

# PLANS D'EXPÉRIENCES ET CODES NUMÉRIQUES: QUELQUES PROBLÈMES EN ADAPTATIF

Luc Pronzato

Laboratoire I3S,  
CNRS/Univ. Nice Sophia Antipolis, France

# Plan

- I) Quel objectif ?
- II) Non séquentiel versus séquentiel/adaptatif
- III) Modèle paramétrique ou non paramétrique  
(structure fixe ou variable)
- IV) Processus gaussien & krigeage
- V) Un objectif peut en cacher un autre ...
- VI) Quelques résultats asymptotiques
- VII) Quelques questions
- GDR MASCOT-NUM : Méthodes d'Analyse  
Stochastique pour les Codes et Traitements  
Numériques

# I) Quel objectif ?

$f(x)$  une fonction (inconnue) définie pour  $x \in \mathcal{X} \subset \mathbb{R}^d$

À partir de couples  $(X_i, f(X_i))$ ,  $i = 1, 2, \dots$  on va construire un modèle  $\eta(\cdot)$  de  $f(\cdot)$ .

On notera  $\eta_n(\cdot)$  la prédiction construite à partir de  $(X_i, f(X_i))$ ,  $i = 1, 2, \dots, n$  ( $n$  n'est pas forcément fixé *a priori*).

Différents objectifs sont possibles :

- ① exploration :  $\eta_n(\cdot)$  doit approximer  $f(\cdot)$  le mieux possible sur  $\mathcal{X}$
- ② optimisation :  
→ utiliser  $\eta_n(\cdot)$  pour déterminer  $\arg \max_{x \in \mathcal{X}} f(x)$
- ③ inversion : on souhaite pouvoir associer à tout  $y$  un  $x$  tel que  $\eta_n(x)$  soit proche de  $y$

**Rq** : on peut avoir  $m$  fonctions  $f_i(\cdot)$ ,  $i = 1, \dots, m$  :  
②  $\rightarrow$  optimisation multicritères ...

**Préciser l'objectif est un préalable indispensable à la construction d'un plan "optimal" !**

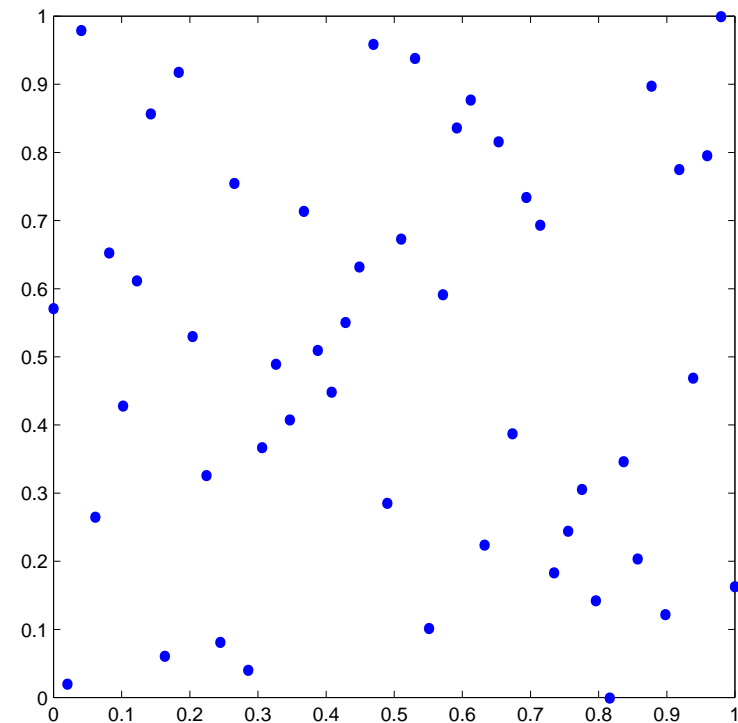
Pour ① le choix des  $X_i$  ne nécessite pas de modèle particulier (remplir  $\mathcal{X}$  suffit), mais ② et ③ ont besoin d'un modèle pour  $f(\cdot)$

Aussi, pour ② et ③ il semble clairement préférable de placer les points séquentiellement

## II) Non séquentiel ou séquentiel/adaptatif ?

Exploration non-séquentielle, sans modèle :

propriété de remplissage  
d'espace ("space filling")  
→ distance maximin ou  
minimax [Johnson *et al.*, 1990], hypercubes  
latins [Morris & Mitchell  
1995]...



**Non-séquentiel, avec modèle** : pour ① (objectif = exploration), **critère = maximum ou moyenne sur  $\mathcal{X}$  de l'erreur quadratique moyenne de prédiction (EQM) ...**

Disposer d'une EQM suppose un modèle  $\eta(\cdot)$   
... et si on dispose d'un modèle, on peut utiliser un plan adaptatif (construit pas à pas)

→ **Problème central : comment prédire l'incertitude en un point  $x$  non encore exploré ?**

## III) Paramétrique/non paramétrique

(Éviter d'utiliser une approche paramétrique dans un contexte non-paramétrique !)

**Modèle paramétrique, contexte stochastique :**

$f(X_i) = \eta(X_i, \bar{\theta}) + \varepsilon_i$ ,  $\theta$  estimé par MC

$$\hat{\theta}^n = \arg \min \frac{1}{n} \sum_{i=1}^n [f(X_i) - \eta(X_i, \theta)]^2$$

et prédiction à étape  $n$   $\eta_n(x) = \eta(x, \hat{\theta}^n)$

Contexte stochastique classique :  $(\varepsilon_i)$  i.i.d. de moyenne 0 et variance  $\sigma^2$ ,  $(X_i)$  i.i.d. avec mesure de probabilité  $\xi$  sur  $\mathcal{X}$   
 $\Rightarrow$  (assez facilement :  $\theta \in \Theta$  compact,  $\eta(x, \theta)$  continue en  $\theta$  pour tout  $x$ )  $\hat{\theta}^n \xrightarrow{\text{p.s.}} \bar{\theta}$  et (régularité de  $\eta(x, \cdot)$ )

$$\sqrt{n}(\hat{\theta}^n - \bar{\theta}) \xrightarrow{d} \zeta \sim \mathcal{N}(0, \sigma^2 \mathbf{M}^{-1}(\xi, \bar{\theta}))$$

avec  $\mathbf{M}(\xi, \theta) = \int_{\mathcal{X}} \frac{\partial \eta(x, \theta)}{\partial \theta} \frac{\partial \eta(x, \theta)}{\partial \theta^\top} \xi(dx)$

Qualité de la prédiction  $\eta_n(x) =$  qualité de l'estimation  $\hat{\theta}^n$ , mesurée par  $\Phi[\mathbf{M}(\xi, \bar{\theta})]$  ( $\Phi(\cdot)$  croissante, concave ...)

Par exemple,  $D$ -optimalité : mesure de probabilité  $\xi_D^*$  sur  $\mathcal{X}$  qui maximise  $\log \det[\mathbf{M}(\xi, \theta)]$

$\xi_D^*$  ?  $\rightarrow$  algorithme d'optimisation, ou séquentiellement :

Soient  $\theta$  fixé et  $X_1, \dots, X_{k_0}$  fixés (tels que  $\mathbf{M}(\xi_{k_0}, \theta)$  de rang plein) et  $\xi_k$  = mesure empirique de  $X_1, \dots, X_k$ . Choisir

$$X_{k+1} = \arg \max_{x \in \mathcal{X}} \frac{\partial \eta(x, \theta)}{\partial \theta^\top} \mathbf{M}^{-1}(\xi_k, \theta) \frac{\partial \eta(x, \theta)}{\partial \theta}, \quad k \geq k_0$$

assure  $\mathbf{M}(\xi_k, \theta) \rightarrow \mathbf{M}(\xi_D^*, \theta), k \rightarrow \infty$  [Wynn, 1970]

**Problème :**  $\xi^*(\theta)$  a  $p(p+1)/2$  points de support au plus (Th. Caratheodory), et souvent seulement  $p$ , avec  $p = \dim(\theta)$

$\Rightarrow \xi_k$  n'est pas "space filling"

$\rightarrow$  On ne détectera pas forcément une erreur de modèle

## Th. équivalence de Kiefer-Wolfowitz (1960) :

$\xi^*$  maximise  $\log \det \mathbf{M}(\xi, \theta)$  (= D-optimalité)

$\Leftrightarrow \xi^*$  minimise  $\max_{x \in \mathcal{X}} \frac{\partial \eta(x, \theta)}{\partial \theta^\top} \mathbf{M}^{-1}(\xi, \theta) \frac{\partial \eta(x, \theta)}{\partial \theta}$

=  $\max_{x \in \mathcal{X}}$  de la variance (asymptotique) de  $\eta(x, \hat{\theta}^n)$   
(= G-optimalité)

Si modèle linéaire : biais = 0

et  $\xi^*$  minimise  $\max_{x \in \mathcal{X}}$  EQM prédiction

## Pour l'intégrale de l'EQM :

$$IEQM(\xi, \theta) = \int_{\mathcal{X}} \frac{\partial \eta(x, \theta)}{\partial \theta^\top} \mathbf{M}^{-1}(\xi, \theta) \frac{\partial \eta(x, \theta)}{\partial \theta} \mu(dx)$$

$$= \text{trace} \left[ \mathbf{M}^{-1}(\xi, \theta) \int_{\mathcal{X}} \frac{\partial \eta(x, \theta)}{\partial \theta} \frac{\partial \eta(x, \theta)}{\partial \theta^\top} \mu(dx) \right]$$

$$= \text{trace} \left[ \mathbf{M}^{-1}(\xi, \theta) \mathbf{M}(\mu, \theta) \right]$$

Minimiser  $IEQM(\xi, \theta) = A$ -optimalité  $\rightarrow$  Algorithme :

$$X_{k+1} = \arg \max_{x \in \mathcal{X}} \frac{\partial \eta(x, \theta)}{\partial \theta^\top} \mathbf{M}^{-1}(\xi_k, \theta) \mathbf{M}(\mu, \theta) \mathbf{M}^{-1}(\xi_k, \theta) \frac{\partial \eta(x, \theta)}{\partial \theta}$$

**Retour à la D-optimalité :**

Il faudrait maximiser  $\log \det[\mathbf{M}(\xi, \bar{\theta})]$  et pas  $\log \det[\mathbf{M}(\xi, \theta)]$   
pour un  $\theta$  ad'hoc  $\rightarrow$  on ré-estime  $\theta$  à chaque étape :

$$X_{k+1} = \arg \max_{x \in \mathcal{X}} \frac{\partial \eta(x, \theta)}{\partial \theta^\top} \Big|_{\hat{\theta}^k} \mathbf{M}^{-1}(\xi_k, \hat{\theta}^k) \frac{\partial \eta(x, \theta)}{\partial \theta} \Big|_{\hat{\theta}^k}, \quad k \geq k_0 \quad (1)$$

Convergence de  $\hat{\theta}^k$  vers  $\bar{\theta}$  et de  $\mathbf{M}(\xi_k, \hat{\theta}^k)$  vers  $\mathbf{M}(\xi^*, \bar{\theta})$  pas prouvée en général, **mais jamais mise en défaut...** (problème toujours difficile :  $(X_k)$  pas i.i.d.)

On sait dire des choses en bayésien (Bayesian imbedding) [Hu, 1998],  
ou pour MC en perturbant la séquence  $(X_k)$  par une sous-séquence déterministe suffisamment excitante [Lai, 1994],  
ou quand  $\mathcal{X}$  est fini [P, soumis]

Pour maximum de vraisemblance  $\hat{\theta}_{MV}^n$ ,  
 $\sum_{i=1}^n \nabla_{\theta} \log \phi(Y_i | X_i, \theta) = \text{martingale}$  [Chaudhuri & Mykland, 1993, 1995; Rosenberger, Flounoy, Durham, 1997]

## Pourquoi est-ce difficile ?

MC pour modèle linéaire  $Y_i = \mathbf{r}_i^\top \bar{\theta} + \varepsilon_i$  ( $\mathbf{r}_i = \mathbf{r}(x_i)$ )

$\mathcal{F}_k = \sigma$ -algèbre engendrée par  $\varepsilon_1, \dots, \varepsilon_k$

$\varepsilon_k$  est  $\mathcal{F}_{k-1}$  mesurable,  $\mathbb{E}\{\varepsilon_k | \mathcal{F}_{k-1}\} = 0$  et  $\mathbb{E}\{\varepsilon_k^2 | \mathcal{F}_{k-1}\} < \infty$

$\mathbf{M}_n = \sum_{i=1}^n \mathbf{r}_i \mathbf{r}_i^\top$  ( $= n\mathbf{M}(\xi_n)$ )

Alors,  $\mathbf{M}_n^{-1} \rightarrow \mathbf{O}$  suffisant pour  $\hat{\theta}^n \xrightarrow{\text{p.s.}} \bar{\theta}$  si  $(\mathbf{r}_i)$  déterministe  
(et N & S si  $(\varepsilon_i)$  i.i.d.) [Lai & Robbins, 1978, 1979]  
mais pas suffisant si  $\mathbf{r}_k$  est  $\mathcal{F}_{k-1}$ -mesurable

CS :  $\lambda_{\min}(\mathbf{M}_n) \xrightarrow{\text{p.s.}} \infty$  et  $[\log \lambda_{\max}(\mathbf{M}_n)]^{1+\delta} = o[\lambda_{\min}(\mathbf{M}_n)]$  p.s.,  
 $\delta > 0$  [Lai & Wei, 1982]

Pour un modèle non linéaire : [Lai 1994] donne une CS  
... qui donnerait en linéaire :

$\lambda_{\min}(\mathbf{M}_n) \xrightarrow{\text{p.s.}} \infty$  et  $\lambda_{\max}(\mathbf{M}_n) = O[\lambda_{\min}^\rho(\mathbf{M}_n)]$  p.s.,  $1 < \rho < 2$

## Modélisation d'un code numérique :

En considérant que modèle non paramétrique = modèle paramétrique avec beaucoup de paramètres, on pourrait penser utiliser (1) dans un cadre déterministe avec une fonction d'approximation pour modèle  $\eta(x, \theta)$  (réseau de neurones, RBF...)

Rq : (1)  $\rightarrow X_{k+1}$  où la variance de la prédiction à l'étape  $k$  est maximale

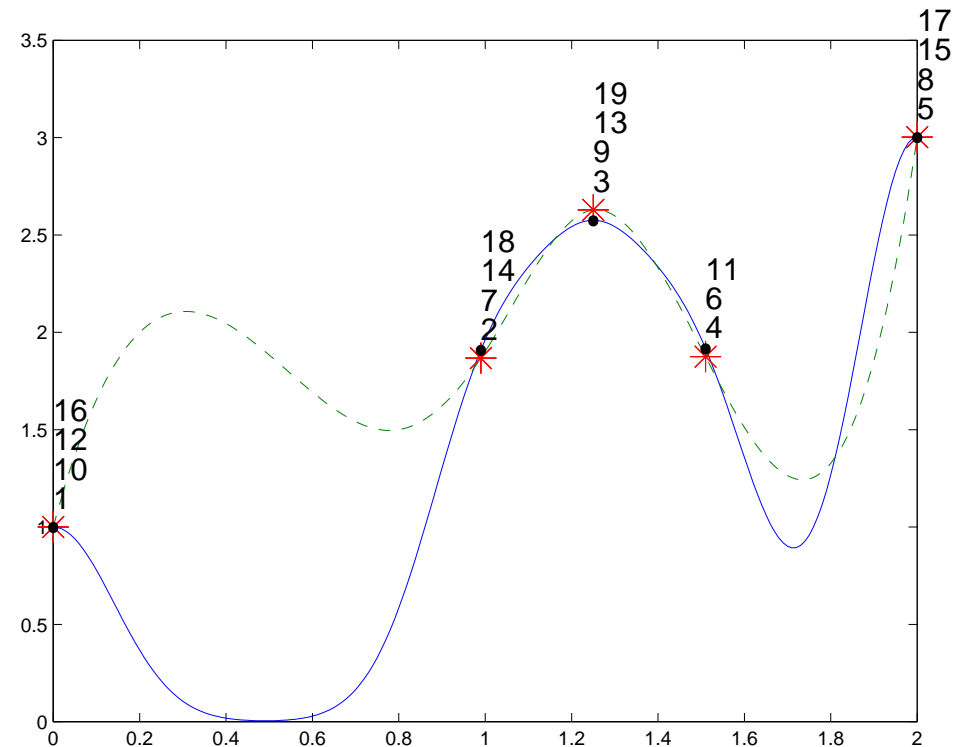
Quelques exemples de difficultés ...

**Ex. 1 : RBF**  $\eta(x) = \sum_{j=1}^p \theta_j K(x - N_j)$  (par ex.  $K(\cdot)$  gaussien)

→ modèle linéaire à  $p$  paramètres

$p = 5$  fixé,  $N_1, \dots, N_5$  fixés,  $(X_1, \dots, X_5) =$  plan  $D$ -optimal  
puis  $X_{i+1}$  là où la variance de la prédiction est maximale  
(ce qui suppose  $f(X_i) = \eta(X_i) + \varepsilon_i \dots$ )

→ pas d'exploration :  
le plan reste concentré en  
 $p$  points. On peut essayer  
de faire croître  $p$  avec  $n$   
...



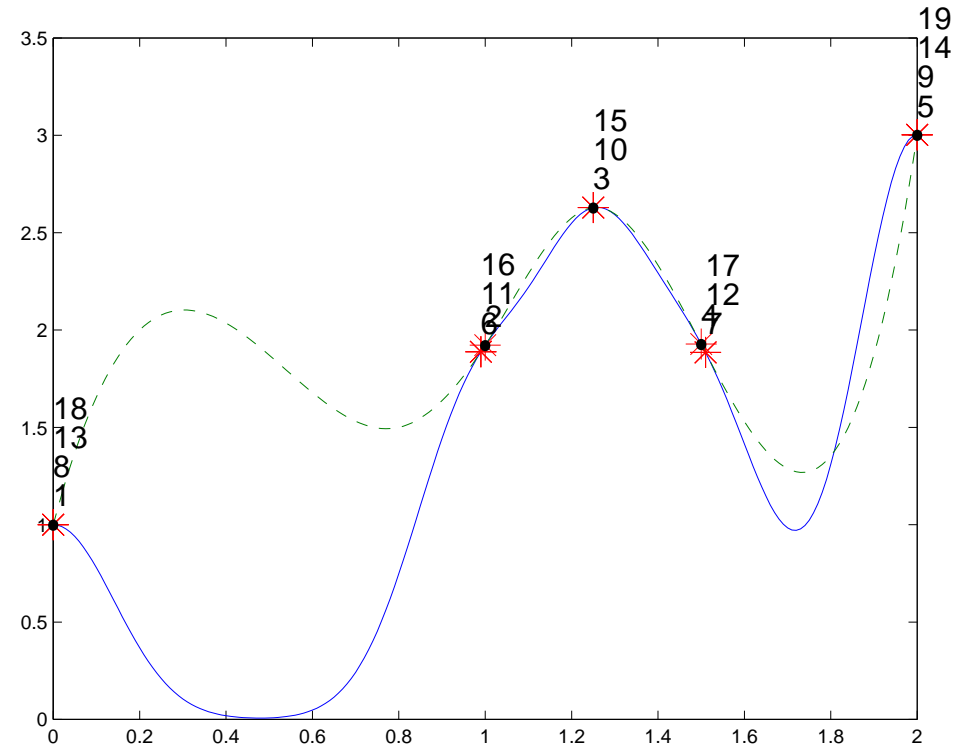
Prenons  $p$  croissant (structure variable) :

$N_1, \dots, N_5$  et  $(X_1, \dots, X_5)$  comme précédemment

puis  $X_{i+1}$  au point de variance de prédiction maximale et

$N_{i+1} = X_{i+1}$  (si ce  $N$  pas déjà présent)

→ pas beaucoup plus  
d'exploration : le plan  
reste concentré



**Le problème n'est pas lié au choix de  $\eta(x, \theta)$  : mêmes difficultés pour polynômes, ou "fonction universelle" du type**

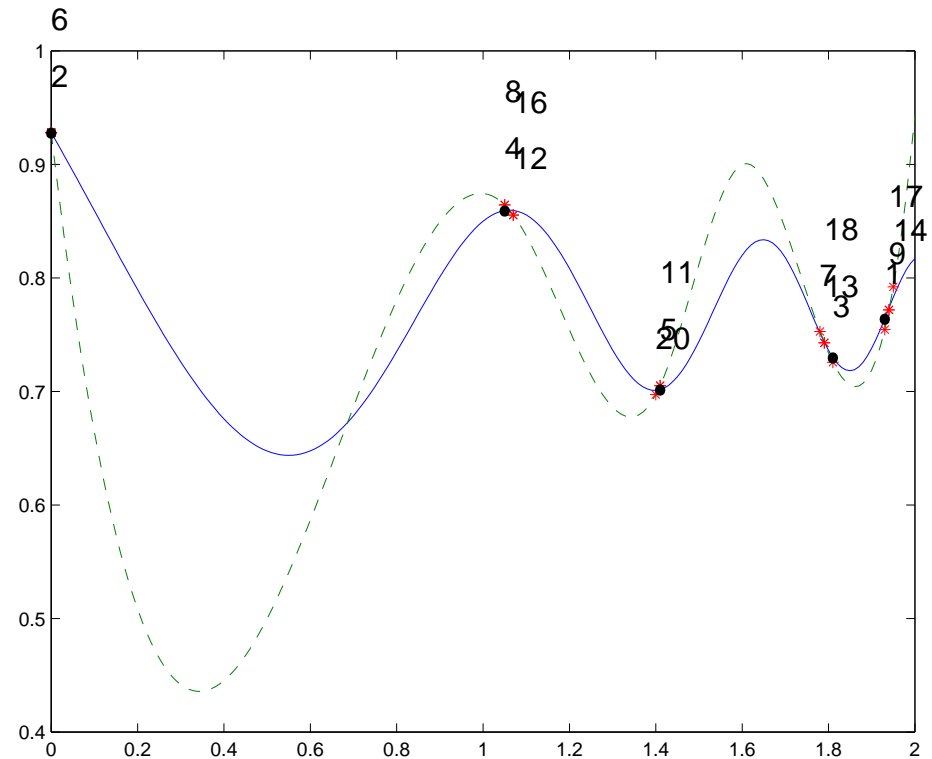
$$\eta(x, \theta) = f_U(x, \theta) = \int_0^{x+\theta_1} \frac{\theta_2 \theta_4}{1 + \theta_4^2 - \cos(\theta_2 t)} \cos[\exp(t)] dt + \theta_3$$

pour  $\theta = (a, b, c, d) \in \mathbb{R}^4, d > 0$

→ famille dense dans  $\mathcal{C}(\mathcal{I})$  pour tout intervalle compact  $\mathcal{I}$  de  $\mathbb{R}$ , avec  $\mathcal{C}(\mathcal{I}) =$  espace de Banach des fonctions continues bornées sur l'intervalle  $\mathcal{I} \subset \mathbb{R}$  pour la norme  $\|f\| = \sup_{x \in \mathcal{I}} |f(x)|$ , **[Boshernitzan, 1996]**

**Ex. 2 :**  $\eta(x, \theta) = \int_0^{x+\theta_1} \frac{\theta_2 \theta_4}{1 + \theta_4^2 - \cos(\theta_2 t)} \cos[\exp(t)] dt + \theta_3$

→ pas vraiment d'exploration : le plan reste concentré



→ **Changeons la façon de prédire l'incertitude en  $x$  !**

**Ex. 1 (suite) : RBF avec Bootstrap et Leave-One-Out**

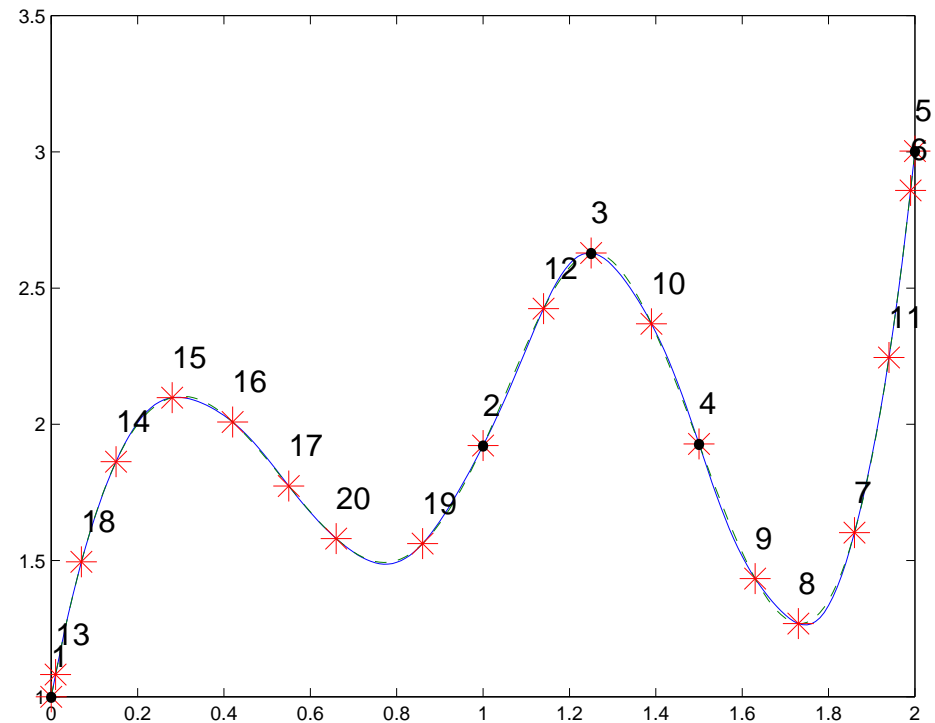
**[Gazut, Martinez, Issanchou 2006] :**

$$\eta^{(i)}(x, \theta_{(i)}) = \sum_{j=1, j \neq i}^n \theta_j K(x - N_j)$$

→ variance  $V_n(x)$  des  $n$  prédictions  $\eta^{(i)}(x, \theta_{(i)})$  (MC)

→ observation en  $X_{n+1} = \arg \max_x V_n(x)$  et  $N_{n+1} = X_{n+1}$   
(si pas déjà présent)

→ exploration ... mais difficile de dire pourquoi ça marche !



## IV) Processus gaussien & krigeage

Modèle :  $f(x) = \beta + Z(x, \omega)$  avec  $Z(x, \omega)$  réalisation d'un processus gaussien stationnaire au second ordre,  
 $\mathbb{E}\{Z(x, \omega)\} = 0$ ,  $\mathbb{E}\{Z(x, \omega)Z(u, \omega)\} = \sigma^2 C((x - u), \theta)$  (ici,  $\beta$  et  $\sigma^2$  inconnus et  $\theta$  fixé)

BLUP en  $x$  :  $\eta_n(x) = \mathbf{v}^\top(x) \mathbf{y}_n$  où

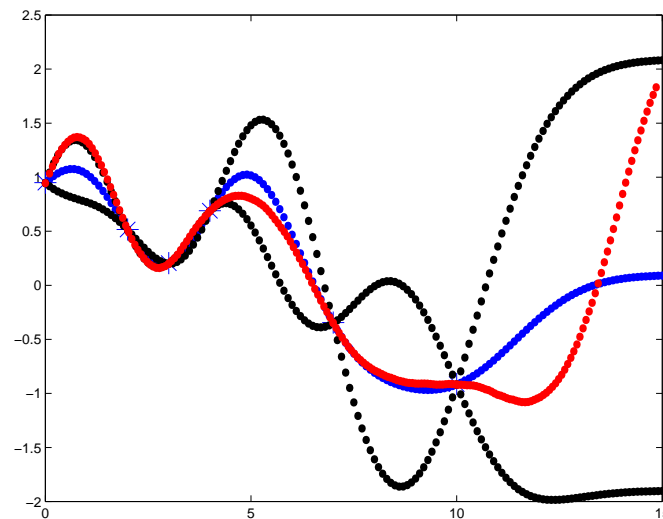
- $\mathbf{y}_n = (f(X_1), \dots, f(X_n))^\top$
- $\mathbf{v}(x)$  minimise  $\mathbb{E}\{(\mathbf{v}^\top \mathbf{y}_n - [\beta + Z(x, \omega)])^2\}$
- sous la contrainte  $\mathbb{E}\{\mathbf{v}^\top \mathbf{y}_n\} = \beta \sum_{i=1}^n v_i = \mathbb{E}\{f(x)\} = \beta$   
(c.a.d.,  $\sum_{i=1}^n v_i = 1$ )

Prédiction :  $\eta_n(x) = \hat{\beta}^n + \mathbf{c}_n^\top(x) \mathbf{C}_n^{-1} (\mathbf{y}_n - \hat{\beta}^n \mathbf{1})$

EQM :

$$\rho_n(x) = \sigma^2 \left( 1 - \begin{bmatrix} \mathbf{c}_n^\top(x) & 1 \end{bmatrix} \begin{bmatrix} \mathbf{C}_n & \mathbf{1} \\ \mathbf{1} & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{c}_n(x) \\ 1 \end{bmatrix} \right)$$

avec  $\{\mathbf{C}_n\}_{i,j} = C((X_i - X_j), \theta)$ ,  $\{\mathbf{c}_n(x)\}_i = C((X_i - x), \theta)$ ,  
 $\hat{\beta}^n = (\mathbf{1}^\top \mathbf{C}_n^{-1} \mathbf{y}_n) / (\mathbf{1}^\top \mathbf{C}_n^{-1} \mathbf{1})$  (WLS) et  $\mathbf{1} = (1, \dots, 1)^\top$



$\Rightarrow \rho_n(X_i) = 0, i = 1, \dots, n \rightarrow$  **interpolateur**

**Extensions :**  $\beta \rightarrow \beta^\top \mathbf{f}(x)$  (krigeage universel)

loi a priori sur  $\beta$  et/ou  $\sigma^2$  (krigeage bayésien)

$\rightarrow$  loi a posteriori pour  $f(x)$  normale  $\mathcal{N}(\eta_n(x), \rho_n(x))$  si  $\sigma^2$  connu et a priori sur  $\beta$  peu informatif

**Pour objectif ① (exploration) :**

**Plan non séquentiel :**

–  $\min_{X_1, \dots, X_n} \max_{x \in \mathcal{X}} \rho_n(x)$  ( $\leftrightarrow$  distance minimax)

–  $\min_{X_1, \dots, X_n} \int_{\mathcal{X}} \rho_n(x) \pi(dx)$

– Max. entropy sampling :  $\max_{X_1, \dots, X_n} \det \mathbf{C}_n$  ( $\leftrightarrow$  distance maximin)

... assez compliqué

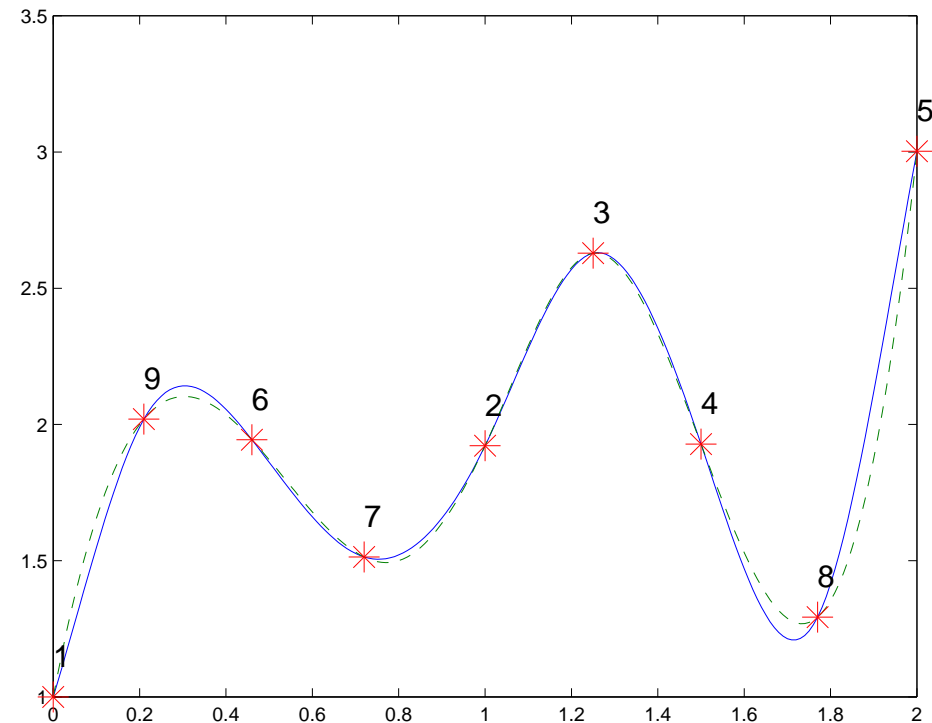
**Plan séquentiel :**  $X_{n+1} = \arg \max_{x \in \mathcal{X}} \rho_n(x)$

## Ex. 3 : Krigage

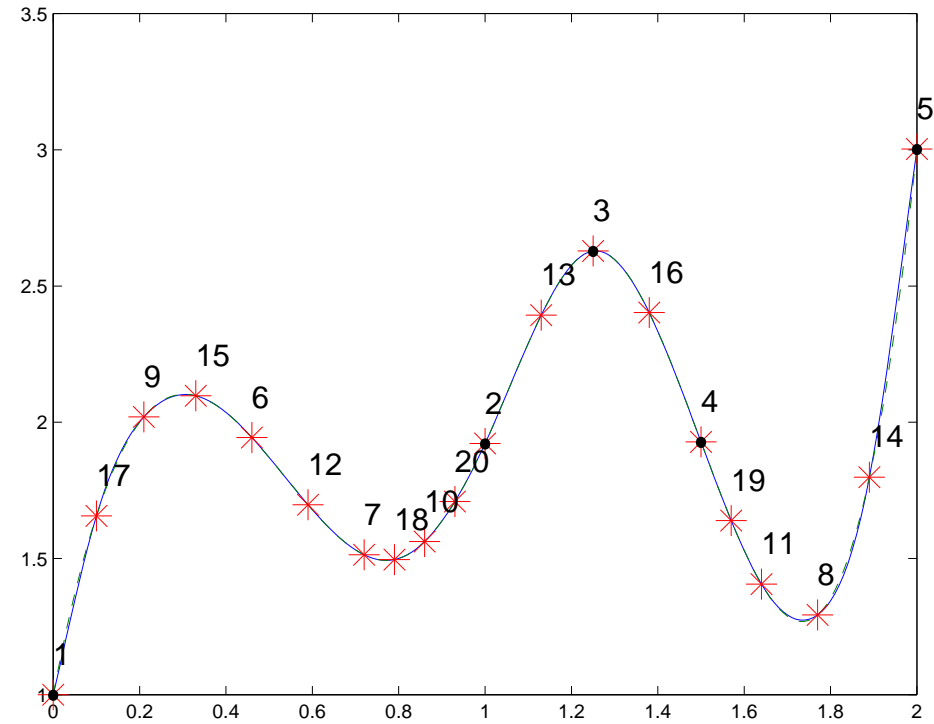
Même données que pour Ex. 1, modèle :  $f(x) = \beta + Z(x, \omega)$   
avec  $\mathbb{E}\{Z(x, \omega)\} = 0$ ,  $\mathbb{E}\{Z(x, \omega)Z(u, \omega)\} = \sigma^2 C(|x - u|, \theta)$  ( $\beta$   
et  $\sigma^2$  inconnus et  $\theta$  fixé)

$X_{n+1}$  au point de variance de krigage maximale

→ exploration  
(9 points seulement)



→ exploration  
(20 points)



**Pour objectif ② (maximisation) :** maximiser l'*expected improvement*  $EI(x) = \int_{\max_i Y_i}^{\infty} [Y - \max_i Y_i] \varphi_n(Y|x) dY$  avec  $\varphi_n(\cdot|x)$  la densité de  $\mathcal{N}(\eta_n(x), \rho_n(x))$  [Mockus 1989, Schonlau *et al* 1998]

(par ex., rechercher le maximum de  $EI(x)$  dans les "triangles" d'une triangulation de Delaunay des  $X_i$  [Bates & P, 2002])

**Pour objectif ③ (inversion) :**

→ associer à tout  $y$  un  $x$  tel que  $\eta_n(x)$  soit proche de  $y$

$$\min_{X_1, \dots, X_n} \mathbb{E}\{\max_y \min_x \mathbb{E}\{\|f(x) - y\|^2 | X_1, Y_1, \dots, X_n, Y_n\}\}$$

OU

$$\min_{X_1, \dots, X_n} \mathbb{E}\{\int \min_x \mathbb{E}\{\|f(x) - y\|^2 | X_1, Y_1, \dots, X_n, Y_n\} \mu(dy)\}$$

... assez compliqué !

→ maximiser une mesure de dispersion des  $Y_i$

1.  $\hat{\varphi}_{n,y}$  estimateur à noyaux (gaussiens) à partir de  $Y_1, \dots, Y_n$  et  $y$ , prochaine observation en  $x$

2. entropy de Tsallis d'ordre 2 de  $\hat{\varphi}_{n,y}$

$$H_{2,n,y} = 1 - \int \hat{\varphi}_{n,y}^2(t) dt$$

3.  $X_{n+1} = \arg \max_x \mathbb{E}\{H_{2,n,y}\}$  pour  $y \sim \mathcal{N}(\eta_n(x), \rho_n(x))$

Tout cela se calcule (car Tsallis d'ordre 2 + gaussien)

[Thèse de R. Bettinger, IFP]

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

- **Modèle paramétrique : prédire  $\Leftrightarrow$  estimer  $\theta$**
- **Modèle non paramétrique (krigeage) : on a supposé  $\theta$  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  $\hat{\theta}_{MV}^n = \arg \min_{\theta} \det[\hat{\sigma}_n^2(\theta) \mathbf{C}_n(\theta)]$

→ le plan d'expériences a une influence sur la précision de l'estimation de  $\theta$

$\mathbb{E}\{Z(x, \omega)Z(u, \omega)\} = \sigma^2 C((x - u), \theta)$ , avec  $C(t, \theta) \searrow 0$  pour  $t \rightarrow \infty$

Soit  $l_n(\nu)$  la log-vraisemblance,  $\nu = (\beta, \alpha) = (\beta, \sigma^2, \theta)$   
( $\alpha = (\sigma^2, \theta)$  et  $\mathbf{C}_{n,\alpha} = \sigma^2 \mathbf{C}_n(\theta)$ )

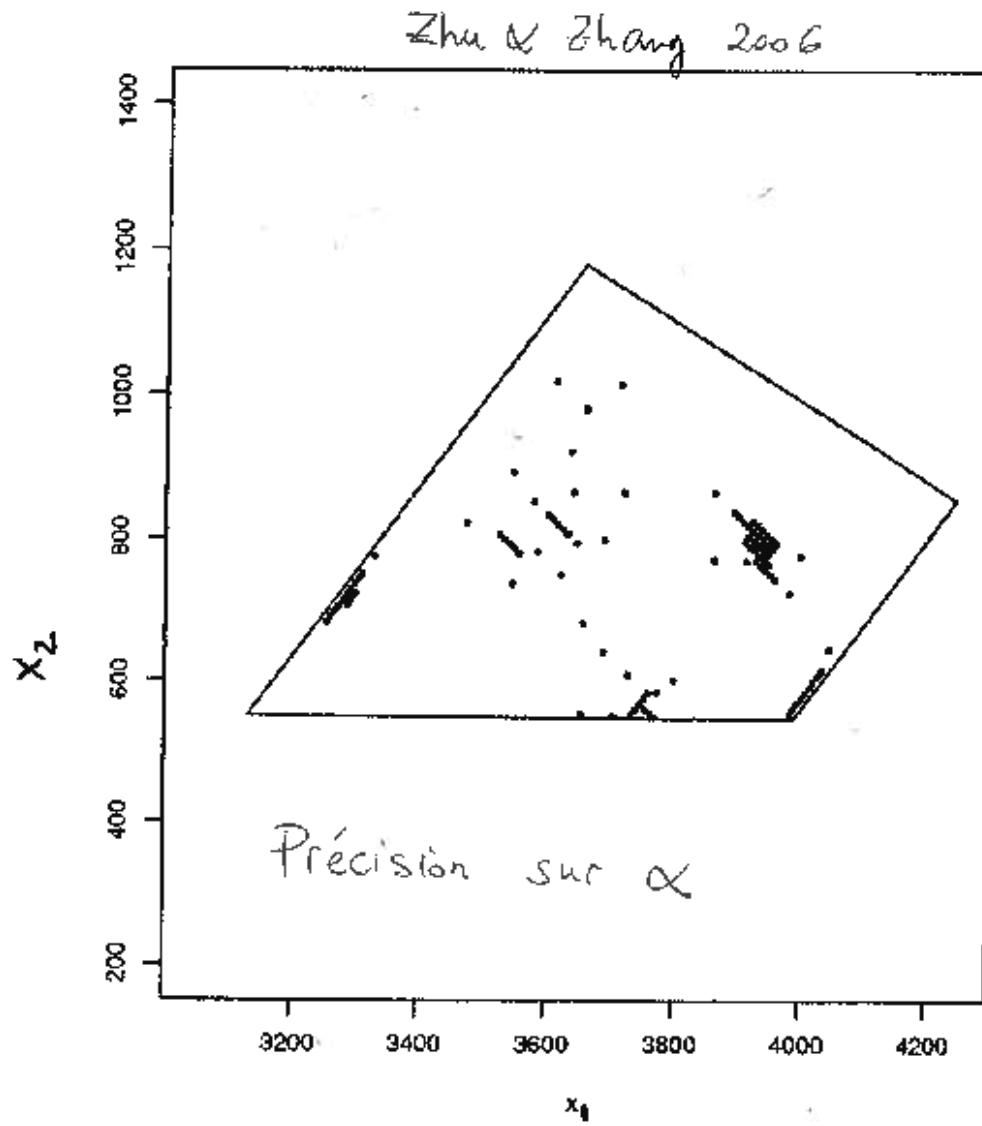
- les  $X_i$  sont dans un domaine de taille croissante

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

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

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

Si on s'intéresse à la précision sur  $\alpha$  : maximiser  $\det \tilde{\mathbf{M}}_n(\alpha)$



## Quelle influence sur la prédiction ?

Prédiction par substitution  $\eta_n(x, \hat{\nu}_{MV}^n) \rightarrow$  quelle EQM ?

Si  $\alpha$  connu,  $\mathbb{E}[(f(x) - \eta_n(x, \nu)]^2 = \rho_n(x) = \rho_n(x, \nu)$

DL au 1er ordre :

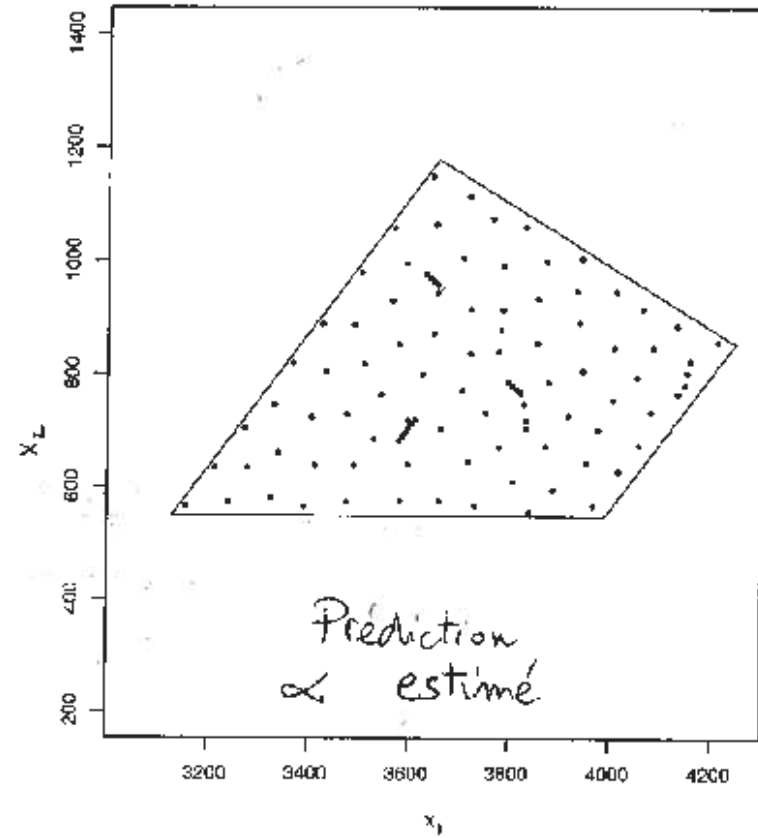
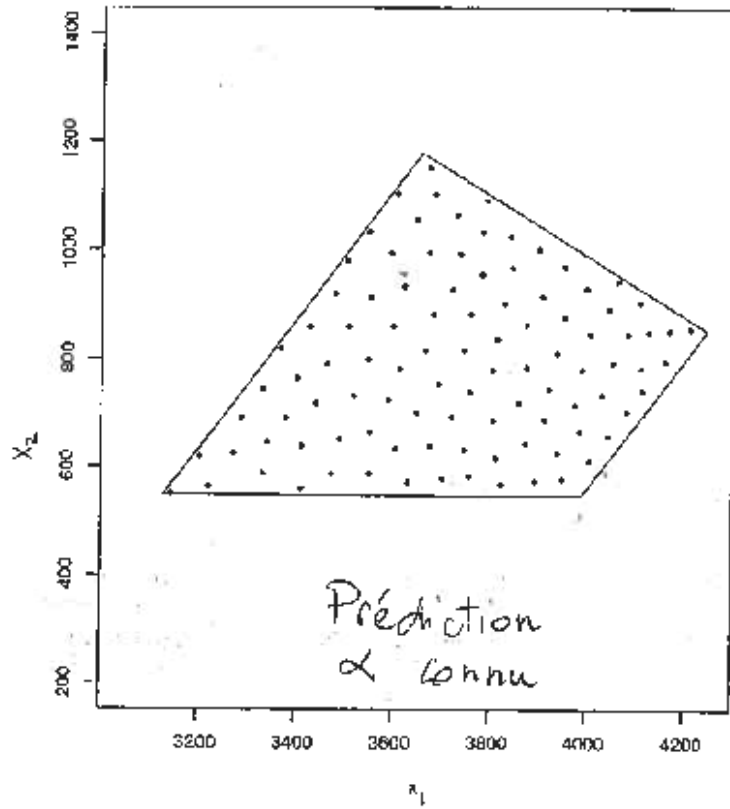
$$\mathbb{E}[(f(x) - \eta_n(x, \nu_{MV}^n)]^2 = \rho_n(x) + \text{trace} \left[ \tilde{\mathbf{M}}_n^{-1}(\alpha) \frac{\partial \mathbf{u}^\top(x)}{\partial \alpha} \mathbf{C}_{n,\alpha} \frac{\partial \mathbf{u}(x)}{\partial \alpha^\top} \right]$$

avec  $\mathbf{u}(x) = \mathbf{C}_n^{-1} \mathbf{c}_n(x)$

On peut aussi prendre en compte l'influence sur  $\rho_n(x, \nu_{MV}^n)$

... [Zhu & Zhang 2006]

Z. ZHU AND H. ZHANG 2006



- les  $X_i$  sont dans un compact (*infill asymptotics*)

$\theta_{MV}^n$  pas toujours consistant (mais effet compensé sur la prédiction — loi de Jeffrey)

→ prendre en compte seulement les paramètres micro-ergodiques

$h(\theta)$  micro-ergodique :  $\forall \theta, \theta', h(\theta) \neq h(\theta') \Rightarrow P_\theta \perp P_{\theta'}$   
(mesures gaussiennes :  $P_\theta \equiv P_{\theta'}$  ssi  $\Delta(t) = C(t, \theta) - C(t, \theta')$  est plus régulière en zéro que  $C(t, \theta)$  et  $C(t, \theta')$  [Stein 1999])

Ex :  $C(t, \alpha) = \alpha_1 \exp(-\alpha_2 |t|)$ ,  $t \in [0, 1]$

$\sqrt{n}(\hat{\alpha}_1 \hat{\alpha}_2 - \alpha_1 \alpha_2) \xrightarrow{d} \mathcal{N}(0, 2(\alpha_1 \alpha_2)^2)$  et  $\alpha_1 \alpha_2$  microergodique  
[Ying, 1991]

## VI) Quelques résultats asymptotiques

**Régression non-paramétrique** :  $Y_i = f(X_i) + \varepsilon_i$

$f \in \Sigma(\beta, L)$  (Hölder) sur  $[0, 1]$  :  $f^{(l)}(\cdot)$  existe,  $l = \lfloor \beta \rfloor$ , et  
 $|f^{(l)}(x) - f^{(l)}(x')| < L|x - x'|^{\beta-l}, \forall x, x'$

Estimateur localement polynomial d'ordre  $l$ :

$$\hat{\theta}_n(x) = \arg \min_{\theta \in \mathbb{R}^{l+1}} \sum_{i=1}^n \left[ Y_i - \theta^\top U \left( \frac{X_i - x}{h} \right) \right]^2 K \left( \frac{X_i - x}{h} \right)$$

$$U(t) = (1, t, t^2/2, \dots, t^l/l!)^\top$$

$$\hat{f}_n(x) = U^\top(0) \hat{\theta}_n(x) = \{\hat{\theta}_n(x)\}_1$$

= généralisation de Nadaraya-Watson (ordre 0)

## Majorations du risque maximal : [Tsybakov, 2004]

$$\limsup_{n \rightarrow \infty} \sup_{f \in \Sigma(\beta, L)} \sup_{x \in [0, 1]} \mathbb{E}\{\Psi_n^{-2} |\hat{f}_n(x) - f(x)|^2\} \leq C < \infty$$

avec  $\Psi_n = n^{-\frac{\beta}{2\beta+1}}$  (pour  $h = h_n = cn^{-\frac{1}{2\beta+1}}$ , un noyau  $K(\cdot)$  convenable, des  $X_i$  déterministes et convenablement répartis sur  $[0, 1]$ , des  $\varepsilon_i$  i.i.d.)

Aussi, pour la norme  $\|\cdot\|_\infty$ ,

$$\limsup_{n \rightarrow \infty} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}\{\Psi_n^{-2} \|\hat{f}_n(x) - f(x)\|_\infty^2\} \leq C < \infty$$

avec  $\Psi_n = (\log n/n)^{\frac{\beta}{2\beta+1}}$  (pour  $h = h_n = c(\log n/n)^{\frac{1}{2\beta+1}} \dots$ )

Idem sur des classes de Sobolev  $W(\beta, L)$  (ou périodiques  $W^{per}(\beta, L)$ ),  $\beta$  entier :

$f^{(\beta-1)}(\cdot)$  absolument continue et  $\int_0^1 (f^{(\beta)}(x))^2 dx < L^2$

(périodique :  $f^{(j)}(1) = f^{(j)}(0)$ ,  $j = 0, 1, \dots, \beta - 1$ )

Estimateur par projection (sur la base trigonométrique  $\rightarrow$  Fourier)

$$\limsup_{n \rightarrow \infty} \sup_{f \in W^{per}(\beta, L)} \mathbb{E}\{\Psi_n^{-2} \|\hat{f}_n(x) - f(x)\|_2^2\} \leq C < \infty$$

avec  $\Psi_n = n^{-\frac{\beta}{2\beta+1}}$  (pour  $N = N_n = cn^{\frac{1}{2\beta+1}} \dots$ )

Idem sur  $W(\beta, L)$  avec base d'ondelettes

En dimension  $d$ ,  $2\beta + 1 \rightarrow 2\beta + d$

## Et en séquentiel (apprentissage actif) ?

[R. Catro *et al.*, 2004] : minoration du risque minimax et méthodes optimales

1) Stratégie de placement des  $X_i$  passive ou active ( $X_i$  peut dépendre de  $X_1, \dots, X_{i-1}$  et  $Y_1, \dots, Y_{i-1}$ )

2) classe PS de fonctions "lisses par morceaux"

En passif :

$$\inf_{(\hat{f}_n, S_n)^{passive}} \sup_{f \in PS} \mathbb{E}\{\|\hat{f}_n(x) - f(x)\|^2\} \geq c \max\left(n^{-\frac{2\beta}{2\beta+d}}, n^{-1/d}\right)$$

(en oubliant les  $\log n$ )

–  $\beta$  petit ( $< 1/(2 - 2/d)$ ) : la discontinuité ne rend pas le problème plus difficile (borne en  $n^{-\frac{2\beta}{2\beta+d}}$ )

–  $\beta$  plus grand : vitesse bornée par  $n^{-1/d}$ , limitation due à la discontinuité

En actif :

$$\inf_{(\hat{f}_n, S_n)^{active}} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}\{\|\hat{f}_n(x) - f(x)\|^2\} \geq cn^{-\frac{2\beta}{2\beta+d}}$$

$\Rightarrow$  actif pas mieux que passif sur  $\Sigma(\beta, L)$

$$\inf_{(\hat{f}_n, S_n)^{active}} \sup_{f \in PS} \mathbb{E}\{\|\hat{f}_n(x) - f(x)\|^2\} \geq c \max\left(n^{-\frac{2\beta}{2\beta+d}}, n^{-1/(d-1)}\right)$$

$d - 1 =$  dimension de la discontinuité

Algorithmes (à base de partitions dyadiques récursives) atteignant la borne (uniformément sur la classe de fonctions en passif). Peut-on faire mieux ? Rui Castro conjecture que non ...

## VII) Quelques questions

- [Zhu, Zhang 2006] : objectif ① (prédiction), que faire pour un autre objectif ? (optimisation)
- approche bayésienne  
$$p(f(x)|\mathbf{y}_n) = \int_{\Theta} p(f(x)|\theta, \mathbf{y}_n) p(\theta|\mathbf{y}_n) d\theta \rightarrow ?$$
- non-stationnarité ?
- problème très général : travailler en séquentiel signifie perdre l'indépendance, peu de résultats asymptotiques ( $\mathcal{X}$  fini simplifie la situation en paramétrique avec erreurs de mesure, mais n'a pas beaucoup de sens en non paramétrique sans erreurs — on va observer partout !)

## Résultats asymptotiques un peu "décevants" :

- le séquentiel (actif) n'apporte rien pour une fonction lisse
- il faut  $n$  grand

Le choix de méthodes performantes pour  $d$  grand et  $n$  petit est encore ouvert

Avec krigeage : placer des points pour remplir l'espace (objectif = prédiction) et d'autres pour estimer  $\theta$  (covariance)

→ peut être fait non-séquentiellement

→ quelles performances asymptotiques, sur quelle classe de fonctions ?