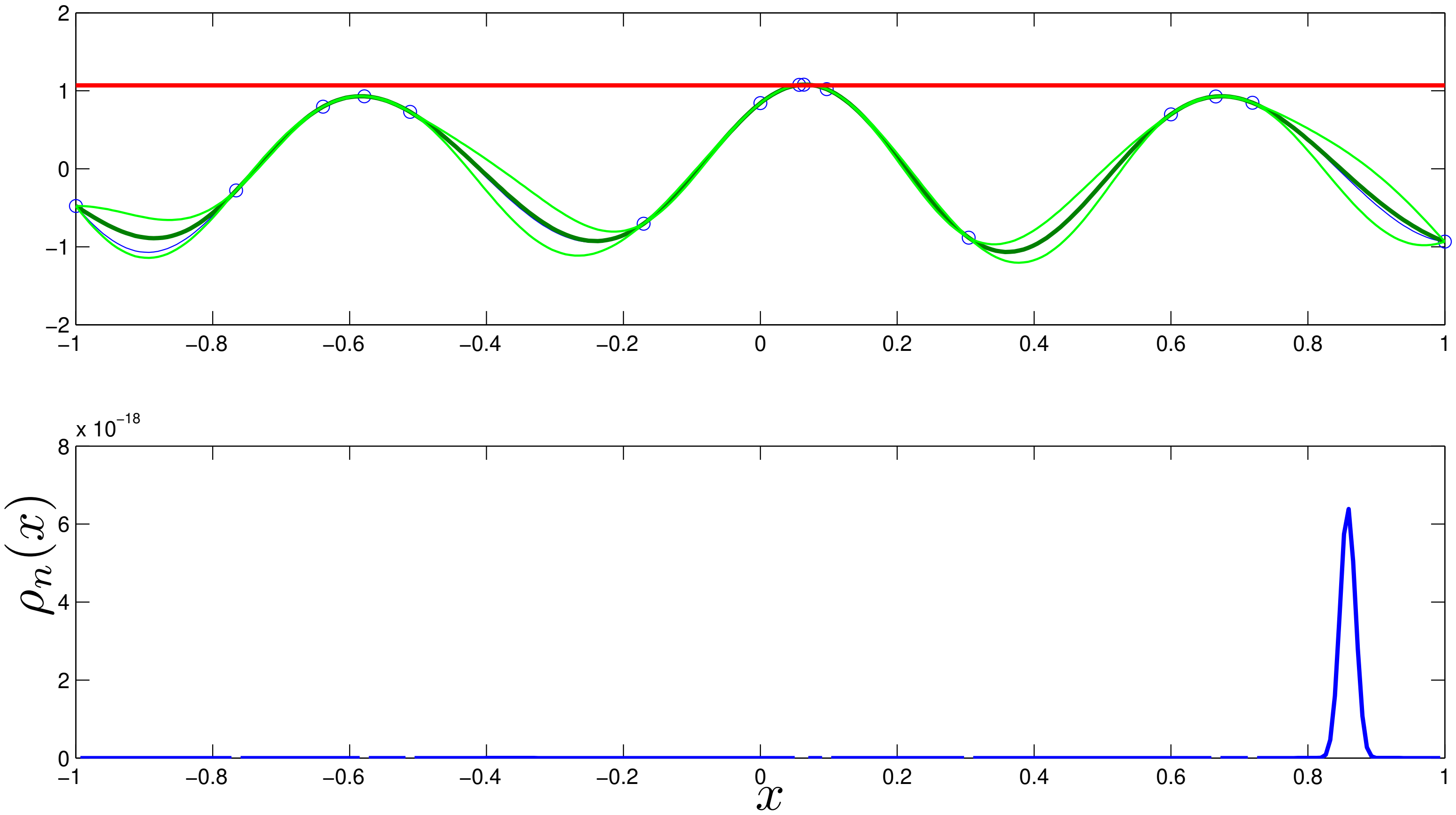
$$x_{n+1} = \arg\max_{x \in \mathbb{X}} \rho_n(x)$$

  $\leftrightarrow$ "stratégie optimiste d'amélioration"

## Exemple




2
1
0
−1
−2
−3
−4
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
$\rho_n(x)$
0.2
0.15
0.1
0.05
0
$x$



2
1
0
−1
−2
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
x 10−18
8
6
4
2
0
$\rho_n(x)$
$x$

[Vazquez, Bect 07] $\rightarrow$ $x^*$ est bien une valeur d'adhérence de $(x_n)_{n\in\mathbb{N}}$ produit par l'algorithme EI (sous certaines conditions de régularité)

**Questions**

- vitesse de convergence ?
- IAGO [Villemonteix et al. 08] :

$$x_{n+1} = \arg\min_{x\in\mathbb{X}_d} \mathrm{E}\left[H(x^*;\underline{\xi}_n,\xi(x)) \mid \underline{\xi}_n\right]$$

$\leftrightarrow$ réduction de l'incertitude sur la position du maximiseur

$\rightarrow$ convergence et vitesse de convergence ?

# Informations pratiques

- Stage M2R souhaité
- Indemnités de stage : 700€
- Thèse financée + monitorat possible
- Lieu : SUPELEC, Gif-sur-Yvette
  Grande Ecole de rang A, dans le domaine des sciences de l'information, de l'énergie et des systèmes (1560 élèves ingénieurs, 200 doctorants, 150 permanents)
- Contacts :

  – Emmanuel Vazquez `emmanuel.vazquez@supelec.fr`
  – Julien Bect `julien.bect@supelec.fr`