

Optimisation bayésienne par méthodes SMC

Romain Benassi Julien Bect Emmanuel Vazquez

SUPÉLEC

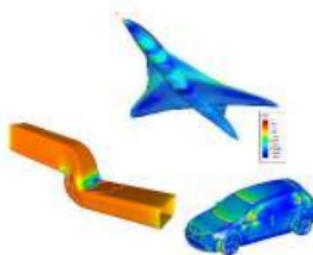
21 mars 2012

Mascot Num 2012

Bruyères-le-Châtel, France

Contexte

- Optimisation globale \Rightarrow **nombreuses applications** dans l'industrie :
 - Conception d'une voilure d'avion [Forrester et al. 2008]
 - Optimisation de la forme des conduits d'admission d'un moteur de voiture [Villemonteix et al. 2007]
 - etc.
- Fonctions à optimiser **coûteuses** à évaluer (appel à des programmes informatiques avec des durées d'exécution importantes)



Exemples d'optimisation de formes en CFD

Optimisation bayésienne 1/2

- Soit

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

avec \mathbb{X} compact et $d \in \mathbb{N}$

- Objectif : trouver $x^* = \operatorname{argmax}_{x \in \mathbb{X}} f(x)$ et $M = f(x^*)$
- f ne peut être évaluée qu'un **nombre limité de fois** N
 - évaluations non bruitées
 - gradient de f non disponible
- Choix d'une **approche bayésienne** pour résoudre le problème
 $\Rightarrow f$ est vue comme une **réalisation d'un processus aléatoire** ξ

Optimisation bayésienne 2/2

- Après n évaluations
 - information disponible** : $\mathcal{F}_n = (X_1, \xi(X_1), \dots, X_n, \xi(X_n))$
 - objectif** : maximiser $M_N = \xi(X_1) \vee \dots \vee \xi(X_N)$
- Nouveau point d'évaluation X_{n+1} choisi à l'aide d'un **critère d'échantillonnage** $\bar{\rho}_n$ (construit à partir de \mathcal{F}_n)

$$X_{n+1} = \operatorname{argmax}_{x \in \mathbb{X}} \bar{\rho}_n(x),$$

- $\bar{\rho}_n$ peut être évalué rapidement
- Exemples de critères :
 - probabilité d'amélioration [Kushner 1964]
 - amélioration moyenne (Expected Improvement) [Mockus et al. 1978]
 - entropie de l'optimiseur (IAGO) [Villemonteix et al. 2008]

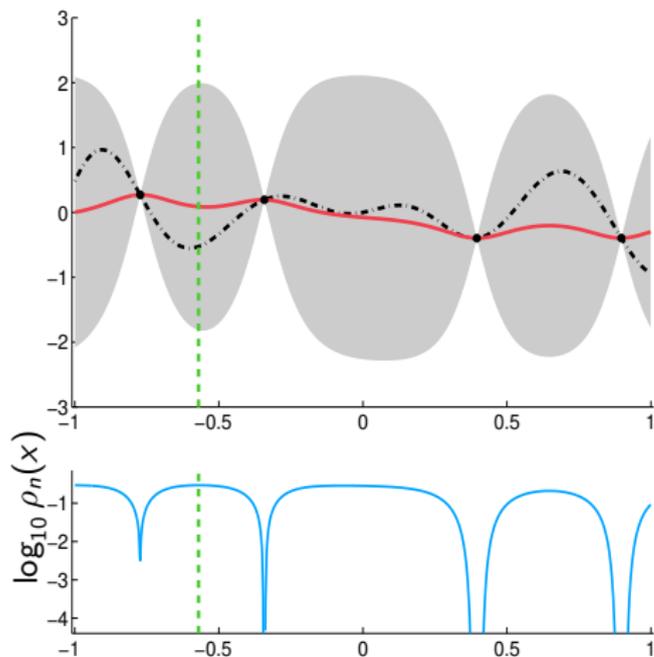
Expected Improvement (EI)

- Stratégie optimale sur un horizon d'un pas :

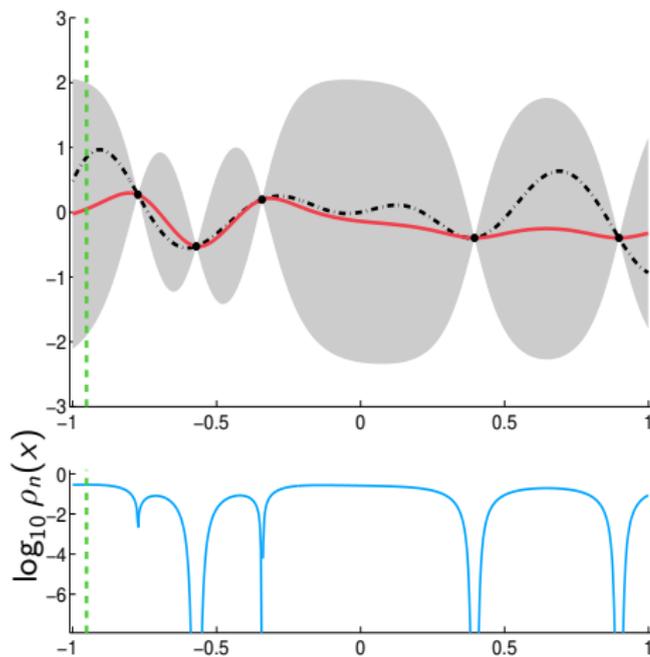
$$\begin{aligned}
 X_{n+1} &= \operatorname{argmin}_{x \in \mathbb{X}} \mathbb{E}_n [M - M_{n+1} \mid X_{n+1} = x] \\
 &= \operatorname{argmax}_{x \in \mathbb{X}} \underbrace{\mathbb{E}_n \left[(\xi(X_{n+1}) - M_n)_+ \mid X_{n+1} = x \right]}_{\text{Expected Improvement (EI) [Mockus et al. 78]}},
 \end{aligned}$$

avec \mathbb{E}_n l'espérance conditionnelle relative à \mathcal{F}_n

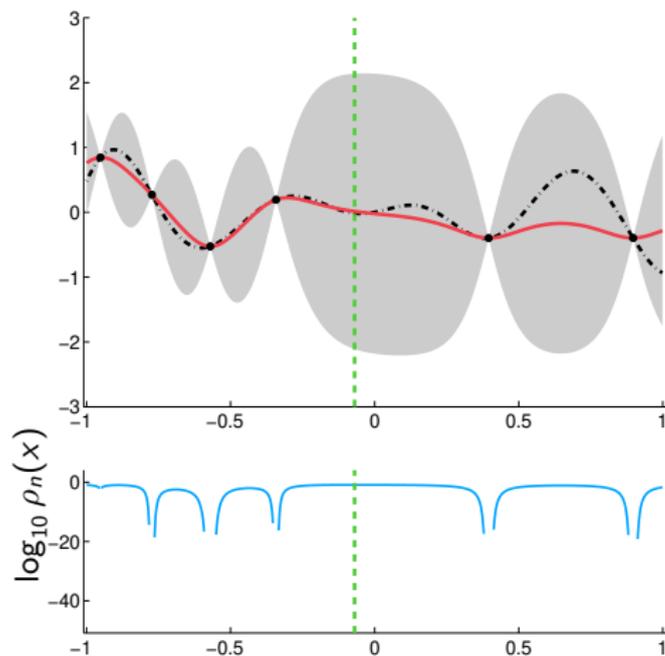
Exemple : itération 1



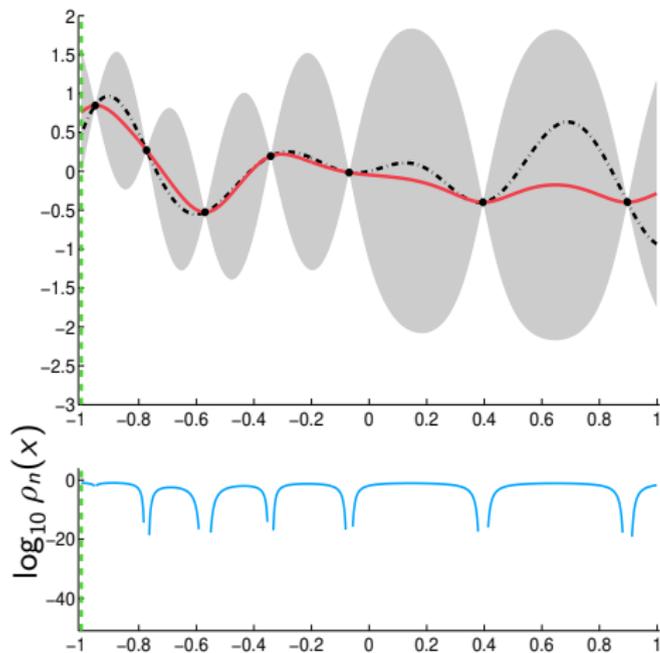
Exemple : itération 2



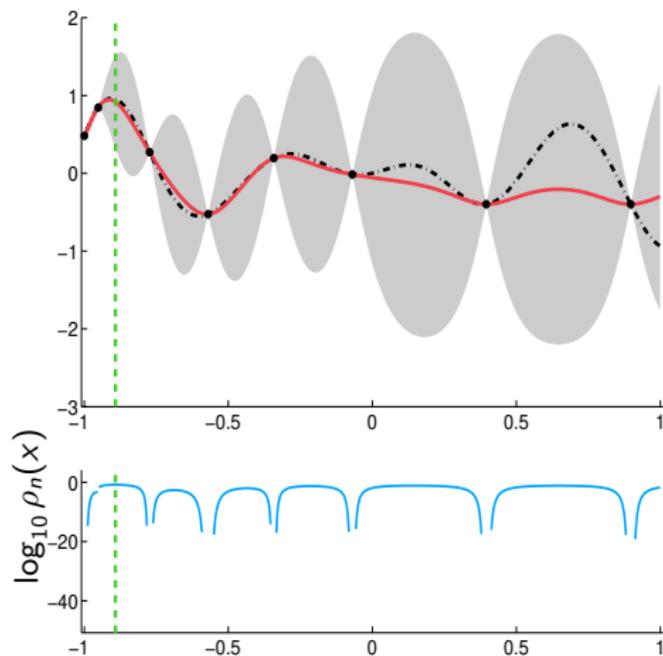
Exemple : itération 3



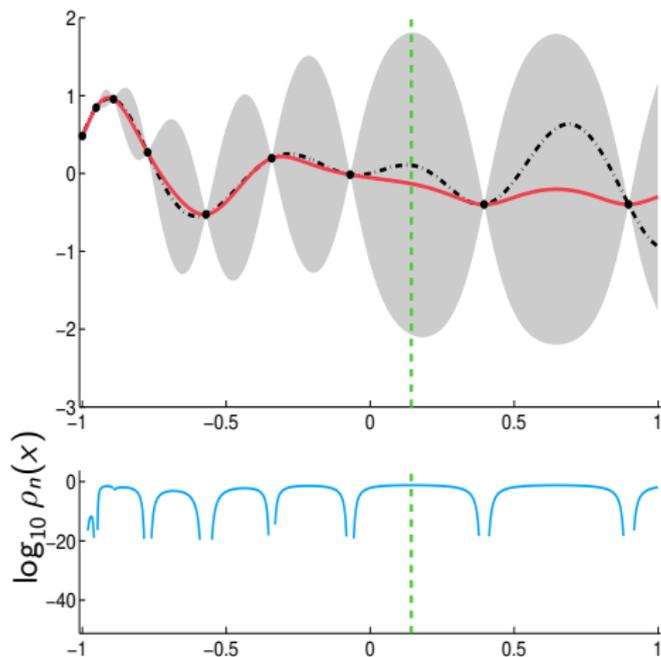
Exemple : itération 4



Exemple : itération 5



Exemple : itération 6



- 1 Introduction
- 2 Approche plug-in
- 3 Approche complètement bayésienne
- 4 Expériences numériques
- 5 Conclusions

- 1 Introduction
- 2 Approche plug-in**
- 3 Approche complètement bayésienne
- 4 Expériences numériques
- 5 Conclusions

Problème : Comment choisir ξ ?

\Rightarrow processus ξ choisi dans une classe de processus gaussiens paramétrés par $\theta \in \Theta$ (p.ex. processus avec moyenne constante mais inconnue et covariance exponentielle avec portée inconnue)

Approche *plug-in* (p.ex. EGO [Jones et al. 1998])

- θ est estimé par **maximum de vraisemblance**
- El *plug-in* :

$$\rho_n(x; \tilde{\theta}) = \mathbb{E}_n \left(\left(\xi(x; \tilde{\theta}) - M_n \right)_+ \right),$$

avec $\tilde{\theta}$ estimé par MV

- Avantage : existence d'une expression **analytique** de l'El pour ξ **gaussien**

El et processus gaussiens

- Si ξ est gaussien (θ fixé) :

$$\rho_n(x; \theta) = \mathbb{E}_n \left((\xi(x) - M_n)_+ \mid \theta \right) = s_n(x) [u\Phi(u) + \Phi'(u)]$$

avec

- $u = (\hat{\xi}_n - M_n) / s_n(x)$
- $\hat{\xi}_n(x) = \mathbb{E}_n [\xi(x; \theta)]$ (prédicteur par krigeage)
- $s_n(x)$ écart type *a posteriori* de $\hat{\xi}_n(x) - \xi(x; \theta)$
- Φ fonction de répartition de la loi normale centrée réduite

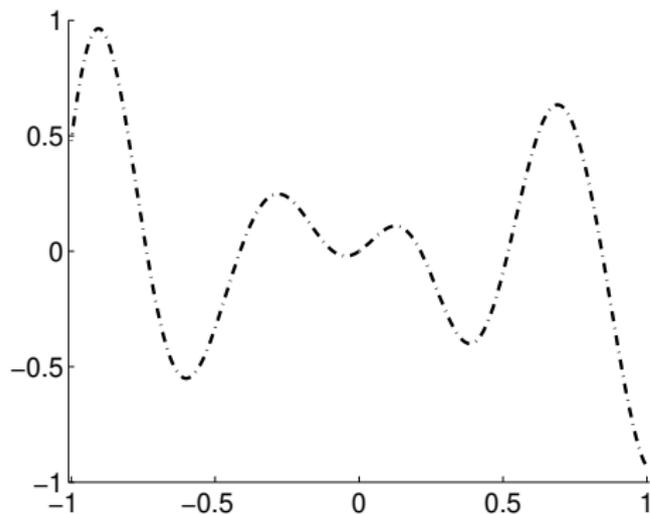
Il est facile de calculer ρ_n !

Problème avec l'approche plug-in

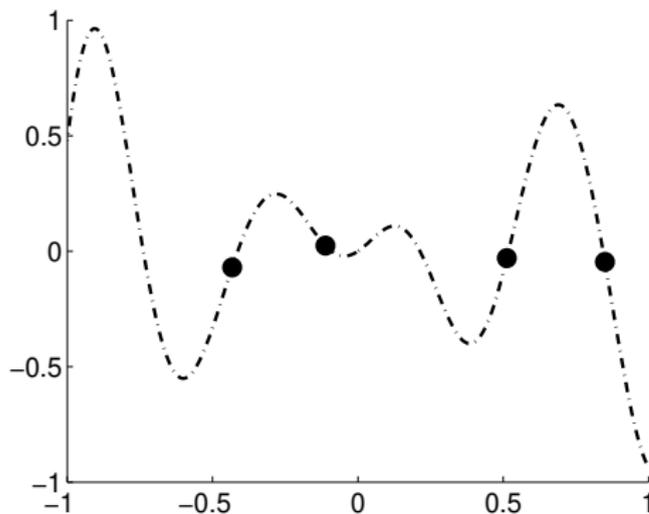
Fonctions trompeuses

- Le terme **fonctions trompeuses** caractérise une fonction pour laquelle un *design initial* mal choisi peut conduire à une très mauvaise estimation des paramètres du processus par MV
- L'utilisation d'une approche *plug-in* conduit parfois à sous-estimer largement l'erreur de prédiction :
 - ⇒ **problème** pour l'optimisation par EI

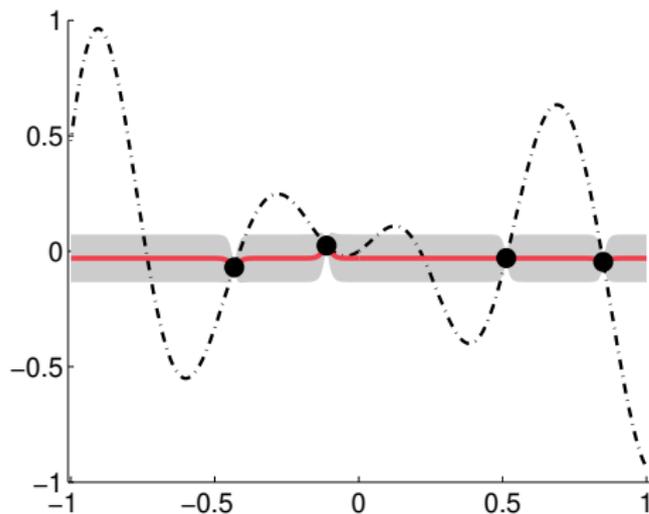
Exemple : Fonction trompeuse



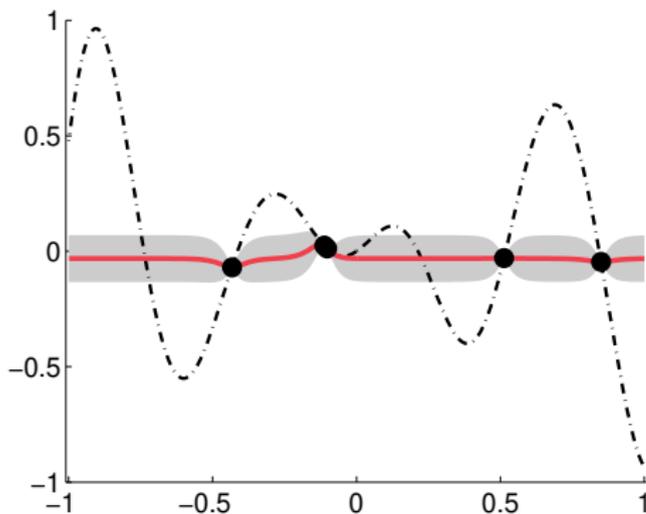
Exemple : Fonction trompeuse



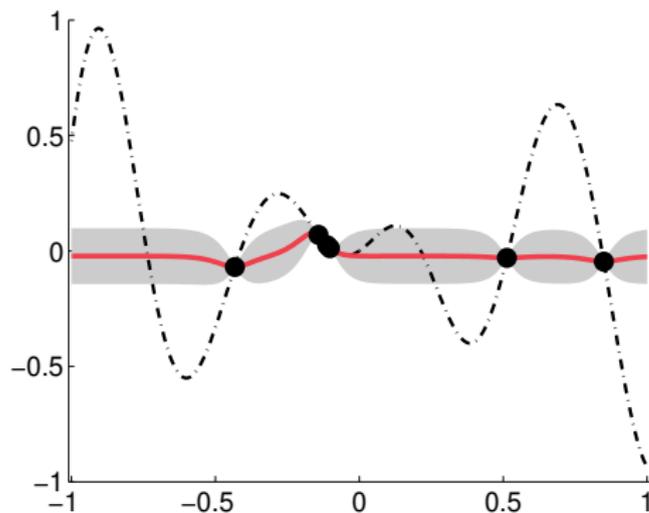
Exemple : Fonction trompeuse



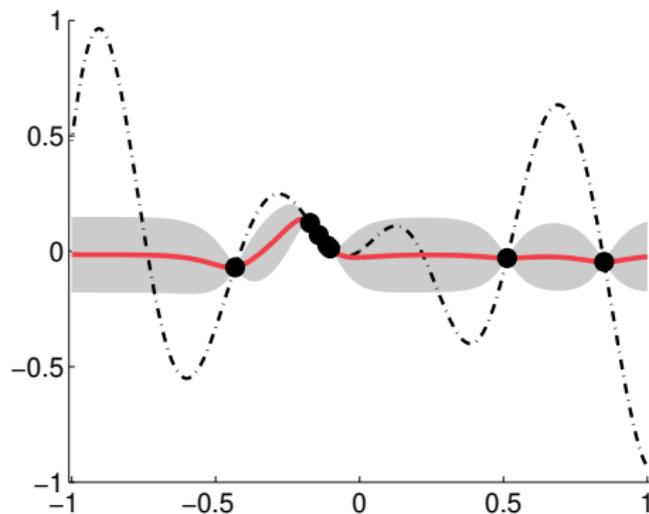
Exemple : Fonction trompeuse



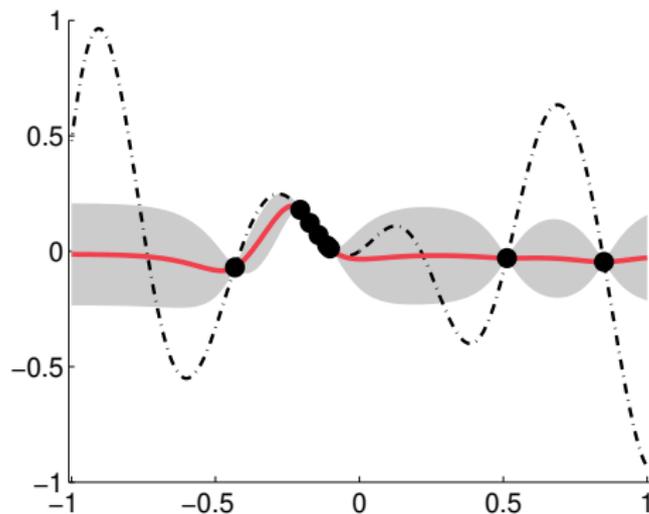
Exemple : Fonction trompeuse



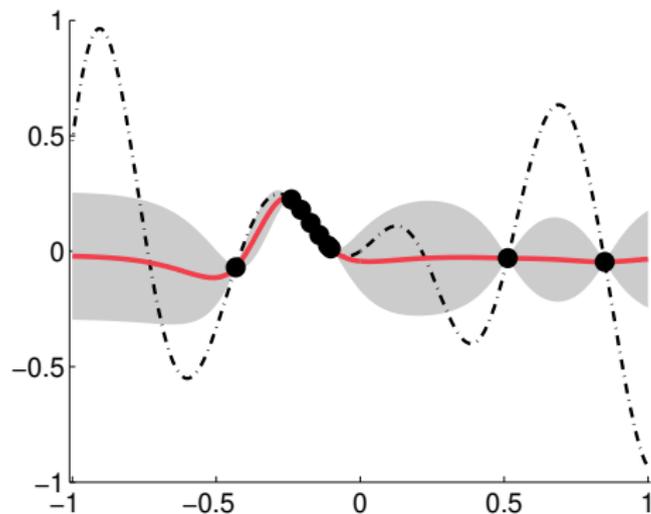
Exemple : Fonction trompeuse



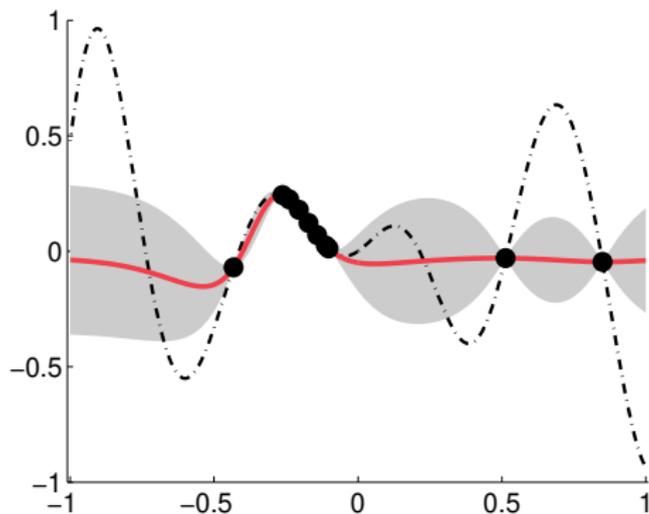
Exemple : Fonction trompeuse



Exemple : Fonction trompeuse



Exemple : Fonction trompeuse



- 1 Introduction
- 2 Approche plug-in
- 3 Approche complètement bayésienne**
- 4 Expériences numériques
- 5 Conclusions

El complètement bayésien

- Utilisons d'un critère El complètement bayésien [Williams et al. 2003, Osborne et al. 2009,...]
 ⇒ choix d'un **a priori** sur θ

- Critère d'échantillonnage El complètement bayésien :

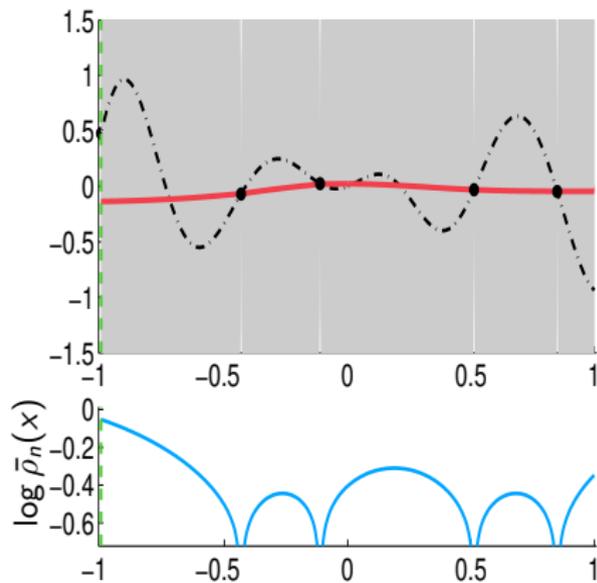
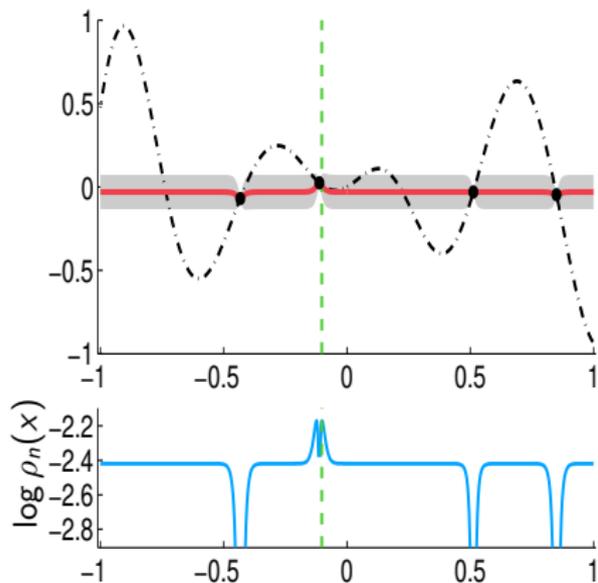
$$X_{n+1} = \operatorname{argmax}_{x \in \mathbb{X}} \bar{\rho}_n := \int_{\theta \in \Theta} \rho_n(x; \theta) d\pi_n(\theta),$$

où π_n est la loi *a posteriori* de θ , et

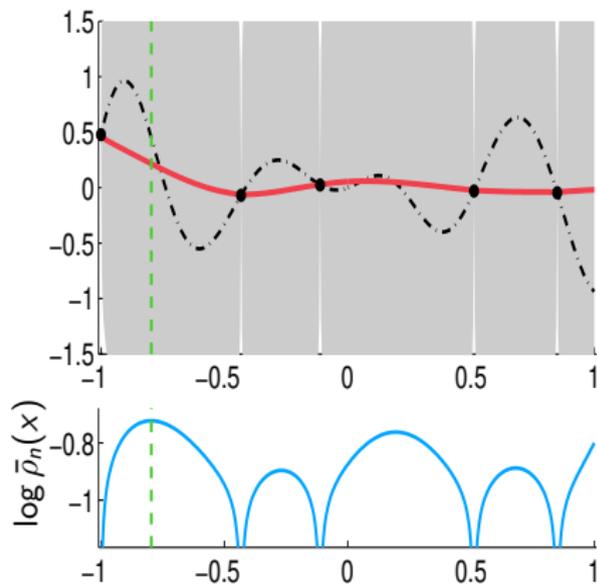
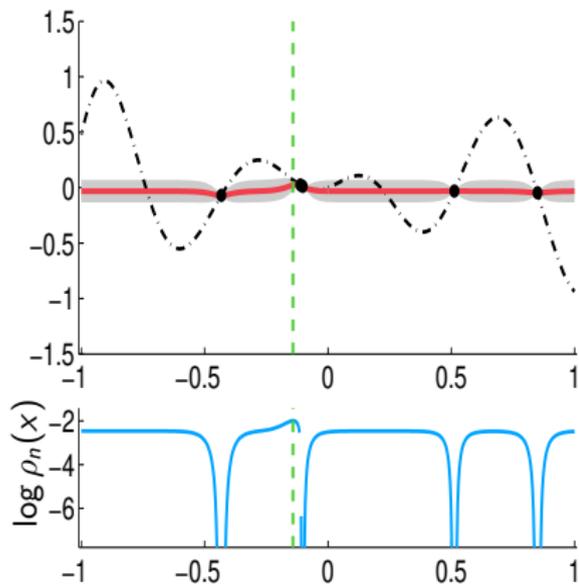
$$\rho_n(x; \theta) := \mathbb{E}_n((\xi(X_{n+1}) - M_n)_+ | X_{n+1} = x; \theta),$$

- Approche plus robuste qu'une approche *plug-in* [Benassi et al. 2011]

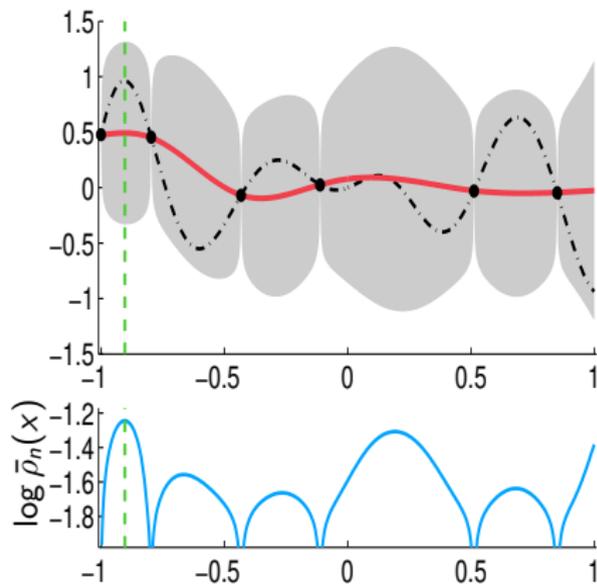
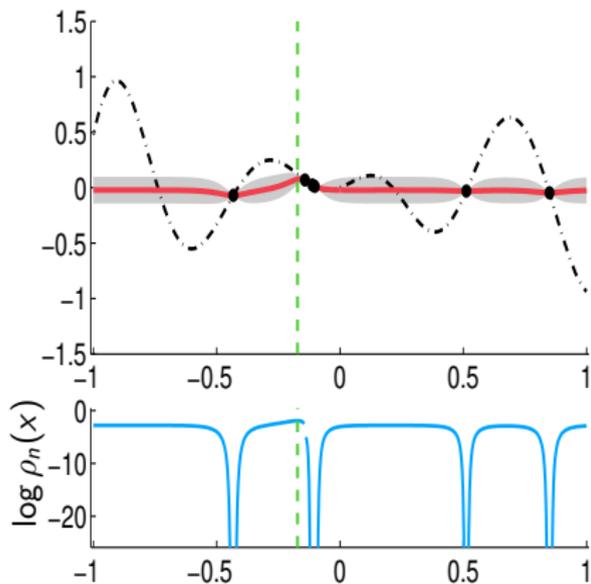
Exemple : itération 1



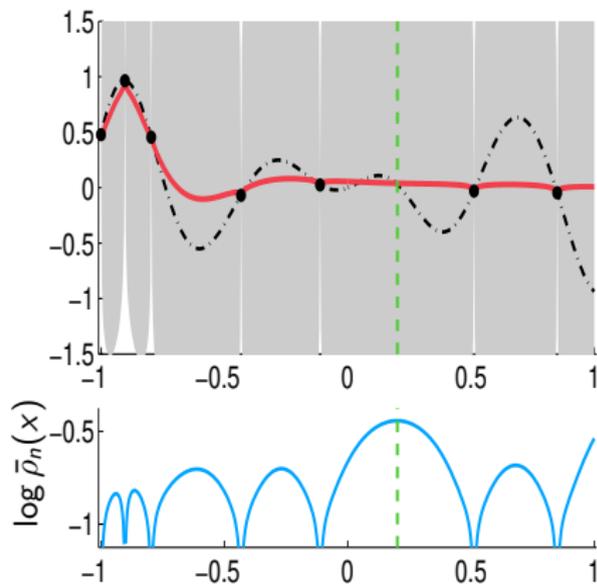
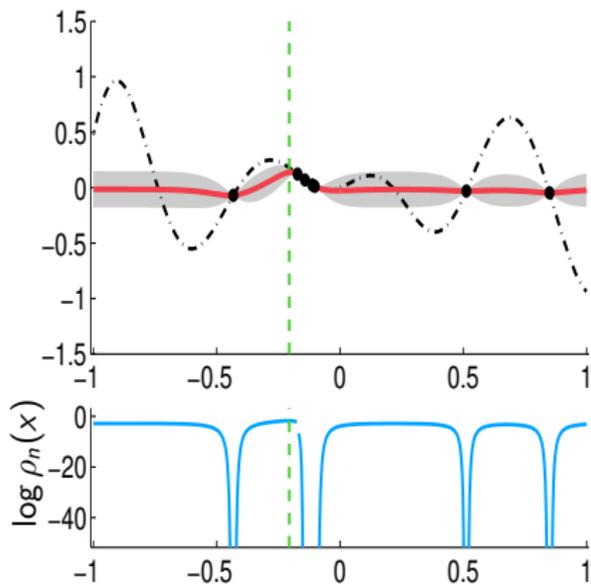
Exemple : itération 2



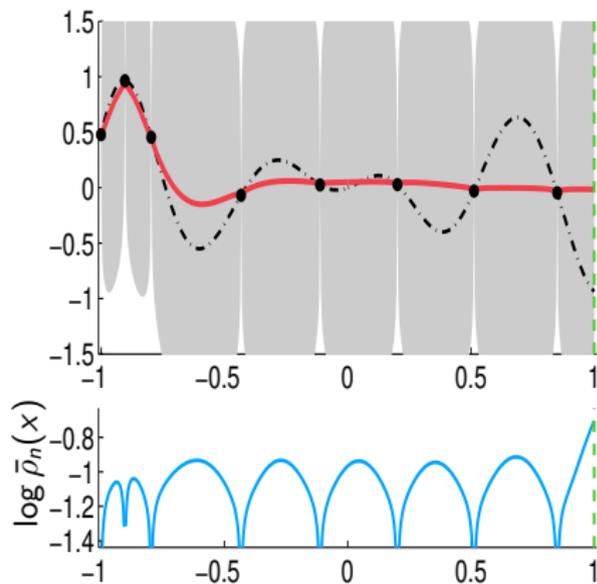
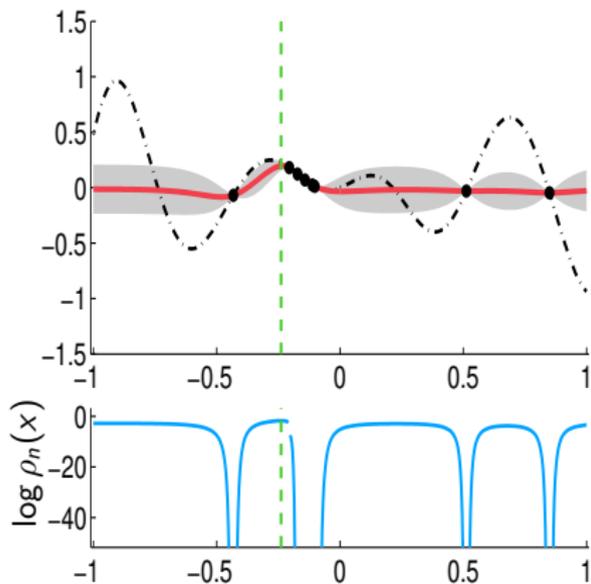
Exemple : itération 3



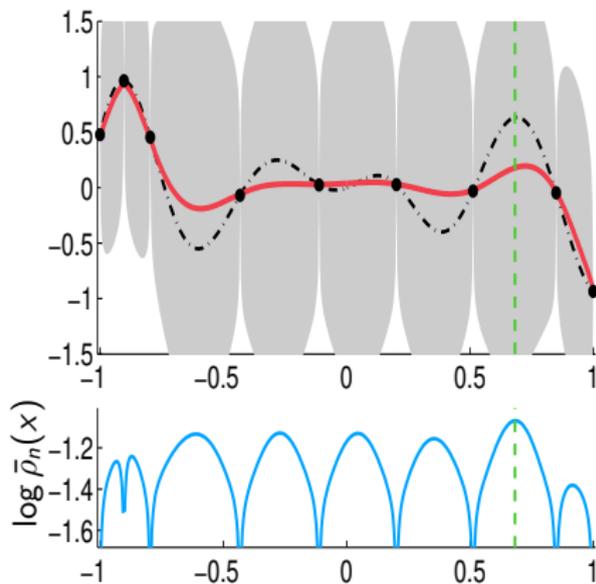
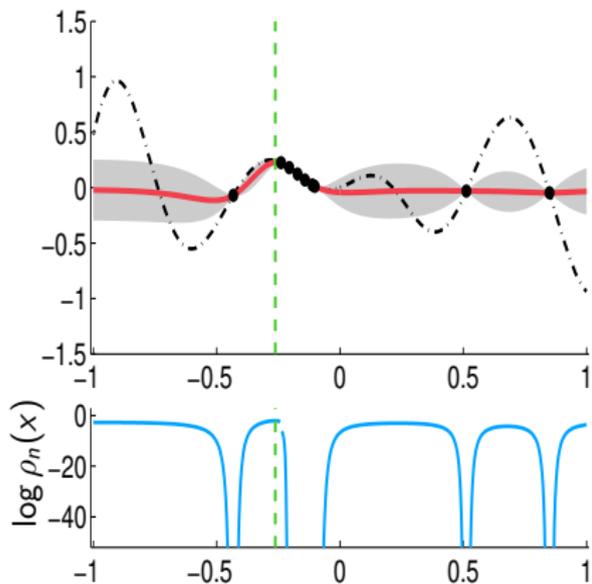
Exemple : itération 4



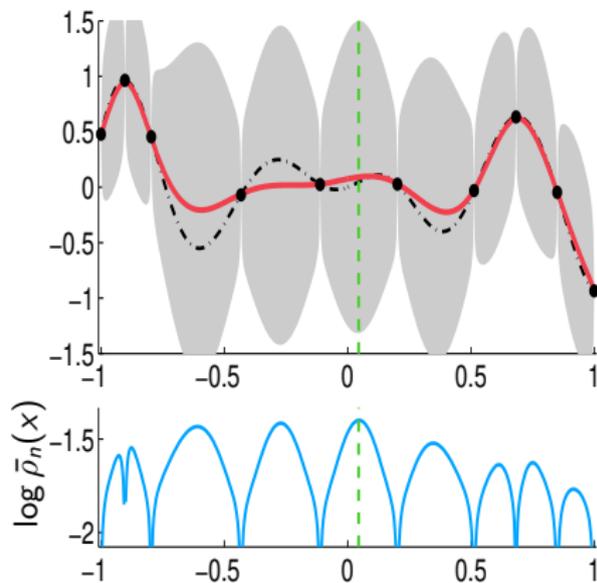
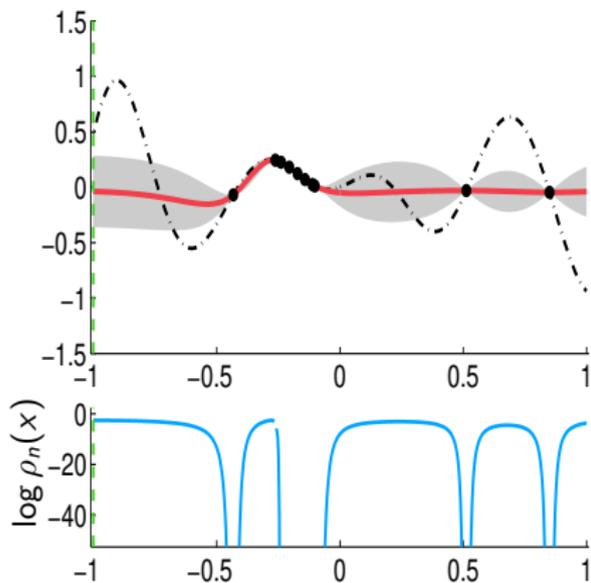
Exemple : itération 5



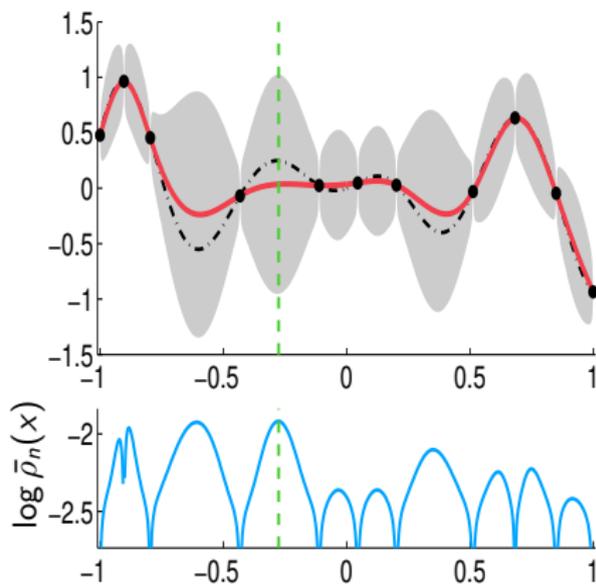
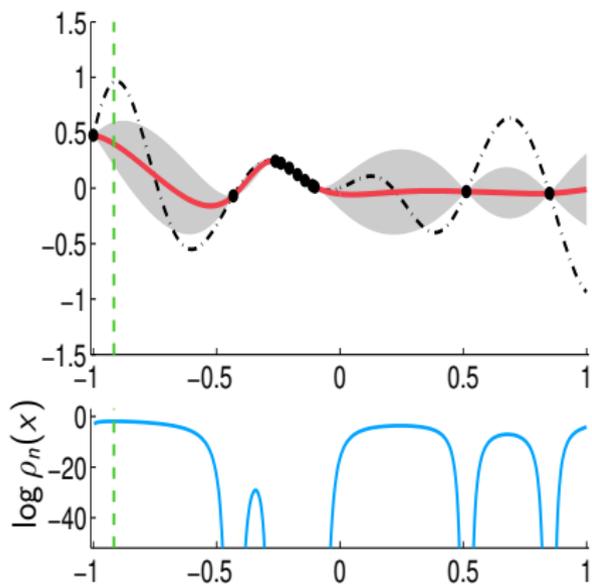
Exemple : itération 6



Exemple : itération 7



Exemple : itération 8



Limites de l'approche complètement bayésienne

À chaque étape, on choisit une nouvelle évaluation

$$X_{n+1} = \operatorname{argmax}_{x \in \mathbb{X}} \bar{\rho}_n := \int_{\theta \in \Theta} \rho_n(x; \theta) d\pi_n(\theta)$$

Questions

- 1 Comment calculer l'intégrale apparaissant dans $\bar{\rho}_n$?
- 2 Comment maximiser $\bar{\rho}_n$?

Nous **proposons** un nouvel algorithme apportant une réponse à ces deux questions, de **façon simultanée**, à l'aide d'une approche **SMC** (*Sequential Monte Carlo*) originale

Limites de l'approche complètement bayésienne

À chaque étape, on choisit une nouvelle évaluation

$$X_{n+1} = \operatorname{argmax}_{x \in \mathbb{X}} \bar{\rho}_n := \int_{\theta \in \Theta} \rho_n(x; \theta) d\pi_n(\theta)$$

Questions

- 1 Comment calculer l'intégrale apparaissant dans $\bar{\rho}_n$?
- 2 Comment maximiser $\bar{\rho}_n$?

Nous **proposons** un nouvel algorithme apportant une réponse à ces deux questions, de **façon simultanée**, à l'aide d'une approche **SMC** (*Sequential Monte Carlo*) originale

Limites de l'approche complètement bayésienne

À chaque étape, on choisit une nouvelle évaluation

$$X_{n+1} = \operatorname{argmax}_{x \in \mathbb{X}} \bar{\rho}_n := \int_{\theta \in \Theta} \rho_n(x; \theta) d\pi_n(\theta)$$

Questions

- 1 Comment calculer l'intégrale apparaissant dans $\bar{\rho}_n$?
- 2 Comment maximiser $\bar{\rho}_n$?

Nous **proposons** un nouvel algorithme apportant une réponse à ces deux questions, de **façon simultanée**, à l'aide d'une approche **SMC** (*Sequential Monte Carlo*) originale

Principe de base de notre algorithme [Benassi et al. 2012]

Idée 1 [voir p.ex. Gramacy et al. 2011]

Un échantillon pondéré $\mathfrak{T}_n = \{(\theta_{n,i}, w_{n,i}) \in \Theta \times \mathbb{R}, 1 \leq i \leq I\}$, distribué selon π_n , est utilisé pour approcher $\bar{\rho}_n$:

$$\sum_{i=1}^I w_{n,i} \rho_n(x; \theta_{n,i}) \rightarrow_I \bar{\rho}_n(x)$$

Idée 2

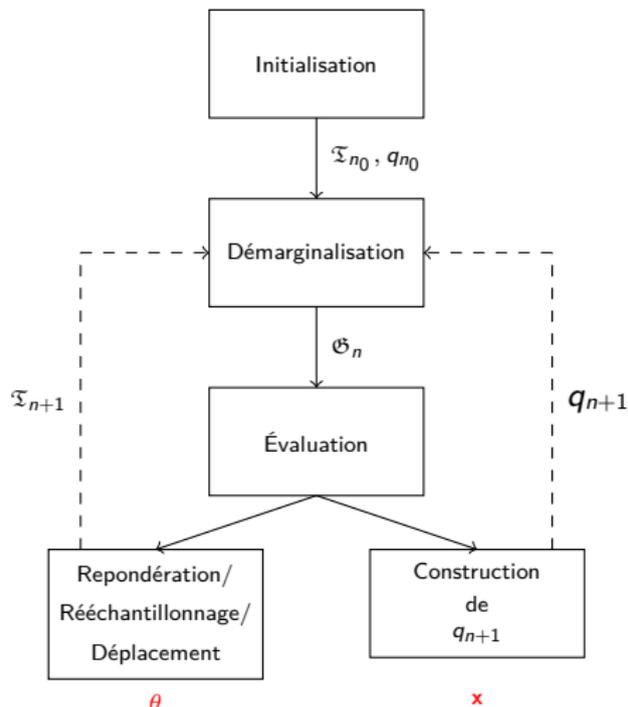
À chaque $\theta_{n,i}$ est associé un (petit) nombre de points candidats $\{x_{n,i,j}, 1 \leq j \leq J\}$ tels que, pour une valeur de i particulière, les $x_{n,i,j}$ soient « bien réparties » (les zones où ρ_n est grand)

Nous considérons un ensemble pondéré \mathfrak{G}_n de couples (θ, x)

Description de l'algorithme

À l'étape n ,

- Choix de X_{n+1} et évaluation de $\xi(X_{n+1})$
- Les θ s sont « déplacés »
- Génération des x s à partir d'une loi instrumentale q_n



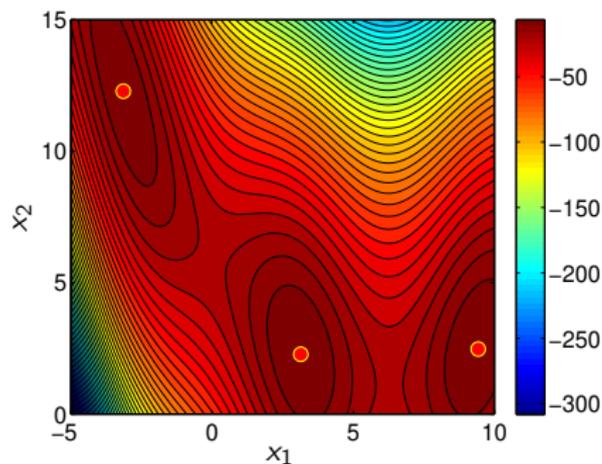
- 1 Introduction
- 2 Approche plug-in
- 3 Approche complètement bayésienne
- 4 Expériences numériques**
- 5 Conclusions

Choix des *a priori*

- ξ est choisi gaussien avec :
 - une moyenne constante mais inconnue
 - une covariance de **Matérn** anisotrope de paramètres $(\sigma^2, \nu, \rho_1, \rho_2, \dots, \rho_d)$
- Approche **complètement bayésienne** \Rightarrow un *a priori* doit être affecté à **tous** les paramètres inconnus
 - une loi **impropre uniforme** sur \mathbb{R} est associée à la moyenne
 - la valeur de régularité ν est **fixée** à $5/2$
 - un ***a priori* de Jeffreys** est utilisé pour σ^2
 - des ***a priori* lognormaux** indépendants sont affectés aux portées ρ_i

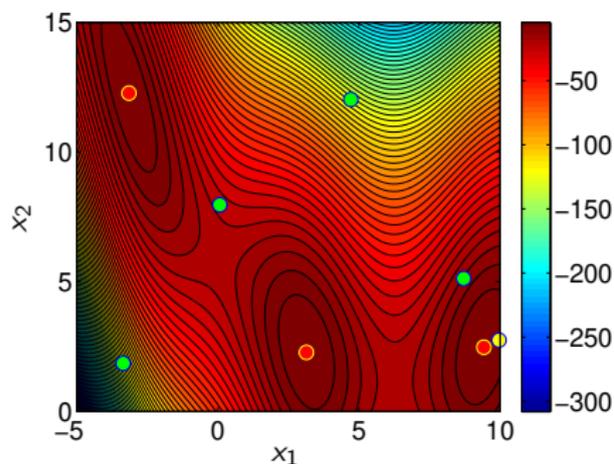
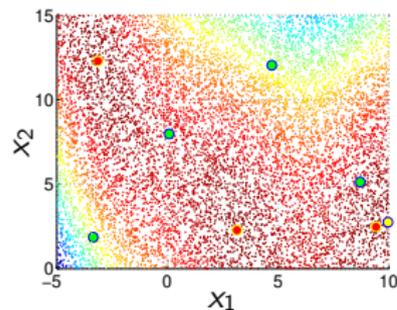
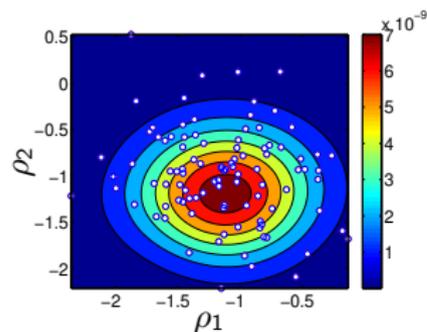
La fonction test

Fonction Branin (optimisation en dimension deux)



Simulations : itération 1

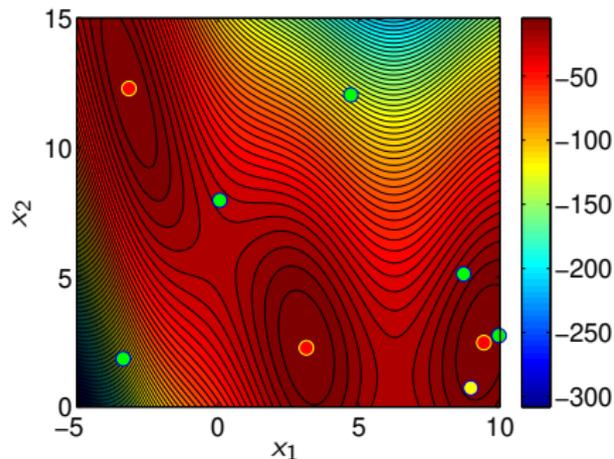
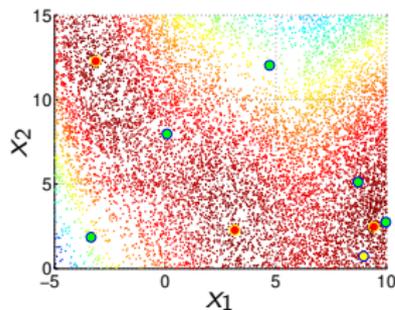
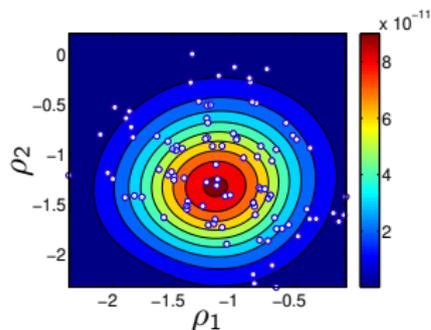
Fonction Branin

répartition des θ s et des x s

- plan d'expérience initial de 4 points
- nombre de θ : 100
- 100 \times pour chaque θ

Simulations : itération 2

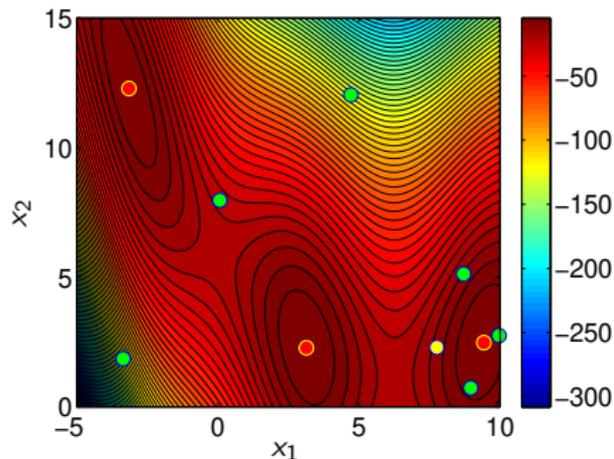
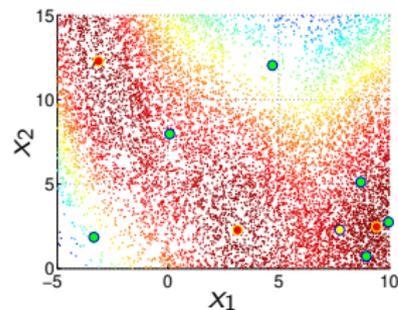
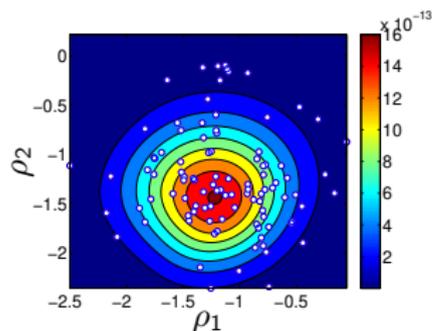
Fonction Branin

répartition des θ s et des x s

- plan d'expérience initial de 4 points
- nombre de θ : 100
- 100 \times pour chaque θ

Simulations : itération 3

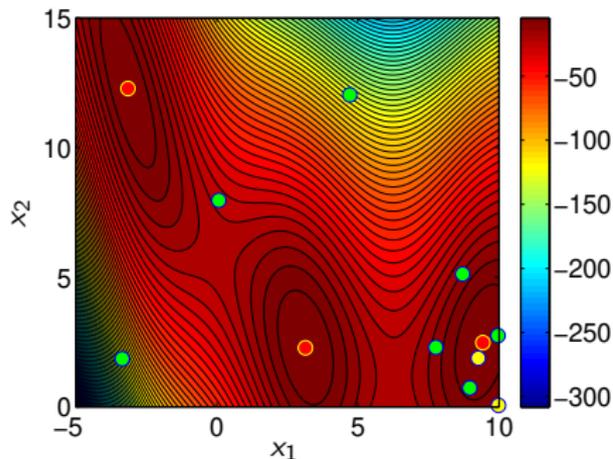
Fonction Branin

répartition des θ s et des x s

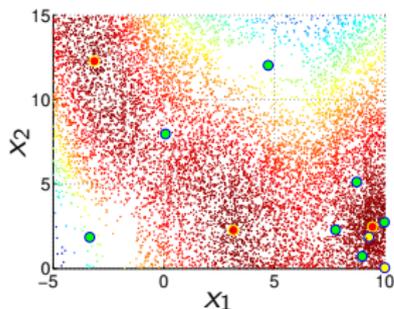
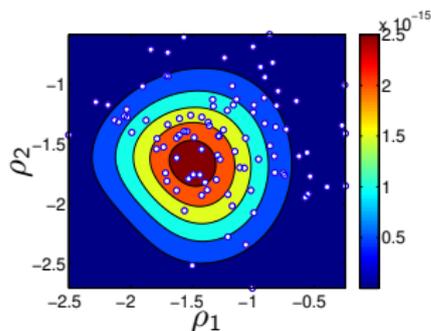
- plan d'expérience initial de 4 points
- nombre de θ : 100
- 100 \times pour chaque θ

Simulations : itération 5

Fonction Branin

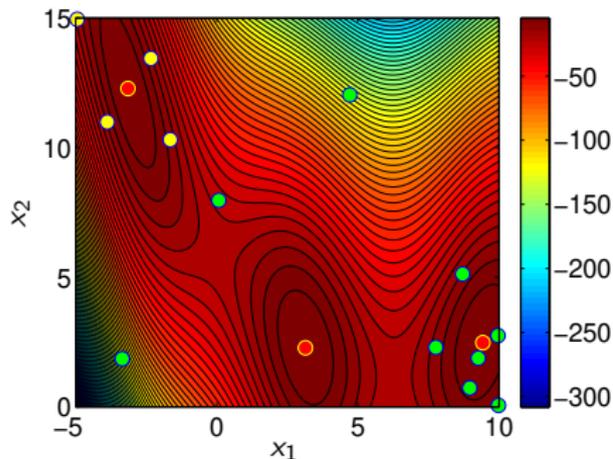
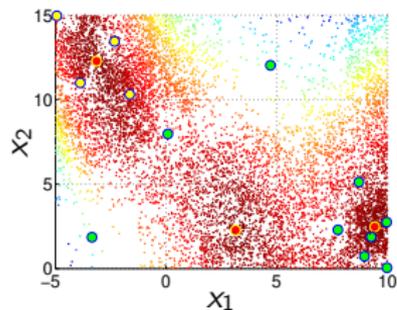
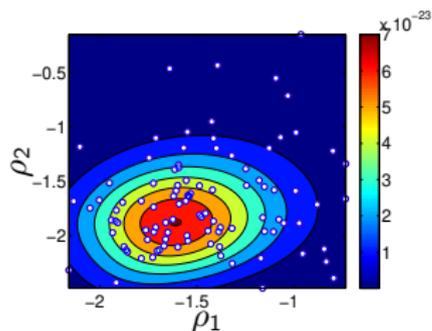


- plan d'expérience initial de 4 points
- nombre de θ : 100
- 100 \times pour chaque θ

répartition des θ s et des x s

Simulations : itération 9

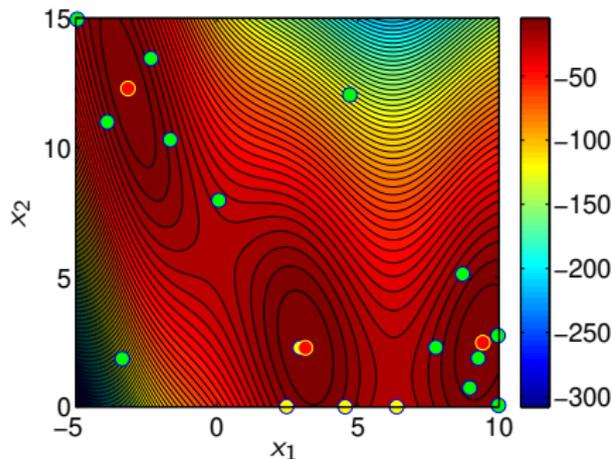
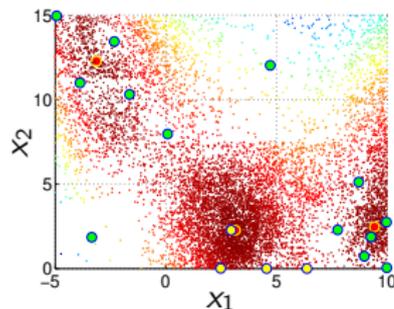
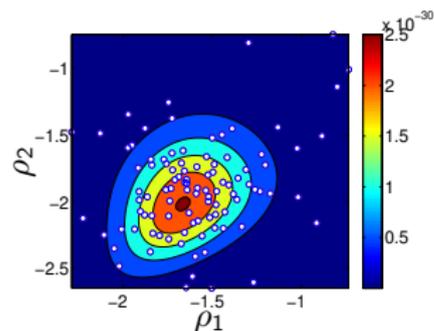
Fonction Branin

répartition des θ s et des x s

- plan d'expérience initial de 4 points
- nombre de θ : 100
- 100 \times pour chaque θ

Simulations : itération 13

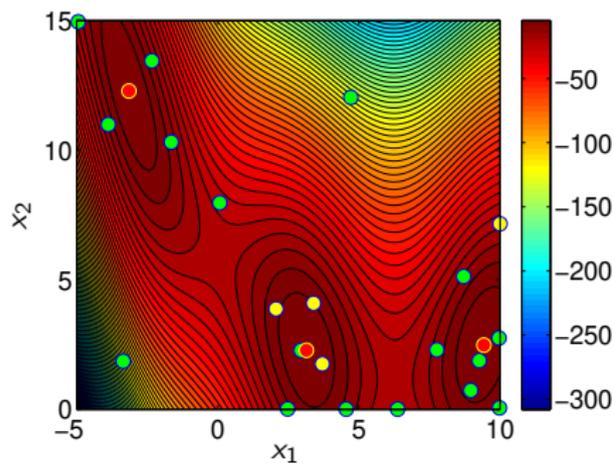
Fonction Branin

répartition des θ s et des x s

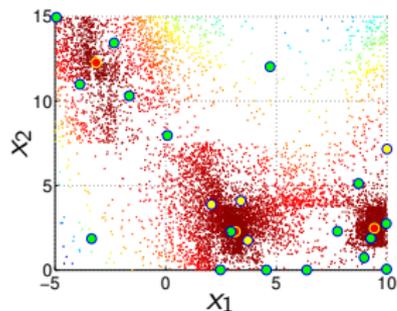
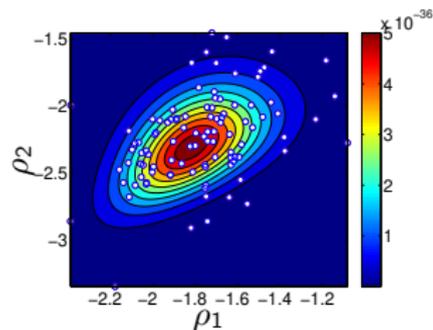
- plan d'expérience initial de 4 points
- nombre de θ : 100
- 100 \times pour chaque θ

Simulations : itération 17

Fonction Branin

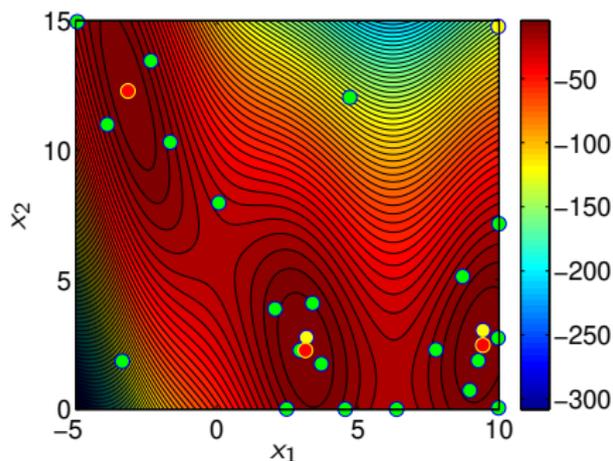


- plan d'expérience initial de 4 points
- nombre de θ : 100
- 100 \times pour chaque θ

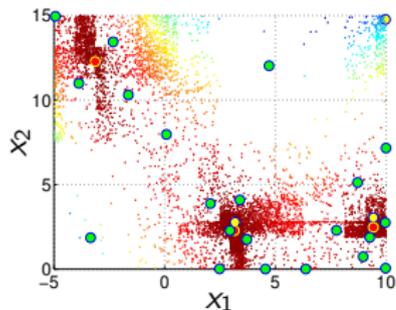
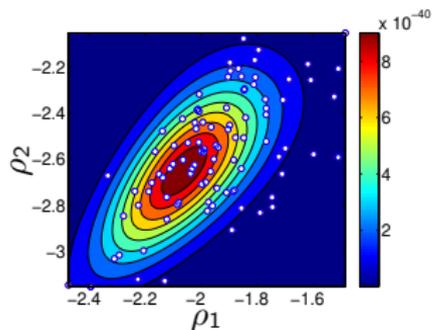
répartition des θ s et des x s

Simulations : itération 21

Fonction Branin



- plan d'expérience initial de 4 points
- nombre de θ : 100
- 100 \times pour chaque θ

répartition des θ s et des x s

Comparaison 1/2

Comparaison entre approche **bayésienne** et approche **plug-in**

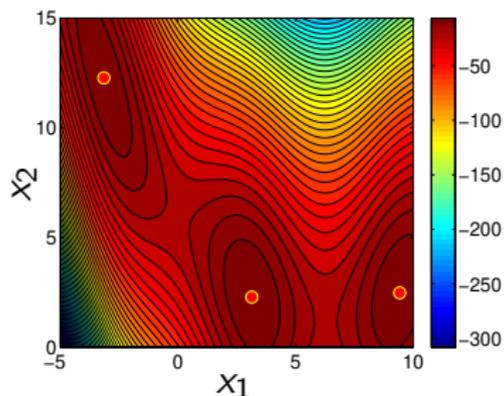
Approche bayésienne

- $I = 100$ valeurs considérées pour les θ
- $J = 100$ valeurs de x pour chaque valeur de θ
- *design initial* de 4 points (deux fois la dimension)

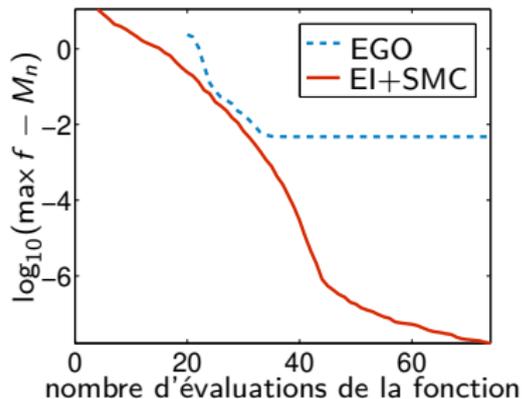
Approche *plug-in*

- θ estimé à chaque itération par MV
- recherche exhaustive sur LHS fixé pour obtenir X_{n+1}
- LHS de taille $I \times J$
- *design initial* de 20 points (dix fois la dimension)

Comparaison 2/2



Fonction Branin



Comparaison de l'erreur moyenne
au maximum (moyenne sur
100 expériences)

- 1 Introduction
- 2 Approche plug-in
- 3 Approche complètement bayésienne
- 4 Expériences numériques
- 5 Conclusions**

Conclusions

Resultats

- Notre nouvel **algorithme complètement bayésien** calcule **efficacement** le critère El
- Le choix, à chaque nouvelle itération, d'un ensemble (petit mais bien choisi) de points candidats permet une maximisation **efficace** du critère

Travail en cours

- Tests en dimensions supérieures (bons résultats pour les dimensions de 1 à 6)
- Applications à des problèmes d'optimisation dans le domaine de la conception de systèmes électroniques