Optimisation de fonctions coûteusesModèles gaussiens pour une utilisation efficace du budget d'évaluations : théorie et pratique industrielle

Julien Villemonteix

Faculté des sciences d'Orsay

UNIVERSITÉ PARIS-SUD 11

N° D'ORDRE : 9278

UNIVERSITÉ PARIS-SUD XI

Faculté des Sciences d'Orsay

THÈSE DE DOCTORAT

Spécialité : physique

École Doctorale « Sciences et Technologies de l'Information, des Télécommunications et des Systèmes »

Présentée par : Julien VILLEMONTEIX

OPTIMISATION DE FONCTIONS COÛTEUSES

Modèles gaussiens pour une utilisation efficace du budget d'évaluations : théorie et pratique industrielle

Soutenue le 10 décembre 2008 devant les membres du jury :

M.	JONES	Donald	Rapporteur
M.	PRONZATO	Luc	Rapporteur
M.	LANGELLE	Régis	Président
M.	TEYTAUD	Olivier	Examinateur
M.	SIDORKIEWICZ	Maryan	Encadrant
M.	VAZQUEZ	Emmanuel	Encadrant
M.	WALTER	Éric	Directeur de thèse

Remerciements

Ca y est ! Le manuscrit est terminé tout comme ces trois années de thèse. Tout s'est déroulé admirablement bien, et il serait dommage de trébucher au dernier obstacle par un oubli malencontreux. Aussi vais-je remercier, d'un bloc, tous ceux que j'ai de près ou de loin approchés durant ces trois années. Ces dernières ont été très agréables, et toutes ces personnes y ont contribué. Je distribuerai néanmoins les mentions particulières suivantes.

Je souhaiterais en premier lieu exprimer ma gratitude à mes trois encadrants qui ont su définir conjointement les objectifs de cette thèse et me créer un petit nid douillet pour y travailler. Merci donc à Maryan pour son soutien sans faille dans la jungle du Technocentre, merci à Emmanuel pour ses sages conseils et sa patience, et enfin merci à Éric pour sa rigueur, son recul et son soutien.

Je remercie les membres de mon jury pour leur lecture attentive de ce manuscrit et leurs appréciations favorables.

Je remercie également le département SSE de Supélec et le groupe optimisation de Renault qui m'ont offert un cadre de travail très sympathique. Au sein de ces derniers, je remercie en particulier tous ceux qui sont devenus plus que des collègues. Jean-François, Jerôme, Julien, Marine, Morgan, Rany, Sylvain, merci à tous !

Enfin, je dédie ce travail à mes amis, à ma famille, et tout particulièrement à Juliette.

Table des matières

Introduction

Concevoir à l'aide de simulations coûteuses

Ce mémoire traite de l'optimisation de systèmes dont l'évaluation, au travers de mesures ou de simulations numériques, est coûteuse. Ce coût, qui peut être de nature financière, temporelle, humaine ou un mélange des trois, implique un budget d'évaluations très réduit pour la résolution du problème. Une fois ce budget dépensé, une solution doit être proposée, indépendamment de la complexité du problème rencontré.

Les exemples de problèmes d'optimisation soumis à cette contrainte ne manquent pas dans la pratique. Ils sont monnaie courante dans les industries automobile et aéronautique, où les systèmes à optimiser le sont généralement à l'aide de simulations numériques. Par exemple, la résistance des véhicules au crash peut être testée par des essais réels sur véhicules, mais aussi simulée, ce qui demande plus d'un millier d'heures CPU. Dans les deux cas, le nombre d'expériences réalisables pour l'optimisation de cette résistance ne dépasse pas quelques dizaines, et il doit permettre de choisir plus d'une centaine de paramètres (principalement des épaisseurs). Au-delà de cet exemple symptomatique, le nombre de paramètres n'excède généralement pas quelques dizaines pour un coût d'évaluation de quelques heures, voir par exemple Huang (2005) ; Dvorak et al. (2003) ; Giannakoglou (2002) ; Huang et Allen (2005) ; Villemonteix et al. (2008a). On retrouve cette problématique en biochimie (Davies et al., 2000 ; Weuster-Botz et C., 1995), lors de la conception de protocoles de mesures (Vaidyanathan et al., 2004 ; O'Hagan et al., 2004) et dans bien d'autres contextes.

Les travaux présentés dans ce mémoire, s'ils sont applicables à tous les problèmes mentionnés précédemment, ont néanmoins été guidés par un problème bien particulier, qui va servir d'exemple dans cette introduction, l'optimisation de la forme de composants de moteurs thermiques dans l'industrie automobile. Dans ce contexte, l'évaluation des fonctions à optimiser requiert plusieurs heures de calcul (simulation de l'écoulement fluide dans ces composants) et le résultat de cette évaluation est sujet à un bruit numérique. On trouve aussi d'autres difficultés liées à la nature du marché automobile. Les moteurs doivent en effet satisfaire à la fois au goût des consommateurs pour des voitures puissantes et à des normes sur les émissions de polluants de plus en plus strictes. Ces besoins simultanés se traduisent par des problèmes d'optimisation multi-objectifs qui, pour

ne rien arranger, doivent être résolus en des temps très courts du fait du rythme toujours croissant avec lequel de nouveaux moteurs doivent être conçus.

Une autre difficulté qu'il convient de mentionner (et ce sera la dernière), inhérente à la production de masse, est qu'il faut pouvoir tenir compte des erreurs commises lors de la fabrication des composants. Ces incertitudes de fabrication peuvent en effet avoir une influence significative sur le fonctionnement du système. Plus précisément, considérons un moteur conçu pour que ses émissions de monoxyde de carbone (CO) soient inférieures à la limite autorisée, sans tenir compte des incertitudes de fabrication. Il est alors très courant de constater *a posteriori* que pour la moitié des moteurs effectivement construits, les émissions de CO sont supérieures à la norme. Pour éviter les coûts subséquents, ces incertitudes doivent être prises en compte dès le commencement du processus de conception, et il pourra alors être plus judicieux de choisir un moteur sous-optimal mais dont les émissions sont peu sensibles aux incertitudes de fabrication. Ce compromis entre l'optimisation de la performance nominale et la minimisation de la variabilité associée est appelée *optimisation robuste* par rapport aux incertitudes de fabrication.

En résumé, les problèmes considérés sont multi-objectifs, de dimension (nombre de paramètres à optimiser) relativement élevée, et doivent être traités de manière robuste en utilisant un nombre très faible d'évaluations. De plus, même si de l'information *a priori* est souvent disponible sur la fonction à optimiser (dérivabilité, majorant et minorant), son gradient n'est en général pas disponible, et elle peut présenter de nombreux optima locaux. L'objectif de ces travaux est de montrer que devant de telles difficultés, qui sont celles que rencontrent quotidiennement les praticiens, il est possible de proposer des méthodes d'optimisation ayant une justification théorique et un intérêt pratique. L'objectif ici n'est pas tant d'estimer précisément l'optimum global et les optimiseurs associés (tâche généralement impossible compte tenu du budget d'évaluations) que d'utiliser chacune des évaluations à bon escient. De plus, le temps passé à choisir la valeur des arguments de la fonction à optimiser lors de la prochaine évaluation n'a que peu d'importance au regard du coût demandé par l'évaluation elle-même.

Utilisation d'un modèle de substitution pour optimiser plus rapidement

Compte-tenu de la complexité des problèmes à traiter, il semble judicieux d'en extraire des sous-problèmes qui pourront être clairement formalisés, faire l'objet de contributions théoriques et servir de base à une solution générale. Commençons par l'optimisation d'une fonction réelle (problème mono-objectif) à partir d'un budget d'évaluations limité.

Supposons, sans perte de généralité, que la fonction à optimiser, par la suite appelée *fonction objectif*, doit être minimisée et appelons *espace de recherche* l'espace où varient les paramètres à optimiser (on parlera aussi de *facteurs* et d'*espace des facteurs* pour éviter la confusion avec

les paramètres des méthodes d'optimisation). La conséquence principale du coût de l'évaluation de la fonction objectif est que les méthodes classiques d'optimisation globale par exploration aléatoire (Monte-Carlo, recuit simulé, algorithmes évolutionnaires...), qui requièrent des milliers d'évaluations, sont inutilisables.

Depuis plus d'un demi siècle (Box et Wilson, 1951), la solution généralement retenue pour diminuer le nombre des évaluations de la fonction objectif est la construction d'un modèle de celle-ci dont l'évaluation soit peu coûteuse. Ce modèle, construit à partir d'un nombre limité de résultats d'évaluations choisies avec soin (on parle alors de *plan d'expériences*), va guider la recherche du minimum global et du ou des minimiseurs associés. Pour éviter la confusion entre ce modèle et la fonction objectif (qui repose elle même souvent sur un modèle physique d'un phénomène), d'autres terminologies ont été proposées. On parle ainsi de surfaces de réponse, de métamodèles, d'émulateurs, de modèles boîte noire, de modélisation comportementale, ou encore de *modèle de substitution* (traduction littérale de *surrogate model*). C'est cette dernière appellation que nous utiliserons par la suite.

Un modèle de substitution, construit à partir des résultats d'évaluation disponibles, peut être utilisé pour construire un critère quantifiant l'intérêt d'une évaluation potentielle. Plus précisément, c'est le résultat de l'optimisation de ce critère (que l'on dénommera par la suite *critère d'échantillonnage*), dépendant des résultats des évaluations précédentes, qui détermine le prochain point de l'espace de recherche où l'évaluation doit être effectuée. On remplace donc un problème d'optimisation coûteux par une série de problèmes d'optimisation moins coûteux. Dans la littérature, et au-delà des seuls problèmes d'optimisation, l'utilisation de critères d'échantillonnage est connue sous le nom de plan d'expériences adaptatif ou encore d'apprentissage actif (traduction littérale d'*active learning*).

Dans ce mémoire, nous laissons de côté les modèles de substitution classiques tels que les modèles polynomiaux et les réseaux de neurones pour nous intéresser aux modèles de substitution probabilistes et en particulier à la modélisation par processus gaussiens. Ces modèles possèdent en effet plusieurs avantages qui les rendent particulièrement adaptés à l'optimisation globale de fonctions coûteuses. Tout d'abord, ils permettent la mise en place d'un cadre bayésien qui facilite le traitement des incertitudes en entrée et en sortie du système ainsi que la prise en compte d'*a priori* (souvent fourni par les experts) sur la fonction inconnue. Ils sont de plus bien adaptés à la prise en compte de contraintes (Schonlau, 1997). Enfin, ils permettent la quantification de l'erreur d'approximation, ce qui permet notamment, comme nous le verrons, de construire des critères d'échantillonnage pertinents pour l'optimisation globale (Jones, 2001).

Plan du mémoire

Dans le premier chapitre de ce mémoire, nous rappellerons le principe de l'optimisation globale de fonctions coûteuses à l'aide de critères d'échantillonnage reposant sur une modélisation par processus gaussiens. Cette partie repose largement sur les travaux de Jones (2001), quoique la présentation de la modélisation par processus gaussiens soit plus générale, et est destinée à présenter les critères d'échantillonnage existants et leurs limitations. Une fois l'état de l'art exposé, nous proposerons un critère d'échantillonnage tirant davantage parti du modèle gaussien. Ce critère, qui consiste à minimiser *l'entropie conditionnelle des minimiseurs globaux* (ECM) n'est qu'une composante de l'algorithme IAGO (pour *Informational Approach to Global Optimization*) d'optimisation globale de fonctions coûteuses qui fait l'objet du chapitre 2. Cet algorithme se démarque assez nettement de ses concurrents, ne serait-ce que parce qu'il s'intéresse davantage aux *minimiseurs* qu'au *minimum*. Une fois son fonctionnement détaillé, nous discuterons des modalités de son utilisation sur des problèmes pratiques, notamment des adaptations qu'il doit subir pour faire face aux difficultés mentionnées précédemment (par exemple l'apparition d'un bruit additif sur le résultat des évaluations).

Le chapitre 3 sera ensuite dédié aux problèmes d'optimisation robuste aux incertitudes sur les facteurs. Nous y discuterons des possibilités d'extension du modèle gaussien à ce contexte, des modifications à apporter aux critères d'échantillonnage et plus particulièrement à la minimisation de l'ECM.

Par la suite, nous chercherons à mettre en avant, dans le chapitre 4, les avantages de la minimisation de l'ECM face aux méthodes d'optimisation globale classiques et aux autres critères d'échantillonnage. Pour ce faire, nous étudierons leurs vitesses de convergence empiriques. Ces vitesses de convergence seront estimées sur des fonctions-tests classiques, ainsi que sur des réalisations du modèle gaussien sous-jacent.

Dans le cinquième et dernier chapitre, nous détaillerons les applications industrielles rencontrées chez Renault au cours de ces travaux, ainsi que les résultats obtenus lors de leur traitement. L'objectif étant de démontrer, comme cela a été dit plus haut, qu'il est tout à fait possible de fournir une solution pertinente à ces problèmes malgré un budget d'évaluation réduit. Nous profiterons aussi de ce chapitre pour mettre en avant les difficultés rencontrées en pratique pour appliquer un algorithme d'optimisation globale dans le contexte industriel qui est le nôtre.

Publications relatives à ces travaux

Les travaux présentés dans ce mémoire ont fait l'objet de deux publications dans le *Journal of Global Optimization* et de trois communications orales dans des conférences internationales avec comités de lecture.

J. Villemonteix, E. Vazquez et É. Walter. « An informational approach to the global optimization of expensive-to-evaluate functions ». *To appear in J. Global Optim.*, 2008b.

J. Villemonteix, E. Vazquez, M. Sidorkiewicz et É. Walter. « Gradient-based IAGO strategy for the global optimization of expensive-to-evaluate functions and application to intake-port design ». Advances in Global Optimization : Method and Applications, Myconos (Greece), June 13–17, 2007a.

J. Villemonteix, E. Vazquez, M. Sidorkiewicz et É. Walter. « Global optimization of expensive-to-evaluate functions : an empirical comparison of two sampling criteria ». *To appear in the Mykonos special issue of the J. Global Optim.*, 2008a.

J. Villemonteix, E. Vazquez et É. Walter. « Identification of expensive-to-simulate parametric models using kriging and stepwise uncertainty reduction ». In *46th IEEE Conference on Decision and Control*, pp. 5505–5510, New Orleans (USA), December 12-14 2007b.

E. Vazquez, J. Villemonteix, M. Sidorkiewicz et É. Walter. « Global optimization based on noisy evaluation : an empirical study of two statistical approaches ». In *6th International Conference on Inverse Problems in Engineering : Theory and Practice*, Dourdan (Paris), France, June 15–19 2008a.

Dans Villemonteix et al. (2008b), nous avons introduit l'algorithme IAGO d'optimisation globale de fonctions coûteuses qui sera détaillé au chapitre 2 et à la fin du chapitre 1. Les idées développées dans cet article ont ensuite été reprises dans Villemonteix et al. (2007a), où deux approximations de IAGO sont testées. Ces approximations n'ayant finalement pas apporté les gains escomptés, elle ne sont que brièvement abordées dans la section 2.2.

Par la suite, nous avons réalisé dans Villemonteix et al. (2008a) une comparaison détaillée entre IAGO et son principal concurrent EGO. Les résultats de cette comparaison, ainsi que la méthode employée (qui est en soi une contribution au domaine) sont présentés et développés dans le chapitre 4. Cette comparaison a aussi été réalisée dans le cas où les résultats des évaluations de la fonction à optimiser sont corrompus par un bruit additif. Ces résultats qui ont fait l'objet d'une publication (Vazquez et al., 2008) sont présentés dans la section 4.4.1.

Enfin, nous présentons dans Villemonteix et al. (2007b) une application de IAGO à un problème d'identification de paramètres. Cette application est rappelée dans la section 4.2.2.

Notations

Symboles alphabétiques

d	dimension de l'espace des facteurs
f	fonction objectif, à minimiser
$\mathbf{f}_n$	résultats des n premières évaluations
$f_{\max}$	$\max_i f(\mathbf{x}_i)$
f^*	minimum global de f
$\hat{f}$	moyenne du prédicteur par krigeage conditionnellement aux résultats d'évaluations
$\mathbf{g}$	vecteur des contraintes d'inégalité
h	$\|\|\mathbf{x}-\mathbf{y}\|\|$
$\mathbf{h}$	vecteur des contraintes d'égalité
k	covariance de F, processus gaussien qui modélise f
k_M	covariance de M
k_{MF}	covariance entre M et F
$k_{\xi F}$	covariance entre ξ et F
$\boldsymbol{k}(\mathbf{x})$	le vecteur des covariances entre $F(\mathbf{x})$ et $\boldsymbol{F}_n$
ℓ	notation générique pour les fonctions de perte bayésienne
$m(\mathbf{x})$	moyenne du modèle gaussien F
m_n	estimation du minimum de f^* donnée par $m_n = \min_i f(\mathbf{x}_i)$
n	nombre d'évaluations réalisées
$\boldsymbol{p}(\mathbf{x})$	vecteur des fonctions de bases définissant la moyenne de F (cf. la section 1.2.1)
s	taille du support de Q
$\mathbf{x}$	un point de $\mathbb{X}$
$\mathbf{x}_c$	facteurs de conception (contrôlables lors de la fabrication ou de la vie du système)
$\mathbf{x}_e$	facteurs d'environnement qui ne sont contrôlables ni au cours de la fabrication ni au cours de la vie du système

$\mathbf{x}_i$	i-ème point d'évaluation de f
$\mathbf{x}^*$	un minimiseur global de f
$\mathbf{y}$	un point de $\mathbb{X}$
y_i	i-ème point du support de Q
$D(\mathbf{x},T)$	$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\mathbb{1}_{F(\mathbf{x}+\boldsymbol{\varepsilon})\geq T}\right]$, modèle de la probabilité de dépassement du seuil T en présence de bruit sur les facteurs
$\mathbb{E}$	notation générique pour l'espérance mathématique
F	processus gaussien qui modélise f
F_n	vecteur aléatoire modélisant les résultats des évaluations aux points dans $\mathbb{S}$
$\mathcal{F}_n$	$\{\boldsymbol{F}_n = \boldsymbol{f}_n\}$
$\hat{F}$	prédicteur par krigeage de F
F^*	minimum global de F
$\mathbb{G}$	sous ensemble fini de $\mathbb{X}$
H	entropie d'une variable aléatoire discrète
$I(\mathbf{x})$	fonction *improvement* définie à la section 1.3.3
J	notation générique pour les critères d'échantillonnage
$\mathbf{K}$	matrice de covariance des résultats d'évaluation
L	notation générique pour une vraisemblance
M	$\mathbb{E}_{\boldsymbol{\varepsilon}}(F(\mathbf{x}+\boldsymbol{\varepsilon}))$, modèle de la valeur moyenne de f en présence de bruit sur les facteurs
M_n	estimateur de F^* donné par $M_n = \min_i F(\mathbf{x}_i)$
$\mathcal{M}_{\mathbb{G}}$	ensemble des minimiseurs globaux de F sur $\mathbb{G}$
$\mathcal{M}_{\mathbb{X}}$	ensemble des minimiseurs globaux de F sur $\mathbb{X}$
N	cardinal de $\mathbb{G}$
$\boldsymbol{P}$	matrice de régression aux points d'évaluation (cf. la section 1.2.2)
Q	opérateur de quantification intervenant dans l'approximation de l'ECM (cf. la section 2.2.2)
$P_{\mathbf{X}^*}(\cdot\|\mathcal{F}_n)$	distribution de probabilité de $\mathbf{X}^*$
$\mathbb{S}_n$	position des n premières évaluations
V	$\mathrm{var}_{\boldsymbol{\varepsilon}}(F(\mathbf{x}+\boldsymbol{\varepsilon}))$, modèle de la variance de f en présence de bruit sur les facteurs
$\mathbf{X}^*$	vecteur aléatoire uniformément distribué sur $\mathcal{M}_{\mathbb{G}}$
$\mathbb{X}$	espace des facteurs
$\mathbb{X}_{\mathrm{e}}$	ensemble de définition des facteurs d'environnement $\mathbf{x}_{\mathrm{e}}$
α	paramètre de réglage du caractère global de la recherche (cf. la section 1.3)
$\boldsymbol{\beta}$	vecteur des coefficients de la combinaison linéaire définissant la moyenne de F
$\boldsymbol{\varepsilon}$	bruit sur les facteurs de f

$\boldsymbol{\varepsilon}_e$	vecteur des variables ou facteurs d'environnement
$\tilde{\varepsilon}$	bruit sur les facteurs de f pendant la phase d'optimisation
Φ	fonction de répartition d'une variable gaussienne centrée réduite
$\boldsymbol{\lambda}(\mathbf{x})$	vecteur des n coefficients de la prédiction par krigeage
ν	régularité de la covariance de Matérn
ρ	portée de la covariance de Matérn
σ	paramètre de la covariance de Matérn correspondant à l'écart type
$\hat{\sigma}^2$	variance de l'erreur de la prédiction par krigeage
$\boldsymbol{\theta}$	vecteur des paramètres de la covariance de F
$\boldsymbol{\mu}(\mathbf{x})$	vecteur des multiplicateurs de Lagrange associés aux contraintes d'universalité
ξ	modèle de $f(\mathbf{x}+\tilde{\varepsilon})$

Abréviations

AEI	Augmented Expected Improvement (défini en section 2.4.1 et tiré de Huang et al., 2006)
ECM	Entropie Conditionnelle des Minimiseurs (définie en section 1.4.3)
EI	Expected Improvement (défini en section 1.3.3)
EIm	Version modifiée de l'expected improvement (cf. la section 2.4.1)
EGO	Efficient Global Optimization (cf. la section 2.2)
IAGO	Informational Approach to Global Optimization (cf. le chapitre 2)
LHS	Latin Hypersquare Sampling

CHAPITRE 1

OPTIMISATION GLOBALE BAYÉSIENNE AVEC UN *a priori* GAUSSIEN

Résumé — Ce chapitre traite de critères d'échantillonnage, reposant sur une modélisation par processus gaussiens, utilisés pour l'optimisation globale de fonctions coûteuses. Après introduction des notations nécessaires, la théorie de la prédiction linéaire sous l'hypothèse gaussienne est brièvement abordée (sous l'angle du krigeage) de manière à fournir les éléments nécessaires à la compréhension du reste du mémoire. Les critères d'échantillonnage les plus représentatifs de la littérature sont ensuite présentés. Nous insistons en particulier sur le compromis entre exploration de l'espace des facteurs et recherche locale qu'ils permettent d'accomplir. Enfin, nous introduisons le critère de maximisation de l'entropie conditionnelle des minimiseurs globaux, principale contribution de nos travaux.

1.1 Objectifs et plan du chapitre

Dans le chapitre introductif nous avons insisté sur le fait qu'étant donnée la complexité des problèmes industriels motivant ces travaux, il était sans doute préférable d'en extraire des problèmes plus simples dont la résolution contribue à la résolution du problème global. Dans ce chapitre, nous nous intéressons au premier d'entre eux, l'optimisation globale d'une fonction scalaire dont l'évaluation est coûteuse.

Considérons une fonction $f : \mathbb{X} \to \mathbb{R}$, avec $\mathbb{X}$, l'*espace des facteurs*, un sous-espace borné de $\mathbb{R}^d$. L'objectif est alors de résoudre le problème de minimisation sous contraintes

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{X}} \quad & f(\mathbf{x}) \\ \text{s.c.} \quad & \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \\ & \mathbf{h}(\mathbf{x}) = \mathbf{0}, \end{aligned}$$

avec, pour $\mathbf{x}$ un point de l'espace des facteurs, $\mathbf{g}(\mathbf{x})$ le vecteur des contraintes d'inégalité, et $\mathbf{h}(\mathbf{x})$

le vecteur des contraintes d'égalité.

Dans ce chapitre, nous considérons simplement le problème non contraint (la prise en compte de contraintes fait l'objet d'une section du chapitre 2), et étudions les moyens de le résoudre avec un budget d'évaluations restreint. Cette restriction, induite par le coût d'une évaluation, implique que le minimum de f et sa position dans l'espace des facteurs ne pourront en général être déterminés avec précision. Elle implique aussi que chacune des évaluations disponibles doit être choisie avec soin de manière à améliorer le plus possible l'estimation du minimum de f ou de sa position. Pour ce faire il est souvent plus avantageux de choisir chaque évaluation en fonction du résultat des précédentes[1]. Ainsi, si n évaluations ont été réalisées aux points de $\mathbb{S}_n = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subset \mathbb{X}$, la $n+1$-ième s'obtient comme le résultat de l'optimisation globale d'un critère d'échantillonnage $J(\mathbf{x}, \mathbb{S}_n, \mathbf{f}_n)$ qui mesure l'intérêt d'une évaluation supplémentaire en $\mathbf{x}$ compte-tenu de $\mathbf{f}_n = [f(\mathbf{x}_1), ..., f(\mathbf{x}_n)]^\mathsf{T}$, les résultats des n premières évaluations. Pour choisir le $n+1$-ième point d'évaluation, il faut donc résoudre

$$\mathbf{x}_{n+1} = \arg\max_{\mathbf{x} \in \mathbb{X}} J(\mathbf{x}, \mathbb{S}_n, \mathbf{f}_n).$$

On trouve surtout dans la littérature des critères d'échantillonnage construits grâce à une approximation $\hat{f}$ de f, que nous avons dénommé modèle de substitution dans le chapitre introductif. Compte-tenu du petit nombre de résultats d'évaluation disponibles, l'erreur associée à l'approximation de f par $\hat{f}$ peut être très importante. C'est pourquoi la plupart des critères reposent sur un modèle probabiliste permettant de quantifier cette erreur et d'en tenir compte pour favoriser les évaluations dans les zones où elle est élevée. Il est ainsi possible d'orienter la recherche vers les zones inexplorées de l'espace des facteurs. Le domaine recouvrant l'utilisation de modèles probabilistes pour l'optimisation globale est connu sous le nom *d'optimisation globale bayésienne* . Un processus stochastique F, indexé sur $\mathbb{X}$, y est utilisé comme a priori sur la fonction à optimiser f.

La nature du modèle F a quelque peu varié depuis l'apparition de cette approche. Initialement (Kushner, 1964), f a été modélisée par un mouvement brownien. La distribution de F conditionnellement aux résultats des évaluations réalisées est alors simple à calculer en dimension un. De nombreuses heuristiques ont ensuite été développées pour le cas multidimensionnel, par exemple dans Perttunen (1991), Elder IV (1992) ou Mockus (1989b). Plus récemment, la *meilleure prédiction linéaire non biaisée* a été utilisée pour prédire F, supposé du second ordre dans Cox et John (1997) puis supposé gaussien dans Jones et al. (1998), avec l'introduction de l'algorithme EGO (pour *Efficient Global Optimization*). Depuis, la majorité des publications en matière d'optimisation globale bayésienne utilise cette hypothèse gaussienne, et traite d'extensions d'EGO (avec par exemple Williams et al. (2000) ou Huang et al. (2006)) ou d'études comparatives (Jones, 2001 ; Sasena et al., 2002). Notre contribution à ce domaine (Villemonteix et al., 2008b) utilise aussi des processus gaussiens.

[1] Ce qui n'est plus nécessairement vrai si l'on peut paralléliser les évaluations, cf. la section 2.4.4

L'hypothèse gaussienne offre davantage de flexibilité pour la modélisation que celle d'un mouvement brownien, et le calcul des lois conditionnelles reste simple. De plus, la modélisation à l'aide de processus gaussiens a fait l'objet de nombreux travaux, notamment dans le domaine des géostatistiques, où cette approche est connue sous le nom de *krigeage*.

La suite de ce chapitre s'organise de la manière suivante. Dans un premier temps, nous donnons quelques éléments de la théorie du krigeage, de manière à introduire les concepts et les notations nécessaires à un tour d'horizon des critères d'échantillonnage classiques. Enfin, après avoir mis en évidence les défauts des approches existantes, nous introduisons le critère de minimisation de l'entropie conditionnelle du minimiseur.

1.2 Processus gaussiens et krigeage

1.2.1 Principe

Le krigeage (Krige, 1951) est une méthode (ou plutôt un ensemble de méthodes) de prédiction reposant sur un modèle probabiliste (généralement gaussien) qui peut être utilisée pour interpoler des données. Cette méthode peut être décrite comme une technique de régression à noyaux, au même titre que les splines (Wahba, 1998) ou la SVR (*support vector regression*, cf. Smola (1998)). Le terme krigeage, *kriging* en anglais, provient du nom de famille de l'ingénieur minier sud-africain Daniel Gerhardus Krige à l'origine de la technique, mais la formalisation mathématique revient à George Matheron (Matheron, 1963). Depuis les années 60, le krigeage, qui n'est au départ qu'une prédiction linéaire, a été largement développé et utilisé en géostatistiques. Depuis les années 90, le krigeage est aussi connu sous le nom de modélisation par processus gaussiens dans le domaine du *machine learning*. Dans ce mémoire, nous nous employons à en décrire simplement les idées principales et leur intérêt pour l'optimisation globale. Pour davantage de détails, nous renvoyons aux ouvrages de référence (Chilès et Delfiner, 1999 ; Stein, 1999 ; Vazquez, 2005).

Le principe sous-jacent au krigeage est de modéliser la fonction inconnue par un processus gaussien de fonction moyenne $m(\mathbf{x}) = E[F(\mathbf{x})]$, et de fonction de covariance

$$k(\mathbf{x},\mathbf{y}) = \mathrm{cov}(F(\mathbf{x}),F(\mathbf{y})).$$

(Dans la suite nous parlerons simplement de moyenne et de covariance.) L'hypothèse gaussienne sera conservée dans tout le mémoire, mais n'est cependant pas nécessaire pour établir les équations de la prédiction.

Le vecteur $\mathbf{f}_n$ des résultats des n premières évaluations, ou observations, est donc considéré comme une réalisation du vecteur aléatoire $\mathbf{F}_n = [F(\mathbf{x}_1),...,F(\mathbf{x}_n)]^\mathsf{T}$. La prédiction par krigeage $\hat{F}(\mathbf{x})$ de $F(\mathbf{x})$ est alors définie comme la prédiction linéaire non biaisée qui minimise la variance de l'erreur de prédiction. En d'autres termes $\hat{F}(\mathbf{x})$ minimise

$$\hat{\sigma}^2(\mathbf{x}) = \mathbb{E}[(\hat{F}(\mathbf{x}) - F(\mathbf{x}))^2]$$

parmi tous les éléments non-biaisés de l'espace vect$\{F(\mathbf{x}_1),...,F(\mathbf{x}_n)\}$.

En tant qu'élément de l'espace engendré par les observations, $\hat{F}(\mathbf{x})$ peut s'écrire comme une combinaison linéaire de celles-ci

$$\hat{F}(\mathbf{x}) = \boldsymbol{\lambda}(\mathbf{x})^\mathsf{T}\mathbf{F}_n. \tag{1.1}$$

La variance de l'erreur s'écrit alors

$$\hat{\sigma}^2(\mathbf{x}) = k(\mathbf{x},\mathbf{x}) + \boldsymbol{\lambda}(\mathbf{x})^\mathsf{T}\mathbf{K}\boldsymbol{\lambda}(\mathbf{x}) - 2\boldsymbol{\lambda}(\mathbf{x})^\mathsf{T}\mathbf{k}(\mathbf{x}), \tag{1.2}$$

avec

$$\mathbf{K} = (k(\mathbf{x}_i,\mathbf{x}_j))\,,\ (i,j) \in [\![1,n]\!]^2$$

la matrice $n \times n$ de covariance des résultats d'évaluations, et

$$\mathbf{k}(\mathbf{x}) = [k(\mathbf{x}_1,\mathbf{x}),\ldots,k(\mathbf{x}_n,\mathbf{x})]^\mathsf{T}$$

le vecteur des covariances entre $F(\mathbf{x})$ et $\mathbf{F}_n$. Il s'agit alors de déterminer le vecteur des coefficients $\boldsymbol{\lambda}(\mathbf{x})$ qui minimise (1.2) tout en assurant une prédiction non biaisée.

Avant de présenter la solution de ce problème d'optimisation sous contrainte, précisons les hypothèses faites sur le modèle gaussien F.

Dans la théorie du krigeage (Matheron, 1969 ; Chilès et Delfiner, 1999), la moyenne de $F(\mathbf{x})$ est une combinaison linéaire finie de fonctions connues et peut donc s'écrire

$$m(\mathbf{x}) = \boldsymbol{\beta}^\mathsf{T}\mathbf{p}(\mathbf{x}), \tag{1.3}$$

avec $\boldsymbol{\beta}$ un vecteur de coefficients fixe (non aléatoire) mais inconnu, et avec

$$\mathbf{p}(\mathbf{x}) = [p_1(\mathbf{x}),\ldots,p_l(\mathbf{x})]^\mathsf{T}$$

un vecteur de fonctions connues. Ces fonctions, en général des monômes dont le degré dépasse rarement deux, reflètent une connaissance *a priori* sur la fonction inconnue f. Dans les exemples qui ponctueront ce mémoire, et sauf mention contraire, nous ne disposons pas d'*a priori* et supposerons simplement la moyenne constante, c'est-à-dire $\mathbf{p}(\mathbf{x}) = 1$.

La covariance de F est, quant à elle, choisie dans une famille de covariances paramétrées et est généralement supposée isotrope. Auquel cas, elle sera notée $k(h) = k(||\mathbf{x}-\mathbf{y}||)$. Le problème du choix de la structure de covariance et de l'estimation de ses paramètres, qui occupe une place centrale dans la théorie et la pratique du krigeage, sera discuté dans l'annexe A.1. Précisons simplement ici que nous suivons les recommandations de Stein (1999) et utilisons la classe des covariances de Matérn, qui permet, à l'inverse de la majorité des autres covariances utilisées en pratique, de jouer simplement sur la *régularité* de la covariance (sa dérivabilité en zéro). En effet, la régularité de la covariance, qui reflète la régularité des réalisations, ou trajectoires de F, a une influence importante sur la qualité de la prédiction par krigeage (Stein, 1999).

La classe des covariances de Matèrn est généralement paramétrée de la manière suivante :

$$k(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}h}{\rho} \right)^{\nu} \mathcal{K}_\nu \left(\frac{2\nu^{1/2}h}{\rho} \right) \ \forall h \geq 0, \tag{1.4}$$

avec Γ la fonction gamma, et $\mathcal{K}_\nu$ la fonction de Bessel de deuxième espèce. Les paramètres $\boldsymbol{\theta} = [\nu, \rho, \sigma]^\mathsf{T}$ s'interprètent alors facilement ; ν contrôle la régularité ; σ^2 est la variance ($k(0) = \sigma^2$) ; et ρ représente la *portée* de la covariance, qui peut être vue comme une distance de corrélation caractéristique. Ces paramètres peuvent être fixés *a priori* d'après les informations disponibles sur la fonction, ou estimés à partir des résultats d'évaluation disponibles. Pour les exemples présentés dans la suite du chapitre, et sauf mention du contraire, cette estimation sera réalisée par maximum de vraisemblance restreint (Stein, 1999).

Remarque 1.1. L'hypothèse d'isotropie de la covariance peut, dans bien des cas, s'avérer maladroite. On choisit alors souvent d'ajouter une matrice $\mathbf{R}$ (de taille $d \times d$, généralement diagonale) aux paramètres des familles de covariances et de considérer $k(\mathbf{x}, \mathbf{y}) = k(\sqrt{(\mathbf{x}-\mathbf{y})^\mathsf{T}\mathbf{R}(\mathbf{x}-\mathbf{y})})$. Dans le reste du mémoire, et sauf mention contraire, les covariances utilisées seront néanmoins isotropes.

1.2.2 Prédiction par krigeage

La prédiction par krigeage en un point $\mathbf{x} \in \mathbb{X}$ est la prédiction linéaire qui possède une variance de l'erreur minimale et qui respecte la contrainte de non-biais $\mathbb{E}[\hat{F}(\mathbf{x})] = m(\mathbf{x})$. Du fait des hypothèses sur la moyenne (1.3), cette contrainte s'écrit

$$\boldsymbol{\beta}^\mathsf{T}\mathbf{P}^\mathsf{T}\boldsymbol{\lambda}(\mathbf{x}) = \boldsymbol{\beta}^\mathsf{T}\mathbf{p}(\mathbf{x}), \tag{1.5}$$

avec

$$\mathbf{P} = \begin{pmatrix} \mathbf{p}(\mathbf{x}_1)^\mathsf{T} \\ \vdots \\ \mathbf{p}(\mathbf{x}_n)^\mathsf{T} \end{pmatrix}.$$

Ainsi, pour satisfaire la contrainte de non-biais, quelle que soit la valeur de $\boldsymbol{\beta}$, les coefficients du krigeage $\boldsymbol{\lambda}(\mathbf{x})$ doivent vérifier les l contraintes

$$\mathbf{P}^\mathsf{T}\boldsymbol{\lambda}(\mathbf{x}) = \mathbf{p}(\mathbf{x}), \tag{1.6}$$

dénommées *contraintes d'universalité* par Matheron. Notons que ces contraintes ne peuvent être satisfaites qu'à la condition[2] $n \geq l$ et qu'elles assurent, en particulier, une prédiction exacte si f est réellement une combinaison linéaire des $p_i(\cdot)$.

[2] Si cette condition n'est pas vérifiée, le modèle appliqué à la moyenne, incompatible avec les données disponibles, ne permet pas la prédiction par krigeage.

Le problème de minimisation de l'erreur de prédiction (1.2) sous les contraintes d'universalité (1.6) peut alors être résolu au travers d'une formulation lagrangienne. Le vecteur $\boldsymbol{\lambda}(\mathbf{x})$ des coefficients du krigeage est ainsi solution du système d'équations

$$\begin{pmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^\mathsf{T} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}(\mathbf{x}) \\ \boldsymbol{\mu}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathbf{k}(\mathbf{x}) \\ \mathbf{p}(\mathbf{x}) \end{pmatrix}, \tag{1.7}$$

avec $\mathbf{0}$ une matrice de zéros, et $\boldsymbol{\mu}(\mathbf{x})$ un vecteur de l multiplicateurs de Lagrange.

Pourvu que la fonction de covariance soit fixée, le vecteur $\boldsymbol{\lambda}$ des coefficients du krigeage peut donc être calculé sans l'aide des résultats d'évaluation de f, tout comme la variance de l'erreur de prédiction

$$\hat{\sigma}^2(\mathbf{x}) = \mathbb{E}\left[F(\mathbf{x}) - \hat{F}(\mathbf{x})\right]^2 = k(\mathbf{x},\mathbf{x}) - \boldsymbol{\lambda}(\mathbf{x})^\mathsf{T}\mathbf{k}(\mathbf{x}) - \mathbf{p}(\mathbf{x})^\mathsf{T}\boldsymbol{\mu}(\mathbf{x}), \tag{1.8}$$

obtenue en remplaçant $\mathbf{k}(\mathbf{x}) - \mathbf{P}\boldsymbol{\mu}(\mathbf{x})$ par $\mathbf{K}\boldsymbol{\lambda}(\mathbf{x})$ dans l'expression de la variance de l'erreur (1.2). Ces quantités ne dépendent en effet que de $k(\cdot,\cdot)$, la covariance de F (nous verrons au paragraphe 2.2 que cette propriété est très avantageuse pour la mise en place algorithmique de IAGO). Notons que cette dernière affirmation est à nuancer en pratique, puisque les résultats des évaluations peuvent être requis pour l'estimation de la covariance de F.

Une fois f évaluée pour tous les $\mathbf{x}_i \in \mathbb{S}_n$, la prédiction par krigeage de $f(\mathbf{x})$ est l'espérance conditionnelle de $F(\mathbf{x})$, c'est-à-dire

$$\hat{f}(\mathbf{x}) = \mathbb{E}[\hat{F}(\mathbf{x})|\mathcal{F}_n] = \boldsymbol{\lambda}(\mathbf{x})^\mathsf{T}\mathbf{f}_n,$$

avec $\mathcal{F}_n = \{\mathbf{F}_n = \mathbf{f}_n\}$ les résultats des évaluations. Il est, en outre,simple de vérifier à partir de (1.7) que

$$\forall\, \mathbf{x}_i \in \mathbb{S},\ \hat{F}(\mathbf{x}_i) = F(\mathbf{x}_i).$$

Si les résultats des évaluations sont considérés comme exacts, la prédiction de f aux points $\mathbf{x}_i \in \mathbb{S}_n$ est alors $f(\mathbf{x}_i)$, le krigeage réalise donc une interpolation des résultats d'évaluation tout en proposant une caractérisation explicite de l'erreur de prédiction qui est gaussienne, centrée et de variance donnée par (1.8).

Il est tout à fait possible d'étendre le principe de prédiction exposé ici au cas où le phénomène observé n'est pas celui que l'on souhaite prédire, mais où l'on dispose de la corrélation entre ces deux phénomènes (cf. l'annexe B). Ceci permet, entre autres, de prédire f à partir de résultats d'évaluations bruitées, ou en utilisant des résultats de l'évaluation de son gradient.

Remarque 1.2. Puisque F est supposé gaussien, le meilleur prédicteur linéaire non-biaisé correspond à l'espérance conditionnelle. Il est donc aussi possible d'obtenir les équations du krigeage dans un cadre bayésien (ces résultats, classiques, peuvent être trouvés, par exemple dans Williams et Rasmussen, 1996), où le modèle F est un *a priori* bayésien sur f. Dans le cas d'un modèle à

moyenne nulle, c'est-à-dire $m(\mathbf{x}) = 0$, la distribution de F conditionnellement aux résultats des évaluations est gaussienne, de moyenne

$$\mathbb{E}[F(\mathbf{x})|\,\mathcal{F}_n] = \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1}\mathbf{f}_n, \tag{1.9}$$

et de variance

$$\mathrm{Var}[F(\mathbf{x})|\,\mathcal{F}_n] = k(\mathbf{x},\mathbf{x}) - \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1}\mathbf{k}(\mathbf{x}),$$

c'est-à-dire exactement la moyenne (1.1) de la prédiction par krigeage et la variance (1.8) de l'erreur associée. Ce point de vue bayésien fonctionne aussi dans le cas où la moyenne de $F(\mathbf{x})$ est inconnue. Sous l'hypothèse gaussienne, la meilleure prédiction linéaire non biaisée est en effet la limite du predicteur bayésien lorsque la variance de l'*a priori* sur les coefficients $\boldsymbol{\beta}$ de la moyenne tend vers l'infini (cf. Parzen, 1963).

1.2.3 Exemple

Lorsque les résultats des évaluations de f sont disponibles, il peut être utile de simuler des trajectoires de F qui interpolent les données $\mathbf{f}_n$. Ces trajectoires, appelées *trajectoires conditionnelles*, sont des réalisations de F conditionnellement à $\mathcal{F}_n$ et constituent des représentations possibles de f compte-tenu du modèle et des évaluations déjà réalisées. La prédiction par krigeage se trouve correspondre à la moyenne de ces trajectoires, et est en tant que telle nettement plus régulière. La simulation de ces trajectoires va se révéler très utile pour l'optimisation globale.

Une illustration des relations entre f, $\hat{f}$, $\hat{\sigma}$ et les trajectoires conditionnelles est présentée sur la figure 1.1. On y distingue la prédiction par krigeage d'une fonction inconnue à partir de quelques résultats d'évaluation, ainsi que des intervalles de confiance obtenus à partir de la variance de l'erreur de prédiction. Pour une valeur de $\mathbf{x}$ donnée, les trajectoires conditionnelles ont 95% de chance d'appartenir à ces intervalles, dont la taille augmente avec la distance aux points d'évaluation.

1.2.4 Complexité calculatoire

Donnons quelques précisions sur les limitations de l'utilisation du krigeage liées à la complexité calculatoire de la prédiction.

Factorisation de la matrice de covariance

Le système (1.7) peut, par exemple, être résolu en factorisant le membre de gauche. Cette factorisation est en $\mathrm{O}(n^3)$, et la prédiction en un point de l'espace des facteurs est ensuite en $\mathrm{O}(n^2)$. Ce coût algorithmique peut rendre cette approche inapplicable à de grands jeux de données. La solution la plus simple pour remédier à ce problème est de ne réaliser la prédiction en un point donné qu'à l'aide des observations situées dans un voisinage de ce point (voisinage dont la taille doit directement dépendre de la portée de la covariance).

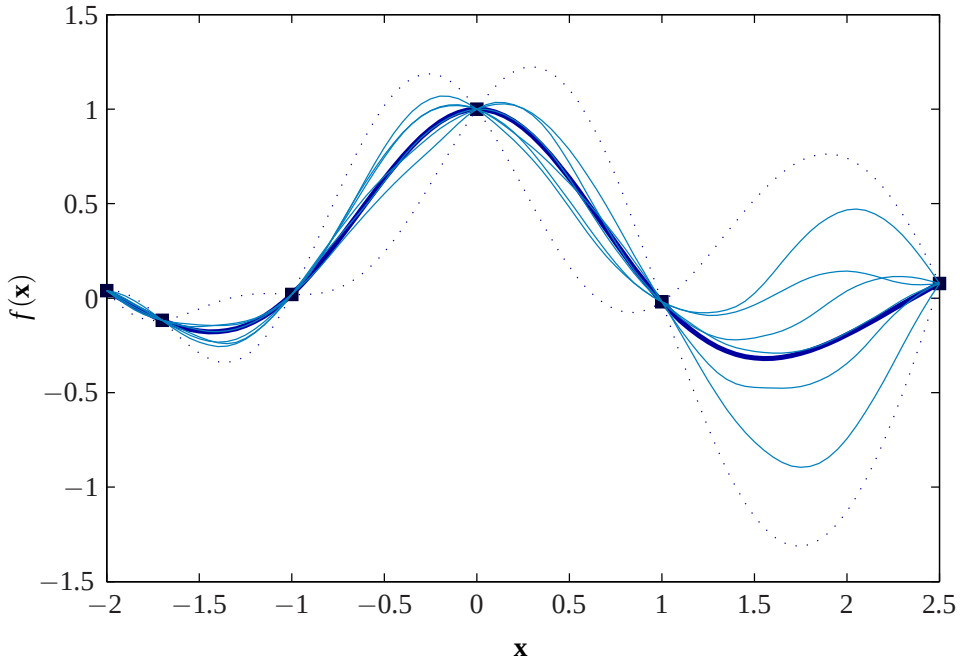


FIG. 1.1: *Exemple de prédiction par krigeage en dimension 1. Les résultats des évaluations (carrés) sont interpolés par la prédiction $\hat{f}$ (en gras). Les courbes en pointillés correspondent aux extrémités des intervalles de confiance à 95% pour la prédiction ($\hat{f} \pm 1.96\hat{\sigma}$). Les courbes en trait fin sont des simulations conditionnelles (nous verrons dans la section 2.2 la méthode utilisée pour les obtenir).*

Krigeage dual

Pour diminuer la complexité calculatoire lorsque l'on souhaite prédire en un grand nombre de points, il est aussi possible d'utiliser la forme *duale* du krigeage (Chilès et Delfiner, 1999) qui permet d'éviter le calcul des coefficients $\boldsymbol{\lambda}$. Pour introduire cette approche, rappelons que le predicteur par krigeage en un point $\mathbf{x}$ s'écrit

$$\hat{F}(\mathbf{x}) = \boldsymbol{\lambda}(\mathbf{x})^\top \mathbf{F}_n = \begin{pmatrix} \mathbf{F}_n^\top & \mathbf{0}^\top \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}(\mathbf{x}) \\ \boldsymbol{\mu}(\mathbf{x}) \end{pmatrix}, \tag{1.10}$$

avec $\mathbf{0}$ un vecteur de zéros de taille l. Or

$$\begin{pmatrix} \boldsymbol{\lambda}(\mathbf{x}) \\ \boldsymbol{\mu}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^\top & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{k}(\mathbf{x}) \\ \mathbf{p}(\mathbf{x}) \end{pmatrix}. \tag{1.11}$$

On peut donc réécrire le prédicteur $\hat{F}(\mathbf{x})$ sous la forme

$$\hat{F}(\mathbf{x}) = \begin{pmatrix} \mathbf{A}^\top & \mathbf{B}^\top \end{pmatrix} \begin{pmatrix} \mathbf{k}(\mathbf{x}) \\ \mathbf{p}(\mathbf{x}) \end{pmatrix}, \tag{1.12}$$

ou encore

$$\hat{F}(\mathbf{x}) = \sum_{i=1}^{n} A_i k(\mathbf{x}, \mathbf{x}_i) + \sum_{i=1}^{l} B_i p_i(\mathbf{x}), \tag{1.13}$$

avec les coefficients $\mathbf{A}$ et $\mathbf{B}$ (aléatoires), obtenus par résolution du système

$$\begin{pmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_n \\ \mathbf{0} \end{pmatrix}. \tag{1.14}$$

Ce système constitue les équations du *krigeage dual*, et permet de mieux comprendre les relations entre le krigeage et les techniques de régression à noyaux. En effet, $\hat{f}(\mathbf{x})$ s'obtient d'après (1.13) comme combinaison linéaire des $k(\mathbf{x}, \mathbf{x}_i)$ et des polynômes qui composent la moyenne de F.

Sur le plan pratique, cette formulation est finalement avantageuse pour diminuer le coût mémoire et CPU de la prédiction. En effet, les vecteurs de coefficients $\mathbf{a}$ et $\mathbf{b}$ (réalisations de $\mathbf{A}$ et $\mathbf{B}$ correspondant aux résultats des évaluations de f) ne dépendent pas du point $\mathbf{x}$ où l'on souhaite prédire. Il suffit donc de les calculer une fois pour toute et d'estimer ensuite f sous la forme

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} a_i k(\mathbf{x}, \mathbf{x}_i) + \sum_{i=1}^{l} b_i p_i(\mathbf{x}). \tag{1.15}$$

Une fois les coefficients $\mathbf{a}$ et $\mathbf{b}$ calculés, la prédiction en un point donné est en $\mathrm{O}(n)$. Ainsi, la prédiction en m points de l'espace des facteurs sera en $\mathrm{O}(n^3 + nm)$ contre $\mathrm{O}(n^3 + n^2 m)$ pour l'approche classique. De plus, l'approche duale permet d'éviter le stockage de la factorisation de la matrice de covariance. Il n'est cependant plus possible de calculer directement $\hat{\sigma}^2(\mathbf{x})$, la variance de l'erreur de prédiction, et les grands jeux de données posent toujours problème.

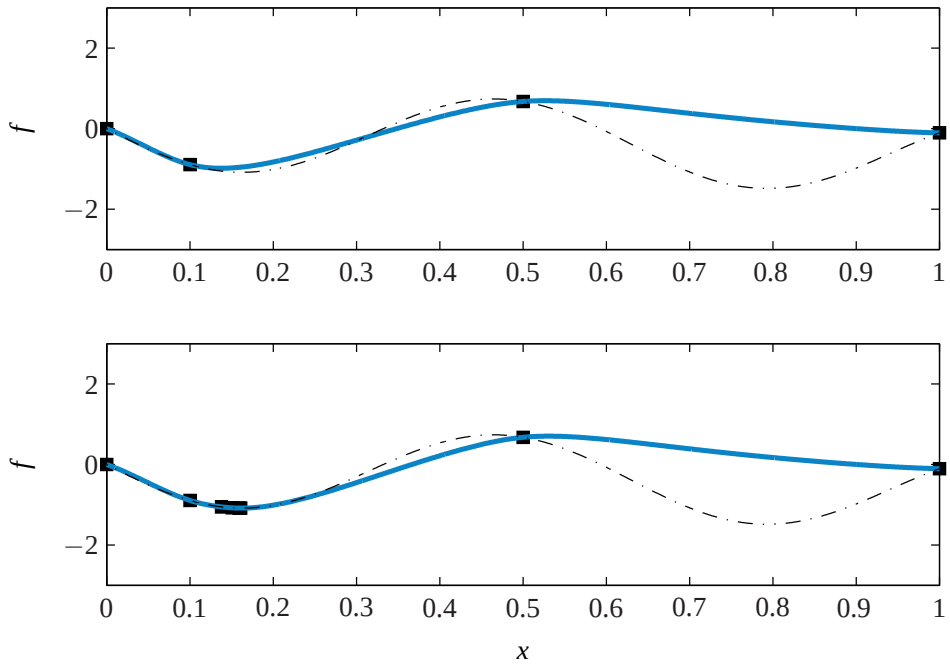


FIG. 1.2: *Approche naïve de l'optimisation par krigeage :* (partie supérieure) *prédiction* $\hat{f}$ *(en trait gras) de la fonction à optimiser* f *(en traits mixtes, considérée comme inconnue), obtenue à partir d'un plan d'expériences initial dont les résultats sont matérialisés par des carrés ;* (partie inférieure) *prédiction obtenue après quatre itérations de l'algorithme qui consiste à évaluer* f *au minimiseur de* $\hat{f}$.

1.3 Krigeage & critères d'échantillonnage

Le principe commun aux méthodes d'optimisation reposant sur le krigeage est d'évaluer f de manière itérative en un point qui optimise un critère d'échantillonnage construit à l'aide des résultats d'évaluation déjà disponibles. L'approche la plus simple serait d'utiliser un minimiseur de la prédiction par krigeage $\hat{f}$ comme nouveau point d'évaluation. Cependant, en procédant de cette manière, on accorderait une confiance trop grande à la prédiction, qui n'est, après tout, construite qu'à l'aide d'un petit nombre d'évaluations de f, et peut être très éloignée de f. L'optimisation risque alors de stagner sur un minimum local, comme c'est le cas pour l'optimisation de la fonction test présentée sur la figure 1.2.

Pour obtenir un compromis satisfaisant entre recherche locale et recherche globale, il est ainsi nécessaire d'utiliser l'erreur de prédiction pour diriger la recherche vers les zones où la confiance dans la prédiction doit être améliorée. Cette idée a conduit à la création de nombreux critères d'échantillonnage utilisant simultanément la prédiction $\hat{f}$ et l'évaluation de l'erreur associée.

Dans cette section, les principaux critères de la littérature sont rappelés. Notre objectif est de mettre en avant leurs avantages et leurs inconvénients de manière à justifier le choix des critères retenus par la suite pour la comparaison avec le critère que nous avons développé et qui sera présenté

au paragraphe 1.4. Cette section n'est pas une analyse détaillée de la littérature en optimisation globale bayésienne ; nous renvoyons pour cela au travaux de Mockus (1989a) et de Jones (2001).

Remarque 1.3. L'exemple de la figure 1.2, qui sera réutilisé pour le reste du chapitre, consiste à optimiser

$$f_{\text{test}}(x) = -\sin(10x) - \exp(x/2) + 1$$

définie pour $x \in [0,1]$, à partir de 4 évaluations initiales utilisées pour estimer les paramètres de la covariance de F. Cette fonction (inspirée de Sasena et al., 2002) présente l'avantage de posséder un minimum local proche du minimum global et à même de piéger les méthodes accordant trop peu d'importance à l'aspect global de la recherche.

1.3.1 Minimiser une borne inférieure

La variance de l'erreur de la prédiction (1.8) par krigeage peut être utilisée pour établir une borne inférieure d'un intervalle de confiance pour f. Par exemple, si le modèle est vérifié, c'est-à-dire si f est une trajectoire de F, $\forall \mathbf{x} \in \mathbb{X},\ f(\mathbf{x}) \geq \hat{f}(\mathbf{x}) - 2\hat{\sigma}(\mathbf{x})$ avec une probabilité de 0.975. Un critère d'échantillonnage simple serait donc

$$\min_{\mathbf{x}\in\mathbb{X}} \hat{f}(\mathbf{x}) - \alpha\hat{\sigma}(\mathbf{x}),$$

avec $\alpha > 0$. Ce critère permettrait, de la même façon que pour un algorithme *branch and bound*, d'écarter rapidement, et avec une certaine confiance en cette décision, des zones entières de l'espace des facteurs. Augmenter α revient à donner davantage d'intérêt à l'échantillonnage dans les zones peu explorées, et renforce ainsi le caractère global.

Utilisons ce critère avec $\alpha = 2$ pour optimiser la fonction f_{test} qui avait mis en défaut l'approche naïve. On constate sur la figure 1.3 que la méthode trouve bien l'optimum global. Cependant, les limitations de cette approche apparaissent clairement. En effet, que se passe-t-il si l'erreur de prédiction est largement sous-estimée ? Ce que nous considérions comme une borne inférieure d'un intervalle de confiance peut alors être très supérieure à f et empêcher toute évaluation au voisinage du minimiseur global (nous en verrons d'ailleurs un exemple à la section 1.3.3). Cette approche, introduite par Cox et John (1997), ne peut donc être globale qu'à la condition que le modèle soit correct (au moins dans ce qu'il implique sur la borne inférieure de l'intervalle de confiance), et sera très peu robuste par rapport à un mauvais choix des paramètres de la covariance (en particulier à une mauvaise estimation de ceux-ci).

1.3.2 Maximiser la probabilité d'amélioration

Historiquement, le premier critère proposé en optimisation bayésienne (Kushner, 1964), fût de maximiser la *probabilité d'amélioration* de la fonction objectif, ou plus précisément la probabilité

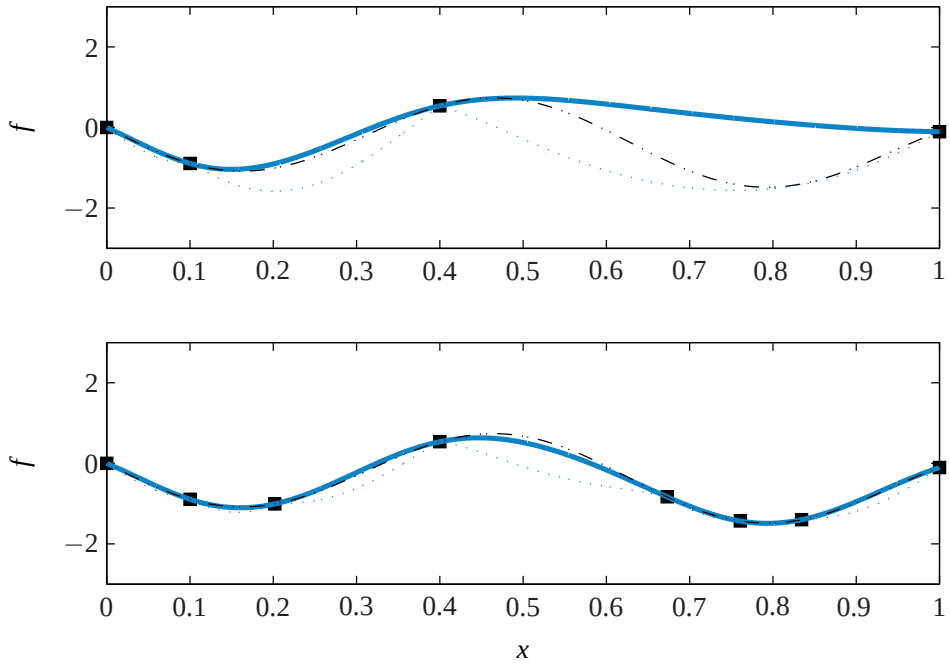


FIG. 1.3: *Utilisation de* $\hat{f}(\mathbf{x}) - 2\hat{\sigma}(\mathbf{x})$ *comme critère d'échantillonnage. Les conventions graphiques sont celles de la figure 1.2.* (Partie supérieure) *Prédiction reposant sur un plan d'expérience initial, et* $\hat{f}(\mathbf{x}) - 2\hat{\sigma}(\mathbf{x})$ *(en pointillé) qui peut être considérée comme la borne inférieure d'un intervalle de confiance pour la valeur de* $f(\mathbf{x})$. (Partie inférieure) *Prédiction par krigeage aprés quatre itérations de l'algorithme qui consiste à évaluer* f *au minimiseur de* $\hat{f} - 2\hat{\sigma}(\mathbf{x})$. *La recherche est plus globale que sur la figure 1.2.*

pour la fonction d'atteindre une valeur inférieure à $f_{\min} = \min_i f(\mathbf{x}_i)$ le minimum des résultats d'évaluation déjà obtenus ou, plus généralement, à un seuil T. Cette probabilité s'exprime en effet simplement puisque, conditionnellement aux résultats des évaluations, le résultat $F(\mathbf{x})$ d'une évaluation en $\mathbf{x}$ suit une loi gaussienne de moyenne $\hat{f}(\mathbf{x})$ et de variance $\hat{\sigma}^2(\mathbf{x})$.

Le critère d'échantillonnage s'écrit donc

$$\arg\max_{\mathbf{x}\in\mathbb{X}} P(F(\mathbf{x}) \leq T | \mathcal{F}_n) = \Phi\left(\frac{T - \hat{f}(\mathbf{x})}{\hat{\sigma}(\mathbf{x})}\right),$$

avec Φ la fonction de répartition d'une variable gaussienne centrée réduite. Initialement, ce critère était calculé analytiquement, pour $\mathbb{X}$ de dimension 1, sous l'hypothèse que F était un processus de Wiener. Par la suite, des extensions au cas multidimensionnel furent développées (cf. par exemple Perttunen, 1991). L'algorithme qui consiste à réitérer l'échantillonnage reposant sur la probabilité d'amélioration fût plus tard baptisé *P-algorithme* par Zilinskas (1992).

Pour choisir T, empruntons la méthode intuitive proposée par Jones (2001), qui consiste à écrire T comme une amélioration désirée en proportion de l'amplitude de variation de f. Ainsi, si $f_{\max}$ et $f_{\min}$ sont le maximum et le minimum des résultats d'évaluation déjà obtenus ($f_{\max} = \max_i f(\mathbf{x}_i)$), T s'écrit sous la forme

$$T = \min_{\mathbf{x}\in\mathbb{X}} \hat{f}(\mathbf{x}) - \alpha(f_{\max} - f_{\min}),$$

avec $\alpha \geq 0$.

Choisir α proche de zéro (par exemple 0.05, cf. la figure 1.4) implique que l'amélioration désirée est faible et que la recherche sera extrêmement locale. Inversement, pour α trop grand (par exemple 0.7, cf. figure 1.4), l'amélioration désirée sera probablement irréaliste et la recherche, excessivement globale, équivaudra à échantillonner le plus loin possible des points précédemment évalués. Pour faire face à cette sensibilité au choix du seuil, il faudrait utiliser plusieurs seuils. Une première approche, proposée par Jones (2001), est de les utiliser simultanément et de paralléliser ainsi les évaluations, ce qui est bien souvent irréalisable en pratique. On pourrait alors envisager de traiter α de la même façon que la « température » dans un recuit simulé, et de diminuer sa valeur après chaque évaluation.

1.3.3 Maximiser l'espérance de l'amélioration

Pour éviter le choix du seuil T et donc le choix d'un compromis entre recherche locale et recherche globale, il semble plus simple de maximiser l'espérance de l'amélioration ou l'EI pour (*expected improvement*) . Proposé initialement par Mockus et al. (1978), le critère EI doit surtout son succès à des publications plus récentes (en particulier Jones et al., 1998) et aux applications qui leur succédèrent (cf. par exemple Huang et Allen, 2005).

Pour présenter le critère de l'EI, utilisons un formalisme bayésien (cf. Mockus, 1989a), qui nous semble plus à même de faire ressortir les mérites de ce critère. Utiliser EI est en effet une

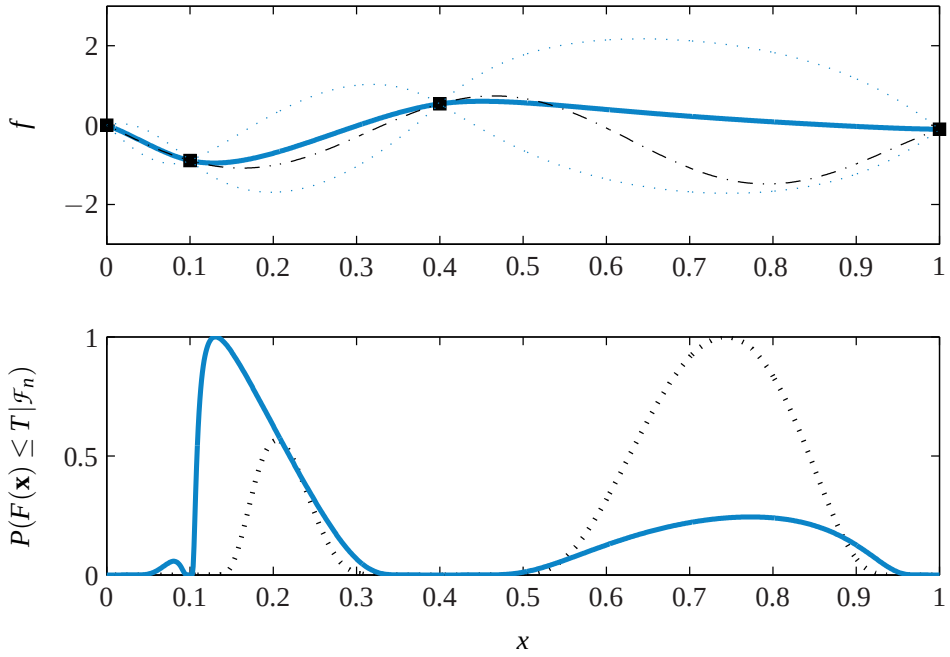


FIG. 1.4: *Influence du seuil sur le caractère global de l'optimisation.* (Partie supérieure) *similaire à la partie supérieure de la figure 1.3.* (Partie inférieure) *Probabilité d'amélioration en fonction des évaluations candidates pour* $\alpha = 0.05$ *(trait plein) et pour un seuil* $\alpha = 0.7$ *(trait pointillé). On peut noter la distance entre les points d'évaluations retenus pour ces deux seuils.*

stratégie optimale à un pas pour l'amélioration de l'estimation du minimum (sachant l'*a priori* gaussien F sur la fonction inconnue f).

Soit $F^* = \min_{\mathbf{x}\in\mathbb{X}} F(\mathbf{x})$ le minimum global de F sur $\mathbb{X}$, $\mathbb{S}_n$ les n premiers points d'évaluation supposés choisis, et soit $M_n = \min_{\mathbf{x}_i\in\mathbb{S}_n} F(\mathbf{x}_i)$ l'estimateur de F^* défini comme le minimum des résultats des évaluations. Alors, avec la fonction de perte

$$\ell(\mathbb{S}_n, F) = |M_n - F^*| = M_n - F^*, \tag{1.16}$$

le risque (ou perte attendue) associé à une évaluation en un point $\mathbf{x} \in \mathbb{X}$ s'écrit conditionnellement aux résultats des évaluations précédentes comme

$$\mathbb{E}(\ell(\mathbb{S}_n \cup \{\mathbf{x}\}, F)|\mathcal{F}_n) = \mathbb{E}(\min\{M_n, F(\mathbf{x})\}|\mathcal{F}_n) - \mathbb{E}(F^*|\mathcal{F}_n). \tag{1.17}$$

Cette quantité n'est pas calculable analytiquement, cependant, la minimiser est équivalent à maximiser le critère de l'EI sous la forme où l'on le retrouve généralement (cf. par exemple Jones, 2001), à savoir,

$$\mathrm{EI}(\mathbf{x}) = \mathbb{E}\left[I(\mathbf{x})|\mathcal{F}_n\right], \tag{1.18}$$

avec

$$I(\mathbf{x}) = \begin{cases} 0 & \text{si} \quad F(\mathbf{x}) \geq M_n \\ M_n - F(\mathbf{x}) & \text{sinon} \end{cases}.$$

En affectant l'évaluation à réaliser au maximiseur d'EI, on assure donc, en moyenne, une diminution optimale sur un coup de $M_n - F^*$.

L'EI s'écrit en outre directement comme

$$\mathrm{EI}(\mathbf{x}) = \hat{\sigma}(\mathbf{x}) \left[u\Phi(u) + \Phi'(u)\right], \tag{1.19}$$

avec

$$u = \frac{f_{\min} - \hat{f}(\mathbf{x})}{\hat{\sigma}(\mathbf{x})},$$

et $f_{\min} = \min_{\mathbf{x}_i\in\mathbb{S}_n} f(\mathbf{x}_i)$ l'estimation courante du minimum.

Ce critère permet donc, sans ajout de paramètre de température plus ou moins arbitraire, de réaliser un compromis entre exploration et recherche locale. L'allure du critère lors de l'optimisation de f_{test}, présentée sur la figure 1.5, semble indiquer que ni la recherche globale ni la recherche locale ne seront négligées. On constate aussi que l'évaluation optimale est la même que celle choisie en minimisant la borne inférieure de l'intervalle de confiance à 95%. Ce n'est pas le cas en général. Pour s'en convaincre, considérons la situation présentée sur la figure 1.6 où les évaluations réalisées ont conduit à une mauvaise estimation des paramètres de la covariance. L'optimisation de l'EI incite à l'exploration, alors que l'optimisation de la borne inférieure de l'intervalle de confiance à 95%, dangereusement fausse, entraînerait une accumulation d'évaluations au voisinage du minimiseur local. La recherche à l'aide de l'EI semble donc plus globale, et plus robuste à un mauvais choix de covariance.

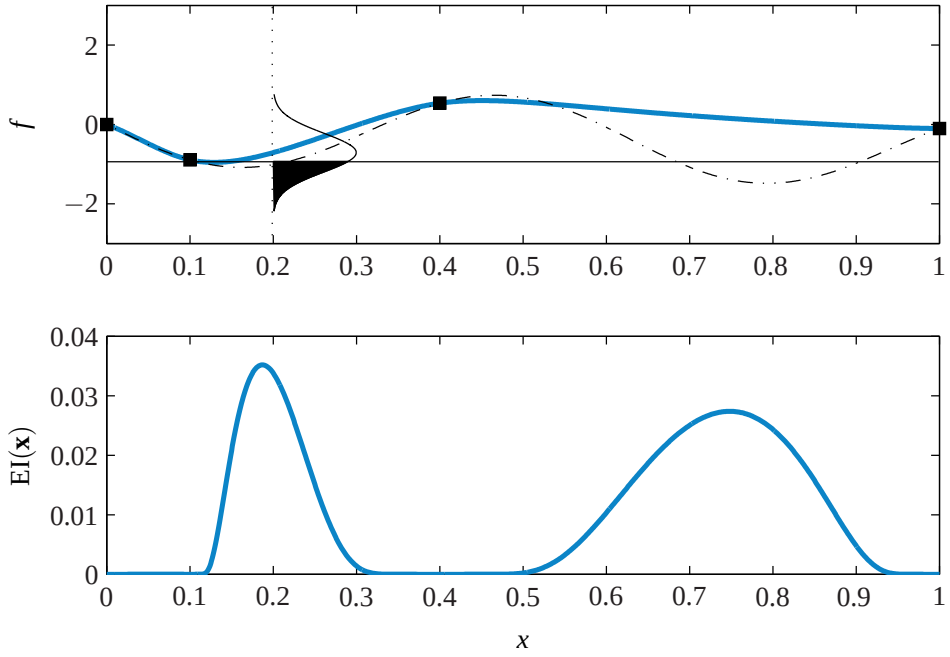


FIG. 1.5: (Partie supérieure) *Prédiction de f (conventions graphiques identiques à celles de la figure 1.3). La loi du résultat d'évaluation en 0.2 est présentée. L'aire noircie représente l'intérêt de l'évaluation dans le cadre du P-algorithme (avec $T = f_{min}$).* (Partie inférieure) *Amélioration attendue en fonction du point d'évaluation.*

1.3.4 Maximiser la crédibilité

Une des conclusions les plus importantes de Jones (2001) est que les critères présentés jusqu'à présent (au même titre que le critère de minimisation de l'ECM présenté par la suite) peuvent se comporter de manière peu efficace si le plan d'expériences initial est trompeur. Un exemple symptomatique en est donné à la page 373 de son article, où une fonction sinus est échantillonnée en utilisant sa propre période. Il s'en suit une prédiction constante et une erreur de prédiction supposée très faible. Pour faire face à ce manque de robustesse à un mauvais choix de covariance, l'idée avancée par Jones est alors de chercher simultanément le prochain point d'évaluation *et* les paramètres de la covariance.

Supposons que la moyenne de F soit nulle[3] et, dans un premier temps, que l'on souhaite non plus trouver le minimum global, mais atteindre une cible T. Le critère proposé par Jones consiste à chercher le point d'évaluation $\mathbf{x}$ et les paramètres de la covariance $\boldsymbol{\theta} = [\nu, \rho, \sigma]^T$ qui maximisent la vraisemblance des résultats d'évaluations disponibles conditionnellement à $\{F(\mathbf{x}) = T\}$. Ce critère s'écrit

$$\max_{\boldsymbol{\theta},\mathbf{x}} L(\mathcal{F}_n | F(\mathbf{x}) = T, \boldsymbol{\theta}), \tag{1.20}$$

[3] On peut, plus généralement, la supposer connue et la soustraire aux observations.

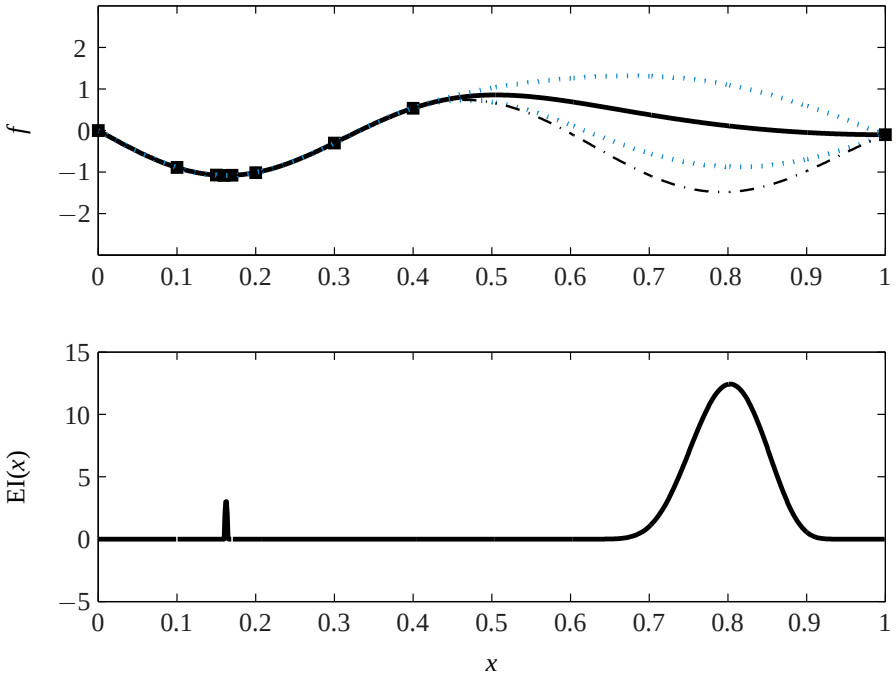


FIG. 1.6: *Même contenu que sur la figure 1.5, avec un plan d'expériences initial plus complet (obtenu par exemple à la suite d'une recherche locale). EI n'est pas perturbé et indique qu'il faut s'éloigner du minimiseur courant dont le voisinage est déjà bien exploré. A contrario, optimiser la borne inférieure* $\hat{f}(\mathbf{x}) - 2\hat{\sigma}(\mathbf{x})$ *conduirait à accumuler davantage d'évaluations au voisinage du minimiseur local.*

où

$$L(\mathcal{F}_n|F(\mathbf{x})=T,\boldsymbol{\theta})=\frac{L(\mathcal{F}_n,F(\mathbf{x})=T|\boldsymbol{\theta})}{L(F(\mathbf{x})=T|\boldsymbol{\theta})},$$

avec

$$L(\mathcal{F}_n,F(\mathbf{x})=T|\boldsymbol{\theta})=\frac{1}{(2\pi)^{(n+1)/2}|\mathbf{K}(\boldsymbol{\theta},\mathbf{x})|^{1/2}}\exp\left[-\frac{1}{2}\mathbf{f}_{n,T}^{\mathsf{T}}\mathbf{K}^{-1}(\boldsymbol{\theta},\mathbf{x})\mathbf{f}_{n,T}\right] \tag{1.21}$$

la vraisemblance des résultats obtenus et de l'apparition de la valeur T en $\mathbf{x}$ et

$$L(F(\mathbf{x})=T|\boldsymbol{\theta})=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{T^2}{2\sigma^2}\right] \tag{1.22}$$

la vraisemblance de l'apparition de la valeur T en $\mathbf{x}$, où

$$\mathbf{K}(\boldsymbol{\theta},\mathbf{x})=\begin{pmatrix}\mathbf{K}(\boldsymbol{\theta}) & \mathbf{k}(\mathbf{x},\boldsymbol{\theta})\\ \mathbf{k}(\mathbf{x},\boldsymbol{\theta})^{\mathsf{T}} & k(0)\end{pmatrix},$$

et $\mathbf{f}_{n,T}^{\mathsf{T}}=\left[\mathbf{f}_n^{\mathsf{T}},T\right]$. Notons la dépendance en $\boldsymbol{\theta}$ de $\mathbf{K}$ et de $\mathbf{k}$ dans (1.21) et (1.22), qui avait été omise jusqu'à présent dans un souci de simplification.

Pour un point d'évaluation candidat $\mathbf{x}$ donné, la vraisemblance optimale ($\max_{\boldsymbol{\theta}} L(\mathcal{F}_n|F(\mathbf{x})=T,\boldsymbol{\theta})$) peut être vue comme une mesure de la confiance (ou *crédibilité* chez Jones) que l'on peut avoir en l'hypothèse d'une valeur T en $\mathbf{x}$. L'évaluation est ensuite réalisée au point qui maximise cette mesure. L'avantage de cette approche est qu'elle ne repose pas sur des paramètres de covariance estimés uniquement à partir des données. Elle semble donc offrir une robustesse plus importante par rapport à un échantillon initial trompeur.

Pour étendre ce principe à l'optimisation, Jones propose, tout comme Gutmann (2001) qui utilise une approche similaire mais reposant sur le formalisme des splines, de procéder comme avec le P-algorithme, et d'utiliser plusieurs cibles $T\leq f_{\min}$ à tour de rôle (ou simultanément si une parallélisation est possible).

Si l'on connaît, même approximativement, le minimum global f^*, et si l'on cherche sa position (par exemple, pour des problèmes d'identification où la fonction objectif est positive et est idéalement nulle à l'optimum), il nous semble plus légitime de tester l'hypothèse d'apparition de la valeur f^* *et* d'annulation simultanée du gradient.

Nous proposons donc de modifier le critère de Jones et de maximiser cette fois la confiance que l'on peut avoir dans l'apparition d'un extremum de valeur f^* en $\mathbf{x}$, ce qui peut s'écrire

$$\max_{\boldsymbol{\theta},\mathbf{x}} L(\mathcal{F}_n|F(\mathbf{x})=f^*,\nabla F(\mathbf{x})=\mathbf{0},\boldsymbol{\theta}), \tag{1.23}$$

où

$$L(\mathcal{F}_n|F(\mathbf{x})=f^*,\nabla F(\mathbf{x})=\mathbf{0},\boldsymbol{\theta})=\frac{L(\mathcal{F}_n,F(\mathbf{x})=T,\nabla F(\mathbf{x})=\mathbf{0}|\boldsymbol{\theta})}{L(F(\mathbf{x})=T,\nabla F(\mathbf{x})=\mathbf{0}|\boldsymbol{\theta})},$$

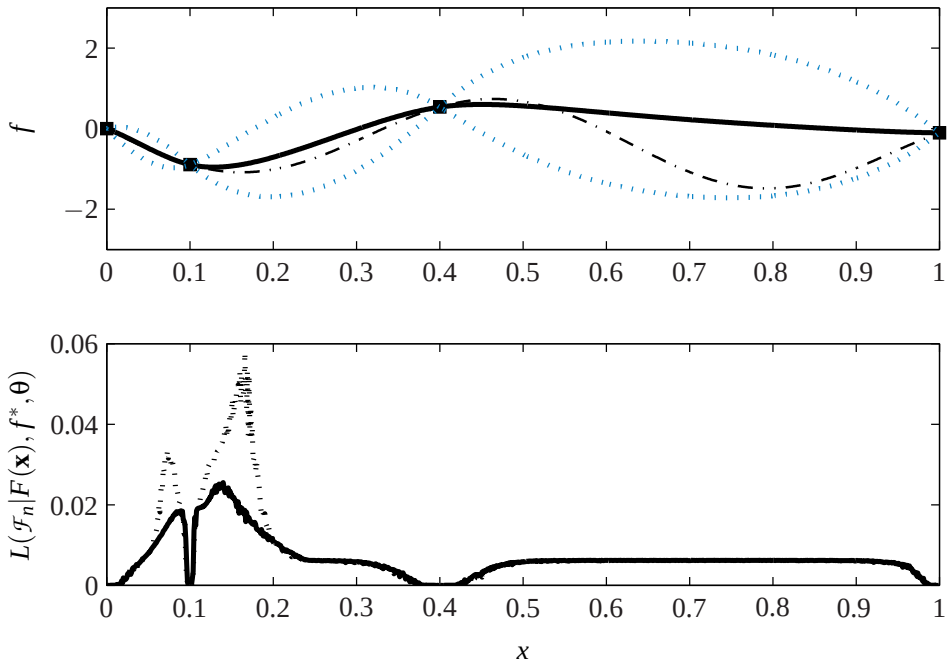


FIG. 1.7: (Partie supérieure) *Prédiction de f_{test} identique à celles présentées précédemment.* (Partie inférieure) *Crédibilité* $\max_{\boldsymbol{\theta}} L(\mathcal{F}_n|F(\mathbf{x}) = f^*, \boldsymbol{\theta})$ *(trait plein) correspondant à la prédiction et version modifiée imposant en outre une annulation du gradient de F (trait en pointillés).*

avec $\nabla F(\mathbf{x})$ le gradient de F. La loi jointe de F et $\nabla F(\mathbf{x})$ est encore gaussienne et se calcule simplement (cf. l'annexe B.3). Quand f^* n'est pas connu, on peut procéder comme pour la maximisation de la probabilité d'amélioration (cf. la section 1.3.2) et utiliser alternativement plusieurs valeurs pour la cible T.

Pour optimiser f_{test}, supposons que l'on connaisse approximativement son minimum, par exemple $f^* = -1.2$ (la vraie valeur est -1.5). La figure 1.7 présente alors les deux mesures de crédibilité présentées ici, c'est-à-dire $\max_{\boldsymbol{\theta}} L(\mathcal{F}_n|F(\mathbf{x}) = f^*, \boldsymbol{\theta})$ et $\max_{\boldsymbol{\theta}} L(\mathcal{F}_n|F(\mathbf{x}) = f^*, \nabla F(\mathbf{x}) = \mathbf{0}, \boldsymbol{\theta})$.

La figure 1.8 présente ensuite la prédiction par krigeage optimale pour chacun des deux critères, ou, en d'autres termes, la prédiction correspondant au couple $(\mathbf{x}, \boldsymbol{\theta})$ qui maximise $L(\mathcal{F}_n|F(\mathbf{x}) = f^*, \boldsymbol{\theta})$ (ou $L(\mathcal{F}_n|F(\mathbf{x}) = f^*, \nabla F(\mathbf{x}) = \mathbf{0}, \boldsymbol{\theta})$). Cette dernière prédiction va donc interpoler les résultats d'évaluation et satisfaire $\hat{f}(\mathbf{x}) = f^*$ (et $\nabla \hat{f} = 0$ pour la version modifiée). Bien que les points choisis pour l'évaluation soient assez proches, les distributions a posteriori sont très différentes. En effet, pour le critère modifié, la prédiction prend en compte l'annulation du gradient, ce qui diminue fortement l'incertitude associée à la prédiction au voisinage du point d'annulation. Remarquons aussi, qu'avec le critère de Jones, f^* n'est pas le minimum de la prédiction.

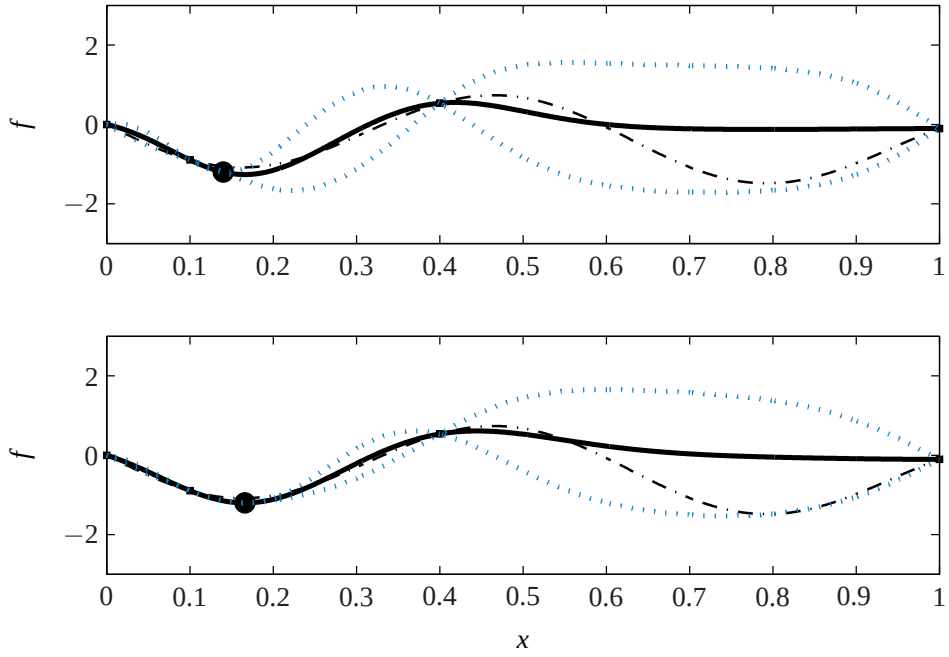


FIG. 1.8: *Prédiction par krigeage reposant sur les évaluations initiales et sur le couple* $(x, \boldsymbol{\theta})$ *maximisant le critère de Jones* (partie supérieure) *et le critère modifié incluant le gradient* (partie inférieure). *La prédiction à l'aide du gradient est présentée dans l'annexe B.3.*

Remarque 1.4. Supposons que les paramètres de la covariance $\boldsymbol{\theta}$ sont connus *a priori*. En remarquant que

$$L(\mathcal{F}_n|F(\mathbf{x})=T,\boldsymbol{\theta}) = \frac{L(F(\mathbf{x})=T|\mathcal{F}_n,\boldsymbol{\theta})L(\mathcal{F}_n|\boldsymbol{\theta})}{L(F(\mathbf{x})=T|\boldsymbol{\theta})},$$

maximiser la crédibilité revient à maximiser $L(F(\mathbf{x})=T|\mathcal{F}_n,\boldsymbol{\theta}) = \Phi'(\frac{T-\hat{f}(\mathbf{x})}{\hat{\sigma}(\mathbf{x})})$, ce qui va ressembler fortement au P-algorithme pour un seuil T donné.

Remarque 1.5. Un des désavantages de cette méthode est qu'elle suppose la moyenne de F connue. Sans cette hypothèse, on pourrait envisager de conserver le principe de ce critère, mais d'utiliser une autre méthode d'estimation des paramètres, comme le maximum de vraisemblance restreint (Stein, 1999) ou la validation croisée.

1.4 Entropie conditionnelle des minimiseurs

Tous les critères présentés précédemment évaluent, plus ou moins explicitement, f en un point dont ils espèrent qu'il soit un minimiseur global. Cependant, dans bien des situations, il peut être plus efficace d'échantillonner la fonction dans le seul but d'écarter des parties de l'espace de recherche où un tel minimiseur n'a que peu de chances d'apparaître.

Dans Villemonteix et al. (2008b), un critère partant de ce constat a été proposé. L'idée est de

mesurer l'attrait d'une évaluation candidate par l'information qu'elle peut apporter sur la position du minimum global, et de réaliser l'évaluation au point qui maximise le gain d'information attendu. Cette information est mesurée par l'entropie conditionnelle des minimiseurs globaux (définie dans la section suivante) et la fonction est évaluée au point qui apporte potentiellement la plus grande réduction de cette entropie. Cette approche est analogue à celle utilisée par Geman et Jedynak (1995) — baptisée SUR pour *Stepwise Uncertainty Reduction* — dans le contexte tout à fait différent de la détection de routes sur des images satellites. Plus récemment (Vergassola et al., 2007), indépendamment des travaux de Geman, cette stratégie a aussi été utilisée, sous le nom d'« infotaxie » pour modéliser la stratégie de déplacement d'une bactérie en quête d'une source de nourriture, et par Bettinger et al. (2008) pour de la planification d'expériences lorsque l'on souhaite inverser le système étudié. Cette stratégie est finalement assez naturelle pour tous les problèmes de recherche à mener avec peu d'information. Cette section détaille son application à l'optimisation globale.

Remarque 1.6. Notons que, plus généralement, l'entropie est un outil classique pour mesurer l'incertitude dans les problèmes de recherche. Mentionnons par exemple les travaux de Hill (1978) sur la recherche d'un plan d'expériences optimal pour le choix d'une structure de modèles, ou encore les travaux plus généraux de Pronzato et al. (1997).

1.4.1 Distribution de probabilité des minimiseurs

Plutôt que de s'intéresser, comme ce fût le cas jusqu'à présent, à un estimateur du minimum, nous préférons considérer la distribution de probabilité des minimiseurs globaux de F. L'*a priori* gaussien sur f implique en effet un *a priori* sur $\mathbf{x}^*$, minimiseur global de f. Plus formellement, soit $\mathcal{M}_{\mathbb{X}}$ l'ensemble (*aléatoire*) des minimiseurs globaux de F sur $\mathbb{X}$, c'est-à-dire

$$\mathcal{M}_{\mathbb{X}} = \{\mathbf{x}^* \in \mathbb{X} | F(\mathbf{x}^*) = \min_{\mathbf{u} \in \mathbb{X}} F(\mathbf{u})\}.$$

A chaque trajectoire de F correspond une réalisation de $\mathcal{M}_{\mathbb{X}}$ contenant un ou plusieurs minimiseurs globaux. Pour s'assurer que $\mathcal{M}_{\mathbb{X}}$ est non vide avec une probabilité 1, il suffit de supposer que F a des trajectoires continues avec une probabilité 1, ce qui est toujours le cas lorsque l'on utilise des covariances classiques et en particulier la classe des covariances de Matèrn (cf. Abrahamsen, 1997).

Nous souhaiterions ensuite étudier la densité de probabilité $p_{\mathbf{X}^*_{\mathbb{X}}}(\mathbf{x}|\mathcal{F}_n)$ d'un vecteur aléatoire $\mathbf{X}^*_{\mathbb{X}}$ uniformément distribué sur $\mathcal{M}_{\mathbb{X}}$ conditionnellement aux résultats d'évaluations disponibles $\mathcal{F}_n$. Cette densité contient en effet tout ce qui a été appris et supposé sur les minimiseurs globaux de f. Cependant, l'existence de cette densité n'est pas simple à établir, et nous ne disposons pas d'une expression analytique utilisable (cf. Adler (2000) ; Sjö (2000) pour une description de l'état de l'art dans l'étude des extrema de processus gaussiens). Nous allons donc travailler sur un sous-ensemble discret de $\mathbb{X}$. La méthode pour l'approximation de la distribution de probabilité des

minimiseurs sera présentée dans la section 2.2.1. Préciserons maintenant l'utilisation qui en est faite.

Soit $\mathbb{G} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ un sous ensemble fini de $\mathbb{X}$, tel que $\mathbb{S}_n \subset \mathbb{G}$, et $\mathcal{M}_{\mathbb{G}}$ l'ensemble des minimiseurs globaux de F sur $\mathbb{G}$. Soit $\mathbf{X}^*$ un vecteur aléatoire uniformément distribué sur $\mathcal{M}_{\mathbb{G}}$, la distribution de probabilité de ce vecteur aléatoire conditionnellement aux résultats disponibles

$$P_{\mathbf{X}^*}(\mathbf{x}|\mathcal{F}_n) = \mathrm{P}(\mathbf{X}^* = \mathbf{x} \,|\, \mathbf{F}_{\mathbb{S}} = \mathbf{f}_{\mathbb{S}}) \,.$$

que nous appellerons désormais *distribution conditionnelle des minimiseurs globaux* ou plus simplement distribution conditionnelle, va permettre non seulement d'estimer les minimiseurs globaux de f (par exemple au travers des modes de cette distribution), mais aussi de caractériser l'incertitude associée à cette estimation.

$P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ résume la connaissance acquise sur les minimiseurs de f compte tenu des hypothèses formulées sur la covariance. Pour illustrer ce point, la figure 1.9 présente non seulement la distribution conditionnelle engendrée par l'exemple utilisé précédemment, mais aussi la distribution conditionnelle engendrée par les mêmes résultats d'évaluation en supposant une covariance différente. Jusqu'à présent, la régularité de la covariance de Matèrn était fixée à $\nu = 2.2$ après estimation à partir des données. Ici nous avons artificiellement augmenté la régularité pour constater son influence sur $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$. Les trajectoires de F sont ainsi plus régulières, ce qui a une influence logique sur la position probable de leurs minimiseurs globaux.

1.4.2 Entropie conditionnelle

L'entropie d'une variable discrète U (exprimée en bits) est définie comme

$$H(U) = -\sum_u \mathrm{P}(U = u) \log_2 \mathrm{P}(U = u).$$

$H(U)$ mesure l'étendue de la distribution de U et décroît à mesure que cette distribution se resserre. En particulier :

- $\forall \mathbf{x} \in \mathbb{G}\, P_{\mathbf{X}^*}(\mathbf{x}|\mathcal{F}_n) = 1/\mathrm{card}(\mathbb{G}) \Rightarrow H(\mathbf{X}^*) = \log_2(N)$,
- $P_{\mathbf{X}^*}(\mathbf{x}|\mathcal{F}_n) = \begin{cases} 0 & \text{si} \quad \mathbf{x} \neq \mathbf{x}_0 \\ 1 & \text{si} \quad \mathbf{x} = \mathbf{x}_0 \end{cases} \Rightarrow H(\mathbf{X}^*) = 0$

De même, pour tout événement $\mathcal{B}$, l'entropie de U conditionnellement à $\mathcal{B}$ est

$$H(U|\mathcal{B}) = -\sum_u \mathrm{P}(U = u|\mathcal{B}) \log_2 \mathrm{P}(U = u|\mathcal{B}).$$

L'entropie de U conditionnellement à la variable continue V de densité p_V vaut

$$H(U|V) = \int_v p_V(v) H(U|V = v),$$

et l'entropie conditionnelle de U sachant $\mathcal{B}$ et V est

$$H(U|\mathcal{B}, V) = \int_v p_V(v|\mathcal{B}) H(U|\mathcal{B}, V = v). \tag{1.24}$$

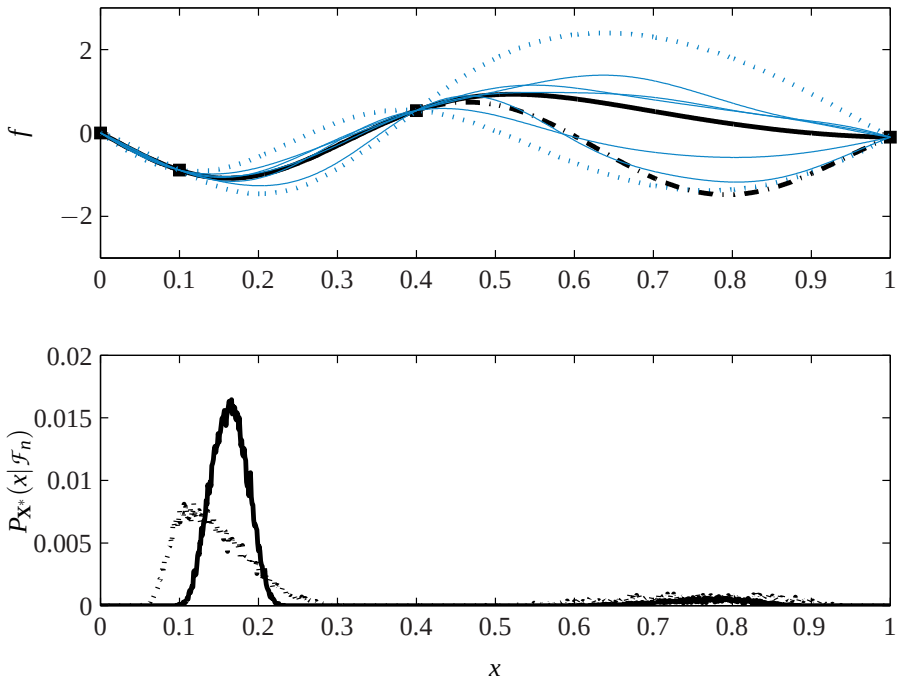


FIG. 1.9: *Influence de la covariance sur l'estimée de la distribution conditionnelle des minimiseurs globaux.* (Partie supérieure) *Prédiction de la fonction test utilisée dans ce chapitre (même conventions graphiques). Les paramètres de la covariance utilisés ici diffèrent de ceux utilisés pour les exemples de cette section (figure 1.4, figure 1.5, etc.). La régularité* ν*, estimée par maximum de vraisemblance à 2.2, a ici été fixée à 8.* (Partie inférieure) *Distribution conditionnelle des minimiseurs globaux correspondant aux résultats des évaluations (les carrés sur la figure de la partie supérieure) pour une régularité* $\nu = 8$ *(trait plein) et pour une régularité* $\nu = 2.2$ *(traits pointillés). La distribution est estimée en utilisant des trajectoires conditionnelles (traits fin sur la figure de la partie supérieure) simulées grâce à la méthode de conditionnement par krigeage présentée à la section 2.2.1.*

Notons que, malgré la similarité de notation avec l'espérance conditionnelle, l'entropie conditionnelle n'est pas une variable aléatoire mais une grandeur déterministe (Cover et Thomas, 1991).

1.4.3 Entropie conditionnelle des minimiseurs

Pour une itération du processus d'optimisation, quantifions la connaissance accumulée sur les minimiseurs globaux par l'entropie de $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$, c'est-à-dire

$$H(\mathbf{X}^*|\mathcal{F}_n) = -\sum_{\mathbf{x}\in\mathbb{G}} P_{\mathbf{X}^*}(\mathbf{x}|\mathcal{F}_n)\log_2(P_{\mathbf{X}^*}(\mathbf{x}|\mathcal{F}_n)).$$

Plus le nombre des évaluations réalisées augmente, plus cette quantité diminue, pour s'annuler si l'incertitude sur le(s) minimiseur(s) a disparue, ce qui sera le cas si $\mathbb{G}$ a été exploré systématiquement. Dans un contexte d'évaluations coûteuses, la résolution exacte du problème, c'est à dire l'annulation de l'entropie n'est pas envisageable. En revanche, nous verrons que l'on peut s'en approcher en réalisant les évaluations dans le seul but de réduire l'entropie de $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$.

Pour traiter ce problème, que l'on peut voir comme un problème de planification d'expériences, nous proposons de mesurer l'intérêt d'une évaluation potentielle en $\mathbf{x} \in \mathbb{X}$ par l'entropie de $\mathbf{X}^*$ sachant $\mathcal{F}_n$ moyennée sur tous les résultats d'évaluation possibles en $\mathbf{x}$, c'est-à-dire

$$H_n(\mathbf{x}) = H(\mathbf{X}^*|\mathcal{F}_n, F(\mathbf{x})).$$

L'évaluation est ensuite réalisée au point qui minimise l'entropie conditionnelle des minimiseurs (ECM), ou en d'autres termes qui maximise le gain d'information attendu sur les minimiseurs

$$\mathbf{x}_{n+1} = \arg\min_{\mathbf{x}\in\mathbb{X}} H_n(\mathbf{x}).$$

A partir de la définition (1.24), on peut écrire l'ECM comme

$$H_n(\mathbf{x}) = \int_{y\in\mathbb{R}} p_{F(\mathbf{x})}(y|\mathcal{F}_n) H(\mathbf{X}^*|\mathcal{F}_n, F(\mathbf{x}) = y)dy, \tag{1.25}$$

avec $p_{F(\mathbf{x})}(\cdot|\mathcal{F}_n)$ la distribution conditionnelle du résultat d'évaluation $F(\mathbf{x})$ (gaussienne de moyenne et de variance obtenues par krigeage) et

$$H(\mathbf{X}^*|\mathcal{F}_n, F(\mathbf{x}) = y) = -\sum_{\mathbf{u}\in\mathbb{G}} P_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n, F(\mathbf{x}) = y)\log_2(P_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n, F(\mathbf{x}) = y)) \tag{1.26}$$

l'entropie de $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n, F(\mathbf{x}) = y)$, la distribution de $\mathbf{X}^*$ conditionnellement à $\mathcal{F}_n$ et $\{F(\mathbf{x}) = y\}$. Le mode de calcul de H_n — qui nécessite entre autres l'approximation de $p_{F(\mathbf{x})}(\cdot|\mathcal{F}_n)$ et de l'intégrale (1.25)— sera traité dans la section 2.2.1.

La figure 1.10 présente l'allure de $H_n(\mathbf{x})$ lors de l'optimisation de f_{test}. Le point retenu pour l'évaluation correspond au maximum de la densité conditionnelle, ce qui n'est pas le cas en général comme en témoigne l'exemple de la figure 1.11, où le critère met en avant son caractère global.

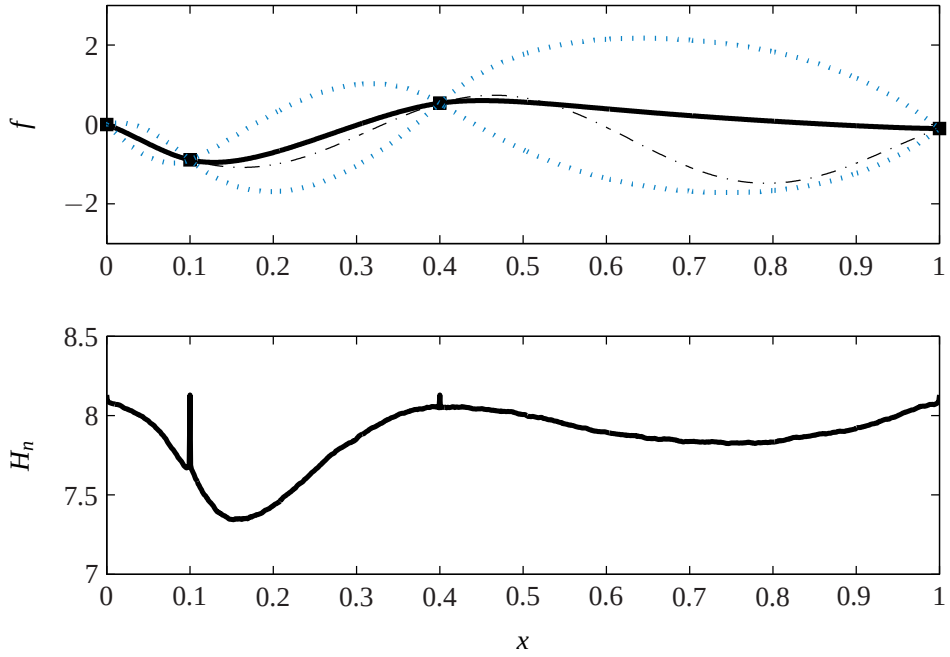


FIG. 1.10: $H_n(\mathbf{x})$ (partie inférieure) *correspondant à la prédiction de* f_{test} *présentée sur la* partie supérieure *(même conventions graphiques que précédement). Le minimiseur de* $H_n(\mathbf{x})$ *est identique au maximiseur de la distribution conditionnelle des minimiseurs (cf. figure 1.9). Ce n'est plus le cas sur la figure 1.11.*

Pour mieux comprendre l'intérêt de l'ECM face à l'EI, considérons la fonction présentée sur la figure 1.12. A partir des trois mêmes points initiaux, l'ECM et l'EI sont optimisés et les points choisis par chacun des deux critères sont présentés sur la figure 1.12 tout comme les prédictions et les distributions conditionnelles à l'issue de la nouvelle évaluation recommandée par chacune des approches. Pour cet exemple, la régularité de la covariance de Matèrn a été fixée *a priori* a une valeur élevée (ν =2.5). L'évaluation proposée par l'ECM permet non seulement d'apprendre sur la zone à gauche de $x = 2$ (du fait de la régularité supposée de la fonction), mais aussi de vérifier que le minimiseur ne se trouve pas à droite de $x = 2$. Ainsi, l'ECM tire partie de la régularité pour conclure plus rapidement que l'EI, qui propose d'évaluer f près du bord du domaine. La distribution conditionnelle résultant de l'évaluation choisie par ECM est donc plus piquée (cf. les figure 1.12(b) et figure 1.12(c)).

Remarque 1.7. L'ECM $H_n(\mathbf{x})$ peut être vue, au même titre que l'EI dans la section 1.3.3, comme une fonction de risque. Dans ce cas, la fonction de perte associée est l'entropie $H(\mathbf{X}^*|\mathcal{F}_n, F(\mathbf{x}) = y)$ de $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n, F(\mathbf{x}) = y)$, la distribution conditionnelle des minimiseurs après $n+1$ évaluations. Cette remarque sera surtout utile dans la section 4.3 pour justifier le choix des mesures de convergence retenues pour la comparaison des critères présentés dans ce chapitre.

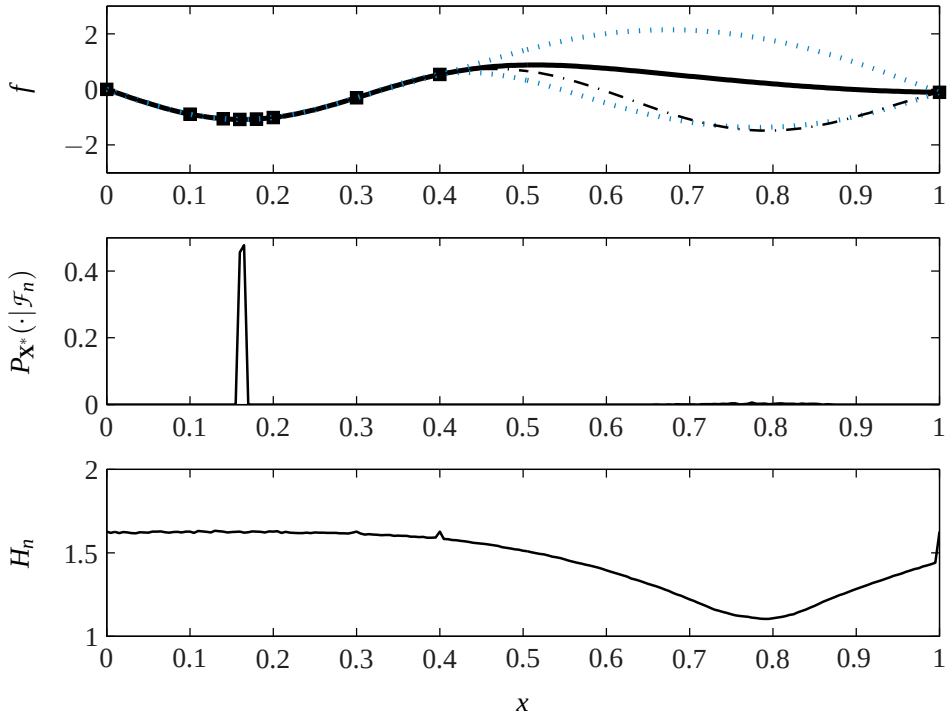


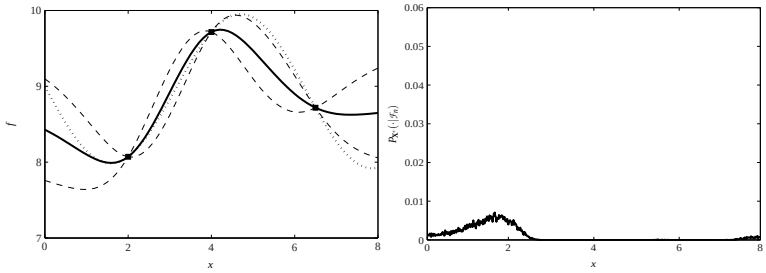
FIG. 1.11: *De haut en bas : prédiction de f_{test} identique à celle de la figure 1.6, distribution conditionnelle correspondante, entropie conditionnelle des minimiseurs (ECM)*

Remarque 1.8. Aux points d'évaluations, l'entropie conditionnelle est égale à l'entropie de la densité conditionnelle. En effet, refaire la même évaluation n'apportera pas d'information sur $\mathbf{x}^*$ (dans le cas où les résultats des évaluations sont considérés comme exacts).

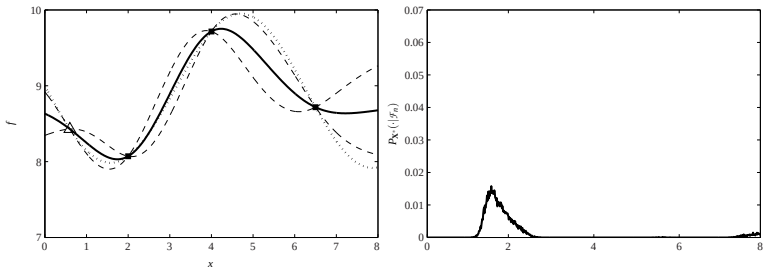
1.5 Discussion

Dans ce chapitre, les critères d'échantillonnage classiques reposant sur la prédiction par krigeage ont été présentés, ainsi qu'un nouveau critère dont le principe nous semble correspondre davantage aux problèmes coûteux qui nous occupent. Au-delà des intuitions, comment rationaliser le choix entre les critères possibles ? Pour décider, deux aspects nous semblent essentiels, à savoir la vitesse de convergence et la robustesse à un mauvais choix de covariance. Précisons ces deux points.

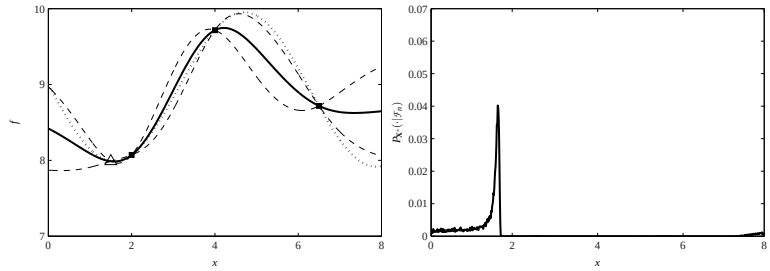
Contrairement à ce que l'on peut souhaiter d'un algorithme d'optimisation classique, ce ne sont ni la vitesse de convergence asymptotique ni l'assurance de déterminer avec précision l'optimum qui nous motivent. En effet, compte-tenu du budget d'évaluations, le minimum global ne pourra être estimé avec précision. En revanche, les vitesses de convergence non-asymptotiques, ou dit autrement, les conséquences des toutes premières évaluations choisies sur l'estimation du minimum et des minimiseurs globaux, sont essentielles pour s'assurer de l'intérêt d'un critère.



(a) Prédiction initiale et distribution conditionnelle



(b) Prédiction et distribution conditionnelle suite à une évaluation choisie par maximisation de l'EI



(c) Prédiction et distribution conditionnelle suite à une évaluation choisie par minimisation de l'ECM

FIG. 1.12: *Comparaison de l'optimisation avec l'ECM et avec l'EI : les figures de* gauche *présentent la prédiction par krigeage avant et après une évaluation choisie avec l'ECM ou l'EI. Les figures de* droite *présentent les distributions conditionnelles correspondantes.*

En pratique, un critère qui permet d'obtenir rapidement une estimation, même grossière, de l'optimum aura un intérêt, même si ce critère se comporte mal asymptotiquement. Aussi laisserons nous de côté la question de la convergence asymptotique de l'algorithme pour insister dans le chapitre 4 sur les taux de convergence non asymptotiques.

Comme mentionné précédemment, le choix de la covariance, ou dans notre cas des paramètres de la covariance de Matèrn, se fait soit *a priori*, soit par estimation à partir des données disponibles. Cependant, dans un contexte où le nombre d'évaluations réalisées est très faible au regard de la dimension du problème, l'incertitude associée à cette estimation sera quoiqu'il arrive très élevée. Ce constat amène deux remarques importantes. Premièrement, les paramètres de la covariance devraient, sauf cas particuliers[4], être choisis *a priori* à partir des connaissances des experts du domaine, ou de manière à assurer une recherche globale (ce qui limite l'intérêt du critère de maximisation de la crédibilité, puisque à paramètres fixés, il est proche du P-algorithme). Deuxièmement, la robustesse des critères par rapport à un mauvais choix de covariance est d'autant plus importante si l'on envisage par souci de simplification que les paramètres de celle-ci peuvent rester fixés. Ainsi, nous pouvons écarter la minimisation d'une borne inférieure, par trop sensible. Pour les autres critères, nous étudierons au chapitre 4 l'influence d'un mauvais choix de covariance sur les vitesses de convergence non asymptotiques. Nous écartons aussi la maximisation de la crédibilité du fait de l'hypothèse d'une moyenne connue pour F qu'elle nécessite.

C'est au chapitre 4 que nous déciderons de l'intérêt du P-algorithme, de l'EI ou de l'ECM.

[4] Par exemple, lorsque l'on dispose d'un grand nombre de données et que l'on souhaite ajouter quelques évaluations pour optimiser, ou encore, lorsque une optimisation a déjà été réalisée sur une fonction suffisamment proche pour que l'on puisse réutiliser les données (phénomène courant dans l'industrie ou de petites modifications des pièces en fin de projet obligent à ré-optimiser, mais conservent l'allure de la fonction objectif).

CHAPITRE 2

ALGORITHME IAGO

Résumé — Ce chapitre discute de la mise en œuvre des critères d'échantillonage présentés au chapitre 1. L'objectif est d'aboutir à un algorithme d'optimisation globale apte à traiter les problèmes présentés dans l'introduction. Une attention plus particulière est accordée au critère de maximisation de l'ECM et à l'algorithme IAGO qui en découle. Après avoir introduit le schéma général de cet algorithme et détaillé les étapes indépendantes du critère d'échantillonnage (plan d'expérience initial, choix de la covariance, critère d'arrêt...), nous présentons plusieurs extensions de IAGO, notamment à l'optimisation sous contrainte ou à l'optimisation en présence de bruit sur les résultats des évaluations.

2.1 Introduction

Nous avons vu au chapitre 1 comment, lorsque le budget d'évaluation est limité, remplacer un problème d'optimisation globale par une série de problèmes plus simples d'optimisation d'un critère d'échantillonnage. Nous nous emploierons ici à décrire plus précisément la mise en place de ce type d'approche et en particulier, l'utilisation pratique du critère de minimisation de l'entropie conditionnelle des minimiseurs (ECM).

Notre algorithme d'optimisation globale utilisant l'ECM, que nous avons dénommé IAGO (pour *Informational Approach to Global Optimization*) est similaire à l'algorithme EGO (pour *Efficient Global Optimization*) proposé par Jones et al. (1998) qui se démarque des travaux l'ayant précédé par un réel souci des aspects pratiques de l'utilisation du critère EI. L'algorithme 1 décrit ci-après, qui constitue la version initiale d'EGO telle que présentée dans Jones et al. (1998), choisit comme nouveau point d'évaluation un maximiseur d'EI, et remet à jour l'estimée des paramètres de la covariance ainsi que la prédiction par krigeage après chaque nouvelle évaluation. Dans ce chapitre, le schéma général de cet algorithme sera conservé, mais l'ECM sera utilisé à la place de l'EI (modification de l'étape 10 de l'algorithme 1). Par la suite, nous remettrons aussi en question le critère d'arrêt (étape 9), la réestimation des paramètres de la covariance (étape 11) et le plan

d'expériences initial (étape 1). Enfin, pour faire face à la variété des problèmes rencontrés dans l'industrie, nous discuterons de plusieurs extentions de IAGO (et de EGO). La première est adaptée à l'optimisation globale en présence de bruit sur les résultats des évaluations. La seconde permet de mettre à profit d'éventuelles évaluations du gradient de f. La troisième est destinée à la prise en compte des contraintes. Nous discuterons aussi des possibilités d'adaptation des algorithmes lorsque la parallélisation des évaluations est envisageable.

Algorithme 1 Algorithme EGO

1: Évaluer f sur un plan d'expériences initial
2: Estimer les paramètres d'une covariance exponentielle (non isotrope) par maximum de vraisemblance
3: Calculer la prédiction par krigeage
4: **Si** La prédiction par krigeage n'est pas satisfaisante **Alors**
5: Transformer les données ou arrêter l'algorithme (transformation logarithme ou inverse)
6: **Fin Si**
7: Calculer $f_{\min}$ le minimum des résultats d'évaluations effectués
8: $\mathrm{EI}_{\max} \leftarrow \max_{\mathbf{x}\in\mathbb{X}} \mathrm{EI}(\mathbf{x})$
9: **Tant que** $\mathrm{EI}_{\max} \geq 0.01 f_{\min}$ **Faire**
10: Évaluer f au maximiseur de EI
11: Réestimer les paramètres de la covariance en tenant compte du résultat de la nouvelle évaluation
12: Mettre à jour la prédiction par krigeage
13: Mettre à jour $\mathrm{EI}_{\max}$ et $f_{\min}$
14: **Fin Tant que**

2.2 Principe de IAGO

Le calcul de l'ECM, au cœur de l'algorithme IAGO, implique une évaluation de l'expression

$$H_n(\mathbf{x}) = \int_{y\in\mathbb{R}} p_{F(\mathbf{x})}(y|\mathcal{F}_n)\left(-\sum_{\mathbf{u}\in\mathbb{G}} P_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n, F(\mathbf{x})=y)\log_2(P_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n, F(\mathbf{x})=y))\right)dy,$$

obtenue à la section 1.4.3 pour $\mathbf{x}\in\mathbb{G}$. Pour approcher cette intégrale, nous utiliserons $F_Q(\mathbf{x})$, une version discrète de $F(\mathbf{x})$, définie par $F_Q(\mathbf{x}) = Q(F(\mathbf{x}))$ avec Q un opérateur de quantification. Q est caractérisé par s nombres réels $\{y_1,\ldots,y_s\}$, et défini $\forall u \in \mathbb{R}$ comme

$$Q(u) = y_k, \text{ avec } k = \min_i |y_i - u|. \tag{2.1}$$

L'ECM va ainsi être approchée par

$$H(\mathbf{X}^*|\mathcal{F}_n, F_Q(\mathbf{x})) = \sum_{i=1}^{s} P_{F_Q(\mathbf{x})}(y_i|\mathcal{F}_n)\left(- \sum_{\mathbf{u}\in\mathbb{G}} P_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n, F(\mathbf{x}) = y_i) \log_2(P_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n, F(\mathbf{x}) = y_i))\right).$$

Pour calculer cette dernière expression, il faut encore être en mesure d'estimer la distribution de $\mathbf{X}^*$ (minimiseur de F sur $\mathbb{G}$) conditionnellement à un ensemble de résultats d'évaluation de f. Un estimateur $\hat{P}_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n)$, utilisant des simulations conditionnelles de F, sera présenté dans la section 2.2.1 et l'ECM sera finalement approchée par

$$\hat{H}_n(\mathbf{x}) = \sum_{i=1}^{s} P_{F_Q(\mathbf{x})}(y_i|\mathcal{F}_n)\left(- \sum_{\mathbf{u}\in\mathbb{G}} \hat{P}_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n, F(\mathbf{x}) = y_i) \log_2(\hat{P}_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n, F(\mathbf{x}) = y_i))\right). \tag{2.2}$$

Cet estimateur de l'entropie n'est sans doute pas le plus satisfaisant en termes de convergence (cf. Beirlant et al., 1997). Nous aurions, par exemple, pu estimer directement, pour $\mathbf{x} \in \mathbb{G}$, $H(\mathbf{X}^*|\mathcal{F}_n, F(\mathbf{x}) = y_i)$ sans utiliser d'estimateur de la distribution conditionnelle des minimiseurs. Cependant, sa mise en œuvre est simple et, comme nous le verrons à la section 2.3.2, une estimation de la distribution conditionnelle des minimiseurs est utile pour l'optimisation de l'ECM. Il conviendra, dans de futurs travaux, d'étudier l'intérêt d'une estimation plus efficace de l'entropie, notamment pour diminuer le nombre de simulations conditionnelles nécessaires.

2.2.1 Approximation de $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ et conditionnement par krigeage

Pour approcher $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$, nous avons choisi de générer des simulations de F conditionnelles à $\mathcal{F}_n$ et de calculer pour chacune d'entre elles, un minimiseur global (s'il en existe plusieurs, on en choisit un au hasard). Si l'on dispose de simulations de F, plusieurs méthodes ont été proposées par les géostatisticiens (Chilès et Delfiner, 1999) pour les transformer en simulations conditionnelles. Parmi celles-ci, c'est le *conditionnement par krigeage* qui sera retenu, car cette méthode permet la factorisation d'une grande partie des calculs lorsque ce conditionnement se répète (cf. la section 2.2.3).

Cette méthode, introduite par G. Matheron, tire parti du caractère non-biaisé de la prédiction par krigeage pour transformer des simulations non conditionnelles en simulations qui interpolent les résultats $\mathbf{f}_n$ des évaluations de f. L'idée principale est de simuler suivant la loi conditionnelle de l'erreur de prédiction $F - \hat{F}$ plutôt que suivant la loi conditionnelle de F. En effet, la loi de l'erreur de prédiction ne dépend ni du résultat des évaluations, ni de la moyenne de F.

Soient Z un processus gaussien de moyenne nulle et de fonction de covariance k (identique à celle de F), $\hat{Z}$ son prédicteur par krigeage à partir des variables aléatoires $Z(\mathbf{x}_i)$, $\mathbf{x}_i \in \mathbb{S}_n$, et

$$T(\mathbf{x}) = \hat{f}(\mathbf{x}) + \left[Z(\mathbf{x}) - \hat{Z}(\mathbf{x})\right], \tag{2.3}$$

où $\hat{f}$ est la prediction par krigeage qui, rappelons le, interpole les données (c'est-à-dire $\forall \mathbf{x}_i \in \mathbb{S}_n$ $\hat{f}(\mathbf{x}_i) = f(\mathbf{x}_i)$).

Le caractère non-biaisé de la prédiction de Z implique que $Z(\mathbf{x}_i) = \hat{Z}(\mathbf{x}_i)$, et donc que $T(\mathbf{x}_i) = f(\mathbf{x}_i)$, $\forall \mathbf{x}_i \in \mathbb{S}$. T est donc tel que ses trajectoires interpolent les valeurs connues de f. Il est alors aisé de vérifier que T possède les mêmes distributions de dimension finie[1] que F conditionnellement aux résultats des évaluations de f (Delfiner, 1977). En effet, l'erreur de prédiction $Z - \hat{Z}$, pour la prédiction de Z, possède la même distribution que l'erreur de prédiction $F - \hat{F}$, pour la prédiction de F. Remarquons enfin que ce sont les mêmes coefficients qui sont utilisés pour la prédiction de F et pour la prédiction de Z. A partir de (1.1), on peut alors réecrire (2.3) comme

$$T(\mathbf{x}) = Z(\mathbf{x}) + \boldsymbol{\lambda}(\mathbf{x})^{\mathsf{T}} \left[\mathbf{f}_n - \mathbf{Z}_n\right], \tag{2.4}$$

avec $\mathbf{Z}_n = [Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n)]^{\mathsf{T}}$.

En résumé, pour simuler F sur $\mathbb{G}$ conditionnellement aux résultats des évaluations de f, il suffit de simuler sur $\mathbb{G}$ un processus gaussien de moyenne nulle, de calculer l'erreur de prédiction pour chacune de ces simulations, et de recentrer la simulation de l'erreur ainsi obtenue autour de $\hat{f}$.

La procédure (illustrée par la figure 2.1) se détaille plus précisément de la manière suivante :

1. Calculer pour chaque point $\mathbf{x}$ de $\mathbb{G}$, le vecteur $\boldsymbol{\lambda}(\mathbf{x})$ des coefficients du krigeage reposant sur les points d'évaluations de $\mathbb{S}_n$.
2. Générer des trajectoires de Z sur $\mathbb{G}$ (si l'on dispose d'un échantillon tiré suivant une loi gaussienne centré réduite, la covariance des données simulées peut être adaptée en utilisant, par exemple, une décomposition de Cholesky).
3. Appliquer (2.4) pour chaque simulation et pour chaque point de $\mathbb{G}$. Ainsi, pour générer $t(\mathbf{x})$, une simulation conditionnelle de $T(\mathbf{x})$ à partir d'une simulation non conditionnelle $z(\mathbf{x})$ de $Z(\mathbf{x})$, il faut appliquer

   $$t(\mathbf{x}) = z(\mathbf{x}) + \boldsymbol{\lambda}(\mathbf{x})^{\mathsf{T}} [\mathbf{f}_n - \mathbf{z}_n], \tag{2.5}$$

   où $\mathbf{z}_n$ contient les valeurs simulées pour Z sur $\mathbb{S}_n$ (dont nous disposons, puisque nous avions supposé au chapitre 1 que $\mathbb{S}_n \subset \mathbb{G}$).

Le conditionnement par krigeage permet alors d'estimer simplement $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$. Considérons pour cela r simulations conditionnelles sur $\mathbb{G}$, et notons, pour chacune, $\mathbf{x}_i^*$ ($i = 1, \ldots, r$) un minimiseur global. Alors, pour tout $\mathbf{x} \in \mathbb{G}$,

$$\hat{P}_{\mathbf{X}^*}(\mathbf{x}|\mathcal{F}_n) = \frac{1}{r} \sum_{i=1}^{r} \delta_{\mathbf{x}_i^*}(\mathbf{x}), \tag{2.6}$$

avec δ le symbole de Kronecker, converge presque sûrement vers $P_{\mathbf{X}^*}(\mathbf{x}|\mathcal{F}_n)$ quand r tend vers l'infini (application directe de la loi des grands nombres). Dans les problèmes traités en pratique

[1] On appelle distribution de dimension finie d'un processus F indexé sur $\mathbb{X}$ la distribution jointe des variables aléatoires $F(\mathbf{x})$, $\mathbf{x} \in \mathbb{S}$, avec $\mathbb{S}$ un sous-ensemble fini de $\mathbb{X}$.

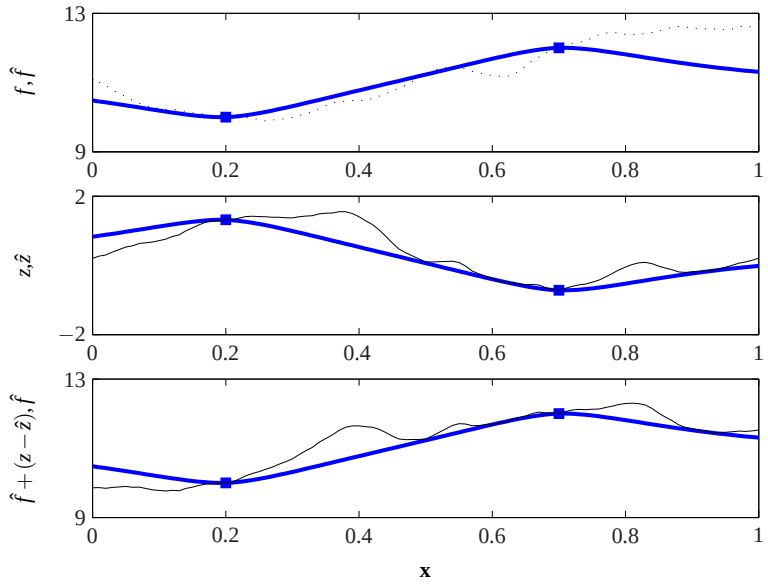


FIG. 2.1: *Conditionnement d'une simulation de F : (*partie supérieure*) fonction inconnue f (trait en pointillés), résultats de deux évaluations (carrés) et prédiction par krigeage associée $\hat{f}$ (trait en gras) ; (*partie médiane*) simulation non conditionnelle z (trait fin), valeurs prises par cette simulation aux deux points d'évaluation de f (carrés), et prédiction par krigeage associée $\hat{z}$ (trait en gras) ; (*partie inférieure*) la simulation de l'erreur de prédiction $z - \hat{z}$ est extraite de la simulation non conditionnelle et ajoutée à la prédiction par krigeage de f pour obtenir une simulation conditionnelle (trait fin). Cette figure est inspirée de Chilès et Delfiner (1999).*

(présentés au chapitre 5), r est fixé à $2N$. Cette valeur, choisie arbitrairement, s'est en pratique révélée suffisante pour une bonne estimation de la distribution.

2.2.2 Choix de Q

Le calcul de (2.2) requiert le choix d'un opérateur de quantification Q. Pour ce choix, utilisons le fait que $F(\mathbf{x})$ soit conditionnellement gaussien de moyenne $\hat{f}(\mathbf{x})$ et de variance $\hat{\sigma}^2(\mathbf{x})$, pour choisir l'ensemble des valeurs possibles $\{y_1(\mathbf{x}),\dots,y_s(\mathbf{x})\}$ de manière à satisfaire

$$P(F_{Q_{\mathbf{x}}}(\mathbf{x}) = y_i | \mathbf{F}_{\mathbb{S}} = \mathbf{f}_{\mathbb{S}}) = \frac{1}{s} \ \forall\, i \in [\![1:s]\!]\,. \tag{2.7}$$

Un opérateur $Q_{\mathbf{x}}$ différent est ainsi utilisé pour chaque valeur de $\mathbf{x}$, de manière à améliorer la précision avec laquelle la moyenne empirique de la réduction d'entropie est approchée dans (2.2). L'ensemble des valeurs possibles doit vérifier, pour satisfaire (2.7),

$$y_1(\mathbf{x}) = 2\hat{\sigma}(\mathbf{x})\Phi^{-1}\left(\frac{1}{s}\right) + \hat{f}(\mathbf{x}) - y_2,$$

$$\forall i \in [\![2:s-1]\!]\ y_i(\mathbf{x}) = \frac{\hat{\sigma}(\mathbf{x})}{2}\left[\Phi^{-1}\left(\frac{i-1}{s}\right) + \Phi^{-1}\left(\frac{i}{s}\right) + 2\hat{f}(\mathbf{x})\right],$$

et

$$y_s(\mathbf{x}) = 2\hat{\sigma}(\mathbf{x})\Phi^{-1}\left(\frac{s-1}{s}\right) + \hat{f}(\mathbf{x}) - y_{s-1}.$$

Pour chacune, $\hat{P}_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n, F(\mathbf{x}) = y_i)$ est approchée à l'aide de simulations conditionnelles. L'entropie conditionnelle des minimiseurs globaux est ensuite approchée par (2.2).

Dans la suite, nous utiliserons un ensemble de dix valeurs possibles ($s = 10$). En effet, on constate empiriquement que cette valeur est suffisante, comme en témoigne, par exemple, la figure 2.2 où l'on peut constater, sur la fonction test déjà utilisée pour la figure 1.11, l'impact d'une valeur trop faible de s.

2.2.3 Schéma général

Le schéma général de IAGO (cf. l'algorithme 2) est similaire à celui de EGO. Nous discuterons dans la section 2.3 les détails de mise en œuvre qui diffèrent, à savoir le plan d'expériences initial (étape 1), le choix d'une fonction de covariance (étape 2) ou encore le critère d'arrêt (étape 4). En revanche, l'optimisation du critère d'échantillonnage ne peut se faire de la même façon. Alors que la structure même du critère EI permettait une optimisation exacte dans EGO, il n'est plus possible d'optimiser exactement l'ECM sur $\mathbb{X}$.

Ce problème de minimisation globale possède en général de nombreux minimiseurs locaux, et on pourrait envisager de le résoudre de manière approchée à l'aide d'une méthode d'optimisation globale ad-hoc (à l'image de l'approche proposée en annexe dans Jones, 2001). Cependant, l'approche proposée ici (et résumée dans la procédure ECM-OPT de l'algorithme 2), essentiellement

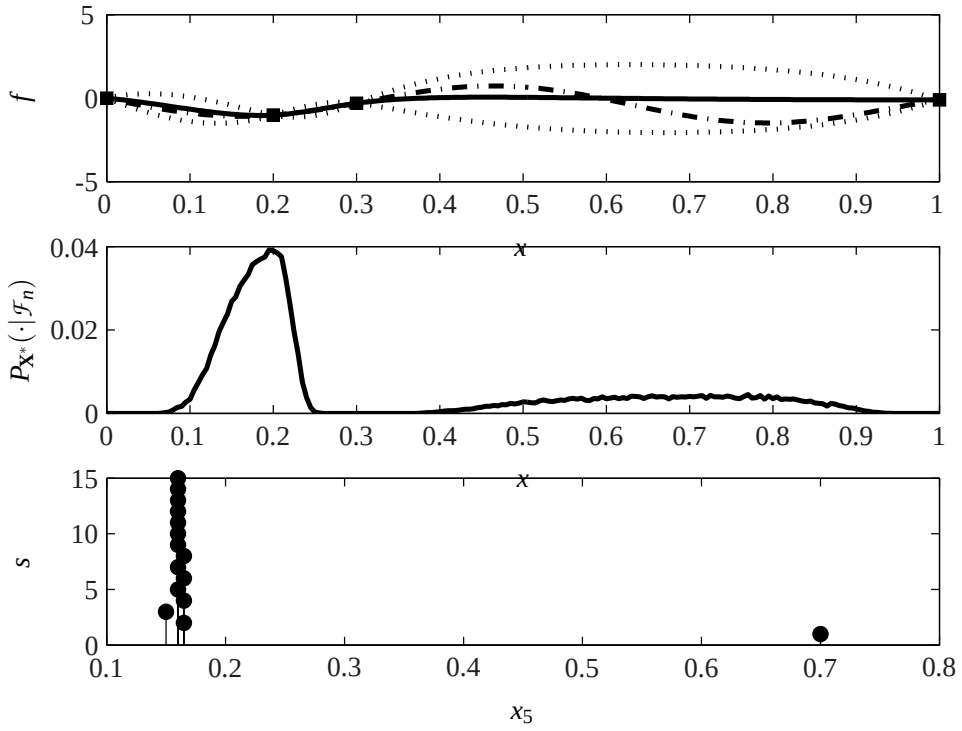


FIG. 2.2: *(*Partie supérieure*) Prédiction par krigeage (en gras) à partir de 4 résultats d'évaluation (carrés) de la fonction test utilisée pour la figure 1.11 (trait en pointillés) et intervalles de confiance associés (traits mixtes). (*Partie médiane*) Distribution conditionnelle des minimiseurs globaux. (*Partie inférieure*) Position de l'évaluation choisie par l'ECM en fonction de s. On constate que pour* $s \leq 8$*, une imprécision apparaît pour devenir flagrante avec* $s = 1$.

pour diminuer la complexité de IAGO, consiste à évaluer exhaustivement $\hat{H}_n$ sur un ensemble fini de points candidats inclus dans $\mathbb{G}$). Ce point sera discuté à la section 2.3.2.

Ainsi, le même jeu de simulation de F, défini sur $\mathbb{G}$, peut être utilisé tout au long de la minimisation de $\hat{H}_n$ (cf. la procédure ECM-OPT dans l'algorithme 2). Par la suite, pour un point candidat $\mathbf{x}_c$ donné, le calcul de $\hat{H}_n$ requiert s jeux de simulations conditionnelles différents (un pour chaque valeur possible $y_i(\mathbf{x}_c)$ de $f(\mathbf{x}_c)$) mais le conditionnement des simulations (2.5) est pour l'essentiel réalisé à l'extérieur de la boucle sur les $y_i(\mathbf{x}_c)$. En effet, si l'on conditionne les simulations par $\mathcal{F}_n$ et un résultat fictif $y_i(\mathbf{x}_c)$ en $\mathbf{x}_c$, les coefficients $\left(\tilde{\boldsymbol{\lambda}}_n^{\mathsf{T}}(\mathbf{x}), \lambda_{n+1}(\mathbf{x})\right)$ de la prédiction dépendent de $\mathbf{x}_c$ mais pas de $y_i(\mathbf{x}_c)$ et la prédiction s'écrit

$$\hat{f}(\mathbf{x}) = \begin{pmatrix} \tilde{\boldsymbol{\lambda}}_n^{\mathsf{T}}(\mathbf{x}) & \lambda_{n+1}(\mathbf{x}) \end{pmatrix} \begin{pmatrix} \mathbf{f}_n \\ y_i(\mathbf{x}_c) \end{pmatrix} \tag{2.8}$$

et l'on peut réécrire l'équation du conditionnement (2.5) comme

$$t(\mathbf{x}) = z(\mathbf{x}) + \tilde{\boldsymbol{\lambda}}_n^{\mathsf{T}}(\mathbf{x})[\mathbf{f}_n - \mathbf{z}_n] + \lambda_{n+1}(\mathbf{x})[y_i(\mathbf{x}_c) - z(\mathbf{x}_c)]. \tag{2.9}$$

Le premier terme de cette équation ne dépend pas de $y_i(\mathbf{x}_c)$ et peut donc être calculé en dehors de la boucle (étape 15). Le second est ensuite ajouté à l'étape 17.

Avec cette approche, la complexité calculatoire de l'approximation de l'entropie conditionnelle après n évaluations (étapes 13 à 20 de l'algorithme 2) est déterminée par les points suivants.

- Calcul des coefficients du krigeage pour tous les points de $\mathbb{G}$ (étape 14) : $\mathrm{O}(n^2N)$. En effet, le système (1.7) doit être résolu N fois. La matrice de covariance peut être factorisée, ce qui limite à $\mathrm{O}(n^2)$ la complexité de la prédiction en un point donné. Notons que la factorisation de la matrice de covariance (en $\mathrm{O}(n^3)$) peut en grande partie être réalisée à l'extérieur de la boucle sur les points candidats.
- Construction des simulations conditionnelles (étape 15 et 17) : $\mathrm{O}(rnN)$.
- Minimisation de toutes les simulations conditionnelles (étape 18) : $O(sNr)$.

Les autres étapes sont en $\mathrm{O}(N)$. Ainsi, la complexité du calcul de $\hat{H}_n$ est en $\mathrm{O}(N)$. A titre de comparaison, l'approximation de l'EI ne requiert que le calcul de la fonction de répartition gaussienne. Nous discuterons à la section 4.5 des conséquences pratiques de cette complexité.

Remarque 2.1. Notons que la génération de r simulations de F sur $\mathbb{G}$ est en $O(rN^2)$. Si $\mathbb{G}$ ne change pas au cours de la procédure, ces opérations peuvent être réalisées pendant l'initialisation de IAGO. Dans le cas contraire, c'est-à-dire si $\mathbb{G}$ est remis à jour après chaque évaluation (comme proposé par la suite), le coût de ces opérations reste négligeable devant celui de l'optimisation de $\hat{H}_n$ ($\mathrm{O}(rnN^2)$ si l'on utilise $\mathbb{G}$ comme ensemble des points candidats).

Algorithme 2 Schéma général de l'algorithme IAGO et mise en œuvre proposée pour l'optimisation de l'ECM

Évaluer f sur un plan d'expériences initial ▷ cf. la section 2.3.1
Choisir ou estimer les paramètres de la covariance de Matèrn ▷ cf. la section 2.3.1
Calculer la prédiction par krigeage
Tant que le critère d'arrêt n'est pas satisfait **Faire** ▷ cf. la section 2.3.3 pour une discussion sur le critère d'arrêt à utiliser
    $\mathbf{x}_{\text{new}} \leftarrow$ ECM-OPT(points d'évaluations, résultats d'évaluation, covariance)
    Evaluer f en $\mathbf{x}_{\text{new}}$
Fin Tant que

procedure ECM-OPT($\mathbb{S}_n$, $\mathbf{f}_n$, k)
    Choisir $\mathbb{G}$ ▷ cf. la section 2.3.2
    Générer r simulations de F sur $\mathbb{G}$
    Calculer $\hat{f}(\mathbf{x})$ et $\hat{\sigma}(\mathbf{x})$ sur $\mathbb{G}$
    Pour tout $\mathbf{x}_c$ dans un ensemble de points candidats (par exemple $\mathbb{G}$) **Faire** ▷ cf. 2.3.2
        Calculer les paramètres $\{y_1(\mathbf{x}_c), \ldots, y_s(\mathbf{x}_c)\}$ de $Q(\mathbf{x}_c)$ (cf. la section 2.2.2)
        Calculer, pour tout $\mathbf{x}$ dans $\mathbb{G}$, les coefficients $\left(\tilde{\boldsymbol{\lambda}}_n^{\mathsf{T}}(\mathbf{x}), \lambda_{n+1}(\mathbf{x})\right)$ de la prédiction par krigeage pour des évaluations aux points de $\mathbb{S}_n \bigcup \{\mathbf{x}_c\}$
        Construire partiellement les simulations conditionnelles en utilisant le premier terme de (2.9)
        Pour $i \leftarrow 1, s$ **Faire**
            Terminer le conditionnement par krigeage des simulations en utilisant le deuxième terme de (2.9) et en supposant $f(\mathbf{x}_c) = y_i(\mathbf{x}_c)$
            Déterminer un minimiseur global pour chacune des r trajectoires conditionnelles
            Estimer $P_{\mathbf{X}^*}(\cdot | \mathcal{F}_n, F(\mathbf{x}_c) = y_i(\mathbf{x}_c))$ (2.6)
            Calculer $H(\mathbf{X}^* | \mathcal{F}_n, F(\mathbf{x}) = y)$ (1.26)
        Fin Pour
        Calculer $\hat{H}_n(\mathbf{x}_c)$ (2.2)
    Fin Pour
    Retourner $\mathbf{x}_{\text{new}} = \min_{\mathbf{x} \in \mathbb{G}} \hat{H}_n(\mathbf{x})$
Fin procedure

2.3 Mise en œuvre de IAGO

2.3.1 Plan d'expériences initial et choix d'une fonction de covariance

Le plan d'expériences initial (étape 1 dans l'algorithme 2 décrit ci-après) est utilisé pour une première recherche globale, mais surtout pour réaliser une première estimation des paramètres de la covariance. Jones et al. (1998) proposent d'utiliser pour cela un plan LHS[2](pour *Latin Hypercube Sampling*, voir par exemple McKay et al., 1979) de taille $10d$. Cette règle empirique est inapplicable ici, puisque le budget *total* en évaluations pour les problèmes motivant ces travaux est en général bien inférieur. En revanche, là où EGO utilise une covariance non isotrope et estime ainsi $2d+1$ paramètres, il nous semble plus raisonnable de se limiter, en l'absence d'information *a priori*, à l'utilisation d'une covariance isotrope. En effet, même à l'aide de cette simplification, l'estimation des paramètres par maximum de vraisemblance reste trop incertaine. Pour s'en convaincre, simulons 10 000 réalisations d'un processus gaussien centré, stationnaire et de covariance de Matérn de paramètres $\nu = 3$, $\rho = 0.3$ et $\sigma^2 = 0.5$. Ces réalisations sont échantillonnées sur un plan LHS, et pour chacune, calculons la dérivée seconde de la log-vraisemblance par rapport au paramètre ν, pour la vraie valeur du vecteur des paramètres. La moyenne de ces dérivées secondes est, au signe près, l'information de Fisher relative à l'estimation par maximum de vraisemblance de ν (les autres paramètres étant supposés connus). Nous pouvons alors calculer la racine carrée de l'inverse de cette quantité pour nous faire une idée grossière de l'écart-type de l'erreur d'estimation. Le tableau 2.1 présente pour différentes dimensions de l'espace de recherche et différentes tailles du plan LHS, les valeurs moyennes de cette quantité. On y constate la difficulté à estimer ne serait-ce qu'un paramètre en utilisant le nombre d'essais préconisés par Jones et al. (1998) pour des dimensions de l'espace des facteurs en accord avec nos objectifs.

Nous proposons donc de n'utiliser en général le plan d'expériences initial que pour effectuer une première recherche et nous préconiserons dans la section 4.3.3 une règle empirique pour le choix du nombre d'évaluations dans ce plan. La covariance devra quant à elle être en général choisie *a priori* (excepté lorsque le budget d'évaluations et la dimension du problème le permettent). Ce choix, qui pourra être révisé au cours de la procédure dès que l'estimation des paramètres devient possible, s'avère souvent relativement aisé en pratique. En effet, des données ou des connaissances métiers sont fréquemment disponibles sur le système étudié et sont d'une grande utilité pour le choix des paramètres de la covariance. Par exemple, lors de la conception d'une nouvelle version d'un système, les paramètres peuvent être directement estimés à partir des données recueillies sur son prédécesseur. En outre, il n'est pas rare que les connaissances sur la physique du système permettent de quantifier sa plage de variation (ce qui peut aider au choix de σ), sa dérivabilité (ce qui guide le choix de ν) ou une allure générale (ce qui influe sur le choix

[2]Notons que des plans d'expériences plus adaptés à l'estimation des paramètres de la covariance existent dans la littérature. Mentionnons en particulier les travaux de Zhu et Zhang (2006).

	Cas de ν	Cas de ρ	Cas de σ^2
$d = 1$, LHS de taille 5	2.58	0.43	0.70
$d = 1$, LHS de taille 10 (recommandé par Jones et al., 1998)	0.32	0.14	0.48
$d = 1$, LHS de taille 15	0.14	0.09	0.38
$d = 6$, LHS de taille 30	66.8	0.67	0.26
$d = 6$, LHS de taille 60 (recommandé par Jones et al., 1998)	16.26	0.30	0.18
$d = 6$, LHS de taille 90	3.90	0.20	0.14
$d = 15$, LHS de taille 50	55.79	6.07	0.14
$d = 15$, LHS de taille 150 (recommandé par Jones et al., 1998)	39.16	3.43	0.11
$d = 15$, LHS de taille 300	14.00	1.58	0.08

TAB. 2.1: *Racines carrées de l'inverse de l'information de Fisher (en proportion de la vraie valeur des paramètres) relative à l'estimation par maximum de vraisemblance des paramètres pris un par un (les autres paramètres étant supposés connus) d'une covariance de Matérn (isotrope de paramètres $\nu = 3$, $\rho = 0.3$, $\sigma^2 = 0.5$) pour des LHS de tailles variées, avec $d = \dim \mathbb{X}$. Cette quantité permet de se faire une idée très grossière (puisque nous ne sommes pas en situation asymptotique) de l'incertitude sur le paramètre considéré. Le plan LHS étant par nature aléatoire, pour chaque taille, 50 plans sont générés, et pour chacun l'information de Fisher est évaluée. Le ratio entre l'inverse de la racine carrée de cette quantité et la vraie valeur du paramètre est ensuite calculée. Enfin, les ratios obtenus pour chacun des 50 plans LHS sont moyennés. On constate, par exemple en dimension six pour l'estimation de ν, que 90 points ne semblent pas suffire pour obtenir une précision équivalente à celle obtenue en dimension un avec seulement cinq points.*

de ρ). Plus généralement, nous verrons dans la section 4.3.2, les conditions à respecter pour qu'un choix *a priori* de paramètres ne handicape pas les algorithmes outre mesure.

2.3.2 Choix de $\mathbb{G}$ et optimisation du critère d'échantillonnage

Le choix le plus simple pour $\mathbb{G}$ est une grille restant fixe au cours de la procédure. Cependant, à mesure que le nombre d'évaluations effectuées augmente, le support de $p_{\mathbf{X}^*_{\mathbb{X}}}(\cdot|\mathcal{F}_n)$ (la distribution conditionnelle des minimiseurs globaux de F sur $\mathbb{X}$) diminue tout comme la capacité de $\mathbb{G}$ à décrire ce support. Ainsi, à moins d'augmenter exponentiellement le nombre d'éléments de $\mathbb{G}$ avec la dimension, l'entropie de $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ pourra s'annuler sans que l'optimisation soit pour autant terminée. Pour conserver une taille raisonnable de $\mathbb{G}$, tout en assurant une précision satisfaisante pour l'approximation de la densité des minimiseurs globaux par $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$, $\mathbb{G}$ peut être rééchan-

tillonné après chaque nouvelle évaluation pour mieux représenter le support de $p_{\mathbf{X}^*_{\mathbb{X}}}(\cdot|\mathcal{F}_n)$. $\mathbb{G}$, noté désormais $\mathbb{G}_n$ pour marquer sa modification à chaque itération, doit donc dépendre des résultats des évaluations disponibles et nous proposons pour cela d'estimer, pour $\mathbf{x} \in \mathbb{X}$, $p_{\mathbf{X}^*_{\mathbb{X}}}(\mathbf{x}|\mathcal{F}_n)$ par

$$\frac{1}{r}\sum_{i=1}^{r} \phi\left(\frac{\mathbf{x}-\mathbf{x}_i^*}{h}\right),$$

l'estimateur à noyaux gaussiens de paramètre h à choisir, avec $\mathbf{x}_i^*, i = 1,...,r$, les minimiseurs globaux des trajectoires conditionnelles sur $\mathbb{G}_{n-1}$. Ainsi, après la n-ième évaluation, un point de $\mathbb{G}_n$ est obtenu par tirage d'une loi gaussienne de variance h^2 centrée sur un point tiré au hasard dans $\{\mathbf{x}_i^*, i = 1,...,r\}$.

Pour conserver une complexité calculatoire acceptable, nous avons vu à la section 2.2.3 que les points candidats à l'évaluation (c'est-à-dire l'espace de recherche associé à l'algorithme d'optimisation du critère d'échantillonnage) doivent être choisis *a priori*. Ainsi il est exclu d'utiliser une méthode d'optimisation classique (par exemple un simplexe de Nelder et Mead utilisant plusieurs points de départs) et le critère doit être optimisé par évaluation extensive sur un ensemble de points candidats inclus dans $\mathbb{G}$. Ici encore, le choix d'un ensemble fixe au cours de la procédure serait peu judicieux puisqu'à mesure que les évaluations sont réalisées, et que la recherche devient plus locale, l'erreur d'estimation de l'optimum du critère risque d'augmenter (à l'extrême, tous les points candidats au voisinage de l'optimum réel du critère ont été évalués). Remarquons qu'un point de l'espace des facteurs où la distribution conditionnelle des minimiseurs est suffisamment faible ne présentera aucun intérêt pour une évaluation supplémentaire. Après quelques évaluations, une large part de l'espace des facteurs possède cette propriété, et le problème de l'optimisation du critère s'en retrouve considérablement simplifié.

Il apparaît finalement que le rééchantillonnage de $\mathbb{G}$ nécessaire à la bonne représentation du support de la densité des minimiseurs globaux sert aussi la qualité de l'optimisation de l'entropie conditionnelle. Nous proposons donc de calculer cette quantité exhaustivement sur $\mathbb{G}_n$ pour choisir la $n+1$-ième évaluation. Avec cette approche, la précision avec laquelle un minimiseur du critère d'échantillonnage est estimé augmente avec le nombre d'évaluations réalisées. Reste à choisir h qui détermine le lissage effectué par l'estimateur à noyaux. Dans notre contexte, une valeur trop faible de h serait préjudiciable à la convergence de l'algorithme, puisqu'elle conduirait à sous estimer le support de $p_{\mathbf{X}^*_{\mathbb{X}}}(\mathbf{x}|\mathcal{F}_n)$ et donc à prendre le risque d'une recherche trop locale. Nous proposons donc de choisir h comme la distance moyenne entre un point de $\mathbb{G}_0$ (en général un plan LHS) et son plus proche voisin.

2.3.3 Critère d'arrêt

Le critère d'arrêt est une composante importante de toute méthode d'optimisation. Dans notre contexte de fonctions coûteuses, il est fréquemment impossible de satisfaire à un autre critère

que celui de l'épuisement du budget d'évaluation. C'est donc ce critère que nous utiliserons le plus souvent. Quand ce budget n'est pas fixé *a priori* de multiples choix deviennent possibles. On peut par exemple utiliser l'estimateur $f_{\min} = \min_{\mathbf{x}_i} f(\mathbf{x}_i)$ du minimum global de F et observer la probabilité pour l'erreur d'estimation d'être plus faible que $\delta > 0$. Schonlau (1997) propose ainsi d'arrêter l'algorithme lorsque

$$\mathrm{P}(F^*_{\mathbb{G}} < f_{\min} + \delta | \mathbf{F}_{\mathbb{S}} = \mathbf{f}_{\mathbb{S}}) < P_{\text{stop}}\,,$$

avec $F^*_{\mathbb{G}} = \min_{\mathbf{x}\in\mathbb{G}} F(\mathbf{x})$, et $P_{\text{stop}} \in [0,1]$ une valeur critique choisie *a priori*. Ce critère est bien adapté ici, car l'estimation de la fonction de répartition de $F^*_{\mathbb{G}}$ ne requiert aucun calcul supplémentaire. Il est en effet possible d'utiliser les simulations conditionnelles réalisées pour calculer $\hat{H}_n$, à condition de conserver pour chacune d'elles non seulement un minimiseur global, mais aussi le minimum associé. L'histogramme ainsi obtenu permet alors d'estimer simplement la fonction de répartition du minimum global conditionnellement aux résultats des évaluations.

L'utilisation de ce critère reste néanmoins peu aisée, puisqu'il dépend de deux paramètres (δ et P_{stop}) qui doivent être choisis *a priori*, et ne peuvent pas en général être déduits de caractéristiques du problème. En revanche, si l'on considère plus simplement l'écart-type de la distribution conditionnelle du minimum global, on peut arrêter IAGO lorsque

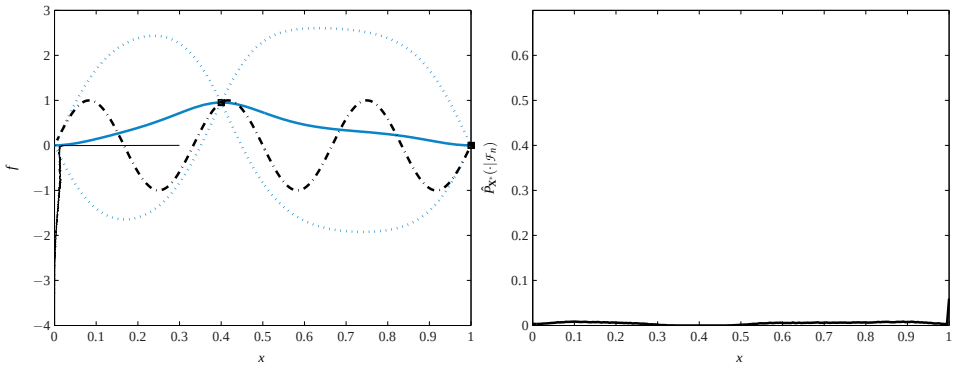
$$\sqrt{\mathrm{var}(F^*_{\mathbb{G}} | \mathcal{F}_n)} < \sigma_{\text{stop}}, \tag{2.10}$$

le paramètre $\sigma_{\text{stop}} > 0$ à choisir *a priori*, peut l'être en fonction de la confiance accordée aux résultats des évaluations.

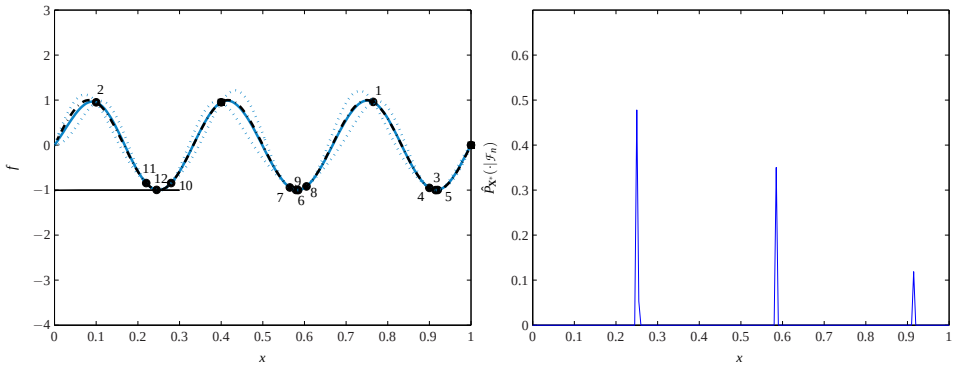
La figure 2.3 présente un exemple d'application de IAGO sur plusieurs périodes d'une fonction sinus. On y constate le caractère global de IAGO qui permet d'identifier avec précision les trois minimiseurs globaux après douze itérations. La figure 2.4 présente le même exemple lorsque IAGO est muni du critère d'arrêt (2.10). Ce critère, satisfait après sept évaluations, permet d'identifier l'un des minimiseurs globaux.

2.4 Extensions

Pour traiter les problèmes pratiques motivant nos travaux, de nombreuses extensions de IAGO et du critère d'échantillonnage sous-jacent sont nécessaires. Nous décrivons dans cette section celles qui se conjuguent bien avec le critère de minimisation de l'ECM. Nous discuterons dans un premier temps, de l'optimisation globale en présence de bruit sur les résultats des évaluations de f, puis de prise en compte de contraintes, et enfin de la prise en compte de résultats d'évaluation du gradient de la fonction objectif. Nous étudions, pour chacune de ces extensions, le comportement des trois critères d'échantillonnage retenus au chapitre précédent, à savoir maximisation de l'EI, minimisation de l'ECM et maximisation de la probabilité d'amélioration. Les versions modifiées de l'ECM ou de l'EI peuvent ensuite être utilisées par IAGO comme par EGO.



(a) Prédiction initiale et distribution conditionnelle



(b) Prédiction et distribution conditionnelle après douze itérations de IAGO

FIG. 2.3: *Optimisation d'une fonction sinus à l'aide de IAGO sans critère d'arrêt. On trouve sur les figures de gauche la prédiction (trait continu) à partir des résultats d'évaluation matérialisés par des carrés (points initiaux) ou des cercles numérotés (points choisis par IAGO) , ainsi que les intervalles de confiance à 95% (traits pointillés) et la fonction sinus à optimiser (traits mixtes). Est aussi représentée, une estimée de la distribution conditionnelle du minimum global (normée à 0.3 pour faciliter la lecture des figures). Les deux figures de droite présentent les distributions conditionnelles des minimiseurs globaux avant et après les douze itérations de IAGO.*

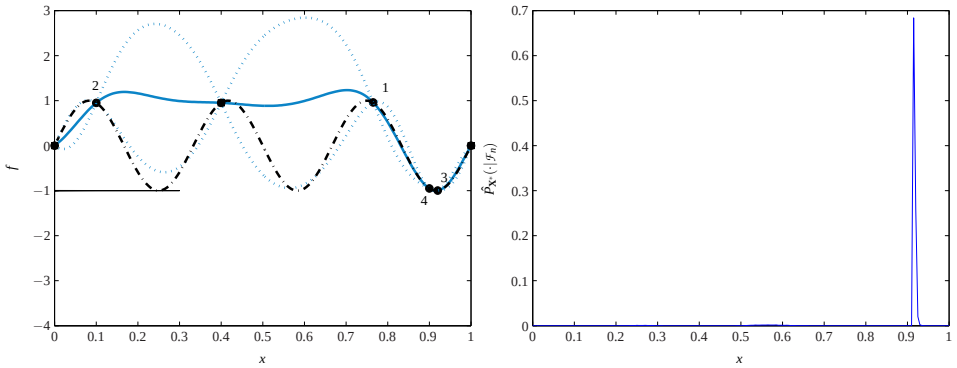


FIG. 2.4: *Optimisation d'une fonction sinus à l'aide de IAGO muni du critère d'arrêt* (2.10) *avec* $\sigma_{\text{stop}} = 0.1$. *Les conventions graphiques sont identiques à celles de la figure 2.3.*

2.4.1 Résultats d'évaluation incertains

Pour de nombreux systèmes à optimiser en pratique, les résultats des évaluations de f sont incertains. Il peut s'agir d'incertitudes de mesure, mais il arrive aussi fréquemment que les résultats de simulations numériques soient eux aussi entachés d'erreurs. Par exemple, lors de simulations d'écoulement fluide, il arrive que pour certaines configurations du dispositif, la solution des équations différentielles modélisant le système ne puisse se stabiliser dans le temps qui lui est imparti (nous en verrons un exemple dans le chapitre 5). Il en résulte, dans le cas où les simulations sont automatisées (de sorte que l'utilisateur ne peut intervenir pour constater la non-convergence de la solution et prolonger la simulation), une incertitude sur le résultat.

Supposons, pour le reste de cette section, les résultats des évaluations corrompus par un bruit additif de moyenne nulle. Ainsi pour $i = 1,...,n$, et d'après notre modèle gaussien F, l'évaluation de f en $\mathbf{x}_i$ produit une réalisation $f^{\text{obs}}(\mathbf{x}_i)$ de la variable aléatoire $F^{\text{obs}}(\mathbf{x}_i) = F(\mathbf{x}_i) + b_i$, où les b_i sont des variables aléatoires, éventuellement corrélées, et de loi connue ou paramétrée (on peut alors estimer les paramètres de cette loi par maximum de vraisemblance conjointement avec les paramètres de k). Le prédicteur par krigeage de F à partir des résultats $\mathcal{F}_n^{\text{obs}}$ des évaluations bruitées de f, s'écrit alors pour $\mathbf{x} \in \mathbb{X}$

$$\hat{F}(\mathbf{x}) = \boldsymbol{\lambda}(\mathbf{x})^{\mathsf{T}} \mathbf{F}_n^{\text{obs}},$$

avec $\mathbf{F}_n^{\text{obs}} = [F^{\text{obs}}(\mathbf{x}_1),...,F^{\text{obs}}(\mathbf{x}_n)]^{\mathsf{T}}$. Le calcul du vecteur $\boldsymbol{\lambda}(\mathbf{x})$, décrit dans l'annexe B.2, se fait de manière tout à fait similaire au cas non bruité et permet d'obtenir la prédiction $\hat{f}$ ainsi que l'écart-type de l'erreur associée $\hat{\sigma}$. A l'aide de cette prédiction, IAGO peut donc s'appliquer directement à l'optimisation globale en présence de bruit sur les résultats d'évaluation. La technique de simulation conditionnelle utilisée pour calculer l'approximation $\hat{P}_{\mathbf{X}^*}(\cdot|\mathcal{F}_n^{\text{obs}})$ de la distribution des minimiseurs doit cependant être légèrement modifiée. En effet, les trajectoires de F conditionnellement à $\mathcal{F}^{\text{obs}}$ n'interpolent plus les résultats des évaluations et il faut, pour conditionner une simulation z sur $\mathbb{G}$, simuler des réalisations du bruit sur les résultats des évaluations et appliquer

$$t(\mathbf{x}) = z(\mathbf{x}) + \boldsymbol{\lambda}(\mathbf{x})^{\mathsf{T}}[\mathbf{f}_n^{\text{obs}} - \mathbf{z}_n^{\text{obs}}], \tag{2.11}$$

avec $\mathbf{z}_n^{\text{obs}}$ les valeurs prises par z sur $\mathbb{S}_n$ corrompues par une réalisation du bruit.

Si le critère de maximisation de l'ECM s'adapte bien au cas bruité, il n'en va pas de même pour la maximisation de l'EI. En effet, l'EI peut toujours être utilisé tel quel, cependant, l'estimateur du minimum sous-jacent, $\hat{M}_n = \min_i F^{\text{obs}}(\mathbf{x}_i)$ ne converge plus vers le minimum global F^* lorsque le nombre d'évaluations augmente (M_n tend à sous-estimer de plus en plus fortement F^*). D'autres estimateurs de F^* peuvent être utilisés, par exemple $\hat{M}_n = \min_{\mathbf{x}} \hat{f}(\mathbf{x})$. La modification du critère ainsi obtenue sera par la suite désignée par EIm. On trouve aussi dans la littérature (Huang et al., 2006) une modification plus heuristique de l'EI utilisant $M_n^{\text{AEI}} = F\left(\arg\min_{\mathbf{x}} \hat{f}(\mathbf{x}) - \hat{\sigma}(\mathbf{x})\right)$ comme estimateur de F^* et reposant sur l'hypothèse, assez restrictive (cf. la section 5.4 pour un problème

ne satisfaisant pas cette contrainte), d'un bruit gaussien de matrice de covariance $\sigma_b^2\mathbf{I}_n$. Dénommée AEI, pour *Augmented Expected Improvement*, cette quantité s'obtient comme

$$\text{AEI}(\mathbf{x}) = \mathbb{E}\left[I^{\text{obs}}(\mathbf{x})|\mathcal{F}_n^{\text{obs}}\right]\left(1-\frac{\sigma_b}{\sqrt{\sigma_b^2+\hat{\sigma}^2(\mathbf{x})}}\right), \tag{2.12}$$

avec

$$I^{\text{obs}}(\mathbf{x}) = \begin{cases} 0 & \text{si} \quad F(\mathbf{x}) \geq M_n^{\text{AEI}} \\ M_n^{\text{AEI}} - F(\mathbf{x}) & \text{sinon} \end{cases}.$$

L'objectif du terme correctif dans (2.12) est de décourager une réplication exagérée des évaluations, ce qui semble être une tendance de l'EI dans le cas bruité. En effet, à mesure que des évaluations sont réalisées en $\mathbf{x}$, $\hat{\sigma}(\mathbf{x})$ diminue tout comme AEI($\mathbf{x}$).

Ces deux modifications seront comparées avec l'original et avec l'ECM dans la section 4.4.1. On peut cependant constater dès à présent sur la figure 2.5 l'impact du bruit sur la prédiction (qui n'interpole plus les données et s'accompagne d'une variance de l'erreur plus forte) et sur la vitesse de convergence de IAGO.

2.4.2 Prise en compte de résultats d'évaluation du gradient

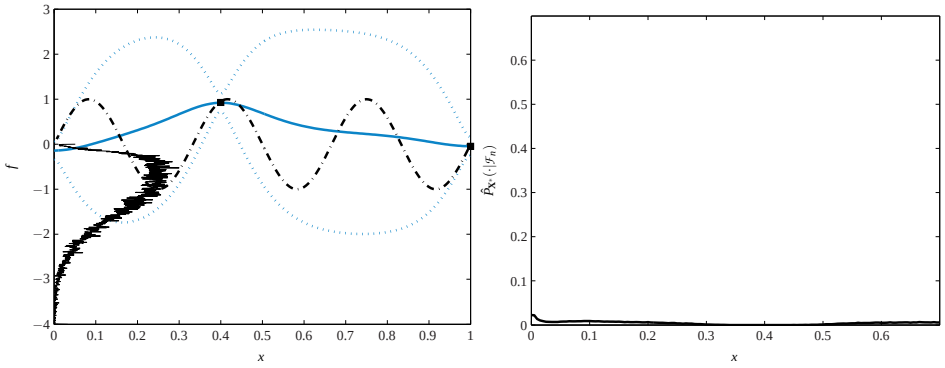
Avec le développement des logiciels de différentiation automatique, il devient plus fréquent de disposer, en sortie de simulateurs numériques, de la valeur de la fonction objectif mais aussi de son gradient. Cette information supplémentaire peut diminuer considérablement la difficulté d'un problème d'optimisation, encore faut-il être en mesure de la prendre en compte dans les critères d'échantillonnage.

Pour ce faire, il suffit de remarquer que la covariance du gradient ∇F de F s'obtient simplement en fonction de k, tout comme la covariance entre F et ∇F, ce qui permet d'incorporer simplement les résultats des évaluations du gradient de f à la prédiction par krigeage (les équations sont présentées dans l'annexe B). L'extension de l'EI ou de la probabilité d'amélioration est alors immédiate et profite de l'amélioration de la prédiction apportée par cette nouvelle information (on peut l'observer en comparant les figures 2.3 et 2.6).

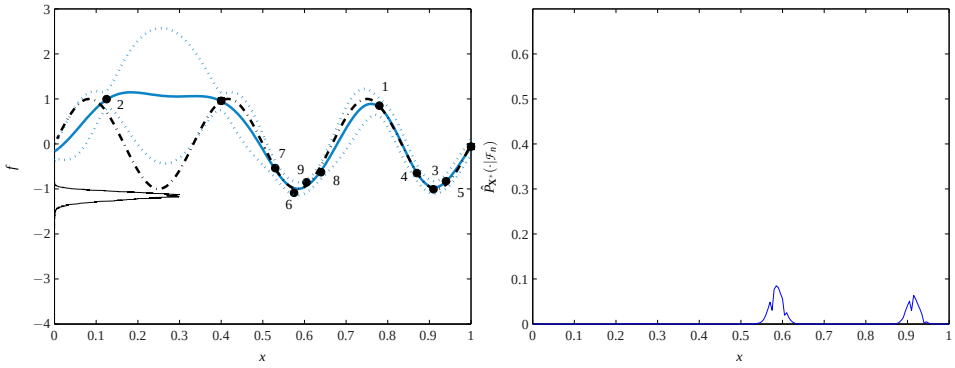
Pour adapter l'ECM, c'est naturellement l'entropie de $\mathbf{X}^*$ conditionnellement à $\mathcal{F}_n$, aux résultats d'évaluation passées $\mathcal{F}_n^{\nabla}$ du gradient et aux résultats $F(\mathbf{x})$ et $\nabla F(\mathbf{x})$ en un point d'évaluation candidat que l'on va chercher à minimiser. Cette entropie s'écrit

$$\begin{aligned} H(\mathbf{X}^*|\mathcal{F}_n,\mathcal{F}_n^{\nabla},F(\mathbf{x}),\nabla F(\mathbf{x})) = \int_{(y,\mathbf{g})\in\mathbb{R}\times\mathbb{R}^d} & p_{F(\mathbf{x}),\nabla F(\mathbf{x})}(y,\mathbf{g}|\mathcal{F}_n,\mathcal{F}_n^{\nabla}) \\ & H(\mathbf{X}^*|\mathcal{F}_n,\mathcal{F}_n^{\nabla},F(\mathbf{x})=y,\nabla F(\mathbf{x})=\mathbf{g})dyd\mathbf{g}, \end{aligned} \tag{2.13}$$

avec $p_{F(\mathbf{x}),\nabla F(\mathbf{x})}(y,\mathbf{g}|\mathcal{F}_n,\mathcal{F}_n^{\nabla})$ la distribution jointe de $F(\mathbf{x})$ et $\nabla F(\mathbf{x})$ conditionnellement aux résultats d'évaluation (gaussienne de moyenne et de matrice de covariance obtenues par krigeage,



(a) Prédiction initiale et distributions conditionnelles



(b) Prédiction et distributions conditionnelles après satisfaction du critère d'arrêt

FIG. 2.5: *Optimisation d'une fonction sinus à l'aide de IAGO (muni du critère d'arrêt* (2.10) *avec* $\sigma_{\text{stop}} = 0.1$*) à partir de résultats d'évaluation corrompus par un bruit gaussien additif d'écart-type 0.1. Les conventions graphiques sont identiques à celles de la figure 2.3. La prise en compte du bruit dans la prédiction par krigeage implique une prédiction non interpolatrice ainsi qu'une convergence plus lente de IAGO.*

cf. l'annexe B.3 pour une partie des équations) et

$$H(\mathbf{X}^*|\mathcal{F}_n, \mathcal{F}_n^{\nabla}, F(\mathbf{x}) = y, \nabla F(\mathbf{x}) = \mathbf{g})$$

l'entropie de $P_{\mathbf{X}^*}(\cdot|\mathcal{F}_n, \mathcal{F}_n^{\nabla}, F(\mathbf{x}) = y, \nabla F(\mathbf{x}) = \mathbf{g})$, la distribution de $\mathbf{X}^*$ conditionnellement aux résultats d'évaluation et aux résultats (fictifs) des évaluations de la fonction *et* du gradient.

L'approximation de (2.13) peut s'effectuer tout comme celle de H_n, par le choix d'un opérateur de quantification et une approximation de la distribution des minimiseurs à l'aide de simulations conditionnelles. L'algorithme qui en résulte est cependant beaucoup plus complexe que la version initiale de IAGO puisque pour chaque point d'évaluation candidat, il faut générer plusieurs résultats potentiels pour la fonction *et* pour son gradient. De plus il est nécessaire de simuler conjointement F et ∇F comme détaillé dans l'annexe B.1.

L'intérêt principal de l'ECM dans ce contexte, est que, contrairement à l'EI ou à la probabilité d'amélioration, cette quantité prend en compte le fait que le gradient de la fonction va aussi être évalué. La figure 2.6 présente l'optimisation par cette version modifiée de IAGO de la fonction sinus déjà utilisée dans ce chapitre. La connaissance des dérivées permet d'obtenir avec seulement 10 évaluations(soit sept itérations de IAGO), une diminution de l'ECM n'étant obtenue qu'au bout de 17 évaluations avec la version standard de IAGO. Notons néanmoins l'augmentation significative de la complexité du calcul de l'ECM qui nécessite désormais $d+1$ prédictions par krigeage, ainsi que l'intégration numérique d'une fonction définie sur $\mathbb{R}^{d+1}$.

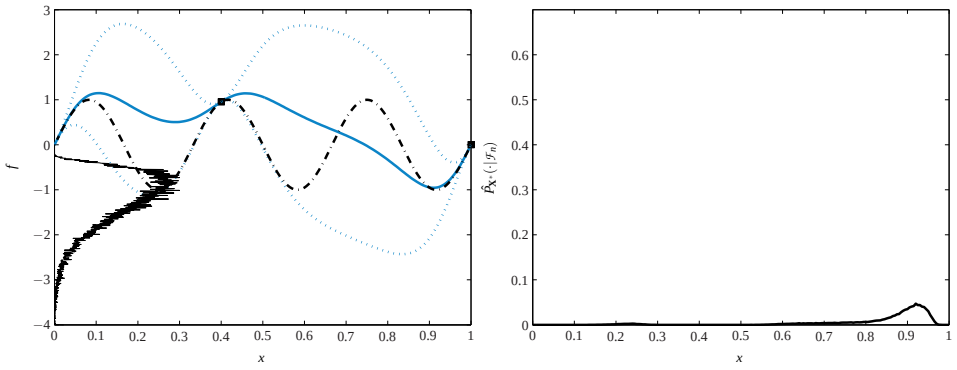
2.4.3 Prise en compte de contraintes

Depuis le début du chapitre 1, le problème de minimisation est supposé non contraint. Considérons maintenant le cas plus général d'un problème contraint par $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$ et $\mathbf{h}(\mathbf{x}) = \mathbf{0}$. Si les contraintes sont peu coûteuses à évaluer, on peut se contenter de sélectionner des points admissibles par un algorithme rejetant ceux qui ne le sont pas. Nous consacrerons ici notre attention au cas des contraintes d'inégalité coûteuses à évaluer et utiliserons un modèle gaussien $\mathbf{G}$ pour leur prédiction. [3]

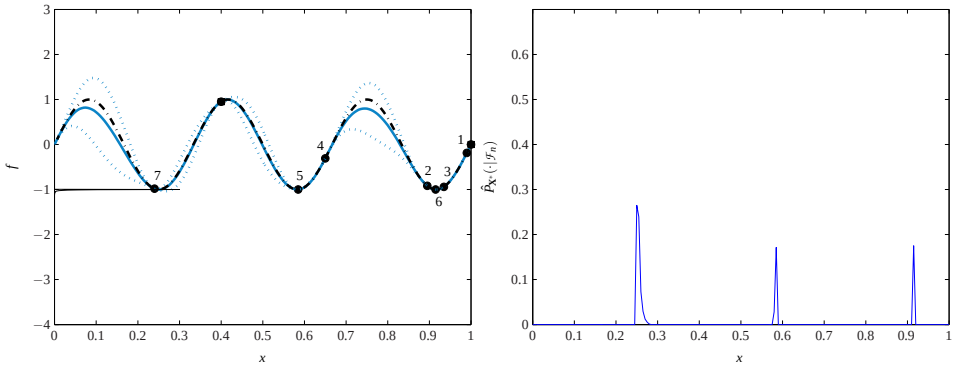
L'extension du critère de maximisation de l'EI à ce problème est abordée dans Schonlau (1997). Dans ce contexte, l'estimateur du minimum sous-jacent devient $M_n^{\mathrm{c}} = \min\{F(\mathbf{x}_i)|\mathbf{G}(\mathbf{x}_i) \leq \mathbf{0}\}$ et la version contrainte de l'EI devient

$$\mathbb{E}\left[I^{\mathrm{c}}(\mathbf{x})|\mathcal{F}_n, \mathcal{G}_n\right], \tag{2.14}$$

[3] La prise en compte, dans notre contexte de budget d'évaluation réduit, de contraintes d'égalité elles aussi coûteuses rend le problème nettement plus difficile. On peut envisager de transformer la contrainte d'égalité en un nouvel objectif (par exemple en considérant la norme de $\mathbf{h}(\mathbf{x})$) et de s'attaquer au problème multi-objectif ainsi défini (cf. la section 5.2.1 pour une méthode de résolution d'un problème multi-objectif coûteux).



(a) Prédiction initiale (incluant les dérivées) et distribution conditionnelle



(b) Prédiction et distribution conditionnelle après 7 itérations de IAGO

FIG. 2.6: *Optimisation d'une fonction sinus à l'aide de IAGO lorsque les dérivées sont disponibles. Les conventions graphiques sont identiques à celles de la figure 2.3. La prise en compte du gradient dans la prédiction améliore sa qualité (pour s'en rendre compte, comparer la prédiction initiale avec celle de la figure 2.3) et permet une convergence plus rapide de IAGO.*

avec

$$I^{\mathrm{c}}(\mathbf{x}) = \begin{cases} M_n^{\mathrm{c}} - F(\mathbf{x}) & \text{si} \quad F(\mathbf{x}) \leq M_n^{\mathrm{c}} \text{ et } \mathbf{G}(\mathbf{x}) \leq \mathbf{0} \\ 0 & \text{sinon} \end{cases},$$

et $\mathcal{G}_n = \{\mathbf{G}(\mathbf{x}_i) = \mathbf{g}(\mathbf{x}_i), \forall \mathbf{x}_i \in \mathbb{S}\}$. Si l'on suppose l'indépendance entre F et $\mathbf{G}$, l'EI contraint s'obtient simplement comme l'EI du cas non contraint multiplié par la probabilité pour les contraintes d'être satisfaites en $\mathbf{x}$ conditionnellement à $\mathcal{F}_n$. En revanche, si cette hypothèse n'est pas satisfaisante, c'est F et $\mathbf{G}$ qui devront être prédits conjointement et le calcul de la version contrainte de l'EI ne peut s'obtenir qu'au prix de l'estimation de la fonction de répartition gaussienne multidimensionnelle.

Comme pour l'EI, la probabilité d'amélioration s'étend aisément au cas contraint à condition de prédire conjointement F et $\mathbf{G}$. Ainsi, le critère dans le cas contraint

$$\arg\max_{\mathbb{X}} P(F(\mathbf{x}) \leq T, \mathbf{G}(\mathbf{x}) \leq \mathbf{0} | \mathcal{F}_n, \mathcal{G}_n),$$

s'obtient en estimant une fonction de répartition gaussienne multivariée.

Pour pouvoir utiliser l'ECM dans ce contexte, quelques adaptations sont nécessaires. Pour les introduire, notons $(\Omega, \mathbb{A}, P)$ l'espace probabilisé sous-jacent, et notons explicitement, dans cette section, la dépendance des variables aléatoires en $\omega \in \Omega$[4]. Considérons, pour $\omega \in \Omega$ l'ensemble $\mathbb{G}^{\mathrm{c}}(\omega) = \{\mathbf{x} \in \mathbb{G} | \mathbf{G}(\mathbf{x}, \omega) \leq \mathbf{0}\}$ des points admissibles de $\mathbb{G}$ et remarquons que cet ensemble possède une probabilité non nulle d'être vide ($P(\mathbb{G}^{\mathrm{c}}(\omega) = \emptyset) > 0$). En d'autres termes, pour une réalisation de F et de $\mathbf{G}$, il n'existe pas nécessairement de solution au problème d'optimisation contraint. Pour procéder de manière similaire au cas non contraint, considérons une valeur fictive $\mathbf{x}_0$ choisie arbitrairement à l'extérieur de $\mathbb{G}$, et définissons l'ensemble des minimiseurs globaux de $F(\cdot, \omega)$ sur $\mathbb{G}^{\mathrm{c}}(\omega)$ comme

$$\mathcal{M}_{\mathbb{G}^{\mathrm{c}}} = \begin{cases} \{\mathbf{x}^* \in \mathbb{X} | F(\mathbf{x}^*) = \min_{\mathbb{G}^{\mathrm{c}}} F(\mathbf{x})\} & \text{si} \quad \mathbb{G}^{\mathrm{c}} \neq \emptyset \\ \mathbf{x}_0 & \text{sinon} \end{cases}.$$

Considérons ensuite, de manière analogue au cas non contraint, $\mathbf{X}^{*,\mathrm{c}}$, un vecteur aléatoire uniformément distribué sur $\mathcal{M}_{\mathbb{G}^{\mathrm{c}}}$. La distribution de probabilité $P_{\mathbf{X}^{*,\mathrm{c}}}(\cdot | \mathcal{F}_n, \mathcal{G}_n)$ de ce vecteur décrit alors les progrès effectués dans la résolution du problème d'optimisation contraint. En particulier, $P_{\mathbf{X}^{*,\mathrm{c}}}(\mathbf{x}_0 | \mathcal{F}_n, \mathcal{G}_n)$ donne la probabilité pour le problème de ne pas avoir de solution.

L'approximation de cette distribution se fait alors, de manière similaire à ce qui a été proposé dans la section précédente, en simulant conjointement F et $\mathbf{G}$. De cette façon, l'ECM dans le cas contraint $H(\mathbf{X}^* | \mathcal{F}_n, \mathcal{G}_n, F(\mathbf{x}), \mathbf{G}(\mathbf{x}))$ peut être approchée simplement.

Un exemple d'application de IAGO à un problème contraint est présenté sur la figure 2.7, où la fonction sinus utilisée depuis le début du chapitre s'accompagne d'une contrainte (une autre fonction sinus décalée sur les deux axes[5]). Pour simplifier la représentation, les intervalles de confiance

[4] Jusqu'à présent, il nous a semblé plus clair de ne pas souligner la dépendance à ω des variables aléatoires. Ce n'est plus le cas dans cette section, notamment pour la définition, à venir, de l'ensemble des points admissibles.

[5] La fonction objectif est $x \rightarrow \sin(6\pi x)$, la contrainte est $\sin(6\pi x + \pi) + 0.2 \geq 0$.

associés à la prédiction de la fonction et de la contrainte ne sont pas représentés. Notons cependant que les paramètres des covariances de la fonction à optimiser et de la contrainte (supposées indépendantes) sont identiques et fixés *a priori*. Les évaluations choisies par IAGO se révèlent pertinente, puisque trois des sept minimiseurs globaux du problème contraint sont identifiés.

Remarque 2.2. Notons que rien n'oblige à évaluer la fonction et les contraintes aux même points (hypothèse néanmoins faite dans cette section pour davantage de clarté). Nous pourrions donc imaginer des stratégies d'échantillonnage plus sophistiquées où l'on tiendrait par exemple compte des écarts de coût entre les évaluations de f et de $\mathbf{g}$ pour évaluer plus souvent la fonction la moins coûteuse.

Remarque 2.3. Si l'on ne suppose plus que F et $\mathbf{G}$ sont indépendants, ces processus doivent être prédits conjointement. Ceci peut se faire aisément grâce au co-krigeage (cf. Section B.1), mais nécessite le choix d'une covariance entre F et $\mathbf{G}$. En l'absence d'information *a priori*, l'estimation de cette covariance se révèle souvent incompatible avec la quantité de données disponible dans le contexte de ce mémoire.

2.4.4 Optimisation à plusieurs pas

Lorsque $l > 1$ évaluations de f peuvent être réalisées simultanément, les critères d'échantillonnage utilisés jusqu'à présent ne sont plus applicables. Nous décrivons dans cette section les quelques développements effectués dans cette direction dans la littérature s'attachant à l'optimisation globale à l'aide d'un budget réduit d'évaluations.

Si n évaluations ont été réalisées aux points de $\mathbb{S}_n$, comment choisir une mesure

$$J(\{\mathbf{y}_1,...,\mathbf{y}_l\},\mathbb{S}_n,\mathbf{f}_n)$$

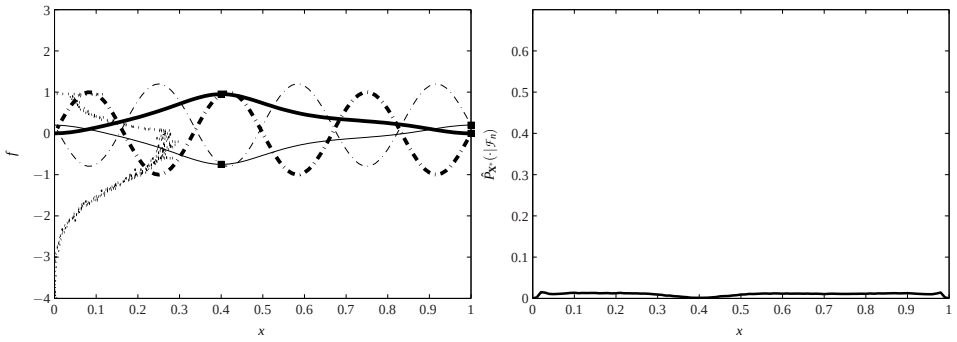
de l'intérêt pour l'optimisation de l évaluations supplémentaires aux points $(\mathbf{y}_1,...,\mathbf{y}_l) \in \mathbb{X}^l$? Et comment faire face à la complexité du problème d'optimisation à résoudre pour décider des points d'évaluations à choisir ? En effet, là où l'EI ou l'ECM étaient optimisés sur $\mathbb{X}$, il faut désormais résoudre

$$\underset{\{\mathbf{y}_1,...,\mathbf{y}_l\}\in\mathbb{X}^l}{\arg\max}\ J(\{\mathbf{y}_1,...,\mathbf{y}_l\},\mathbb{S}_n,\mathbf{f}_n).$$

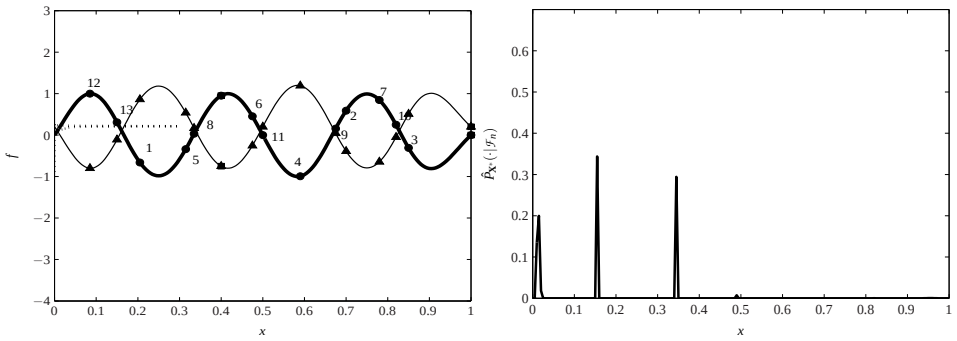
Schonlau (1997) puis Ginsbourger et al. (2007), proposent d'étendre EI à ce cas de figure. La définition de l'EI à l pas (ou l-EI chez Ginsbourger et al., 2007) est immédiate. L'expression (1.17) devient

$$\underset{F(\mathbf{y}_1)...F(\mathbf{y}_l)}{\mathbb{E}}(\min\{M_{n+l},F(\mathbf{x})\}-F^*|\mathcal{F}_n),$$

et maximiser le l-EI revient à minimiser l'erreur moyenne d'estimation du minimum après l évaluations. Cependant, le calcul direct de cette quantité implique l'estimation de fonctions de



(a) Prédiction initiale et distribution conditionnelle des minimiseurs globaux du problème contraint



(b) Prédiction et distribution conditionnelle après satisfaction du critère d'arrêt par IAGO

FIG. 2.7: *Optimisation d'une fonction sinus à l'aide de IAGO en présence d'une contrainte (trait mixte fin sur la partie supérieure gauche). Les figures de gauche présentent la prédiction de la fonction à optimiser (trait continu gras) et de la contrainte (trait continu fin) à partir des résultats des évaluations, ainsi que la distribution conditionnelle du minimum du problème contraint (traits pointillés). Les résultats d'évaluation initiaux de la contrainte et de la fonction sont représentés par des carrés. Les points obtenus par IAGO sont représentés par des cercles (dans l'ordre où les choisi IAGO) pour la fonction et des triangles (pour la contrainte). Sont aussi présentées sur la partie supérieure gauche, la fonction sinus considérée (trait gras mixte) et la contrainte (trait fin mixte). Les figures de droite présentent les distributions conditionnelles des minimiseurs globaux du problème contraint compte tenu des résultats d'évaluation disponibles. Les résultats présentés sont obtenus suite à la satisfaction du critère d'arrêt par IAGO (l'écart-type de la densité conditionnelle du minimum global du problème contraint plus faible que 0.1).*

répartition gaussienne en dimension l et devient rapidement impraticable (cf. Ginsbourger et al., 2007). Pour sortir de cette impasse, des heuristiques ont été proposées. Le principe général en est de choisir itérativement chacun des points d'évaluation en affectant une valeur fictive au résultat associé (par exemple la prédiction par krigeage ou le minimum courant dans Ginsbourger et al., 2007). Ainsi, l'optimisation du l-EI s'effectue de manière approchée par l maximisations de l'EI standard.

Ces heuristiques pourraient aussi s'appliquer à la mise au point d'un critère de minimisation de l'ECM à l-coups, mais il est probable que l'augmentation de la charge de calcul qui en résulte soit prohibitive. Rappelons en effet la méthode peu coûteuse proposée par Jones (2001) qui consiste à mener l P-algorithmes en parallèle et à utiliser pour chacun un seuil différent (cf. la section 1.3.2).

2.5 Conclusions

Nous avons décrit dans ce chapitre comment l'entropie conditionnelle des minimiseurs globaux peut être estimée à l'aide de simulations conditionnelles, puis optimisée par un choix adaptatif de l'ensemble des points candidats. Ces développements ont été regroupés dans l'algorithme d'optimisation IAGO dont nous avons discuté chacune des étapes, ainsi que l'extension à des problèmes d'optimisation contraints, à la prise en compte de résultats d'évaluation du gradient, et à la prise en compte de résultats d'évaluation bruités. Au chapitre 3, nous verrons comment étendre IAGO à l'optimisation robuste par rapport à une incertitude sur les facteurs.

CHAPITRE 3

OPTIMISATION ROBUSTE DE FONCTIONS COÛTEUSES

Résumé — A l'issue d'une procédure d'optimisation, une solution aussi proche de l'optimum que possible est retenue. Par la suite, cette solution est exploitée pour faire face à une réalité souvent plus complexe que la simulation qui en est faite. Il arrive alors souvent que les erreurs de mise en œuvre de la solution (dues par exemple à l'imprécision d'outils d'usinage) dégradent considérablement les performances. C'est la prise en compte de ces erreurs dans le processus d'optimisation, ou optimisation robuste par rapport aux incertitudes sur les facteurs, qui nous intéresse dans ce chapitre. Dans le domaine des fonctions coûteuses à évaluer, ce problème a peu été abordé. En revanche, le problème de prédiction de la robustesse est mieux connu en géostatistique et plus généralement en statistique. Dans ce chapitre, nous résumons les travaux existants et présentons l'intérêt des idées sous-jacentes pour une extension de IAGO à l'optimisation robuste.

3.1 Introduction

Dans la pratique, l'optimisation pour la conception de systèmes conduit au choix d'un système dit *nominal* qu'il faudra ensuite s'efforcer de réaliser. Les valeurs souhaitées pour les facteurs de ce système, appelés facteurs nominaux, résultent de l'estimation, à l'issue du processus d'optimisation, d'un optimiseur global de la performance f. La réalisation de ce système nominal est souvent sujette à des dispersions de sorte que les valeurs des facteurs obtenues différent des valeurs nominales. Si l'on note $\boldsymbol{\varepsilon}_{\mathbf{x}^*}$ la différence entre les facteurs du système effectivement réalisé et les facteurs nominaux, la performance obtenue s'écrit $f(\mathbf{x}^* + \boldsymbol{\varepsilon}_{\mathbf{x}^*})$ et s'avère généralement moins bonne que celle du système nominal.

La question centrale ici est de choisir des facteurs nominaux tels que la performance du système soit peu dégradée par l'introduction de $\boldsymbol{\varepsilon}_{\mathbf{x}^*}$. Il est clair que, dans bien des cas, le choix d'un optimum global n'est pas du tout adapté. Pour s'en convaincre, considérons l'exemple de la fi-

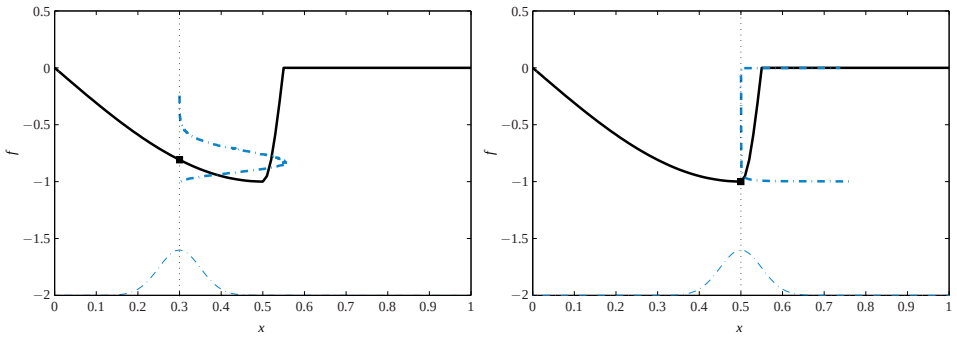


FIG. 3.1: *Influence d'un bruit gaussien sur la performance (traits pleins) au nominal. Deux nominaux sont testés (représentés par des carrés), 0.3* (partie gauche) *et 0.5* (partie droite). *Pour chacun, la distribution en sortie est présentée (traits mixtes). Le choix du minimiseur global comme nominal se révèle catastrophique puisque la performance est nulle dans 50% des cas.*

gure 3.1, pour lequel on a modélisé l'erreur sur les facteurs $\boldsymbol{\varepsilon}_{\mathbf{x}}$ par un bruit blanc gaussien. La fonction à optimiser sur cette figure est définie sur $[0,1]$ par

$$f_{\text{test}}^{(2)}(x) = \mathbb{1}_{[0,0.5]}(x)\sin(\pi x+\pi) + \mathbb{1}_{[0.5,0.55]}(x)\sin(10\pi(x-0.5)+\frac{3\pi}{2}),$$

et sera utilisée tout au long de ce chapitre. La performance du système réalisé $f_{\text{test}}^{(2)}(x+\varepsilon_x)$ est donc une grandeur aléatoire et sa distribution pour $x = x^* = 0.5$ et pour $x = 0.3$ est représentée sur la figure 3.1. Il apparaît alors clairement que le choix d'un minimiseur global pour les facteurs nominaux peut se révéler catastrophique en termes de sensibilité de la performance à une incertitude sur les facteurs.

Au lieu de chercher un minimiseur global de $f(\mathbf{x})$, nous nous intéresserons donc dans ce chapitre à une optimisation de f qui soit *robuste* à une incertitude sur les facteurs, c'est-à-dire que nous souhaitons prendre en compte non seulement la performance, mais aussi la robustesse (ou le manque de sensibilité) de cette performance par rapport à une erreur sur les facteurs. Nous verrons dans la section suivante plusieurs définitions de l'optimisation robuste, qui relèvent chacune d'un compromis entre performance et robustesse.

Ces formulations de l'optimisation robuste reposent sur des *mesures de robustesse* et se traduisent à nouveau par des problèmes d'optimisation globale de fonctions coûteuses. Mais cette fois, la fonction évaluée n'est plus celle que l'on cherche à optimiser. L'application de IAGO et des autres algorithmes d'optimisation globale va donc passer par la prédiction de ces mesures de robustesse à partir des résultats d'évaluation de f.

Commençons par préciser la nature des perturbations sur les facteurs à prendre en compte. Jusqu'à présent, nous avons considéré des perturbations additives sur les facteurs de conception. Elles permettent en effet de décrire de nombreuses sources d'incertitudes, comme des incertitudes de fabrication ou encore des erreurs de positionnement de capteurs. Cependant, comment rendre

compte de perturbations comme des variations de la température de fonctionnement du système, et plus généralement, de l'influence de tout ce qui relève de l'environnement du système et n'est donc pas contrôlable lors de sa réalisation ? Notons $\boldsymbol{\varepsilon}_{\mathrm{e},\mathbf{x}}$ le vecteur de ces variables d'environnement et $\mathbb{X}_{\mathrm{e}}$ son ensemble de définition. La performance $f(\mathbf{x}+\boldsymbol{\varepsilon}_{\mathbf{x}},\boldsymbol{\varepsilon}_{\mathrm{e},\mathbf{x}})$ du système s'écrit désormais en fonction de deux types de facteurs. L'optimisation robuste de f consistera donc à optimiser la robustesse du nominal choisi par rapport aux incertitudes représentées par $\boldsymbol{\varepsilon}_{\mathbf{x}}$ et par $\boldsymbol{\varepsilon}_{\mathrm{e},\mathbf{x}}$. Pour simplifier la présentation, nous ne considérerons pas, dans un premier temps, les perturbations dues à l'environnement du système. Nous verrons à la section 3.4.1 que la prise en compte de ces dernières dans un algorithme d'optimisation robuste dérivé de IAGO peut être réalisée très simplement.

Nous verrons aussi à la section 3.3.4 comment généraliser les techniques d'optimisation robuste au cas où les incertitudes sur les facteurs sont aussi présentes au cours de la phase d'optimisation. Ce cas de figure n'a, *a priori*, pas d'intérêt pour des problèmes d'optimisation reposant sur des simulations numériques, mais dès que l'optimisation est conduite à l'aide de réalisations de prototypes, le point où est effectivement réalisée l'évaluation peut être mal connu.

3.2 Formulations du problème

On peut distinguer les formulations de la robustesse suivant qu'elles utilisent un cadre déterministe ou probabiliste. Les approches déterministes reposent en général sur des critères minimax. Par exemple, on cherchera à minimiser la valeur maximale de f sur un voisinage $\mathcal{X}_{\mathbf{x}}$ du nominal $\mathbf{x}$, c'est-à-dire à minimiser $\max_{\mathbf{u}\in\mathcal{X}_{\mathbf{x}}} f(\mathbf{u})$ par rapport à $\mathbf{x}$.

Ce type de critère, qui s'intéresse surtout au *pire cas*, risque de conduire à une solution robuste mais de piètre performance. Il est souvent plus utile de considérer des mesures de robustesse reposant sur une distribution de probabilité du bruit, qui permettent d'explorer simplement différents compromis entre performance nominale et variabilité de cette performance. Ce choix est encore plus judicieux lorsque le budget d'évaluation est très limité, le manque d'information sur la fonction risquant en effet de renvoyer une vision très pessimiste du pire cas. En outre, le cadre probabiliste correspond bien mieux aux méthodes que nous proposons.

Nous allons donc modéliser les perturbations sur les facteurs de conception par une grandeur aléatoire $\boldsymbol{\varepsilon}_{\mathbf{x}}$. Nous supposerons que $\forall(\mathbf{x},\mathbf{y})\in\mathbb{X}^2$, $\boldsymbol{\varepsilon}_{\mathbf{x}}$ et $\boldsymbol{\varepsilon}_{\mathbf{y}}$ sont indépendants et que la distribution de $\boldsymbol{\varepsilon}_{\mathbf{x}}$, notée $p_{\boldsymbol{\varepsilon}}$, ne dépend pas de $\mathbf{x}$ (cette dernière hypothèse n'est faite que pour simplifier l'exposé). Notons que pour un facteur non affecté par une perturbation, la composante correspondante de $\boldsymbol{\varepsilon}_{\mathbf{x}}$, que l'on notera désormais simplement $\boldsymbol{\varepsilon}$, est nulle.

La performance du système après réalisation pour un vecteur de facteurs nominaux $\mathbf{x}$ est donc décrite par la variable aléatoire $f(\mathbf{x}+\boldsymbol{\varepsilon})$. Deux questions se posent alors.

– Étant donnés n résultats d'évaluation $\mathbf{f}_n$ de f, comment mesurer la robustesse et estimer

cette mesure pour un nominal $\mathbf{x}$ donné ?

– Comment choisir de nouvelles évaluations de f pour optimiser la performance et sa robustesse ?

3.2.1 Mesures de robustesse

L'hypothèse probabiliste faite sur le bruit nous conduit à mesurer la robustesse par une statistique de $f(\mathbf{x}+\boldsymbol{\varepsilon})$. Suivant la situation, on s'intéressera par exemple à

– la moyenne $\mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}}[f(\mathbf{x}+\boldsymbol{\varepsilon})]$,

– l'écart-type $\sqrt{\mathop{\mathrm{var}}_{\boldsymbol{\varepsilon}}(f(\mathbf{x}+\boldsymbol{\varepsilon}))}$,

– la probabilité pour la performance d'être supérieure à un seuil T : $\mathrm{P}_{\boldsymbol{\varepsilon}}(f(\mathbf{x}+\boldsymbol{\varepsilon})>T)$,

– ou encore à plusieurs de ces mesures simultanément.

Dans un contexte où le nombre d'évaluations de f est limité, il est exclu d'évaluer l'une ou l'autre de ces quantités par simulation directe. Nous allons donc chercher à les estimer à partir d'un petit nombre d'évaluations de f en exploitant le modèle gaussien F. Ainsi, une mesure de robustesse en $\mathbf{x}$ sera modélisée par une variable aléatoire. Par exemple, la performance moyenne $\mathbb{E}_{\boldsymbol{\varepsilon}}[f(\mathbf{x}+\boldsymbol{\varepsilon})]$ sera modélisée par $\mathbb{E}_{\boldsymbol{\varepsilon}}[F(\mathbf{x}+\boldsymbol{\varepsilon})]$. Dans la section 3.3, nous verrons comment prédire ces critères de robustesse à partir d'évaluations de f.

3.2.2 Formulations du problème d'optimisation

Si l'on est capable d'estimer ces mesures de robustesse en tout point de l'espace des facteurs, il reste encore à décider de la formulation du problème d'optimisation robuste à résoudre. Ce choix doit dépendre du problème considéré et de l'information disponible. Par exemple, alors que la minimisation de $\mathbb{E}_{\boldsymbol{\varepsilon}}(f(\mathbf{x}+\boldsymbol{\varepsilon}))$ peut s'avérer suffisante lorsque les dispersions n'affectent pas outre mesure le système (Wiesmann et al., 1998), la prise en compte de $\sqrt{\mathrm{var}_{\boldsymbol{\varepsilon}}(f(\mathbf{x}+\boldsymbol{\varepsilon}))}$ peut être essentielle dans le cas contraire (Das, 1997). Intéressons nous à la résolution du problème multi-objectif de minimisation simultanée de l'espérance et de l'écart-type de la performance. Si l'on dispose du *front de Pareto*, c'est-à-dire du sous ensemble de $\mathbb{X}$ dont chaque élément $\mathbf{x}$ est tel que

$$\forall \mathbf{y} \in \mathbb{X} \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}}(f(\mathbf{x}+\boldsymbol{\varepsilon})) \leq \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}}(f(\mathbf{y}+\boldsymbol{\varepsilon})) \text{ ou } \sqrt{\mathop{\mathrm{var}}_{\boldsymbol{\varepsilon}}(f(\mathbf{x}+\boldsymbol{\varepsilon}))} \leq \sqrt{\mathop{\mathrm{var}}_{\boldsymbol{\varepsilon}}(f(\mathbf{y}+\boldsymbol{\varepsilon}))},$$

il est possible de choisir parmi tous les compromis possibles entre performance et sensibilité de cette performance.

Pour éviter la difficulté supplémentaire apportée par le caractère multi-objectif (qui pose problème quand le budget d'évaluation est réduit), de nombreuses simplifications ont été proposées. On peut citer de manière non exhaustive (nous reviendrons sur les différentes possibilités à la section 3.4) :

– l'optimisation d'une agrégation des objectifs (cf. par exemple, Apley et al., 2006) :

$$\mathbf{x}^* \in \arg\min_{\mathbf{x}} \mathbb{E}_{\boldsymbol{\varepsilon}}[f(\mathbf{x}+\boldsymbol{\varepsilon})] + c\sqrt{\operatorname{var}_{\boldsymbol{\varepsilon}}[f(\mathbf{x}+\boldsymbol{\varepsilon})]}, \tag{3.1}$$

avec c une constante positive à choisir,

– la minimisation de la probabilité de défaillance

$$\mathbf{x}^* \in \arg\min_{\mathbf{x}} \mathrm{P}_{\boldsymbol{\varepsilon}}[f(\mathbf{x}+\boldsymbol{\varepsilon}) > T], \tag{3.2}$$

– ou encore la minimisation de la moyenne avec une contrainte sur la variance (définie comme la recherche de la M-robustesse par Lehman et al., 2004))

$$\begin{aligned} \mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathbb{X}} \quad & \mathbb{E}_{\boldsymbol{\varepsilon}}[f(\mathbf{x}+\boldsymbol{\varepsilon})], \\ \text{s.c.} \quad & \sqrt{\operatorname{var}_{\boldsymbol{\varepsilon}}[f(\mathbf{x}+\boldsymbol{\varepsilon})]} < \delta \end{aligned} \tag{3.3}$$

avec δ une constante positive à choisir.

Remarquons dès à présent que, malgré sa complexité, la formulation multi-objectif ne requiert pas de choix *a priori* de constantes par l'utilisateur, à l'inverse des simplifications proposées.

3.2.3 Choix d'une formulation

Ce choix est en fait double, puisque la formulation utilisée pour orienter le choix des évaluations de f peut être différente de celle utilisée pour le choix final. On peut en effet estimer l'optimum robuste pour une formulation à partir des résultats d'évaluation de f quelle que soit la façon dont ces évaluations ont été choisies. Cette distinction est importante pour la mise en pratique de l'optimisation robuste sur des fonctions coûteuses.

Pour de nombreux problèmes, et en particulier ceux considérés dans ce mémoire, la distribution du bruit est très mal connue. En pratique, ce manque d'information est dû

– aux difficultés rencontrées pour mesurer les facteurs en sortie de chaîne (par exemple un diamètre de conduit dans la culasse d'un moteur) ;
– à l'externalisation de la conception et de la construction de nombreuses pièces, qui rend plus complexe l'obtention des informations ;
– à la difficulté de spécifier la distribution des variables d'environnement (par exemple la température de fonctionnement d'un moteur).

En outre, le choix *a priori* d'une formulation du problème d'optimisation robuste pose de réelles difficultés aux praticiens qui souhaiteraient idéalement disposer du front de Pareto dans le plan moyenne – écart-type. Pour toutes ces raisons, il semble pour le moins hasardeux d'utiliser un critère probabiliste pour optimiser une fonction coûteuse. Cependant, la formulation du problème de robustesse et la loi du bruit ne sont utilisées que pour orienter l'échantillonnage de la fonction. Et, quand bien même une modification de ces dernières déplacerait l'optimum, au premier

ordre, une zone de l'espace des facteurs robuste pour un critère et une distribution du bruit le reste pour d'autres (nous vérifierons cela empiriquement à la fin de ce chapitre). Il semble légitime de considérer que, quelque soit la complexité de la fonction objectif, cette hypothèse est réaliste à la lumière du petit nombre d'évaluations disponibles. Une fois les évaluations réalisées plusieurs critères et plusieurs distributions du bruit peuvent de toute façon être utilisés pour choisir la solution finale.

3.3 Prédiction des mesures de robustesse

Nous nous intéressons ici à la prédiction des mesures de robustesse présentées à la section 3.2.1 à partir des résultats d'évaluations de f. Les mesures considérées sont des statistiques de $f(\mathbf{x}+\boldsymbol{\varepsilon})$ (moyenne, variance...), elle même modélisée par $F(\mathbf{x}+\boldsymbol{\varepsilon})$. L'estimation directe de la loi conditionnelle de ces statistiques en un point $\mathbf{x} \in \mathbb{X}$ est conceptuellement simple à mettre en place, mais très coûteuse en pratique. Elle implique en effet la simulation d'un échantillon à partir de la loi du bruit, puis la simulation de trajectoires conditionnelles de F sur cet échantillon et enfin le calcul de la statistique empirique pour chacune des trajectoires. Par exemple, la loi conditionnelle de l'écart-type de la performance en $\mathbf{x}$ pourra être estimée à partir de l'échantillon $s_{(1)},\dots,s_{(N)}$ obtenu comme

$$s_{(i)} = \sqrt{\frac{1}{r-1}\sum_{k=1}^{r}(f_{(i)}(\mathbf{x}+\boldsymbol{\varepsilon}_{(k)}) - \overline{m}_{(i)})^2},$$

avec $f_{(1)},\dots,f_{(N)}$ des simulations conditionnelles de F (générées comme décrit au chapitre 2), $\boldsymbol{\varepsilon}_{(1)},\dots,\boldsymbol{\varepsilon}_{(r)}$ des échantillons tirés suivant la loi du bruit et

$$\overline{m}_{(i)} = \frac{1}{r}\sum_{k=1}^{r} f_{(i)}(\mathbf{x}+\boldsymbol{\varepsilon}_{(k)})$$

la moyenne empirique de la *i*-ème trajectoire sous la loi du bruit. La figure 3.2 présente la distribution estimée de l'écart-type après cinq évaluations de $f_{\text{test}}^{(2)}$ sous l'hypothèse d'un bruit gaussien centré d'écart-type 0.05.

Une simulation directe de la loi semble inadaptée à l'optimisation puisqu'elle doit être renouvelée après chaque évaluation. Nous verrons dans cette section qu'il est possible de prédire plus efficacement la loi (ou au moins ses deux premiers moments) de la moyenne, de la variance et de la fonction de répartition de $F(\mathbf{x}+\boldsymbol{\varepsilon})$.

3.3.1 Prédiction de la moyenne

On cherche à prédire $M(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\varepsilon}}(F(\mathbf{x}+\boldsymbol{\varepsilon})) = F * p_{\boldsymbol{\varepsilon}}(\mathbf{x})$ à partir des évaluations de f. L'opérateur de convolution étant linéaire, la loi de cette quantité est gaussienne et l'on peut dériver des expressions analytiques pour sa moyenne et sa variance, conditionnellement aux résultats des évaluations de f, à partir des équations du krigeage (cf. par exemple Haylock et O'Hagan, 1996 ;

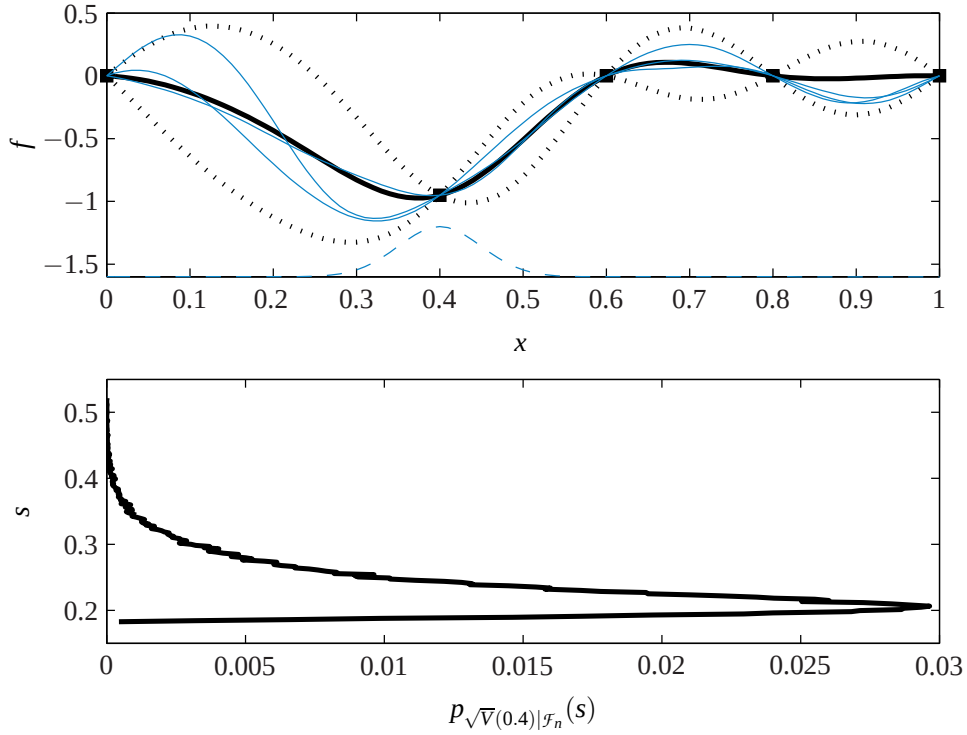


FIG. 3.2: *Estimation directe de la loi de l'écart-type en* $\mathbf{x} = 0.4$ *(*partie inférieure*). La* partie supérieure *de la figure présente la moyenne de la prédiction par krigeage (en gras) ainsi que les intervalles de confiance associés (en pointillés). Pour obtenir un échantillon suivant la loi de* $\sqrt{V}(0.4)$*, des trajectoire conditionnelles* $f_{(i)}$ *sont générées (trois exemples en traits fins) et l'écart-type de* $f_{(i)}(0.4+\boldsymbol{\varepsilon})$ *est estimé empiriquement en générant un échantillon suivant* $p_{\boldsymbol{\varepsilon}}$ *(normale, centrée et d'écart-type 0.05).*

Williams et al., 2000 ; Lehman et al., 2004 ; Apley et al., 2006) ou du krigeage dual (cf. Girard, 2004, pour la moyenne uniquement). L'application de EGO ou du P-algorithme à l'optimisation de la performance moyenne se fait donc très simplement en remplaçant la prédiction de f et la variance de l'erreur associée par la moyenne et la variance conditionnelle de M.

On peut aussi calculer la loi du processus M conditionnellement aux résultats des évaluations de f. En effet, M est un processus gaussien, de moyenne

$$\mathbb{E}(M(\mathbf{x})) = m * p_{\boldsymbol{\varepsilon}}(\mathbf{x}) \tag{3.4}$$

et de covariance

$$\begin{aligned} k_M(\mathbf{x},\mathbf{y}) &= \underset{F}{\mathbb{E}}[(F * p_{\boldsymbol{\varepsilon}}(\mathbf{x}) - m * p_{\boldsymbol{\varepsilon}}(\mathbf{x}))(F * p_{\boldsymbol{\varepsilon}}(\mathbf{y}) - m * p_{\boldsymbol{\varepsilon}}(\mathbf{y}))] \\ &= \int_{\mathbf{u}}\int_{\mathbf{u}'} \underset{F}{\mathbb{E}}[(F(\mathbf{x}+\mathbf{u}) - m(\mathbf{x}+\mathbf{u}))(F(\mathbf{y}+\mathbf{u}') - m(\mathbf{y}+\mathbf{u}'))] p_{\boldsymbol{\varepsilon}}(\mathbf{u}) p_{\boldsymbol{\varepsilon}}(\mathbf{u}') d\mathbf{u} d\mathbf{u}', \end{aligned} \tag{3.5}$$

ce qui peut aussi s'écrire comme un produit de convolution $k_M(\mathbf{x},\mathbf{y}) = [k(\cdot,\cdot) * \tilde{p}_{\boldsymbol{\varepsilon}}(\cdot,\cdot)](\mathbf{x},\mathbf{y})$, avec $\tilde{p}_{\boldsymbol{\varepsilon}}(\mathbf{x},\mathbf{y}) = p_{\boldsymbol{\varepsilon}}(\mathbf{x}) p_{\boldsymbol{\varepsilon}}(\mathbf{y})$. De plus, la covariance entre M et F s'exprime simplement[1] comme

$$k_{MF}(\mathbf{x},\mathbf{y}) = \underset{F}{\mathbb{E}}[(F * p_{\boldsymbol{\varepsilon}}(\mathbf{x}) - m * p_{\boldsymbol{\varepsilon}}(\mathbf{x}))(F(\mathbf{y}) - m(\mathbf{y}))] = k(\cdot,\mathbf{y}) * p_{\boldsymbol{\varepsilon}}(\mathbf{x}). \tag{3.6}$$

La prédiction de M est ainsi un cas particulier du *cokrigeage* (cf. Annexe B) qui permet de prédire tout processus G à partir des observations de F (moyennant quelques hypothèses sur la moyenne de G), du moment que sont connues la moyenne et la covariance de G, ainsi que la covariance entre G et F. Ce rapprochement est connu en géostatistique depuis les années 70 où la présence de bruit sur les facteurs est connue sous le nom d'*erreur de positionnement* (Chilès, 1976, 1977).

Les équations du cokrigeage dans le cas général sont données en annexe B.1. Cependant, pour mieux préciser les calculs à effectuer pour la prédiction de M, nous présentons ici le système à résoudre dans ce cas particulier. La prédiction par krigeage $\hat{M}$ à partir des résultats d'évaluation de f (que l'on peut assimiler à $\mathbb{E}[M(\mathbf{x})|\mathcal{F}_n]$ d'après la remarque 1.2) s'obtient comme combinaison linéaire des résultats d'évaluation de F,

$$\hat{M}(\mathbf{x}) = \boldsymbol{\lambda}_M(\mathbf{x})^{\mathsf{T}} \mathbf{f}_n,$$

avec $\boldsymbol{\lambda}_M$ solution de

$$\begin{pmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^{\mathsf{T}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_M(\mathbf{x}) \\ \boldsymbol{\mu}_M(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathbf{k}_{MF}(\mathbf{x}) \\ \mathbf{p}_M(\mathbf{x}) \end{pmatrix}, \tag{3.7}$$

où $\boldsymbol{\mu}_M(\mathbf{x})$ est un vecteur de coefficients de Lagrange,

$$\mathbf{k}_{MF}(\mathbf{x}) = [k_{MF}(\mathbf{x}_1,\mathbf{x}),\ldots,k_{MF}(\mathbf{x}_n,\mathbf{x})]^{\mathsf{T}}$$

[1] Notons que la convolution entraîne la perte de la propriété d'isotropie. Ainsi, à moins que l'on ne puisse écrire $\forall \mathbf{u} \in \mathbb{X}\ p_{\boldsymbol{\varepsilon}}(\mathbf{u}) = p_{\boldsymbol{\varepsilon}}(|\mathbf{u}|)$, k_M et k_{MF} ne sont pas isotropes.

et

$$\mathbf{p}_M(\mathbf{x}) = [p_1 * p_{\boldsymbol{\varepsilon}}(\mathbf{x}), \ldots, p_l * p_{\boldsymbol{\varepsilon}}(\mathbf{x})]^\mathsf{T}.$$

La variance de l'erreur de prédiction de M (qui peut être vue comme la variance de M conditionnellement à $\mathcal{F}_n$, cf. la section 1.2.2) s'obtient ensuite comme

$$\mathbb{E}\left[\left(\hat{M}(\mathbf{x}) - M(\mathbf{x})\right)^2\right] = k_M(\mathbf{x},\mathbf{x}) - \boldsymbol{\lambda}_M(\mathbf{x})^\mathsf{T}\mathbf{k}_{MF}(\mathbf{x}) - \mathbf{p}_M(\mathbf{x})^\mathsf{T}\boldsymbol{\mu}_M(\mathbf{x})\,. \tag{3.8}$$

Si l'on peut calculer ou approcher les intégrales (3.4), (3.5) et (3.6), il est donc possible de calculer, pour tout $\mathbf{x} \in \mathbb{X}$, la moyenne et la variance de $M(\mathbf{x})$ conditionnellement aux observations de f. Ceci était déjà possible par la dérivation directe de la moyenne et de la variance de F. Cependant, connaissant k_M et k_{MF}, il est maintenant possible de simuler conjointement M et F conditionnellement aux résultats des évaluations de f (cf. la section B.1) et d'étendre directement IAGO à l'optimisation de la performance moyenne[2].

Dans le cas où le bruit $\boldsymbol{\varepsilon}$ est supposé gaussien, $\mathbb{E}(M(\mathbf{x}))$, k_M et k_{MF} peuvent être calculées analytiquement pour nombre de covariances usuelles (gaussiennes, polynomiales...). Ces calculs sont en partie réalisés dans Girard (2004) dans le cas où la moyenne de F est supposée nulle.

L'utilisation d'une covariance de Matèrn présente une difficulté ici puisque sa forme analytique plus complexe ne permet pas de calculer explicitement le produit de convolution avec la loi gaussienne. En revanche, ce résultat peut être approché de manière assez précise par une covariance de Matèrn avec d'autres paramètres (cf. annexe A.2). Cette approximation a permis d'obtenir la figure 3.3 qui présente simultanément des simulations conditionnelles de F et de M reposant sur cinq évaluations de $f^{(2)}_{\text{test}}$ ($p_{\boldsymbol{\varepsilon}}$ est supposée gaussienne centrée d'écart-type 0.05). On peut y constater l'effet de régularisation du produit de convolution (cet effet apparaît aussi dans le paramètre ν de la covariance de Matèrn modélisant M qui passe de 1.5 pour k à 2.9 pour k_{MF} puis 4.4 pour k_M). La prédiction par krigeage ainsi que les trajectoires de M sont plus régulières que leurs équivalents pour F. Cela signifie que l'optimisation de $\mathbb{E}_{\boldsymbol{\varepsilon}}(f(\mathbf{x}+\boldsymbol{\varepsilon}))$ sera moins difficile que celle de f et d'autant plus si l'on utilise IAGO qui tire parti de la régularité des trajectoires pour accélérer la convergence.

Malgré les difficultés supplémentaires qu'elle entraîne, l'utilisation de la covariance de Matèrn dans ce contexte reste donc judicieuse, puisqu'elle permet de rendre compte de l'impact, sur la derivabilité du processus, du lissage effectué par la convolution avec la loi du bruit. En comparaison, les covariances infiniment dérivables en zéro, telles les covariances gaussiennes, impliquent des trajectoires analytiques (Stein, 1999) insensibles à ce phénomène.

Remarque 3.1. Notons que si l'on ne s'intéresse pas à la variance de l'erreur de prédiction, la prédiction de la performance moyenne s'obtient directement comme la convolution de $\hat{f}(\cdot)$ avec la loi du bruit.

[2] C'est le même principe que l'extention de IAGO à la prise en compte d'évaluations du gradient (cf. la section 2.4.2).

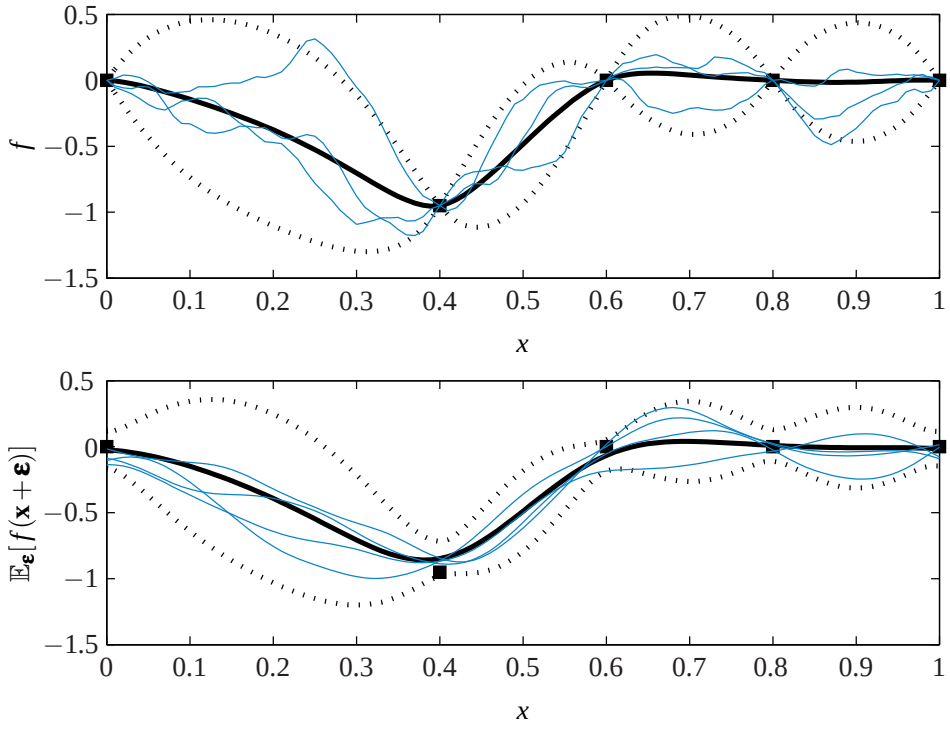


FIG. 3.3: *Prédiction de la performance moyenne* $\mathbb{E}_{\boldsymbol{\varepsilon}}[f^{(2)}_{test}(\mathbf{x}+\boldsymbol{\varepsilon})]$ *à partir de cinq évaluations de* $f^{(2)}_{test}$ *(carrés). La* partie supérieure *présente la prédiction de* $f^{(2)}_{test}$ *(en gras), les intervalles de confiance associés (en pointillés) et des simulations conditionnelles (traits fins). La* partie inférieure *présente, avec les mêmes conventions graphiques la prédiction de* $\mathbb{E}_{\boldsymbol{\varepsilon}}[f^{(2)}_{test}(\mathbf{x}+\boldsymbol{\varepsilon})]$ *(avec* $\boldsymbol{\varepsilon}$ *gaussien centré d'écart-type 0.05), les intervalles de confiance associés et des simulations* $M(\mathbf{x})$ *conditionnellement aux résultats des évaluations de* $f^{(2)}_{test}$.

3.3.2 Prédiction de la variance

Il s'agit maintenant de prédire

$$V(\mathbf{x}) = \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}}\left([F(\mathbf{x}+\boldsymbol{\varepsilon}) - M(\mathbf{x})]^2\right)$$

à partir des résultats des évaluations de f. On ne peut plus utiliser le principe du cokrigeage puisque V ne dépend plus linéairement de F. Il reste cependant possible d'obtenir directement une expression analytique pour la moyenne et la variance de $V(\mathbf{x})$ conditionnellement à $\mathcal{F}_n$. Ces résultats ont été présentés par Williams et al. (2000) et par Lehman et al. (2004) sous l'hypothèse d'un bruit à support discret, ainsi que par Apley et al. (2006) dans le cas général, mais avec un formalisme bayésien n'autorisant que peu de simplifications. Dans la suite de cette section, nous allons proposer des formes analytiques pour ces quantités, puis discuter de leur utilité. Pour simplifier les notations, l'espérance pour la loi de F conditionnellement à $\mathcal{F}_n$ sera noté $\mathbb{E}_n$.

La moyenne de V conditionnellement aux résultats des évaluations s'écrit

$$\begin{aligned}\mathop{\mathbb{E}}_{F}[V(\mathbf{x})|\mathcal{F}_n] &= \mathop{\mathbb{E}}_{n}\mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}}\left(\left[F(\mathbf{x}+\boldsymbol{\varepsilon}) - \mathop{\mathbb{E}}_{n}[M(\mathbf{x})] + \mathop{\mathbb{E}}_{n}[M(\mathbf{x})] - M(\mathbf{x})\right]^2\right) \\ &= \mathop{\mathbb{E}}_{n}\mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}}\left(\left[F(\mathbf{x}+\boldsymbol{\varepsilon}) - \mathop{\mathbb{E}}_{n}[M(\mathbf{x})]\right]^2\right) - \mathrm{var}\left[M(\mathbf{x})|\mathcal{F}_n\right],\end{aligned} \tag{3.9}$$

puis, en inversant les espérances,

$$\begin{aligned}\mathop{\mathbb{E}}_{F}[V(\mathbf{x})|\mathcal{F}_n] &= \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}}\mathop{\mathbb{E}}_{n}\left(\left[F(\mathbf{x}+\boldsymbol{\varepsilon}) - \mathop{\mathbb{E}}_{n}[F(\mathbf{x}+\boldsymbol{\varepsilon})] + \mathop{\mathbb{E}}_{n}[F(\mathbf{x}+\boldsymbol{\varepsilon})] - \mathop{\mathbb{E}}_{n}[M(\mathbf{x})]\right]^2\right) - \mathrm{var}[M(\mathbf{x})|\mathcal{F}_n] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}}\left[\hat{\sigma}^2(\mathbf{x}+\boldsymbol{\varepsilon})\right] + \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}}\left[\mathop{\mathbb{E}}_{n}\left(F(\mathbf{x}+\boldsymbol{\varepsilon}) - M(\mathbf{x})\right)^2\right] - \mathrm{var}[M(\mathbf{x})|\mathcal{F}_n].\end{aligned} \tag{3.10}$$

Enfin, en remplaçant $\mathbb{E}_n[F(\mathbf{x}+\boldsymbol{\varepsilon})]$ par $\hat{f}(\mathbf{x}+\boldsymbol{\varepsilon})$ dans le deuxième terme, l'espérance conditionnelle de $V(\mathbf{x})$ s'écrit

$$\mathop{\mathbb{E}}_{F}[V(\mathbf{x})|\mathcal{F}_n] = \hat{\sigma}^2 * p_{\boldsymbol{\varepsilon}}(\mathbf{x}) + \mathop{\mathrm{var}}_{\boldsymbol{\varepsilon}}[\hat{f}(\mathbf{x}+\boldsymbol{\varepsilon})] - \mathrm{var}[M(\mathbf{x})|\mathcal{F}_n]. \tag{3.11}$$

Le troisième terme est donné par (3.8), mais les deux premiers doivent en général être approchés. Girard (2004) les calcule explicitement pour un bruit *et* une covariance gaussiens et propose, pour les autres covariances, une approximation analytique reposant sur un développement à l'ordre deux de la covariance de F.

Le calcul de

$$\mathrm{var}[V(\mathbf{x})|\mathcal{F}_n] = \mathop{\mathbb{E}}_{n}\left[V(\mathbf{x})^2\right] - \mathop{\mathbb{E}}_{n}[V(\mathbf{x})]^2,$$

s'effectue de la même façon que celui de la moyenne, cependant, l'inversion de l'ordre des espérances ne peut se faire que terme à terme, une fois $V(\mathbf{x})^2$ développé. Bon nombre de termes

s'annulent ou se simplifient, et on arrive finalement à

$$\begin{aligned}(3.12)\quad \operatorname{var}[V(\mathbf{x})|\mathcal{F}_n] &= 2\operatorname*{var}_n[M(\mathbf{x})]^2 + k_n^2(\cdot,\cdot) * (p_{\boldsymbol{\varepsilon}}(\cdot)p_{\boldsymbol{\varepsilon}}(\cdot))\,(0,0) \\ &+ 4\int_{\mathbf{u}}\int_{\mathbf{u}'} k_n(\mathbf{x}+\mathbf{u},\mathbf{x}+\mathbf{u}')\left[\hat{f}(\mathbf{x}+\mathbf{u}) - \mathop{\mathbb{E}}_n[M(\mathbf{x})]\right]\left[\hat{f}(\mathbf{x}+\mathbf{u}') - \mathop{\mathbb{E}}_n[M(\mathbf{x})]\right] p_{\boldsymbol{\varepsilon}}(\mathbf{u})p_{\boldsymbol{\varepsilon}}(\mathbf{u}')d\mathbf{u}d\mathbf{u}' \\ &- 2\int_{\mathbf{u}} \mathop{\mathbb{E}}_n\left(\left[F(\mathbf{x}+\mathbf{u}) - \hat{f}(\mathbf{x}+\mathbf{u})\right]\left[M(\mathbf{x}) - \mathop{\mathbb{E}}_n[M(\mathbf{x})]\right]\right) p_{\boldsymbol{\varepsilon}}(\mathbf{u})d\mathbf{u},\end{aligned}$$

avec k_n, la fonction de covariance de l'erreur de prédiction $F - \hat{F}$. Cette dernière s'obtient simplement en fonction des coefficients du krigeage sous la forme :

$$k_n(\mathbf{x},\mathbf{y}) = \mathbb{E}\left[F(\mathbf{x}) - \hat{F}(\mathbf{x})\right]\left[F(\mathbf{y}) - \hat{F}(\mathbf{y})\right] = k(\mathbf{x},\mathbf{y}) + \boldsymbol{\lambda}(\mathbf{x})^{\mathsf{T}}\mathbf{K}\boldsymbol{\lambda}(\mathbf{y}) - \boldsymbol{\lambda}(\mathbf{x})^{\mathsf{T}}\mathbf{k}(\mathbf{x}) - \boldsymbol{\lambda}(\mathbf{y})^{\mathsf{T}}\mathbf{k}(\mathbf{y})$$

(rappelons que $k_n(\mathbf{x},\mathbf{x}) = \hat{\sigma}^2(\mathbf{x})$).

L'équation (3.12) peut être approchée numériquement. Son premier terme est donné par (3.8), l'approximation du second peut se faire indépendamment de $\mathbf{x}$, l'approximation du troisième implique un produit de convolution en $2d$ dimensions, quant au dernier, il s'agit là encore d'un produit de convolution en dimension d. Il est à noter que ces deux produits de convolution impliquent chacun une fonction dont l'évaluation requiert le calcul des coefficients de la prédiction par krigeage de M et de F. Malgré sa complexité, cette méthode de prédiction reste compétitive face à une approximation directe par Monte-Carlo (détaillée au début de cette section) qui implique autant de convolutions en dimension d que l'on utilise de simulations conditionnelles. Elle ne propose cependant que l'approximation de la moyenne et de la variance de $\operatorname{var}_{\boldsymbol{\varepsilon}} F(\mathbf{x}+\boldsymbol{\varepsilon})$. L'application de EI ou du P-algorithme à la minimisation de la variance de $\operatorname{var}_{\boldsymbol{\varepsilon}} f(\mathbf{x}+\boldsymbol{\varepsilon})$ ne peut donc se faire qu'au prix d'approximations supplémentaires (sans parler de l'utilisation de IAGO qui ne paraît pas envisageable). On peut aussi citer Apley et al. (2006) qui approchent la loi de $\operatorname{var}_{\boldsymbol{\varepsilon}} F(\mathbf{x}+\boldsymbol{\varepsilon})$ par une gaussienne de moyenne et de variance données par (3.11) et (3.12). Pour simplifier encore la prédiction, il est aussi envisageable de proposer une approximation analytique de (3.11) et de (3.12) (cf. Pronzato et Thierry (2003), qui proposent une approximation obtenue par développement de Taylor à l'ordre deux de F), avant d'appliquer EI ou le P-algorithme sous l'approximation gaussienne.

Cependant, comme le montre la figure 3.2, la loi de $\operatorname{var}_{\boldsymbol{\varepsilon}} F(\mathbf{x}+\boldsymbol{\varepsilon})$ est, en général, assez différente d'une gaussienne. C'est pourquoi Lehman et al. (2004) proposent d'approcher directement le critère EI par Monte-Carlo (sans l'hypothèse gaussienne, l'expression analytique pour EI vue au chapitre 1 n'est plus valable) sans utiliser les expressions analytiques pour la moyenne et la variance. Malheureusement, la complexité calculatoire de cette approche exclut *a priori* toute application en dimension élevée.

3.3.3 Prédiction de la probabilité de défaillance

Intéressons nous maintenant à la prédiction de la probabilité de dépasser un seuil, ou de manière équivalente, à la fonction de répartition de $F(\mathbf{x}+\boldsymbol{\varepsilon})$

$$D(\mathbf{x},T) = P_{\boldsymbol{\varepsilon}}\left[F(\mathbf{x}+\boldsymbol{\varepsilon}) \geq T\right] = \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}}\left[\mathbb{1}_{F(\mathbf{x}+\boldsymbol{\varepsilon})\geq T}\right]. \tag{3.13}$$

Comme pour la prédiction de la variance, l'absence de relation linéaire entre F et D empêche la prédiction de D par co-krigeage.

Dans ce contexte, où l'on s'intéresse davantage à une probabilité de dépassement qu'à l'amplitude de ce dépassement, les géostatisticiens ont recours au *krigeage des indicatrices* (Journel, 1982 ; Chilès et Delfiner, 1999). Il s'agit simplement d'appliquer le principe du krigeage à $\mathbb{1}_{f(\cdot)\geq T}$ en lieu et place de f. Ainsi, c'est $\mathbb{1}_{f(\cdot)\geq T}$ qui est modélisée par un processus gaussien, et la prédiction s'effectue à l'aide des seules valeurs des indicatrices $\left[\mathbb{1}_{f(\mathbf{x}_1)\geq T},...,\mathbb{1}_{f(\mathbf{x}_n)\geq T}\right]$. Le modèle gaussien est peu satisfaisant, puisqu'il ne rend pas compte du caractère discontinu de $\mathbb{1}_{f(\cdot)\geq T}$ et qu'il ne contraint pas les valeurs de la prédiction entre 0 et 1. De plus, une grande partie de l'information apportée par les évaluations est ainsi inutilisée, mais cette méthode permet d'estimer simplement la probabilité de dépasser le seuil T ainsi qu'un intervalle de confiance associé. La figure 3.4 présente un exemple où la prédiction et les intervalles de confiance à 95% sont contraints entre 0 et 1 par simple seuillage.

L'équation (3.13) fournit une relation linéaire entre D et les indicatrices de F identique à celle qui liait F et M. On peut donc procéder de la même façon qu'à la section 3.3.1, et prédire D à partir des valeurs des indicatrices. En effet, la moyenne de D, sa covariance et la covariance entre F et D s'obtiennent par les relations (3.4), (3.5) et (3.6) en remplaçant m et k, par la moyenne et la covariance utilisées pour modéliser $\mathbb{1}_{f(\cdot)\geq T}$. La figure 3.4 présente un exemple de prédiction de D.

Cette approche simple possède malheureusement un handicap majeur, si toutes les valeurs prises par les indicatrices sont égales à 1, c'est-à-dire si $\forall i \in [\![1;n]\!]\ f(\mathbf{x}_i) \geq T$ (ce qui risque fortement d'arriver en grande dimension), la prédiction de D sera constante sur $\mathbb{X}$ et égale à 1. Il deviendra alors impossible d'orienter efficacement l'échantillonnage.

Pour pallier ce problème, il est toujours possible de calculer directement la moyenne et la variance de D conditionnellement aux résultats des évaluations de f. Cependant, tout comme pour la prédiction de la variance, $D(\mathbf{x},T)$ n'est pas une variable aléatoire gaussienne, et l'on perd la possibilité d'appliquer EI ou IAGO sans approximation supplémentaire.

On obtient pour la moyenne

$$\begin{aligned}\mathbb{E}\left[D(\mathbf{x},T)|\mathcal{F}_n\right] &= \int_{\mathbf{u}} \mathop{\mathbb{E}}_{n}\left[\mathbb{1}_{F(\mathbf{x}+\mathbf{u})\geq T}\right] p_{\boldsymbol{\varepsilon}}(\mathbf{u})d\mathbf{u} \\ &= \int_{\mathbf{u}} \Phi\left(\frac{T-\hat{f}(\mathbf{x}+\mathbf{u})}{\hat{\sigma}(\mathbf{x}+\mathbf{u})}\right) p_{\boldsymbol{\varepsilon}}(\mathbf{u})d\mathbf{u},\end{aligned} \tag{3.14}$$

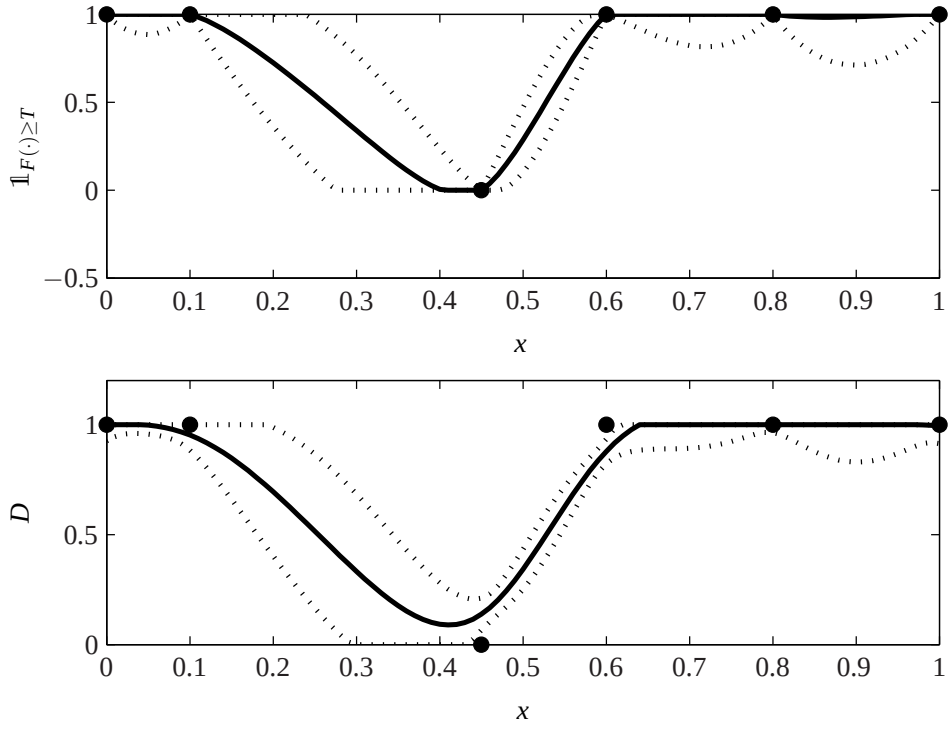


FIG. 3.4: *Prédiction de la probabilité pour $f_{test}^{(2)}$ de dépasser le seuil -0.65, à partir de cinq évaluations de f.* Partie supérieure *: prédiction par krigeage de $\mathbb{1}_{f(\cdot)\geq T}$ (trait gras) à partir des indicatrices (disques), et intervalles de confiance associés seuillés entre 0 et 1 (traits pointillés).* Partie inférieure *: prédiction par krigeage de D à partir des indicatrices sous l'hypothèse d'un bruit gaussien centré d'écart-type 0.05 (même conventions graphiques que pour la partie supérieure).*

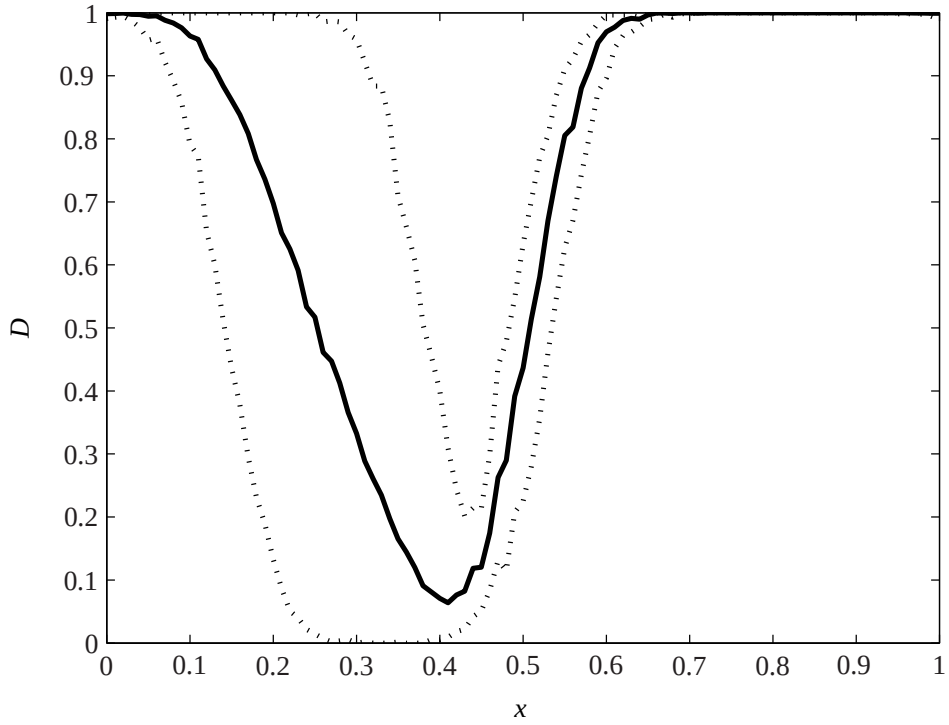


FIG. 3.5: *Estimation de la moyenne (trait gras) et des intervalles de confiance à 95% (traits pointillés) de D conditionnellement à cinq résultats d'évaluations de $f_{test}^{(2)}$ (identiques à ceux de la figure 3.4). Le bruit est ici encore supposé gaussien centré d'écart-type 0.05.*

et pour la variance

(3.15)

$$\begin{aligned}\mathrm{var}\left[D(\mathbf{x},T)|\mathcal{F}_n\right] &= \int_{\mathbf{u}}\int_{\mathbf{u}'} \mathop{\mathbb{E}}_{n}\left[\mathbb{1}_{F(\mathbf{x}+\mathbf{u})\geq T}\right]\mathop{\mathbb{E}}_{n}\left[\mathbb{1}_{F(\mathbf{x}+\mathbf{u}')\geq T}\right] p_{\boldsymbol{\varepsilon}}(\mathbf{u})p_{\boldsymbol{\varepsilon}}(\mathbf{u}')d\mathbf{u}d\mathbf{u}' - \mathbb{E}\left[D(\mathbf{x},T)|\mathcal{F}_n\right]^2 \\ &= \int_{\mathbf{u}}\int_{\mathbf{u}'} P(\{F(\mathbf{x}+\mathbf{u})\geq T\}\cap\{F(\mathbf{x}+\mathbf{u}')\geq T\}|\mathcal{F}_n)p_{\boldsymbol{\varepsilon}}(\mathbf{u})p_{\boldsymbol{\varepsilon}}(\mathbf{u}')d\mathbf{u}d\mathbf{u}' \\ &\quad - \mathbb{E}\left[D(\mathbf{x},T)|\mathcal{F}_n\right]^2.\end{aligned}$$

Ces deux expressions doivent être approchées numériquement, et il ressort de Oakley et O'Hagan (1998), qu'il est en réalité plus pratique d'estimer directement la loi de D par Monte Carlo comme détaillé au début de cette section. Un exemple de ce que l'on peut ainsi obtenir est présenté sur la figure 3.5 où la loi de D est résumée par sa moyenne, son quantile à 2.5% et son quantile à 97.5% (l'exemple est identique à celui de la figure 3.4).

Finalement, l'utilisation de EI ou du P-algorithme pour l'optimisation de $P_{\boldsymbol{\varepsilon}}[f(\mathbf{x}+\boldsymbol{\varepsilon})\geq T]$ ne peut se faire qu'au prix d'approximations importantes et peu satisfaisantes (en particulier le modèle gaussien des indicatrices) ou de simulations très coûteuses.

3.3.4 Position incertaine des évaluations

Il arrive en pratique, que l'incertitude sur les facteurs ne puisse être éliminée au cours de la phase d'optimisation. Par exemple, l'incertitude de fabrication d'un prototype réalisé à la main, pourra être plus faible que l'incertitude en sortie de chaîne de montage, tout en restant difficilement négligeable. Le problème de la prédiction en présence d'incertitude sur les facteurs est bien connu en géostatistique et a aussi été abordé dans Girard (2004).

Notons $\tilde{\boldsymbol{\varepsilon}}$ le bruit sur les facteurs au cours de la phase d'optimisation (supposé blanc), et $p_{\tilde{\boldsymbol{\varepsilon}}}$ sa densité de probabilité (supposée indépendante de $\mathbf{x}$), *a priori* différente de $p_{\boldsymbol{\varepsilon}}$. On cherche maintenant à prédire les critères de robustesse mentionnés précédemment à partir de réalisations de $f(\mathbf{x}+\tilde{\boldsymbol{\varepsilon}})$ pour $\mathbf{x} \in \mathbb{S}_n$.

Pour étendre à ce nouveau cas les expressions analytiques obtenues pour la moyenne et la variance de V et de D, il suffit de remplacer la moyenne et la covariance de l'erreur de la prédiction par krigeage de f à partir de $\mathbf{F}_n$ (c'est-à-dire $\hat{f}$ et k_n) par leur équivalent lorsque la prédiction repose sur les réalisations de $f(\mathbf{x}+\tilde{\boldsymbol{\varepsilon}})$. Pour ce faire, on peut une fois de plus utiliser le principe du cokrigeage, puisque la moyenne de $\xi(\mathbf{x}) = F(\mathbf{x}+\tilde{\boldsymbol{\varepsilon}})$ est connue

$$\mathbb{E}[\xi(\mathbf{x})] = m * p_{\tilde{\boldsymbol{\varepsilon}}}(\mathbf{x})$$

(notons que l'espérance porte à la fois sur F et sur le bruit $\boldsymbol{\varepsilon}$), tout comme sa covariance

$$k_\xi(\mathbf{x},\mathbf{y}) = \left[k(\cdot,\cdot) * \left(p_{\tilde{\boldsymbol{\varepsilon}}}(\cdot)p_{\tilde{\boldsymbol{\varepsilon}}}(\cdot)\right)\right](\mathbf{x},\mathbf{y})$$

ainsi que la covariance entre $F(\mathbf{x})$ et $F(\mathbf{y}+\tilde{\boldsymbol{\varepsilon}})$

$$k_{F\xi}(\mathbf{x},\mathbf{y}) = \mathbb{E}\left(\left[F(\mathbf{x}) - m(\mathbf{x})\right]\left[\xi(\mathbf{x}) - \mathbb{E}[\xi(\mathbf{x})]\right]\right) = k(\cdot,\mathbf{y}) * p_{\tilde{\boldsymbol{\varepsilon}}}(\mathbf{x}).$$

Le prédicteur par krigeage de $F(\mathbf{x})$ à partir des résultats $\boldsymbol{\xi}_n = [\xi(\mathbf{x}_1),...,\xi(\mathbf{x}_n)]^\mathsf{T}$ des évaluations de ξ s'écrit donc

$$\boldsymbol{\lambda}_F(\mathbf{x})^\mathsf{T}\boldsymbol{\xi}_n,$$

avec $\boldsymbol{\lambda}_F$ solution de

$$\begin{pmatrix} \mathbf{K}_\xi & \mathbf{P}_\xi \\ \mathbf{P}_\xi^\mathsf{T} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_F(\mathbf{x}) \\ \boldsymbol{\mu}_F(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathbf{k}_{F\xi}(\mathbf{x}) \\ \mathbf{p}(\mathbf{x}) \end{pmatrix}, \tag{3.16}$$

où $\boldsymbol{\mu}_F(\mathbf{x})$ est un vecteur de coefficients de Lagrange,

$$\mathbf{k}_{F\xi}(\mathbf{x}) = [k_{F\xi}(\mathbf{x}_1,\mathbf{x}),\ldots,k_{F\xi}(\mathbf{x}_n,\mathbf{x})]^\mathsf{T},$$

$$\mathbf{K}_\xi = \left(k_\xi(\mathbf{x}_i,\mathbf{x}_j)\right)_{1\le i,j\le n},$$

$$\mathbf{P}_\xi = \begin{pmatrix} \mathbf{p}_\xi(\mathbf{x}_1)^\mathsf{T} \\ \vdots \\ \mathbf{p}_\xi(\mathbf{x}_n)^\mathsf{T} \end{pmatrix},$$

et où

$$\mathbf{p}_\xi(\mathbf{x}) = [p_1 * p_{\tilde{\boldsymbol{\varepsilon}}}(\mathbf{x}), \ldots, p_l * p_{\tilde{\boldsymbol{\varepsilon}}}(\mathbf{x})]^\mathsf{T}.$$

La variance de l'erreur associée à cette prédiction s'obtient ensuite comme

$$\text{var}\,[M(\mathbf{x})] = k(\mathbf{x},\mathbf{x}) - \boldsymbol{\lambda}_F(\mathbf{x})^\mathsf{T}\mathbf{k}_{F\xi}(\mathbf{x}) - \mathbf{p}_F(\mathbf{x})^\mathsf{T}\boldsymbol{\mu}_F(\mathbf{x})\,. \tag{3.17}$$

Pour la prédiction de la moyenne, le cokrigeage s'applique là encore en utilisant le fait que la covariance entre $M(\mathbf{x})$ et $\xi(\mathbf{x})$ s'exprime comme

$$k_{M\xi}(\mathbf{x},\mathbf{y}) = \mathbb{E}\,[M(\mathbf{x}) - \mathbb{E}[M(\mathbf{x})]]\,[F(\mathbf{x}+\tilde{\boldsymbol{\varepsilon}}) - \mathbb{E}[F(\mathbf{x}+\tilde{\boldsymbol{\varepsilon}})]] = k(\cdot,\cdot) * (p_{\tilde{\boldsymbol{\varepsilon}}}(\cdot)p_{\boldsymbol{\varepsilon}}(\cdot))\,(\mathbf{x},\mathbf{y}).$$

On obtient ainsi la prédiction en modifiant le terme de droite dans (3.16).

3.4 Optimisation robuste

En fonction de la formulation choisie pour le problème d'optimisation robuste, une ou plusieurs des mesures de robustesse considérées dans la section précédente interviennent. Nous avons vu, comment, à partir des résultats des évaluations de f, elles peuvent être prédites. Prédiction qui peut être utilisée par les critères d'échantillonnage vus au chapitre 1 pour choisir les positions de nouvelles évaluations dans le but d'estimer un optimum robuste. Il faut donc choisir une formulation de l'optimisation robuste et un critère d'échantillonnage.

Dans la littérature, ces questions ont été assez peu traitées. Williams et al. (2000) s'intéressent les premiers à l'optimisation robuste de fonctions coûteuses à l'aide de processus gaussiens, et font le choix d'optimiser la performance moyenne M grâce à l'EI. Dans ces travaux, la loi du bruit sur les facteurs possède un support discret, ce qui facilite encore la prédiction de M. Par la suite, Lehman et al. (2004), toujours avec la même hypothèse sur le bruit, utilisent l'EI pour résoudre les problèmes d'optimisation de la M-robustesse (définie par (3.3)) et de la V-robustesse, définie comme

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{X}} \quad & \sqrt{\text{var}_{\boldsymbol{\varepsilon}}[f(\mathbf{x}+\boldsymbol{\varepsilon})]}, \\ \text{s.c.} \quad & \mathbb{E}_{\boldsymbol{\varepsilon}}[f(\mathbf{x}+\boldsymbol{\varepsilon})] < \delta \end{aligned} \tag{3.18}$$

avec δ une constante positive à choisir. La prise en compte des contraintes y est effectuée comme le propose Schonlau (1997) (cf. la section 2.4.3), c'est-à-dire en insérant la probabilité pour la contrainte d'être satisfaite comme facteur multiplicatif dans l'EI. Plus récemment, Apley et al. (2006) proposent, toujours grâce à l'EI, d'optimiser une combinaison linéaire de M et $\sqrt{V}$ sous l'hypothèse que ces deux processus sont conjointement gaussiens.

Outre ces possibilités, nous avons vu qu'il pouvait aussi être légitime, compte-tenu des informations disponibles sur le système, de s'intéresser au front de Pareto entre moyenne et écart-type ou encore à une probabilité de défaillance.

Plutôt que de vanter les mérites comparés de chacune des approches, une remarque importante nous semble s'imposer. Dans un contexte de fonctions coûteuses, la formulation du problème d'optimisation robuste n'a que peu d'importance sur l'échantillonnage effectué. Pour s'en convaincre, considérons la figure 3.6 où les prédictions et intervalles de confiance à 95% de la moyenne M de la performance, de sa variance V et de la probabilité de défaillance D, conditionnellement à deux résultats d'évaluation de f sont présentés. On y constate que les comportements de M et D sont similaires (et V semble inutilisable sans évaluations supplémentaires). Ainsi, pour les problèmes en grande dimension que l'on pourra rencontrer en pratique, les critères d'échantillonnage reposant sur M ou D indiqueront au premier ordre les mêmes zones, alors que les critères reposant sur V ne pourront servir que si le nombre des évaluations réalisées est important.

Or nous avons vu que la prédiction de M se trouve être, et de beaucoup, la moins coûteuse, tout en autorisant l'utilisation de l'EI et de l'ECM sans approximation ou simulation supplémentaire. Nous formulerons donc le problème d'optimisation robuste comme la minimisation de la performance moyenne. Il est bien évident qu'à l'issue du processus d'échantillonnage, il sera possible et sans doute même nécessaire de revenir aux autres mesures de la robustesse pour estimer, par exemple, une probabilité de défaillance ou le front de Pareto moyenne écart-type. En résumé, la vision simple et pratique de la robustesse qu'est la minimisation de la performance moyenne peut être utilisée pour l'échantillonnage, mais le choix de la solution retenue peut se faire à l'aide de mesures plus complexes.

Avec cette vision de la robustesse, on retrouve pour l'utilisation de l'EI, de la probabilité d'amélioration ou de l'ECM, les mêmes conditions qu'en présence d'un bruit sur les résultats des évaluations (cf. la section 2.4.1). Il est donc nécessaire de spécifier pour l'EI un estimateur du minimum. Nous reviendrons sur cette question à la section 4.4.2.

3.4.1 Prise en compte des variables d'environnement

Si l'on souhaite prendre en compte, dès la conception, des perturbations extérieures au système considéré, telles que la variation de la température de fonctionnement ou le vieillissement, il peut être nécessaire de différencier ces perturbations des facteurs de conception effectivement contrôlables lors de la réalisation du système. Dans ce paragraphe, nous souhaitons donc résoudre le problème de minimisation

$$\min_{\mathbf{x}\in\mathbb{X}} \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}_e} (f(\mathbf{x}+\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}_e), \tag{3.19}$$

avec $\boldsymbol{\varepsilon}_e$ un bruit non corrélé avec $\boldsymbol{\varepsilon}$, et de distribution $p_{\boldsymbol{\varepsilon}_e}$ connue.

On suppose qu'au cours de la conception, tous les facteurs sont contrôlables. Considérons n résultats d'évaluations $f(\mathbf{x}_1,\mathbf{x}_{e,1}),...,f(\mathbf{x}_n,\mathbf{x}_{e,n})$, avec $\mathbf{x}_{e,1},...,\mathbf{x}_{e,n}$ les valeurs choisies pour les facteurs environnementaux. La prédiction de $M(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}_e}(f(\mathbf{x}+\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}_e)$ à partir de ces données se

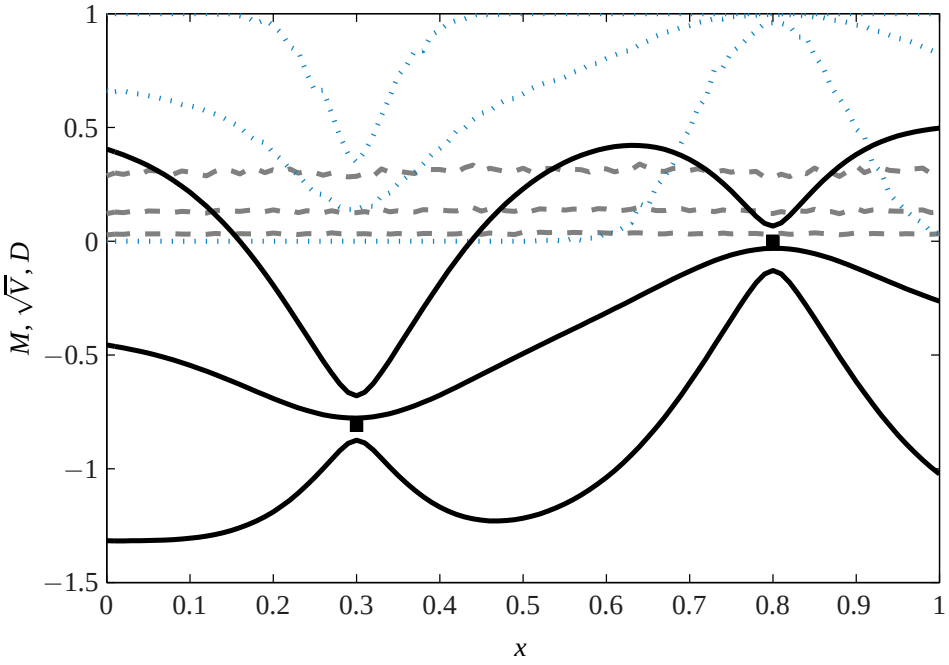


FIG. 3.6: *Prédiction de $M(\mathbf{x})$ (trait plein), $\sqrt{V(\mathbf{x})}$ (traits pointillés longs) et $D(\mathbf{x})$ (avec $T = -0.65$, trait pointillés courts) en 1D à partir de deux évaluations (carrés). Pour chacune de ces trois mesures de robustesse et en chacun des points de $[0,1]$, la moyenne, le quantile à 2.5% et le quantile à 97.5 % sont estimés par Monte-Carlo (excepté pour M, prédit directement par krigeage comme détaillé à la section 3.3.1).*

fait par krigeage comme présenté au paragraphe 3.3.1. En revanche, comment choisir une évaluation supplémentaire pour progresser dans la résolution de (3.19) ? Le choix de $\mathbf{x}_{n+1}$ est toujours motivé par l'amélioration de l'estimation du minimum, et peut donc se faire à l'aide des critères d'échantillonnage du chapitre 1. Par contre, le choix de $\mathbf{x}_{\mathrm{e},n+1}$ n'est motivé que par l'amélioration de la prédiction. Lehman et al. (2004) proposent, une fois choisi $\mathbf{x}_{n+1}$ d'utiliser un critère maximin classique en planification d'expériences. $\mathbf{x}_{\mathrm{e},n+1}$ est ainsi choisi de manière à maximiser la distance minimale entre $(\mathbf{x}_{n+1}, \mathbf{x}_{\mathrm{e},n+1})$ et les points d'évaluations précédents, c'est-à-dire

$$\mathbf{x}_{\mathrm{e},n+1} = \underset{\mathbf{x}_\mathrm{e} \in \mathbb{X}_\mathrm{e}}{\arg\max} \min_{i \in [\![1,n]\!]} |f(\mathbf{x}_{n+1}, \mathbf{x}_\mathrm{e}) - f(\mathbf{x}_i, \mathbf{x}_{\mathrm{e},i})|. \tag{3.20}$$

Il semble cependant plus légitime d'utiliser ici une technique de planification plus spécifique à la prédiction par krigeage, par exemple de minimiser l'intégrale de l'erreur quadratique moyenne (IMSE pour *Integrated Mean Square Error*)

$$IMSE(\mathbf{x}_\mathrm{e}) = \int_{\mathbf{x} \in \mathbb{X}} \mathbb{E}\left[\left(\hat{M}_{\mathbf{x}_\mathrm{e}}(\mathbf{x}) - M(\mathbf{x})\right)^2\right] d\mathbf{x}, \tag{3.21}$$

avec $\hat{M}_{\mathbf{x}_\mathrm{e}}(\mathbf{x})$ le prédicteur par krigeage de $M(\mathbf{x})$ à partir de $F(\mathbf{x}_1, \mathbf{x}_{e,1}), \ldots, F(\mathbf{x}_n, \mathbf{x}_{e,n})$ et de $F(\mathbf{x}_{n+1}, \mathbf{x}_e)$. Le terme à intégrer dans (3.21), à savoir la variance de l'erreur de prédiction, est donné par (3.8) et ne dépend pas du résultats de ces évaluations. Ce critère a été proposé par Sacks et al. (1989) avec l'objectif plus simple d'améliorer la prédiction de f.

Remarque 3.2. Sacks et al. (1989) proposent aussi de minimiser le maximum de l'erreur quadratique moyenne, mais, au dire des auteurs, la résolution de ce problème est significativement plus complexe, pour un gain en efficacité non prouvé.

3.5 Conclusions

Le cadre probabiliste utilisé dans ce mémoire se prête naturellement à une formalisation probabiliste du problème d'optimisation robuste. Pour adapter IAGO à ce contexte, il s'agit ensuite d'étendre la prédiction par krigeage à la prédiction de mesures de robustesse. Parmi celles envisagées (moyenne, variance ou fonction de répartition de la performance), c'est la performance moyenne M qui est la plus simple à prédire, du fait de la relation linéaire qui la lie au processus observé F. Le principe du cokrigeage permet ensuite de simuler conjointement M et F et d'estimer ainsi l'entropie conditionnelle des minimiseurs globaux de M.

Remarquons que la vision simple et pratique de la robustesse qu'est la minimisation de la performance moyenne peut être utilisée pour l'échantillonnage, mais le choix de la solution retenue peut se faire à l'aide de mesures plus complexes à estimer, mais plus adaptées au problème considéré.

CHAPITRE 4

COMPARAISON DES CRITÈRES

Résumé — Dans ce chapitre, nous nous intéressons à la rapidité de la convergence non-asymptotiques des algorithmes d'optimisation reposant sur les critères d'échantillonnage décrits au chapitre 1. Ces vitesses sont estimées sur des fonctions-tests classiques, mais aussi sur des trajectoires du processus gaussien sous-jacent aux critères d'échantillonnage. Ce protocole pour le test des algorithmes d'optimisation globale bayésienne constitue une contribution importante de nos travaux. Il nous permet non seulement de comparer les critères, mais aussi d'évaluer l'impact d'un mauvais choix de covariance sur leur efficacité. Nous en déduisons des recommandations pour le choix *a priori* de la covariance ainsi que pour la taille du plan d'expériences initial.

4.1 Objectifs et méthode

Nous ne chercherons pas ici à démontrer la convergence asymptotique des algorithmes reposant sur les critères d'échantillonnage vu à la section 1.5, car ces questions sont complexes et n'ont finalement qu'un intérêt limité pour le praticien (mentionnons toutefois les travaux récents de Bect et Vazquez (2007), sur la convergence de l'optimisation avec EI, lorsque la covariance est connue). Par contre, l'optimisation globale à l'aide d'un budget réduit d'évaluations requiert l'utilisation de méthodes qui exploitent efficacement les évaluations disponibles et ce dès la première d'entre elles. Les taux de convergences *non-asymptotiques* sont donc essentiels pour l'évaluation des mérites comparés de chacun des critères et ils peuvent de plus être évalués simplement.

Des trois critères d'échantillonnage retenus au chapitre 1, à savoir la maximisation de l'EI, la maximisation de la probabilité d'amélioration et la minimisation de l'ECM, c'est ce dernier qui devrait conduire aux convergence les plus rapides et ce pour trois raisons principales. Premièrement, l'EI et la probabilité d'amélioration cherchent à estimer le *minimum*, alors que l'ECM se concentre sur l'estimation des *minimiseurs*. Deuxièmement, l'EI et la probabilité d'amélioration cherchent à améliorer l'estimation du minimum de la fonction en échantillonnant là où son apparition est la plus probable. Il semble plus raisonnable de tenter de diminuer l'incertitude associée

à cette probabilité. Par exemple, affiner l'estimation dans un petit voisinage d'un minimum global potentiel (qui peut être simplement local) peut s'avérer très coûteux en évaluations, alors que quelques évaluations choisies grâce à l'ECM pourraient indiquer qu'une large part de l'espace de recherche n'a que peu de chance de contenir un minimiseur global (cette idée sera confirmée dans la section 4.3.1). Enfin, le calcul de l'ECM implique les propriétés statistiques des trajectoires de F, alors que le calcul de l'EI ou de la probabilité d'amélioration n'implique que la moyenne et la variance conditionnelle de f au point considéré. Cette utilisation plus complète de l'information disponible sur la fonction semble cruciale dans un contexte de budget limité d'évaluations. Les pages qui suivent ont pour objectif de vérifier la validité de ces intuitions, en observant en particulier les taux de convergence non-asymptotiques de l'optimisation à l'aide de chacun de ces critères.

Dans la section 4.2, ces taux de convergence sont évalués sur des fonctions-tests classiques et sur un problème d'identification de manière à démontrer l'intérêt des critères d'échantillonnage face aux méthodes classiques d'optimisation globale. Nous utilisons ensuite des trajectoires du modèle gaussien sous-jacent pour estimer les taux de convergence moyens dans la section 4.3.

Pour assurer le caractère équitable de la comparaison entre critères, le comportement de l'ECM, de l'EI ou de la probabilité d'amélioration doit, dans un premier temps, être étudié indépendamment de l'estimation des paramètres de la covariance et de la technique d'optimisation utilisée pour chacun des critères. C'est pourquoi, les expériences numériques qui suivent utilisent la même covariance de Matérn, choisie *a priori* pour chacun des trois critères. L'ensemble $\mathbb{G}$, sur lequel s'effectue le choix du prochain point d'évaluation par calcul extensif, est lui aussi commun à l'ECM, à l'EI et à la probabilité d'amélioration. Ainsi, nous sommes en mesure de répondre à la question suivante : étant donné un *a priori* commun sur la fonction à optimiser, quel critère offre la meilleure vitesse de convergence sur un ensemble fini de points d'évaluation potentiels (différentes mesures de la convergence seront utilisées, cf. la section 4.3) ? Ainsi, les critères d'échantillonnage peuvent-ils être comparés toutes choses égales par ailleurs. Nous ne parlons donc pas ici des algorithmes EGO et IAGO, mais bien des critères d'échantillonnage que sont la maximisation de l'EI ou la minimisation de l'ECM. Nous étudierons cependant dans les sections 4.3.4 et 4.3.3 l'influence du paramètrage des algorithmes (plan d'expériences initial, choix de la covariance) sur ces vitesses de convergence.

4.2 Comparaison à l'aide de fonctions-tests

Dès l'introduction, nous avons supposé que les critères d'échantillonnage probabilistes constituaient la meilleure réponse à un budget d'évaluations limité. Dans cette section, nous nous proposons de vérifier *a posteriori* cette hypothèse et de mettre en évidence l'intérêt de ces critères face aux autres approches de l'optimisation globale. Cette validation va s'effectuer par comparai-

son avec les résultats du simplexe de Nelder-Mead[1] et de l'algorithme DIRECT[2], qui serviront de référence face aux algorithmes reposant sur l'EI et l'ECM. Ces derniers auraient pu aussi être comparés aux approches dites *métaheuristiques* (on pourrait d'ailleurs considérer que IAGO ou EGO rentrent dans cette catégorie), par exemple les recuits simulés ou les algorithmes génétiques. Ces approches métaheuristiques sont intéressantes en ce qu'elles ne requièrent que très peu d'hypothèses sur la fonction à optimiser, pas même la continuité, mais il est généralement admis que leur efficacité est inférieure à celles des approches précédentes lorsque la fonction à optimiser est continue et régulière (ce qui est le cas de l'ensemble des problèmes industriels rencontrés au cours de ces travaux).

4.2.1 Exemples classiques en optimisation globale

Les quatre fonctions-tests proposées dans cette section ont été largement utilisées lors d'étude comparatives en optimisation globale, mais elles sont plus particulièrement tirées de Huang et al. (2006), où une comparaison entre l'EI, DIRECT et Nelder-Mead est menée. La dimension de l'espace des facteurs varie, suivant les fonctions, de trois à cinq, et toutes présentent des minima locaux (cf. le tableau 4.1).

Pour l'EI et l'ECM, les paramètres de la covariance sont estimés *a priori* à partir des résultats de 200 évaluations choisies aléatoirement dans l'espace de recherche à l'aide d'un plan LHS[3], et les deux critères sont optimisés sur un LHS contenant mille points rééchantillonnés après chaque évaluation.

Un point $\mathbf{x}_1$ de l'espace de recherche est choisi aléatoirement pour l'initialisation des algorithmes et l'optimisation est répétée 50 fois pour annuler la dépendance à ce choix (sauf pour DIRECT qui ne requiert pas de point d'initialisation et qui n'a donc été lancé qu'une seule fois par fonction). Après la i-ème évaluation de la fonction-test f, la convergence de chaque algorithme est mesurée par

$$G_i = \frac{f(\mathbf{x}_1) - m_i}{f(\mathbf{x}_1) - f^*},$$

avec $m_i = \min_{\mathbf{x} \in \{\mathbf{x}_1, \dots \mathbf{x}_i\}} f(\mathbf{x})$ l'estimateur courant du minimum global. G_i (version modifiée de la mesure de convergence utilisée dans Barton, 1984) décrit ainsi la réduction, après i itérations, de l'erreur initiale d'estimation du minimum global. En d'autres termes, c'est une mesure du chemin parcouru vers la résolution du problème. Le tableau 4.2 présente, pour chacun des algorithmes,

[1] Dans la version disponible dans la toolbox optimization de Matlab.

[2] Proposé initialement dans Jones et al. (1993), cet algorithme déterministe permet de réaliser un compromis efficace entre recherche globale et recherche locale. L'implémentation utilisée est celle de Finkel et Kelley (2004) qui peut être trouvée à l'adresse www4.ncsu.edu/ definkel/research/index.html.

[3] Ces évaluations réalisées *a priori* risquent de donner un avantage à l'EI et l'ECM, face à DIRECT où Nelder-Mead, en plaçant les critères dans une situation favorable. Elles permettent en effet, pour chacune des fonctions, de choisir une covariance adaptée. Nous verrons à la section 4.3.4 que cet avantage n'est pas significatif.

TAB. 4.1: *Fonctions-tests à minimiser (Huang et al., 2006).*

Nom	Description
Six-Hump Camel Back (Branin, 1972)	$d = 2$ $f(\mathbf{x}) = 4x_1^2 - 2.1x_1^4 + 1/3x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4$ $-1.6 \leq x_1 \leq 2.4, -0.8 \leq x_2 \leq 1.2$ $N_{\text{local}} = 6, N_{\text{global}} = 2$ $\mathbf{x}^* = [0.089, 0.713]^\mathsf{T}$ et $[0.089, 0.713]^\mathsf{T}$, $f^* = -1.03$
Tilted Branin (Huang et al., 2006)	$d = 2$ $f(\mathbf{x}) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos x_1 + 10 + 0.5x_1$ $-5 \leq x_1 \leq 10, 0 \leq x_2 \leq 15$ $N_{\text{local}} = 3, N_{\text{global}} = 1$ $\mathbf{x}^* = [-3.2, 12.3]^\mathsf{T}$ $f^* = -1.17$
Hartman 3 (Hartman, 1973)	$d = 3$ $f(\mathbf{x}) = -\sum_{i=1}^4 d_i \exp\left[-\sum_{j=1}^3 \alpha_{ij}(x_j - p_{ij})\right]$, avec $\alpha = \begin{pmatrix} 3 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3 & 10 & 30 \\ 0.1 & 10 & 35 \end{pmatrix}$ $\mathbf{d} = \begin{pmatrix} 1 \\ 1.2 \\ 3 \\ 3.2 \end{pmatrix}$ $\mathbf{p} = \begin{pmatrix} 0.3689 & 0.1170 & 0.2673 \\ 0.4699 & 0.4387 & 0.7470 \\ 0.1091 & 0.8732 & 0.5547 \\ 0.03815 & 0.5743 & 0.8828 \end{pmatrix}$ $0 \leq x_i \leq 1$, $i = 1, 2, 3$ $N_{\text{local}} > 1, N_{\text{global}} = 1$ $\mathbf{x}^* = (0.114, 0.556, 0.852)^\mathsf{T}$, $f^* = -3.86$
Ackley 5 (Ackley, 1987)	$d = 5$ $f(\mathbf{x}) = -20\exp[-0.2\sqrt{\frac{1}{d}\sum_{i=1}^d x_i^2}] - \exp[\frac{1}{d}\sum_{i=1}^d \cos(2\pi x_i)] + 20 + e$ $\forall i \in [\![1:3]\!] \; -32.8 \leq x_i \leq 32.8$ $N_{\text{local}} > 1, N_{\text{global}} = 1$ $\mathbf{x}^* = \mathbf{0}$, $f^* = 0$

le critère G_i pour $i = 20, 50$ et 100 évaluations. Notons que DIRECT, de par son principe même, n'admet pas le nombre d'évaluations comme un critère d'arrêt exact dans le sens où ce nombre sera toujours légèrement dépassé. Les résultats du tableau 4.2 concernant DIRECT sont donc obtenus avec un nombre d'évaluations légèrement supérieur à ce qui est annoncé (de 10 dans le pire des cas).

DIRECT converge très rapidement pour la fonction Six-Hump Camel Back, mais pour le reste les algorithmes reposant sur l'EI ou l'ECM convergent plus rapidement (avec une préférence pour l'ECM excepté pour la fonction Ackley 5). Il apparaît aussi que pour la fonction Ackley 5, qui possède de nombreux minima locaux et une grande variabilité, les méthodes DIRECT et Nelder Mead sont largement moins efficaces.

TAB. 4.2: *Comparaison entre les algorithmes reposant sur l'EI, l'ECM, Nelder-Mead et DIRECT. Pour chacune des fonctions du tableau 4.1, et pour chaque algorithme, la mesure de convergence G_i présentée est obtenue (sauf pour DIRECT) comme la moyenne des résultats obtenus sur les 50 répétitions de l'optimisation (L'écart-type de l'erreur d'estimation pour les chiffres présentés dans ce tableau est toujours plus faible que 0.01).*

	G_i pour Nelder-Meald			G_i pour DIRECT			G_i pour l'EI			G_i pour l'ECM		
$i=$	20	50	100	20	50	100	20	50	100	20	50	100
Six-Hump Camel Back	0.82	0.83	0.86	**1**	1	1	0.65	1	1	0.76	1	1
Tilted Branin	0.84	0.89	0.90	0.47	0.65	0.77	0.83	0.92	0.98	**0.89**	**0.95**	0.97
Hartman 3	0.61	0.80	0.85	0.31	0.8	0.98	0.64	0.98	1	**0.82**	0.99	1
Ackley 5	0.03	0.03	0.08	0	0.03	0.19	0.36	**0.73**	0.75	0.34	0.59	0.72

4.2.2 Un problème d'identification

Une application caractéristique des méthodes considérées dans ce mémoire est l'estimation des paramètres d'un modèle dont la simulation est coûteuse. Utilisons un problème de ce type pour prolonger la comparaison débutée à la section précédente. Considérons pour cela un exemple suffisamment simple pour que les détails de la mise en œuvre puissent être exposés simplement, mais qui va cependant mettre en défaut les approches classiques, et insistons sur le fait que la même méthodologie peut être appliquée à l'estimation des paramètres de modèles dont la simulation est véritablement complexe.

Soit donc un modèle dont le vecteur d'état $\mathbf{q} = [q_1, q_2]^\mathsf{T}$ correspond aux quantités de matière dans deux compartiments. Ces quantités sont gouvernées par les équations d'évolution

$$\left\{ \begin{array}{rcl} \dot{q}_1 & = & -(x_1+x_3)q_1 + x_2 q_2, \\ \dot{q}_2 & = & x_1 q_1 - x_2 q_2. \end{array} \right. \tag{4.1}$$

A l'instant $t = 0$, une injection de matière a lieu dans le compartiment 1 pour assurer $\mathbf{q}(0) = (1,0)^\mathsf{T}$. Des mesures $y(t_i)$ de la quantité de matière dans le compartiment 2 sont ensuite réalisées aux instants t_i, $i = 1, ..., 15$.

Pour les besoins de cet exemple illustratif à données simulées, un vecteur de mesures (non bruitées) est généré à l'aide du solveur ODE45 de Matlab pour le vecteur de paramètres $\mathbf{x}_0 = (0.6, 0.15, 0.35)^\mathsf{T}$. L'optimisation est ensuite menée sur $[0,1]^3$ et repose sur une fonction objectif quadratique

$$f(\mathbf{x}) = \sum_{i=1}^{15} (q_2(\mathbf{x}, t_i) - y(t_i))^2.$$

Cet exemple présente deux difficultés. Premièrement, comme suggéré par les lignes de niveaux (courbes noires) de la figure 4.1, les zones où f est très faible sont assez larges au regard de la taille de l'espace des facteurs. En d'autres termes, la fonction est très plate au voisinage de l'optimum global. La seconde difficulté provient du fait que les paramètres du modèle ne sont

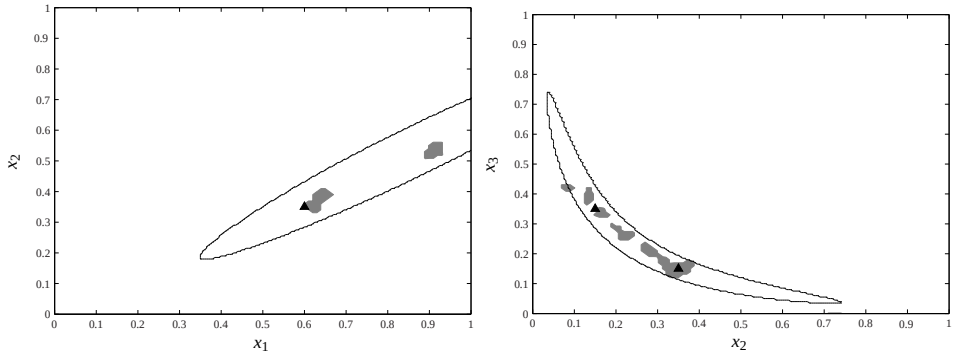


FIG. 4.1: *Coupes (en gris) du support de $\hat{P}_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ dans le plan défini par $x_3 = 0.15$ dans la* partie gauche *et par $x_1 = 0.6$ dans la* partie droite*, lorsque 40 évaluations ont été choisies par l'ECM. Les points pour lesquels $\hat{P}_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ est non nulle se trouvent dans la partie grisée. La courbe noire est la ligne de niveau à 0.03 de f. $\mathbf{x}_0$ et $\mathbf{x}_1$ sont matérialisés par des carrés.*

pas identifiables. En effet, les valeurs de x_2 et de x_3 peuvent être échangées sans modifier pour autant la sortie du système (Walter et Pronzato, 1997). La fonction objectif f possède en fait deux minimiseurs globaux, à savoir $\mathbf{x}_0$ et $\mathbf{x}_1 = (0.6, 0.35, 0.15)^\mathsf{T}$.

Pour cet exemple, les paramètres de la covariance utilisée pour l'EI et l'ECM sont réestimés après chaque évaluation par maximum de vraisemblance. L'initialisation des algorithmes (excepté DIRECT) se fait, de la même façon que pour les simulations de la section précédente, à l'aide d'un point choisi aléatoirement dans l'espace des facteurs. Ici encore, cent répétitions sont effectuées et les erreurs d'estimation ($||\arg\min_i f(\mathbf{x}_i) - \mathbf{x}^*||$ pour le minimiseur et m_n pour le minimum) après 40 et 80 évaluations, présentées dans le tableau 4.3 (en partie repris de Villemonteix et al., 2007b), sont moyennées. On constate que l'algorithme Nelder-Mead est nettement moins performant que ses trois concurrents. A titre d'exemple, il faut en moyenne 215 évaluations à Nelder-Mead pour atteindre la même précision que celle atteinte par IAGO après 40 évaluations. Sans parler du fait qu'en tant que méthode locale, Nelder-Mead n'identifie qu'un seul minimiseur global. L'algorithme DIRECT se comporte en revanche très bien sur cet exemple et offre, après 80 évaluations, des performances comparables à celles de l'algorithme reposant sur l'EI. C'est au final l'algorithme reposant sur l'ECM qui permet de converger le plus rapidement.

Pour mieux s'en convaincre et mettre en avant l'intérêt de la distribution conditionnelle des minimiseurs pour estimer l'avancement de la résolution du problème, considérons le support de $\hat{P}_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ (c'est-à-dire l'ensemble des points pour lesquels cette distribution est non nulle) à l'issue de 40 évaluations de f. Lorsque les évaluations sont choisies par l'ECM, le support de la distribution correspond aux lignes de niveaux à 0.03 de f (cf. la figure 4.1) et, de plus, les minimiseurs globaux $\mathbf{x}_0$ et $\mathbf{x}_1$ appartiennent au support estimé de la distribution conditionnelle des minimiseurs globaux. En revanche lorsque ces 40 évaluations sont choisies avec l'EI, aucun de ces deux minimiseurs globaux n'appartient au support de $\hat{P}_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ (cf. la figure 4.2).

TAB. 4.3: *Avancement de l'optimisation de la fonction de coût par Nelder-Mead, DIRECT, EGO et IAGO après 40 et 80 évaluations. Pour EGO et IAGO, les deux résultats proposés correspondent chacun à l'estimation des deux minimiseurs globaux (comme mode de $\hat{P}_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$). Pour le simplexe de Nelder-Mead, méthode locale, une unique valeur est présentée relativement au minimiseur global le plus favorable. Pour ces trois algorithmes, la recherche est répétée pour 100 points de départ choisis aléatoirement dans l'espace de recherche. Les chiffres présentés sont ainsi les résultats d'une moyenne sur ces 100 répétitions et sont accompagnés de l'écart-type de l'erreur d'estimation (entre parenthèses). Pour DIRECT, qui ne recquiert pas d'initialisation, aucune répétition n'est effectuée. L'erreur d'estimation du minimiseur est mesurée par $||\arg\min_i f(\mathbf{x}_i) - \mathbf{x}^*||$ et l'erreur d'estimation du minimum par m_n.*

Algorithme	Nelder-Mead	DIRECT	EGO	IAGO
Erreur d'estimation du minimiseur après 40 évaluations	0.44 (0.27)	0.27	0.14 (0.15) 0.14 (0.11)	**0.06** **(0.1)** **0.025** **(0.03)**
Erreur d'estimation du minimum après 40 évaluations	0.135 (0.23)	0.1	0.03 (0.1) 0.06 (0.02)	$\mathbf{10^{-3}}$ **(0.01)** $\mathbf{10^{-3}}$ **(0.04)**
Erreur d'estimation du minimiseur après 80 évaluations	0.35 (1.04)	0.09	0.08 (0.01) 0.09 (0.01)	**0.01** **(0.01)** **0.01** **(0.02)**
Erreur d'estimation du minimum après 80 évaluations	5.10^{-2} (0.18)	10^{-3}	10^{-4} (10^{-4}) 10^{-3} (10^{-4})	$\mathbf{10^{-5}}$ **(10^{-4})** $\mathbf{10^{-5}}$ **(10^{-4})**

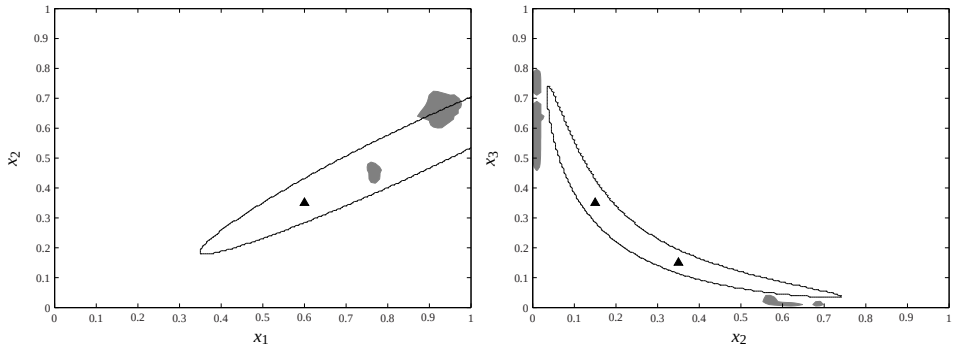


FIG. 4.2: *Coupes du support de la densité de $\hat{P}_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$ dans le plan défini par $x_3 = 0.15$ dans la partie* partie gauche *et par $x_1 = 0.6$ dans la* partie droite, *lorsque 40 évaluations ont été choisies par l'EI. Les conventions graphiques sont identiques à celles de la figure 4.1.*

4.3 Estimation des vitesses de convergence

La section précédente a mis en avant l'intérêt de l'ECM et de l'EI face à DIRECT, ainsi que la supériorité de l'ECM sur l'EI. Cependant la variabilité des résultats d'une fonction-test à l'autre est souvent significative et rarement explicable. Il est ainsi difficile de déduire de ces comparaisons la méthode à utiliser *a priori* pour un problème donné.

Il serait plus utile de disposer de vitesses de convergence pour chacune des méthodes sous des hypothèses raisonnables sur la fonction à optimiser (continuité, constante de Lipschitz ou autres). Dans le contexte qui nous préoccupe, ces vitesses devraient être non-asymptotiques, mais nous n'avons pas connaissance de résultats de ce type dans la littérature.

En revanche, étant donné notre *a priori* gaussien sur f, il semble légitime de considérer les taux de convergence des algorithmes lorsque cette hypothèse est satisfaite, c'est-à-dire lorsque les algorithmes sont confrontés à des réalisations d'un processus gaussien. Ainsi, même si l'obtention d'expressions analytiques paraît impossible, il est relativement aisé d'estimer des taux de convergence non-asymptotiques moyens pour une covariance donnée. Nous allons pour cela simuler des trajectoires de F et optimiser chacune à l'aide de l'EI, de l'ECM, de la probabilité d'amélioration et de DIRECT.

Les taux de convergence relevant de cette expérimentation numérique dépendent de la covariance et il est impossible d'assurer qu'un classement des méthodes valable pour une covariance sera valable pour toutes les autres. Nous pensons cependant que les taux de convergence sur deux processus Gaussiens, l'un avec des trajectoires lisses et l'autre avec des trajectoires irrégulières, offrent un bon aperçu de l'intérêt de chacune des méthodes. Les trajectoires de ces processus présentent en effet une grande variété de comportements. La figure 4.3 présente des exemples de simulations d'un processus Gaussien de covariance de Matérn pour deux régularités différentes.

4.3.1 Vitesses de convergence à covariance connue

Deux jeux de 1000 simulations sont générés sur une grille régulière de 1500 points dans $[0,1]^2$. Le premier est créé avec une covariance de Matérn régulière ($\nu = 5$), et le second avec une covariance de Matérn peu régulière ($\nu = 1.5$). Trente évaluations sont ensuite choisies pour chacune des simulations à l'aide de l'EI, de l'ECM ou de DIRECT (une exploration aléatoire est aussi réalisée à l'aide d'une loi uniforme pour fournir un point de comparaison).

Pour comparer les critères, les progrès effectués sur chacune des trajectoires, après chaque évaluation et pour chaque méthode, doivent être quantifiés. Il faut alors veiller à ne pas biaiser la comparaison par le choix du critère de progrès. L'entropie de $\hat{P}_{\mathbf{X}^*}(\cdot|\mathcal{F}_n)$, par exemple, est la fonction de perte qui fonde l'ECM (cf. la section 1.4). L'erreur d'estimation du minimum par $m_n = \min_i f(\mathbf{x}_i)$ est la fonction de perte sur laquelle repose l'EI. Ces deux critères de progrès doivent donc être complétés par des critères plus neutres.

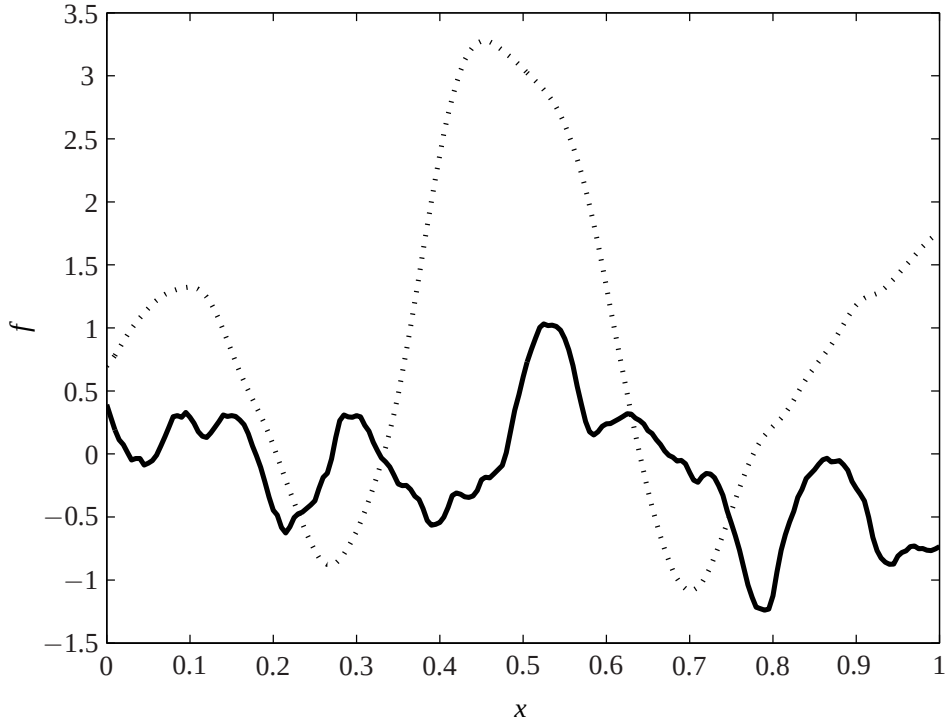


FIG. 4.3: *Exemples de simulations d'un processus Gaussien indexé sur* $[0,1]$ *avec* $\rho = 0.3$, $\sigma = 1.5$ *et* $\nu = 5$ *(pour le trait en pointillés) ou* $\nu = 1.5$ *(pour le trait continu).*

Considérons ce qui se passe en pratique pour un budget de n évaluations. Une fois $n-1$ évaluations réalisées, une estimation d'un minimiseur global est calculée, par exemple un maximiseur de la distribution conditionnelle des minimiseurs $P_{F^*}(\cdot|\mathcal{F}_n)$ et une évaluation dite *de validation* est réalisée en ce point. La meilleure solution obtenue parmi les n évaluations réalisées est ensuite conservée. Le minimum global est ainsi estimé par

$$\tilde{m}_n = \min\{m_{n-1}, f(\arg\max_{\mathbf{x}\in\mathbb{G}} P_{F^*}(\mathbf{x}|\mathcal{F}_{n-1}))\}$$

et un minimiseur global par $\tilde{\mathbf{x}}_n^*$, la position de $\tilde{m}_n$. Finalement, une mesure de convergence légitime vis-à-vis de l'utilisation pratique des critères d'échantillonnage semble donc être l'erreur d'estimation du minimum par $\tilde{m}_n$ (et dans une moindre mesure l'erreur d'estimation $||\tilde{\mathbf{x}}_n^* - \mathbf{x}^*||$ d'un minimiseur global $\mathbf{x}^*$ par $\tilde{\mathbf{x}}_n^*$). Dans la suite, nous utilisons les quatres critères de progrès proposés ici à savoir, l'entropie de la distribution conditionnelle des minimiseurs, l'erreur d'estimation du minimum commise par m_n et $\tilde{m}_n$ et enfin l'erreur d'estimation du minimiseur commise par $\tilde{\mathbf{x}}_n^*$.

L'EI, l'ECM et la probabilité d'amélioration sont tout trois optimisés par calcul exhaustif sur le support des simulations et utilisent la même covariance, identique à celle utilisée pour générer les simulations. Le seuil T utilisé pour la probabilité d'amélioration varie d'une itération à l'autre suivant la méthode proposée par Jones (2001), où

$$T = \min_{\mathbf{x}\in\mathbb{X}} \hat{f}(\mathbf{x}) - \alpha(f_{\max} - f_{\min}),$$

avec $\alpha \geq 0$ évoluant de manière cyclique dans $\{0, 0.2, 0.4, 0.6, 0.8\}$. Nous avons aussi testé une suite décroissante de α telle que proposée dans Calvin (2001), où une preuve de convergence est fournie dans le cas où F est un processus de Wiener. Les trois approches sont initialisées par un unique point choisi aléatoirement dans l'espace de recherche.

La figure 4.4 présente, pour les quatres mesures de convergence mentionnées précédemment, les taux de convergence moyens lorsque chacune des méthodes est confrontée à des trajectoires d'un processus Gaussien de covariance de Matèrn de paramètres $\nu = 1.5$, $\rho = 0.3$, $\sigma^2 = 1.5$.

De manière prévisible, l'utilisation de l'ECM est significativement plus efficace que celle des autres approches en termes d'entropie de la densité conditionnelle des minimiseurs (cf. la figure 4.4(a)). L'incertitude sur la position des minimiseurs globaux diminue donc plus rapidement si les points sont choisis à l'aide de l'ECM. Cette supériorité était garantie pour la première évaluation, puisque la minimisation de l'ECM est un critère optimal à un coup pour cette fonction de perte, mais elle n'allait pas de soi à l'issue de plusieurs évaluations. De manière similaire, si la convergence est mesurée par l'erreur d'estimation du minimum (avec m_n comme estimateur), EI doit être plus efficace si l'on ne considère que la première évaluation, puisque l'erreur d'estimation du minimum est la fonction de perte sur laquelle repose l'EI. On observe cependant que cette supériorité n'est pas significative au regard des performances de l'ECM ou de la probabilité d'amélioration. En termes d'erreur commise par $\tilde{m}_n$, les trois algorithmes sont équivalents (cf. la figure 4.4(c)). L'ECM, en moyenne, permet une meilleure estimation du minimiseur par $\tilde{\mathbf{x}}_n^*$ (cf. la figure 4.4(d)). Si l'on considère le comportement de l'ECM et de l'EI sur la trajectoire particulière associée aux résultats de la figure 4.5, les inconvénients liés à l'utilisation de l'EI sont clairement mis en avant. En effet, comme pressenti au début de cette section, l'EI reste bloqué sur un minimum local car, avec l'erreur d'estimation du minimum comme fonction de risque, il est plus avantageux d'assurer une petite amélioration au voisinage d'un point déjà détecté comme intéressant que de vérifier que ce point est effectivement un minimum global. Il n'en va pas de même avec l'ECM. Cependant, même si le minimum identifié par l'EI est local, sa valeur est proche de celle du global et en moyenne sa performance est équivalente à celle de l'EI en ce qui concerne l'erreur d'estimation du minimum.

Lorsque la régularité ν de la covariance augmente (cf. la figure 4.6), les trajectoires du processus considéré deviennent plus régulières et les minima locaux se font plus rares. La convergence des algorithmes est alors plus rapide, mais les conclusions sont les mêmes que lorsque la régularité est faible.

Notons finalement que les performances des deux méthodes de choix de seuil pour la probabilité d'amélioration semblent comparables (cf. la figure 4.4(a)).

Remarque 4.1. On constate sur les figures 4.4(c) et 4.6(c), que l'erreur d'estimation du minimum par $\tilde{m}_n$ est *significativement* plus faible que l'erreur d'estimation du minimum par m_n.

Remarque 4.2. La convergence de DIRECT (présentée uniquement sur la figure 4.4(b)) est nettement plus lente que celle des autres algorithmes.

Remarque 4.3. Il apparaît sur la figure 4.4(d) que dès la première itération des méthodes (la seconde évaluation après celle effectuée en un point choisi aléatoirement), l'erreur d'estimation du minimiseur est en moyenne plus faible pour l'ECM que pour les autres approches. On constate aussi, de manière plus surprenante, que l'EI et la probabilité d'amélioration se comportent, pour les premières évaluations, moins bien qu'une recherche aléatoire. Ceci s'explique par le fait qu'à la suite de la première évaluation choisie aléatoirement, l'EI et la probabilité d'amélioration sont maximales sur les bords du domaine de recherche. Or il semble qu'une évaluation au centre du domaine soit optimale pour minimiser l'erreur d'estimation du minimiseur à l'issue de la première évaluation.

Les simulations effectuées dans cette section semblent donc indiquer que, si la covariance du processus considéré est connue, les efficacités des trois critères d'échantillonnage sont comparables, avec néanmoins un avantage pour l'ECM qui estime mieux les minimiseurs globaux. Pour simplifier la présentation, nous ne retiendrons, pour la suite des comparaisons, que l'EI et l'ECM (la probabilité d'amélioration s'avérant, au sens de toutes les mesures de progrès, moins efficace que l'EI pour toutes les covariances testées).

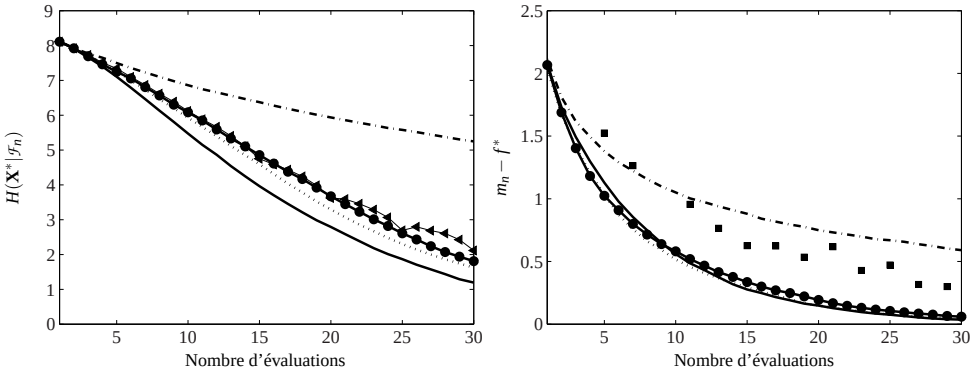
Considérons maintenant ce qui se passe en pratique lorsque les paramètres de la covariance sont mal choisis.

4.3.2 Robustesse par rapport à une erreur d'estimation de la covariance

Étudions maintenant les performances de l'EI et de l'ECM quand ils sont confrontés à des trajectoires d'un processus gaussien dont la covariance diffère de celle choisie pour l'optimisation. Pour l'EI (cf. la figure 4.7) et pour l'ECM (cf. la figure 4.8), quatre scénarios d'erreurs sont considérés. Nous simulons en effet séparément une sous-estimation de la régularité ($\nu = 0.5$ face à $\nu = 5$) puis de la portée ($\rho = 0.05$ face à $\rho = 0.3$), une surestimation de la régularité ($\nu = 100$ face à $\nu = 5$) et enfin une surestimation de la portée ($\rho = 0.6$ face à $\rho = 0.3$). Pour chacun de ces scénarios, les références sont les vitesses de convergence obtenues pour l'EI et l'ECM présentées sur la figure 4.6 et obtenues pour $\nu = 5$, $\rho = 0.3$ et $\sigma^2 = 1.5$.

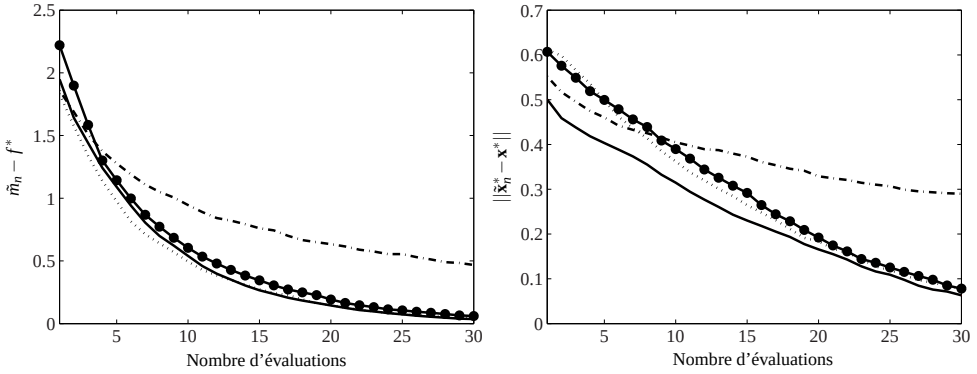
A l'issue des simulations, il apparaît très nettement qu'une surestimation des paramètres de régularité ou de portée est bien moins dommageable qu'une sous-estimation[4]. Ce constat a des conséquences pratiques importantes. En effet, il va suffire d'une connaissance réduite du problème considéré pour être en mesure de choisir *a priori* (par surestimation) des paramètres de covariance assurant un comportement satisfaisant.

[4]Ceci est vrai pour d'autres paramètres de référence (par exemple $\nu = 1.5$ et $\rho = 0.1$), mais les résultats ne sont pas présentés ici pour ne pas ajouter à une liste de courbes déjà longue.



(a) Évolution de l'entropie de la distribution conditionnelle des minimiseurs globaux

(b) Évolution de l'erreur d'estimation par m_n du minimum



(c) Évolution de l'erreur d'estimation par $\tilde{m}_n$ du minimum

(d) Évolution de l'erreur d'estimation pour le minimiseur

FIG. 4.4: *Évolution des quatre mesures de progrès des méthodes d'optimisation confrontées à des trajectoires d'un processus Gaussien peu régulier (covariance de Matèrn* $\nu = 1.5$, $\rho = 0.3$, $\sigma^2 = 1.5$*). Les traits pleins sont relatifs à l'ECM, les traits pointillés à l'EI, les traits mixtes à la recherche aléatoire (avec une loi uniforme) et les disquesx à la probabilité d'amélioration (version de Calvin, 2001). Sur la figure (a), on trouve aussi (triangles) le comportement de la probabilité d'amélioration lorsque le seuil est choisi suivant les préconisations de Jones (2001). La vitesse de convergence de DIRECT en termes d'erreur moyenne d'estimation du minimum est présentée sur la figure (b) (carrés). Notons que cette dernière n'est pas disponible pour chaque nombre d'évaluations compte-tenu du fonctionnement même de DIRECT.*

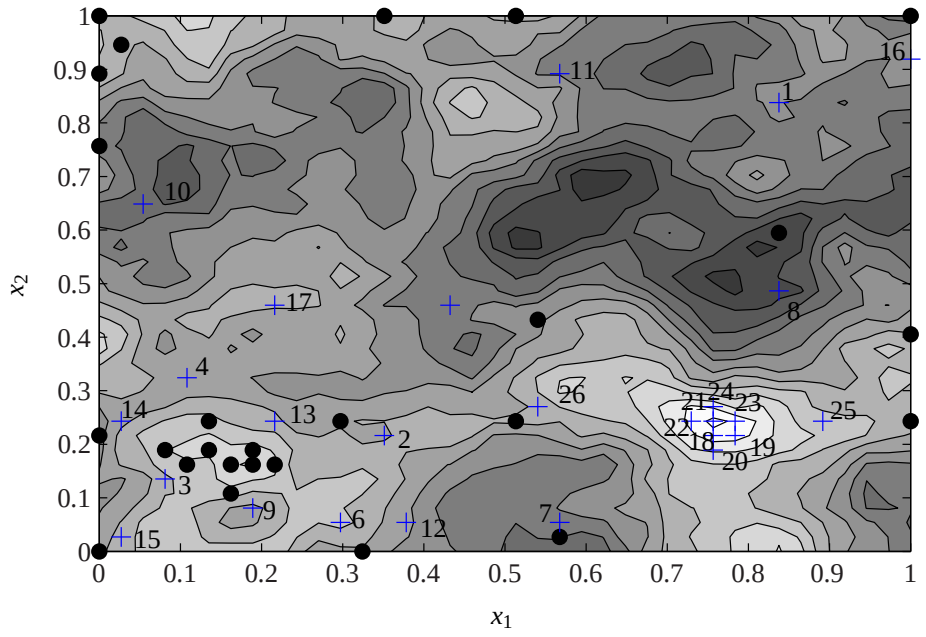


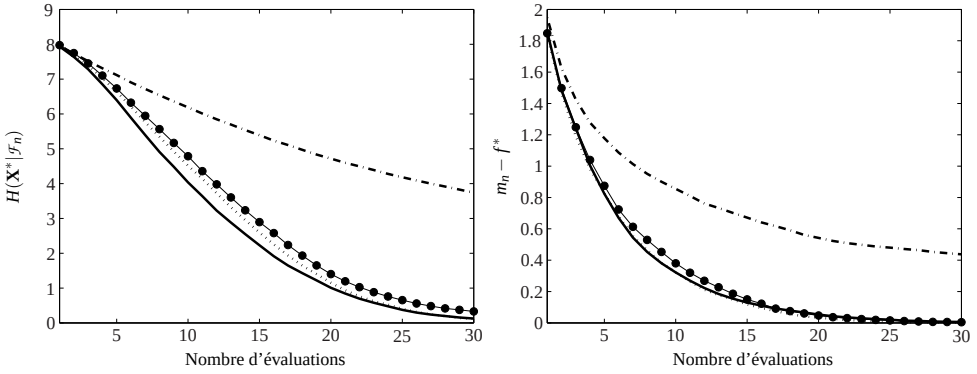
FIG. 4.5: *Minimisation d'une des simulations du processus Gaussien utilisé pour obtenir les taux de convergence de la figure 4.4. Les points décrivent les points d'évaluation choisis par maximisation de l'EI. Les croix indiquent les points choisis par minimisation de l'ECM. L'ordre dans lequel les évaluations sont réalisées est aussi indiqué pour l'ECM.*

Remarque 4.4. Nous n'avons pas considéré de scénarios où la variance était mal estimée. En effet, la variance est souvent le paramètre sur lequel l'information *a priori* est la plus détaillée, et c'est aussi le plus simple à estimer par maximum de vraisemblance, comme en témoignent les résultats du tableau 2.1.

Nous allons maintenant utiliser les mêmes mesures de progrès empiriques pour répondre aux questions posées par l'utilisation pratique de IAGO et de EGO. En particulier, quelle taille donner au plan d'expérience initial ? Comment influe le rééchantillonnage (proposé à la section 2.3.2 pour le choix de $\mathbb{G}$) sur la convergence de l'optimisation avec l'ECM ?

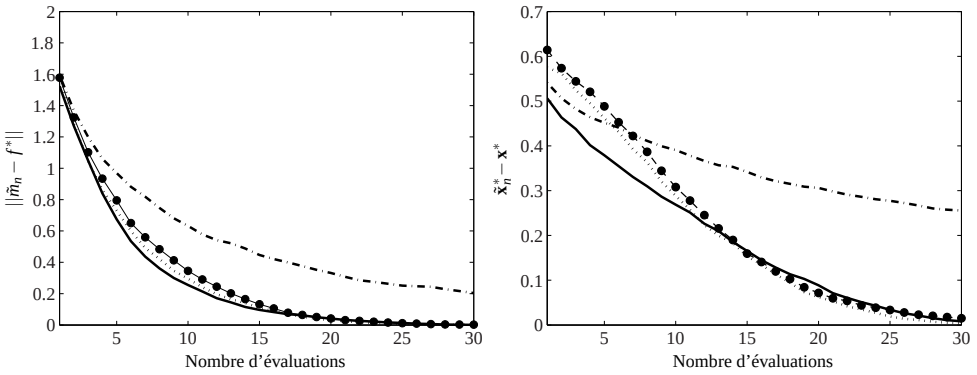
4.3.3 Influence de la taille du plan d'expériences initial

Les mesures de progrès empiriques utilisées pour comparer les critères d'échantillonnage peuvent aussi servir à la mise au point des algorithmes. Intéressons-nous par exemple à l'influence de la taille du plan d'expériences initial sur la vitesse de convergence des méthodes d'optimisation reposant sur l'EI ou l'ECM. La figure 4.9 présente trois courbes de convergence moyenne de l'optimisation avec l'EI, lorsque ce dernier est confronté au processus Gaussien déjà utilisé pour la figure 4.4. Chacune de ces courbes présente, après chaque évaluation, l'erreur d'estimation moyenne du minimum, mais elles se différencient par la taille du plan d'expériences LHS



(a) Évolution de l'entropie de la distribution conditionnelle des minimiseurs globaux

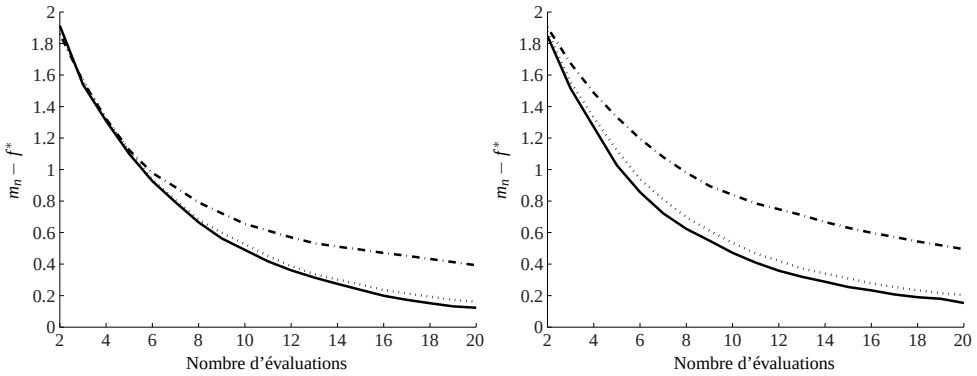
(b) Évolution de l'erreur d'estimation par m_n du minimum



(c) Évolution de l'erreur d'estimation par $\tilde{m}_n$ du minimum

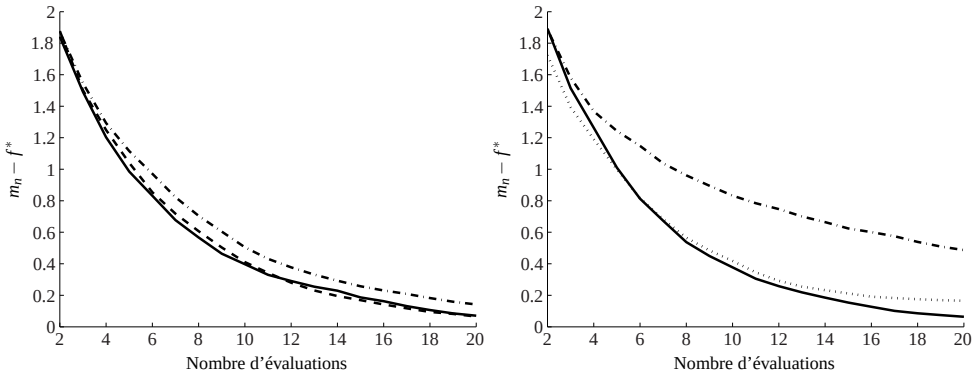
(d) Évolution de l'erreur d'estimation du minimiseur

FIG. 4.6: *Évolution de quatre mesures de progrès des méthodes d'optimisation confrontées à des trajectoires d'un processus Gaussien régulier (covariance de Matèrn* $\nu = 5$, $\rho = 0.3$, $\sigma^2 = 1.5$*). Les conventions graphiques sont identiques à celles de la figure 4.4.*



(a) Erreur sur la régularité (b) Erreur sur la portée

FIG. 4.7: *Évolution de l'erreur d'estimation du minimum ($m_n - f^*$) lorsque l'EI est confronté à des trajectoires d'un processus Gaussien (covariance de Matèrn $\nu = 5$, $\rho = 0.3$, $\sigma^2 = 1.5$) de covariance différente de celle utilisée pour calculer le critère d'échantillonnage. Sur la figure de gauche, c'est la régularité qui est mal estimée, alors que sur la figure de droite, c'est la portée. Les traits pleins représentent les vitesses de convergence de référence, c'est-à-dire lorsque les paramètres utilisés par l'EI sont ceux qui ont servi à simuler les trajectoires. Les traits en pointillés correspondent à une surestimation des paramètres ($\nu = 100$, ou $\rho = 0.6$) et les traits mixtes à une sous-estimation ($\nu = 0.5$, ou $\rho = 0.05$). Pour chaque courbe (excepté pour les courbes de référence en traits pleins), un seul des paramètres utilisés par l'EI est différent de celui utilisé pour les simulations.*



(a) Erreur sur la régularité (b) Erreur sur la portée

FIG. 4.8: *Évolution de l'erreur d'estimation du minimum ($m_n - f^*$) lorsque l'ECM est confrontée à des trajectoires d'un processus Gaussien (covariance de Matèrn $\nu = 5$, $\rho = 0.3$, $\sigma^2 = 1.5$) de covariance différente de celle utilisée pour calculer le critère d'échantillonnage. La figure se lit comme la figure 4.7, et a été obtenue de la même manière, à la différence que l'ECM a été utilisée à la place de l'EI.*

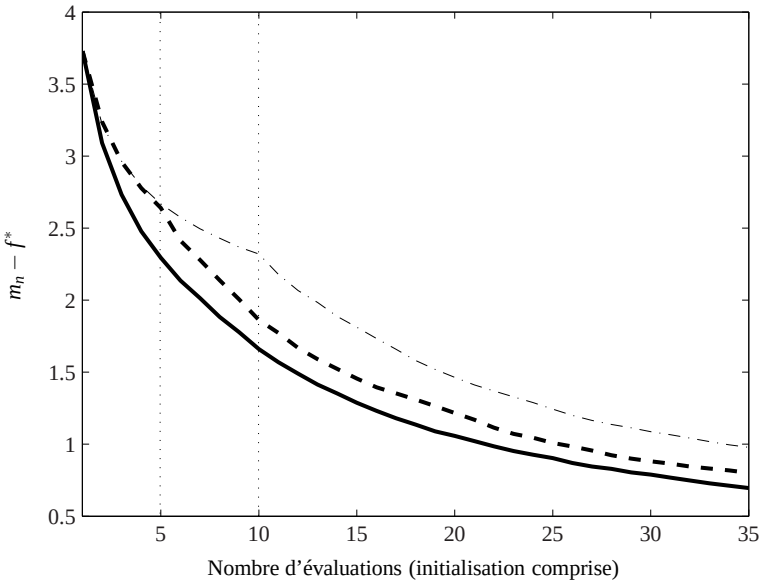


FIG. 4.9: *Influence de la taille du plan d'expériences initial sur le progrès (quantifié par $m_n - f^*$) de l'optimisation avec l'EI confrontée à des trajectoires d'un processus Gaussien peu régulier (covariance de Matèrn $\nu = 1.5$, $\rho = 0.3$, $\sigma^2 = 1.5$). Pour chacune des 1000 trajectoires simulées pour obtenir ces courbes, trois plans d'expériences LHS ont été générés. Le trait plein présente la vitesse de convergence lorsque le plan d'initialisation se limite à un point. Le trait en pointillés correspond à un plan LHS de cinq points et le trait mixte à un plan LHS de dix points.*

utilisé pour initialiser l'optimisation. Il apparaît que la convergence moyenne la plus rapide est obtenue lorsque ce plan est limité à un point. Ceci nous conduit à recommander que IAGO (et EGO compte tenu de la similarité entre l'EI et l'ECM) soient initialisés en un unique point. Cette conclusion doit néanmoins être nuancée puisqu'elle présuppose que la covariance utilisée par les algorithmes représente bien le phénomène observé. Sous cette hypothèse, le caractère global de la recherche effectuée par chacun des critères est suffisant, et il n'est pas nécessaire de le renforcer par un plan initial. En revanche, si la covariance doit être estimée, ou si la confiance dans un choix effectué *a priori* est limitée, un plan d'expériences initial demeure indispensable.

Remarque 4.5. Si la surestimation des paramètres de la covariance de Matérn est trop importante, on se retrouve confronté à des problèmes numériques dans la résolution du système (1.7).

4.3.4 Comparaison des critères dans des conditions d'utilisation réalistes

Jusqu'à présent, les critères d'échantillonnage ont été comparés sur la base de leur efficacité dans des conditions d'utilisation idéales. En pratique, le problème de leur optimisation est bien souvent résolu de manière approchée, et il importe de connaître l'impact de cette approximation sur les mérites comparés de l'EI et de l'ECM.

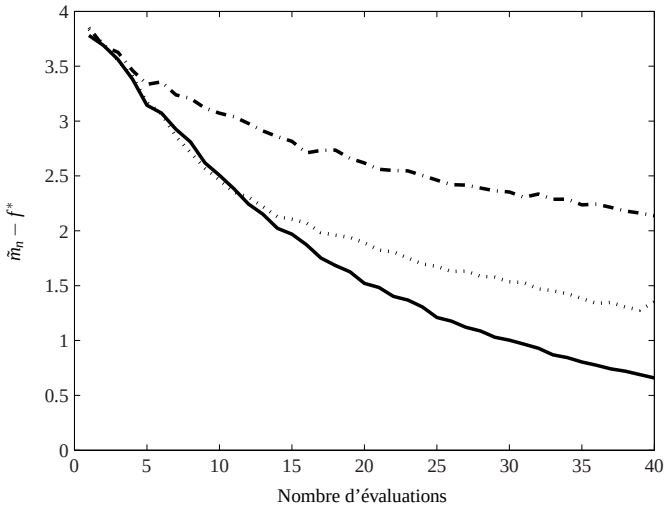


FIG. 4.10: *Impact du rééchantillonage de $\mathbb{G}$ sur les mérites comparés de l'EI et de l'ECM. L'EI et l'ECM sont confrontés à des trajectoires d'un processus gaussien, indexé sur $[0,1]^6$, de covariance de Matèrn ($\nu = 1.5$, $\rho = 0.5$, $\sigma^2 = 1.5$). Les progrès de l'optimisation sont quantifiés par $\tilde{m}_n - f^*$ et les résultats moyens obtenus sur 1000 trajectoires sont matérialisés par un trait plein pour l'ECM, un trait pointillé pour l'EI et un trait mixte pour la recherche aléatoire (avec une loi uniforme). L'EI est maximisé par évaluation exhaustive sur un plan LHS de taille 45000 alors que l'ECM est minimisée par évaluation exhaustive sur un ensemble de 400 points rééchantillonnés aprés chaque évaluation (avec $p_{\mathbf{X}^*_{\mathbb{X}}}(\mathbf{x}|\mathcal{F}_n)$ comme* a priori, *cf. la section 2.3.2).*

Dans la section 2.3.2, nous avons décrit en quoi le rééchantillonnage de $\mathbb{G}$ se présente comme une solution prometteuse pour l'optimisation de l'ECM mais aussi pour son approximation. Dans cette section, nous allons comparer les performances de l'ECM lorsque $\mathbb{G}$ contient 400 points rééchantillonnés après chaque évaluation avec celles de l'EI optimisé sur un LHS à 45000 éléments eux aussi rééchantillonnés après chaque évaluation.

Cette comparaison est effectuée à l'aide d'un processus Gaussien indexé sur $[0,1]^6$, et il apparaît que l'ECM fonctionne dans ces conditions en moyenne mieux que l'EI (cf. la figure 4.10). Ces résultats préliminaires doivent être relativisés, puisque la méthode d'optimisation de l'EI est assez naïve. Ils indiquent cependant l'efficacité de la procédure d'échantillonnage qui permet de diminuer considérablement le cardinal de $\mathbb{G}$. Dans de futurs travaux, une comparaison équitable entre l'ECM, muni du rééchantillonnage, et l'EI ne pourra se faire qu'à la condition de mettre en place un méthode d'optimisation adaptée à l'EI (cf. par exemple Bates et Pronzato, 2001).

4.4 Influence du bruit sur les vitesses de convergence

4.4.1 Influence d'un bruit sur le résultat des évaluations

La prise en compte d'un bruit additif sur les résultats des évaluations pose les questions suivantes. Comment évoluent les taux de convergence observés dans les sections précédentes ? L'ECM reste-t-il toujours préférable à l'EI ou à l'une de ses modifications présentées à la section 2.4.1 ?

Pour y répondre, utilisons les deux jeux de simulations déjà utilisés à la section 4.3.1, mais ajoutons à chaque simulation un bruit blanc gaussien de variance 0.5. Soixante évaluations sont alors réalisées à l'aide de l'EI, de l'EIm, de l'AEI et de l'ECM. Après chaque évaluation, la performance est mesurée par la moyenne (sur toutes les simulations) de la distance

$$|f^* - \min_i \hat{f}(x_i)|$$

entre le minimum global f^* de la simulation et le minimum des valeurs prédites aux points d'observation. Nous calculons aussi la moyenne de l'entropie de la distribution $\hat{P}_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n^{\text{obs}})$ des minimiseurs globaux conditionnellement aux résultats bruités des évaluations.

Pour ces deux mesures de progrès et pour les deux régularités testées (cf. les figures 4.11 et 4.12), l'ECM fournit les meilleurs résultats et ce dès les premières évaluations. On peut d'ailleurs constater que cette conclusion ne semble pas dépendre de la régularité.

L'EI offre de meilleurs résultats qu'une recherche aléatoire (avec une loi uniforme) en termes d'erreur moyenne d'estimation du minimum, mais ne présente aucun intérêt en termes de diminution de l'entropie de la distribution des minimiseurs. Les deux modifications d'EI n'apportent pas d'amélioration significative. En effet, EIm est bien moins efficace qu'une recherche aléatoire, et les progrès offerts par l'AEI ne suffisent pas à concurrencer l'ECM.

4.4.2 Performances des versions robustes

En présence d'un bruit sur les facteurs, nous avons vu au chapitre 4, qu'il était avantageux d'optimiser la performance moyenne, qui peut être prédite directement par krigeage. La situation est alors similaire au cas où les résultats des évaluations sont bruitées, à savoir que la prédiction de la fonction à optimiser n'interpole plus les données et que la variance de l'erreur de prédiction ne s'annule plus aux points d'évaluations. L'adaptation de l'ECM est alors immédiate, mais celle d'EI requiert le choix d'un estimateur du minimum tout comme dans le cas d'un bruit sur les résultats des évaluations (cf. la section 2.4.1). En effet, le minimum des observations ne converge pas vers le minimum de la performance moyenne. On se retrouve donc dans des conditions très similaires à celles déjà présentes dans la section précédente, et les taux de convergence empiriques devraient conduire aux mêmes conclusions (ce qui est semble être confirmé par des résultats préliminaires non présentés ici).

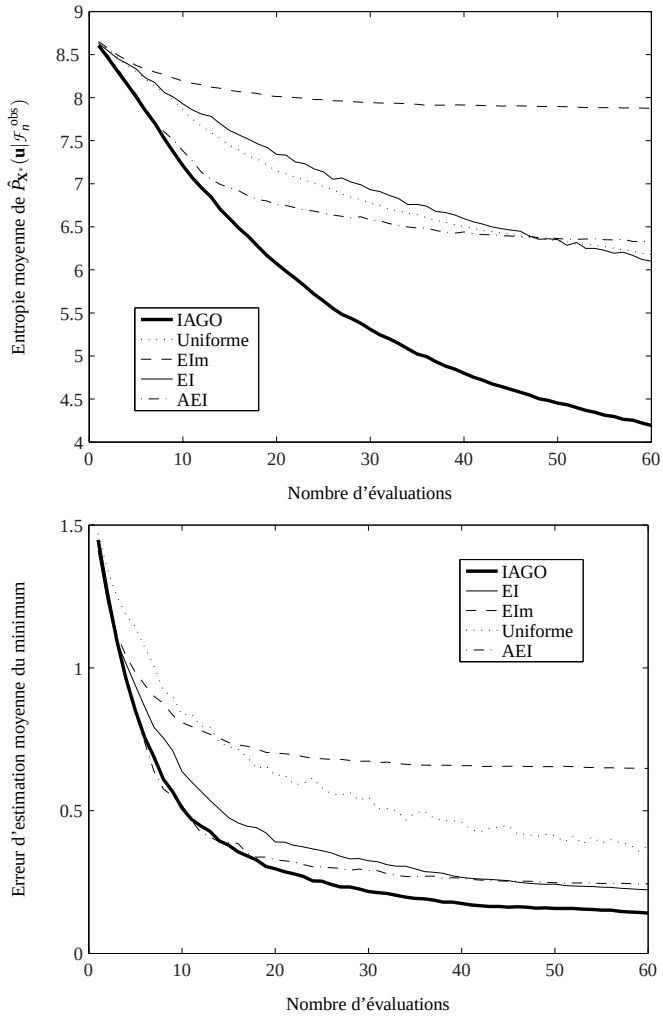


FIG. 4.11: *Performances en présence de bruit de l'optimisation reposant sur l'EI, l'EIm, l'AEI, l'ECM et une recherche aléatoire avec une loi uniforme, mesurée par l'entropie de* $\hat{P}_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n^{\text{obs}})$ *(*partie supérieure*), et par* $|f^* - \min_i \hat{f}(x_i)|$ *(*partie inférieure*). Les trajectoires du processus gaussien simulé sont plus régulières que celles du processus utilisé pour la figure 4.12 (les paramètres de la covariance de Matèrn sont* $\nu = 5$, $\rho = 0.3$ *et* $\sigma^2 = 1.5$*). La variance du bruit gaussien qui corrompt les résultats des évaluations est de 0.5.*

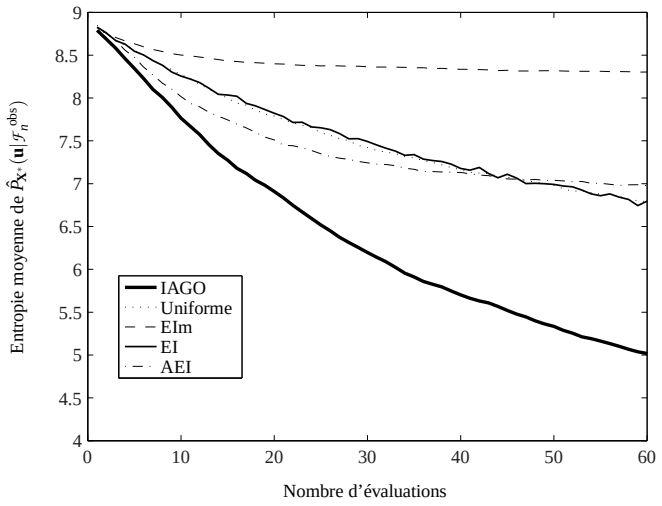


FIG. 4.12: *Performance en présence de bruit (gaussien, de variance 0.5), mesurée par l'entropie de $\hat{P}_{\mathbf{X}^*}(\mathbf{u}|\mathcal{F}_n^{\text{obs}})$, de l'optimisation reposant sur l'EI, l'EIm, l'AEI, l'ECM et une recherche aléatoire avec une loi uniforme. Les trajectoires sont simulées à l'aide d'une covariance de Matérn de paramètres $\nu = 1.5$, $\rho = 0.3$ et $\sigma^2 = 1.5$. Notons, après comparaison avec la figure 4.11, que la régularité a peu d'influence.*

4.5 Discussion

La conclusion principale de ce chapitre est que les taux de convergence empiriques obtenus sur des trajectoires de processus gaussiens apparaissent comme un outil important pour la comparaison des méthodes d'optimisation globale bayésiennes, mais aussi pour leur mise au point. Ainsi, nous avons pu mettre en avant l'intérêt de l'ECM face à l'EI (en particulier en présence de bruit sur les résultats des évaluations) et constater qu'il semble possible d'utiliser EGO ou IAGO sans estimer les paramètres de la covariance de Matèrn, du moment que l'on est en mesure de les surestimer.

Ces résultats doivent être nuancés, et ce pour deux raisons. Premièrement, même si les taux de convergences empiriques des méthodes sur des trajectoires du modèle sous-jacent nous paraissent plus pertinents pour la comparaison que l'optimisation d'un ensemble de fonctions-tests, les résultats dépendent assez fortement de la covariance considérée. Ainsi, bien que nous nous soyons efforcés de choisir un ensemble représentatif de covariances, les résultats présentés ne valent que pour cet ensemble. Deuxièmement, nous n'avons pas pris en compte dans la comparaison la complexité calculatoire de chacun des critères d'échantillonnage.

L'EI est en effet plus simple à calculer que l'ECM, puisque son évaluation ne nécessite que la moyenne et la variance de la prédiction par krigeage au point considéré, alors que le calcul de l'ECM est en O(N), avec N la taille d'une approximation discrète de $\mathbb{X}$ utilisée pour l'estimation de la distribution conditionnelle des minimiseurs. En pratique, le choix par l'ECM d'un nouveau

point d'évaluation pour les trajectoires de la section 4.3.4 prend environ 40s sur un serveur AMD Opteron 285 (le point est choisi par évaluation exhaustive de l'entropie conditionnelle sur un ensemble de 1500 points candidats rééchantillonné après chaque évaluation). En comparaison, le choix d'un point avec l'EI ne nécessite que quelques secondes dans les mêmes conditions.

Pour passer outre cette limitation et élargir le champ d'application potentiel pour IAGO, nous avons tenté de diminuer la complexité calculatoire de IAGO par l'utilisation d'autres approximations de la distribution conditionnelle des minimiseurs (la manipulation des simulations conditionnelles étant responsable de la plus grande partie des calculs nécessaires). Nous avons par exemple proposé (Villemonteix et al., 2007a) d'estimer les dérivées de f par krigeage (cf. l'annexe B.3) et de calculer la probabilité pour un point de l'espace de recherche d'être à la fois un optimiseur global et de correspondre à une valeur de la fonction objectif inférieure à un seuil donné. Il est alors aisé de construire une approximation relativement bonne de la distribution conditionnelle des minimiseurs. Cependant, cette approximation nuit fortement à la convergence et EI devient alors plus efficace. IAGO reste donc réservé à l'optimisation de fonctions dont le coût de l'évaluation est significatif devant celui du choix du point d'évaluation. L'existence de telles fonctions est de toute façon la motivation première pour l'introduction de IAGO, et l'on en rencontre de nombreuses dans les applications industrielles, notamment dans celles présentées dans le prochain chapitre.

CHAPITRE 5

APPLICATIONS INDUSTRIELLES

Résumé — Ce chapitre traite d'applications industrielles de EGO et de IAGO, ainsi que des difficultés rencontrées à leur contact. Nous insistons en particulier sur le fait que l'automatisation des simulations numériques, qui permettent l'évaluation de la fonction à optimiser, pose bien souvent problème. L'automatisation, lorsqu'elle est effective, est souvent synonyme d'une baisse de la qualité de la simulation qui implique un bruit numérique sur les résultats des évaluations et parfois même, un échec de la procédure de simulation. Nous présentons l'approche retenue pour faire face à ce cas de figure, ainsi qu'au contexte multi-objectif présent dans trois des quatre applications considérées ici. Plus généralement, ces applications démontrent l'intérêt pratique de l'optimisation globale bayésienne pour la conception en milieu industriel, et ce pour des dimensions importantes de l'espace des facteurs (jusqu'à 35).

5.1 Introduction

Dans ce chapitre, nous détaillons les applications de IAGO ou de EGO [1] à quatre problèmes d'optimisation globale rencontrés chez Renault. Ces problèmes (résumés dans le tableau 5.1) ont pour point commun le coût élevé de l'évaluation de la fonction à optimiser et possèdent chacun des particularités ayant nécessité une adaptation des algorithmes. L'objectif de ce qui suit est triple. Il s'agit en effet de montrer l'efficacité pratique des méthodes proposées, de détailler les quelques adaptations nécessaires à la résolution de problèmes réels, mais aussi d'insister sur les difficultés techniques rencontrées lors de la mise en œuvre d'algorithmes d'optimisation dans l'industrie, et en particulier lors de l'automatisation de l'évaluation de la fonction à optimiser. Ces problèmes, que l'on peut aisément sous-estimer, représentent en effet un frein majeur à la généralisation des techniques d'optimisation et sont actuellement au cœur des préoccupations industrielles.

Pour chacun des problèmes considérés ici, le choix entre EGO et IAGO est fait en fonction des

[1] EGO et IAGO désignent simplement ici l'optimisation séquentielle de l'EI ou de l'ECM. Les autres détails algorithmiques sont spécifiés pour chaque application.

Problème considéré	*Dimension de l'espace des facteurs*	*Budget d'évaluations*	*Particularités du problème*	*Méthodes utilisées*	*Gain constaté en nombre d'évaluations* a posteriori
Optimisation de la forme du conduit d'admission	6	20	Multi-objectif	IAGO, EGO, LHS	
Optimisation de la forme de la chasse combustion	4	16	Multi-objectif, résultats d'évaluation bruités et parfois manquants	IAGO, plan orthogonal	
Optimisation du contrôle de la direction assistée	32	300	Grande dimension	EGO, LHS, Nelder-Mead	85%
Optimisation de la masse d'un absorbeur de choc	35	180	Grande dimension, présence d'une contrainte	EGO, plan d'expériences suivi d'une optimisation locale	85%

TAB. 5.1: *Récapitulatif des problèmes traités pour Renault. La colonne* gains en évaluation *est renseignée uniquement lorsque le problème traité a déjà été étudié à l'aide d'autres approches chez Renault. Pour des raisons de confidentialité, la nature des paramètres n'est pas communiquée.*

préconisations décrites à la section 4.5. Notons que les méthodes de référence utilisées ici pour la comparaison avec EGO ou IAGO ne sont pas nécessairement les plus pertinentes. Ce sont celles qui étaient utilisées chez Renault, et il était important de démontrer aux praticiens l'intérêt des approches proposées face aux méthodes qu'ils connaissent et utilisent.

5.1.1 Contraintes propres à la conception dans l'industrie automobile

Les problèmes d'optimisation rencontrés chez Renault sont profondément marqués par des contraintes propres au marché automobile. La première de ces contraintes tient au caractère contradictoire du cahier des charges d'un véhicule. On souhaitera par exemple un véhicule puissant mais consommant moins que ses concurrents tout en satisfaisant aux normes de pollution, ou encore un véhicule résistant au crash, mais dont le poids ne détériore pas outre-mesure la consommation. Ces contradictions se déclinent sur chacune des parties du véhicule, et les problèmes d'optimisation rencontrés possèdent donc bien souvent plusieurs objectifs. Il est alors possible de se ramener à un problème mono-objectif si un compromis peut être choisi *a priori* (ce sera le cas pour l'application de la section 5.5), ou encore à un problème contraint (cf. la section 5.6), si le cahier des charges est suffisamment précis. Cependant il n'est pas rare que les praticiens souhaitent disposer du *front de Pareto* [2] dans son ensemble de manière à faire face facilement à un changement de cahier des

[2] Considérons le problème de minimisation de L fonctions objectifs $f_1, ..., f_L$ (définies sur $\mathbb{X}$) ; un point $\mathbf{x}$ est dit *optimal au sens de Pareto* (ou non dominé) si $\forall \mathbf{y} \in \mathbb{X}\ \exists i \in [1:L]$ t.q. $f_i(\mathbf{x}) \leq f_i(\mathbf{y})$. Le front de Pareto est l'ensemble des points optimaux au sens de Pareto. Nous y reviendrons à la section 5.2.1.

charges survenant au cours du projet (un tel changement n'est pas rare, compte-tenu de l'interdépendance des cahiers des charges de chacune des pièces du véhicule). En effet, pour choisir entre deux points du front, il est nécessaire de spécifier une préférence entre les objectifs. Si une telle préférence est exprimée *a priori*, l'optimisation se concentre sur une zone du front de Pareto. En cas de changement de cahier des charges et donc de ces préférences, il sera alors nécessaire de reprendre l'optimisation pour explorer une nouvelle zone du front. En résumé, la connaissance du front de Pareto permet de s'adapter instantanément à une reformulation du cahier des charges.

La deuxième contrainte propre au marché automobile, et qui affecte directement ces travaux, concerne le coût et les délais du développement d'un nouveau véhicule. En effet, pour de nombreux organes, la conception s'effectue à l'aide d'essais coûteux ou de longues simulations numériques. Les budgets d'évaluation associés aux problèmes d'optimisation rencontrés sont donc bien souvent très limités.

Une troisième contrainte provient du mode de fabrication des véhicules. La fabrication à la chaîne s'accompagne en effet de dispersions de fabrication importantes qui compliquent bien souvent le processus de conception. Nous n'avons cependant pas encore eu l'occasion d'appliquer les méthodes présentées au chapitre 4 à un problème réel. Aussi n'insisterons-nous pas sur ce point.

5.1.2 Automatisation des évaluations

Avec le développement et la baisse des coûts des serveurs de calcul, il est devenu possible, pour l'ensemble du monde industriel, de remplacer de nombreux essais par des simulations numériques. Lorsque ces simulations sont utilisées pour la conception, il devient maintenant envisageable de coupler directement une méthode d'optimisation avec le programme de simulation pour accélérer le processus et améliorer la validité des solutions retenues, l'utilisateur n'intervenant plus que pour l'analyse des résultats. Cependant, la réalisation de cette interface entre le programme d'optimisation et le programme de simulation est complexe, et l'on préfère parfois l'éviter, ce qui implique pour l'utilisateur de mettre en place lui-même chacune des simulations programmées.

Dans la situation la plus favorable, ces difficultés à réaliser le couplage optimisation – simulation relèvent simplement de la différence entre les langages ou les logiciels utilisés. Elles peuvent alors être résolues, éventuellement au prix de développements longs et coûteux. En revanche, pour beaucoup de problèmes d'optimisation, les difficultés proviennent du programme de simulation numérique qui n'est pas *a priori* automatique et nécessite l'intervention de l'utilisateur. On retrouve en particulier cette difficulté pour les problèmes d'optimisation de forme. En effet, la procédure standard dans ce contexte est de créer une représentation numérique de la forme (on parle aussi de *conception assistée par ordinateur* ou CAO) puis de simuler son comportement à l'aide d'un solveur aux éléments finis qui résout numériquement les équations de la physique du phénomène étudié (écoulement fluide, vibratoire, acoustique...), et enfin d'extraire la grandeur d'intérêt des résultats de la simulation (étape dite de *post traitement*). Cette approche nécessite

l'intervention humaine pour la création de la CAO, la création d'un maillage de cette dernière, puis la spécification de conditions aux limites, le démarrage du calcul et enfin le post traitement.

Ces cinq actions peuvent en théorie être automatisées, mais en pratique cette automatisation s'accompagne d'erreurs qui peuvent survenir à chaque étape. En particulier, le choix du paramétrage (définition de $\mathbf{x}$) et de l'espace de recherche pour ces paramètres (définition de $\mathbb{X}$) est une étape difficile. En effet, certaines valeurs des paramètres peuvent entraîner des erreurs dans la génération de la CAO (nous en verrons un exemple à la section 5.3). Ces erreurs sont, pour des formes simples, prévisibles et l'on est alors en mesure de les éviter. Cependant, pour des formes complexes, il est très probable que de telles erreurs subsistent malgré les efforts du concepteur. Une autre difficulté majeure tient à l'automatisation du maillage. Cette dernière est de plus en plus proposée par les logiciels de CAO, mais l'adaptation du maillage à la forme est souvent imparfaite et entraîne alors une diminution de la précision et parfois même l'interruption de la procédure, interruption qui peut aussi être causée par la divergence de la solution calculée par le solveur.

En résumé, l'automatisation de l'évaluation de la fonction à optimiser est une tâche complexe et son résultat n'est pas totalement fiable. Il faut donc s'attendre à une incertitude significative sur le résultat des évaluations, et parfois même, à une absence de résultat contre laquelle il convient de se protéger.

5.2 Mise en place des algorithmes d'optimisation

Pour traiter les problèmes d'optimisation rencontrés chez Renault, il faut donc être en mesure de traiter des problèmes multi-objectifs et de faire face à d'éventuelles défaillances de la chaîne de simulation numérique. Dans cette section, nous présentons les techniques utilisées pour adapter les critères d'échantillonnage à ces deux cas de figure.

5.2.1 Extension des critères d'échantillonnage aux problèmes multi-objectifs

Considérons L fonctions $f_1, ..., f_L$ définies sur $\mathbb{X}$. Pour choisir entre deux solutions au problème de minimisation de ces fonctions, on fait souvent appel à une relation de dominance entre les points de l'espace des facteurs. Ainsi, $\mathbf{x} \in \mathbb{X}$ domine $\mathbf{y} \in \mathbb{X}$ au sens de Pareto si et seulement si

$$\forall i \in [\![1 : L]\!] \ f_i(\mathbf{x}) \leq f_i(\mathbf{y}),$$

ou en d'autres termes si $\mathbf{x}$ est meilleur que $\mathbf{y}$ pour tous les objectifs. On définit alors la notion d'optimalité au sens de Pareto par

$\mathbf{x} \in \mathbb{X}$ est optimal au sens de Pareto si et seulement si $\forall\, \mathbf{y} \in \mathbb{X}\ \exists\, i \in [\![1, L]\!]$ t.q. $f_i(\mathbf{x}) \leq f_i(\mathbf{y})$.

Autrement dit, un point est optimal au sens de Pareto s'il n'est dominé par aucun autre point.

La résolution d'un problème d'optimisation multi-objectif passe alors par le calcul de l'ensemble des points optimaux au sens de Pareto, aussi appelé *front de Pareto*. Cet ensemble représente l'ensemble des solutions entre lesquelles il est impossible de décider sans exprimer de préférence entre les objectifs. Il contient en particulier, les minimiseurs de chacune des fonctions et présente souvent, dans l'espace d'arrivée, une forme convexe similaire à celle présentée sur la figure 5.1 pour $L = 2$. Pour estimer le front de Pareto, il est alors d'usage de faire appel à des méthodes d'optimisation mono-objectif. En effet, sous l'hypothèse de convexité du front[3], chacun des points $\mathbf{x}$ de celui-ci est solution d'un problème de minimisation d'une combinaison linéaire $z_{\mathbf{x}}$ des objectifs (cf. la figure 5.1 pour une représentation graphique de cette équivalence),

$$z_{\mathbf{x}}(\cdot) = \sum_{i=1}^{L} a_{\mathbf{x},i} f_i(\cdot),$$

avec $a_{\mathbf{x},i} \geq 0 \, \forall \, i \in [\![1 : L]\!]$ les coefficients de pondération des objectifs, qui vérifient $\sum_{i=1}^{L} a_{\mathbf{x},i} = 1$ (on parle aussi d'agrégation des objectifs). Il y a alors correspondance (bijection) entre le front de Pareto et l'ensemble des pondérations possibles qui correspondent à autant de préférences pour les objectifs. Par exemple, si $a_{\mathbf{x},i} = \delta_{i,1}$, seul le premier objectif importe et l'on cherche à résoudre le problème de minimisation de f_1 qui va fournir une des extrémité du front de Pareto. En pratique, plusieurs jeux de pondérations sont donc choisis et les problèmes correspondants résolus. Il est alors possible d'estimer l'allure du front de Pareto dans l'espace d'arrivée par interpolation des points obtenus.

Si le budget d'évaluation est limité, l'exploration séquentielle de chaque zone du front semble impraticable. Il a donc été proposé dans Knowles (2003) d'utiliser avant chaque nouvelle évaluation un nouveau jeu de pondérations, choisi aléatoirement à l'aide d'une loi uniforme. Ainsi, une estimation de *l'ensemble* du front de Pareto est construite graduellement. Dans ces travaux, les auteurs introduisent l'algorithme parEGO, extension de l'algorithme EGO aux problèmes multi-objectifs. Cet algorithme choisit les évaluations à réaliser à l'aide de l'EI appliqué sur une combinaison linéaire des objectifs (que l'on a pris soin de normer, c'est-à-dire de ramener entre 0 et 1 [4]) choisie à l'aide d'une loi uniforme. Pour appliquer IAGO aux problèmes rencontrés chez Renault, nous avons choisi de procéder de la même façon et de traiter de manière séquentielle une série de problèmes mono-objectifs extraite aléatoirement du problème multi-objectif.

Remarque 5.1. En pratique, on choisit souvent de limiter le support de la loi uniforme utilisée pour choisir les pondérations. Il est ainsi possible d'éviter l'exploration des extrémités du front qui ne présentent en général que peu d'intérêt.

[3] Cet hypothèse est satisfaite, au moins approximativement, par les problèmes rencontrés au cours de cette thèse. Dans le cas contraire, il est encore possible d'utiliser des approches similaires (Miettinen, 1999).

[4] Dans le cas où le minimum et le maximum d'une fonction objectif f ne sont pas connus *a priori*, la transformation appliquée à f évolue après chaque évaluation. Par exemple, on peut considérer $(f(\cdot) - f_{\min})/(f_{\max} - f_{\min})$ en lieu et place de $f(\cdot)$ (avec $f_{\min}$ et $f_{\max}$ le minimum et le maximum des résultats des évaluations réalisées).

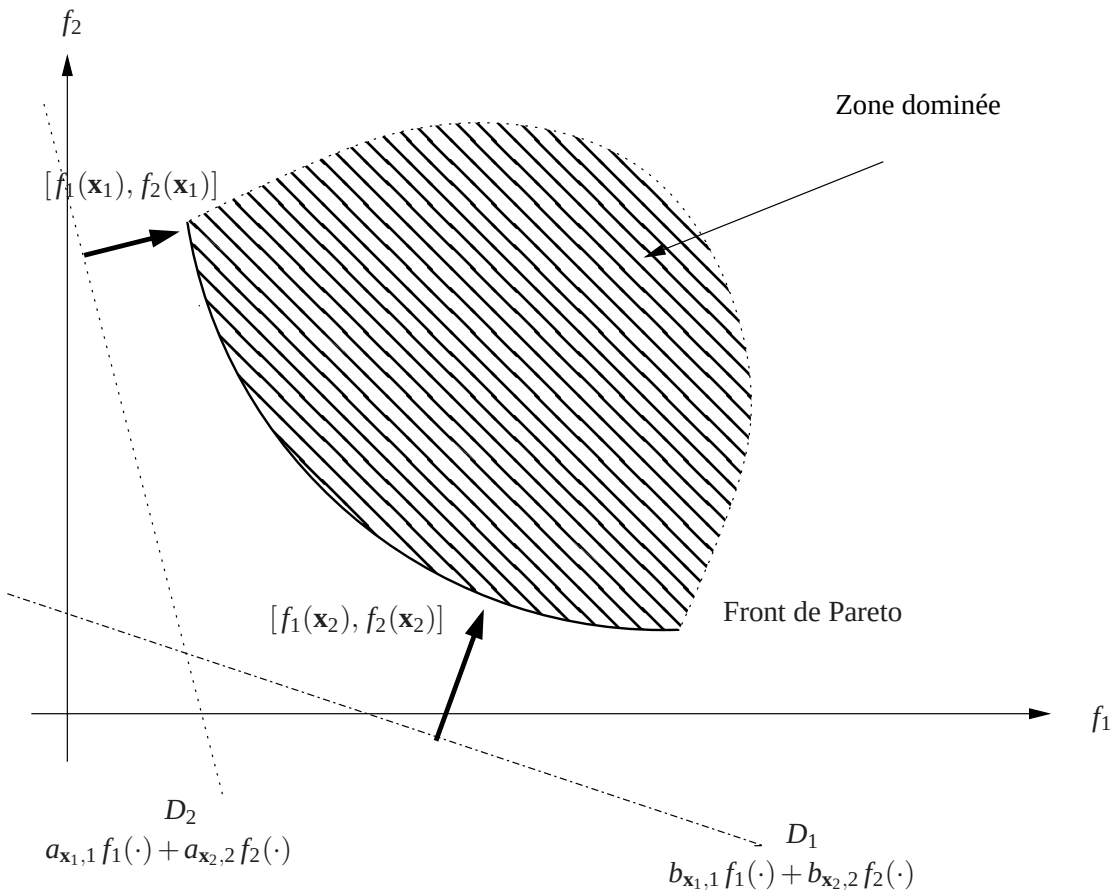


FIG. 5.1: *Allure classique, dans l'espace d'arrivée, du front de Pareto d'un problème bi-objectif et correspondance entre un problème multi-objectif et une série de problèmes mono-objectifs. Si* $\mathbf{x}_1$ *et* $\mathbf{x}_2$ *sont optimaux au sens de Pareto pour le problème bi-objectif de minimisation de* f_1 *et* f_2*, alors ils minimisent respectivement* $a_{\mathbf{x}_1,1} f_1(\cdot) + a_{\mathbf{x}_1,2} f_2(\cdot)$ *et* $b_{\mathbf{x}_2,1} f_1(\cdot) + b_{\mathbf{x}_2,2} f_2(\cdot)$.

Remarque 5.2. Nous avons envisagé d'autres approches du choix des pondérations et en particulier, étant donné un ensemble de jeux de pondérations candidats, de choisir la pondération associée au sous-problème mono-objectif possédant l'entropie de la distribution conditionnelle des minimiseurs globaux la plus forte. Ainsi, chaque évaluation est consacrée à l'estimation de la zone du front de Pareto la plus incertaine. Cependant, les mesures de progrès empiriques ne nous ont pas permis de constater de gain significatif par rapport à l'approche proposée par Knowles (2003).

5.2.2 Problème des données manquantes

Comme nous le verrons au cours de la description du problème d'optimisation d'un conduit d'admission (section 5.3), il convient de se prémunir contre les erreurs de la chaîne de simulation automatique qui se traduisent par l'absence de résultat pour certains points d'évaluation. Si les points d'évaluation sont choisis à l'aide d'un plan d'expériences décidé *a priori*, ces erreurs ne posent pas de problème, si l'on excepte le gaspillage de ressources associé. En revanche, si l'on choisit la nouvelle évaluation en fonction des résultats des précédentes à l'aide de l'EI ou de l'ECM, il faut affecter une valeur fictive au résultat de l'évaluation qui a échoué.

Il serait aussi souhaitable de prendre en compte l'échec de l'évaluation pour estimer progressivement le sous-ensemble de $\mathbb{X}$ sur lequel les évaluations échouent. Cependant, l'expérience montre que cet ensemble peut être non connexe et de forme complexe. Il apparaît en fait que le seul principe à respecter est d'éviter d'évaluer la fonction au voisinage d'un point qui a précédemment échoué. Pour cela, nous proposons d'affecter aux évaluations qui ont échoué leur prédiction par krigeage à partir des résultats disponibles, et de les mettre à jour à chaque fois qu'une évaluation réussit. Nous proposons aussi, même en présence de bruit sur les résultats d'évaluation (ce sera le cas à la section 5.4) d'affecter un bruit nul aux résultats fictifs. Cette approche, inspirée de Forrester et al. (2006), permet ainsi de détourner l'échantillonnage des points qui sont traduits par des échecs, tout en permettant d'utiliser, sans modification, les critères de maximisation de l'EI et de minimisation de l'ECM.

5.2.3 Choix des paramètres de la covariance

Résumons ce qui a été dit sur le choix des paramètres de la covariance depuis le début de ce mémoire. Nous avons abordé au chapitre 1 l'importance de ce choix (nous y reviendrons dans l'annexe A) pour la qualité de la prédiction par krigeage, ainsi que son impact sur les critères d'échantillonnage. Ces paramètres sont généralement estimés par maximum de vraisemblance à partir des résultats d'évaluation disponibles. Cependant, nous avons vu au chapitre 2 que, lorsque le budget d'évaluations est aussi réduit que pour les applications considérées dans ce chapitre, il est illusoire d'espérer en obtenir une estimée valable. Ce problème n'a, à notre connaissance, pas été abordé dans le littérature, où l'on considère généralement des budgets d'évaluations suffisamment

importants pour réaliser un plan d'expériences initial dans le seul but d'estimer les paramètres de la covariance[5]. Notre objectif est de montrer qu'il est souvent tout à fait possible en pratique de choisir *a priori* les paramètres de la covariance.

Le premier élément motivant cette recommandation, est que l'on dispose souvent de données issues de la conception d'un système similaire à celui que l'on cherche à optimiser. En effet, il arrive fréquemment, qu'un système déjà optimisé subisse de petites modifications qui obligent à ré-optimiser. Bien que la fonction obtenue diffère de la fonction initiale, il peut rester légitime d'utiliser le même jeu de paramètres pour la covariance (l'optimisation de la commande d'une direction assistée, présentée à la section 5.5, utilise des paramètres de covariance choisis de cette manière).

Au delà de cette situation favorable, nous avons vu au chapitre 4 que l'optimisation semble plus robuste que la prédiction par rapport à un mauvais choix des paramètres de la covariance. En effet, il semble qu'une surestimation des paramètres de la covariance de Matérn ne diminue que faiblement les performances de IAGO et EGO. Il semble ainsi possible de procéder comme nous le proposions à la section 2.3.1, et d'utiliser les connaissances sur la physique du système pour choisir *a priori* les paramètres de la covariance.

Cependant, d'une part ces connaissances physiques (que l'on souhaiterait porter sur l'amplitude de variation, la dérivabilité, ou encore sur une constante de Lipschitz) ne sont pas disponibles dans les applications présentées dans ce chapitre, et d'autre part leur prise en compte pour guider le choix des paramètres nécessite un recul important sur la nature des paramètres de la covariance et leur impact sur l'optimisation. Ainsi, pour assurer que EGO et IAGO soient utilisables simplement et dans une large gamme de situations, nous proposons d'utiliser un jeu de paramètres standard, qui ne sera modifié qu'une fois le nombre de données suffisant pour justifier une estimation par maximum de vraisemblance, ou en cas de mauvais comportement des algorithmes[6]. Laissons de côté les règles empiriques pour la modification de ces paramètres, mais notons que ces règles devront faire l'objet de travaux futurs, et sont un préalable à une utilisation réellement automatique de IAGO et d'EGO.

Intéressons nous au choix de ces paramètres standards, et choisissons les, malgré le manque d'arguments théoriques pour ce choix, de manière à satisfaire l'intuition et à vérifier le bon fonctionnement des méthodes sur des applications pratiques, dans un contexte défavorable. Ainsi, même si ce choix se révèle naïf et peu approprié, nous pourrons conclure sur la validité de l'hypothèse que les paramètres peuvent être choisis *a priori*.

Le préalable à l'utilisation d'un jeu de paramètres standard est de normer les données. Il faut

[5] Rappelons que les recommandations de Jones et al. (1998) conduisent à des plans d'expériences initiaux de taille 10d, avec d le dimension de l'espace de recherche.

[6] Un exemple de mauvais fonctionnement serait l'accumulation de points dans une petite zone de l'espace de recherche. On pourrait alors envisager d'augmenter la régularité si la distance entre deux points d'évaluation est trop faible.

pour cela ramener l'espace de recherche à $[0,1]^d$, et s'assurer que la fonction objectif f varie entre 0 et 1. Là encore, si le minimum et le maximum de la fonction sont connus *a priori*, la transformation à appliquer à f est immédiate. Dans le cas contraire, il faut modifier la fonction à optimiser après chaque évaluation et considérer $(f(\cdot) - f_{\min})/(f_{\max} - f_{\min})$ en lieu et place de $f(\cdot)$ (avec $f_{\min}$ et $f_{\max}$ le minimum et le maximum des résultats des évaluations réalisées). Une fois cette transformation réalisée, nous proposons de choisir pour paramètres de la covariance de Matérn, $\nu = 5$, $\rho = 0.3\sqrt{d}$ et $\sigma^2 = 0.1$. La régularité nous semble suffisante pour éviter une sous-estimation trop importante, tout en nous affranchissant des problèmes numériques dûs à un mauvais conditionnement de la matrice de covariance. La portée dépend de la dimension pour s'adapter à la taille de l'espace de recherche, et la variance nous semble à même d'assurer le caractère global de la recherche.

Ces paramètres ont été utilisés pour toutes les applications de ce chapitre, excepté celle de la section 5.5 pour laquelle les paramètres sont estimés à partir de données issues d'un problème similaire.

5.3 Optimisation du conduit d'admission

Le conduit d'admission (cf. la figure 5.3 pour une vue globale du conduit d'admission d'un moteur à essence considéré ici) est la pièce du moteur par laquelle le mélange air – carburant pénètre dans la chambre de combustion. L'importance de ce composant tient aux propriétés turbulentes du flux qu'il induit dans la chambre de combustion. Ces turbulences ont en effet un impact direct sur les émissions de polluants. L'objectif est de générer dans la chambre un vortex d'axe normal au cylindre appelé mouvement de *tumble*. Ce mouvement est communément quantifié par une grandeur scalaire, homogène à une vitesse de rotation, appelée *taux de rotation tumble*, ou plus simplement *tumble* (cf. Lumley, 1999, pour une définition et une description des enjeux associés). Le problème pour les concepteurs est que la génération de cet écoulement turbulent, et par la même l'augmentation du tumble, va souvent de paire avec une diminution de la *perméabilité*[7] du conduit. Or la quantité de mélange introduite dans la chambre est directement proportionnelle à la puissance développée par le moteur. La conception des conduits d'admission est donc un problème bi-objectif avec la maximisation de la perméabilité (et donc de la puissance) et la maximisation du tumble (et donc la minimisation des émissions de polluants).

Dans l'étude qui nous a été soumise, il s'agissait de démontrer sur une pièce déjà terminée (cf. la figure 5.3, partie droite) les gains potentiels apportés par les approches proposées dans ce mémoire.

[7] La perméabilité est une mesure de la capacité d'un système à transmettre un fluide. Homogène. à une surface, cette quantité est proportionnelle au débit traversant le système.

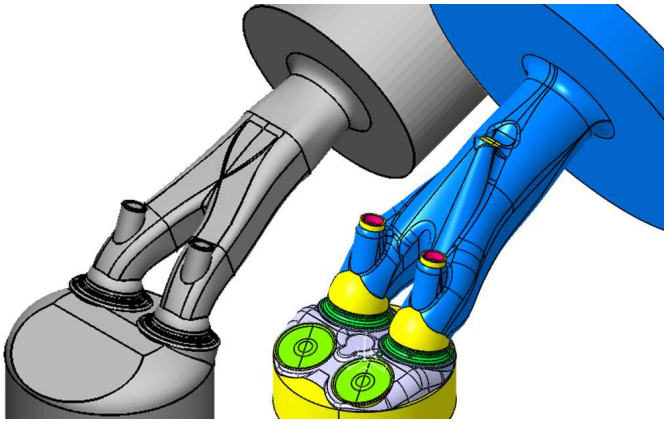


FIG. 5.2: *Vue globale du conduit simplifié et du conduit réel.*

5.3.1 Automatisation des simulations numériques

L'optimisation de la forme du conduit repose sur des simulations numériques qui demandent plusieurs heures sur des serveurs dédiés (une série d'essais en soufflerie est ensuite réalisée pour valider les résultats numériques). Ces simulations, outre leur temps d'exécution, nécessitent plusieurs jours de mise en place. Il faut en effet créer une CAO à l'aide de CATIA[8], son maillage surfacique à l'aide d'ANSA[9], et enfin son maillage volumique et la simulation aux éléments finis à l'aide de STAR-CD[10]. Le problème d'optimisation de fonctions coûteuses est donc simple à identifier, mais sa résolution à l'aide des algorithmes présentés dans ce mémoire requiert l'automatisation de cette chaîne de calcul. On retrouve alors les difficultés d'interfaçage mentionnées précédemment.

Pour y remédier, nous avons choisi d'utiliser le logiciel STAR-design, qui permet d'effectuer automatiquement et dans le même langage la génération de la CAO et du maillage, ainsi que le lancement du calcul. Cette approche, nouvelle pour Renault, nous a permis d'automatiser la simulation et d'appliquer directement nos méthodes d'optimisation, mais elle a aussi ouvert la voie vers une utilisation généralisée de la CAO paramétrée et du maillage automatique. Grâce à cette approche, nous avons pu réaliser, pour les besoins de l'étude, une CAO paramétrée à partir de la CAO réellement retenue. Ceci s'est fait au prix d'une simplification de la géométrie (cf. figure 5.3), qui d'après les experts n'enlève rien à la validité de l'approche.

Nous avons finalement retenu six paramètres de forme pour lesquels l'évaluation des fonctions tumble et débit requiert approximativement une heure. En dépit des simplifications apportées, la chaîne de calcul n'est pas *a priori* fiable. Nous avons en effet été confronté aux problèmes d'erreurs de maillage mentionnées à la section 5.2.2. Par exemple, l'intersection de volumes pour la création

[8] Logiciel de CAO.
[9] Logiciel de maillage.
[10] Solveur aux éléments finis.

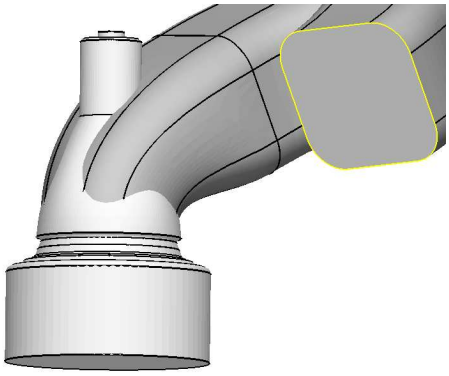


FIG. 5.3: *Assemblage de volumes au cours de la création automatique de la CAO du conduit. L'intersection entre la forme en ogive (en clair) et le conduit (en cours de construction en grisé) peut, pour certaines configurations, mettre en défaut la procédure de maillage automatique.*

de la CAO peut faire apparaître des formes complexes (cf. la figure 5.3) que les techniques de maillage automatique ne supportent pas. Dans le cadre de cette étude, nous avons été en mesure d'identifier tous les problèmes de ce type et d'y apporter une solution, mais il apparaît que pour des CAO plus complexes et dans des délais plus restreints, une élimination totale des erreurs n'est pas envisageable à court terme.

5.3.2 Résultats

Pour traiter ce problème, IAGO et EGO ont tout deux été utilisés. Les paramètres de la covariance ont été fixés *a priori* suivant les recommandations de la section précédente. L'optimisation des deux critères d'échantillonnage s'est effectuée sur un LHS à 1500 points rééchantillonnés après chaque évaluation.

Le budget a été fixé à vingt évaluations. Cette quantité correspond en effet au nombre de calculs qui avaient été réalisés lors d'une étude similaire. Avec l'automatisation des simulations, il sera sans doute possible d'en réaliser davantage, mais l'objectif était ici de démontrer l'intérêt de l'optimisation globale bayésienne pour un budget d'évaluation représentatif des situations rencontrées par les praticiens. Outre IAGO, EGO et un plan LHS ont également été utilisés avec ce même budget. Les points jugés optimaux au sens de Pareto parmi chacun des trois ensembles de points obtenus à partir des résultats fournis par chacune des méthodes, sont présentés sur la figure 5.4. La comparaison entre estimées de fronts de Pareto est généralement une tâche complexe, qui nécessite l'utilisation de multiples mesures de qualité (Knowles et al., 2006). Cependant, pour cette application, la comparaison est nettement en faveur de IAGO pour trois critères de comparaison souvent utilisés. En effet,

– parmi toutes les simulations réalisées, le point le plus proche de la solution idéale (c'est-

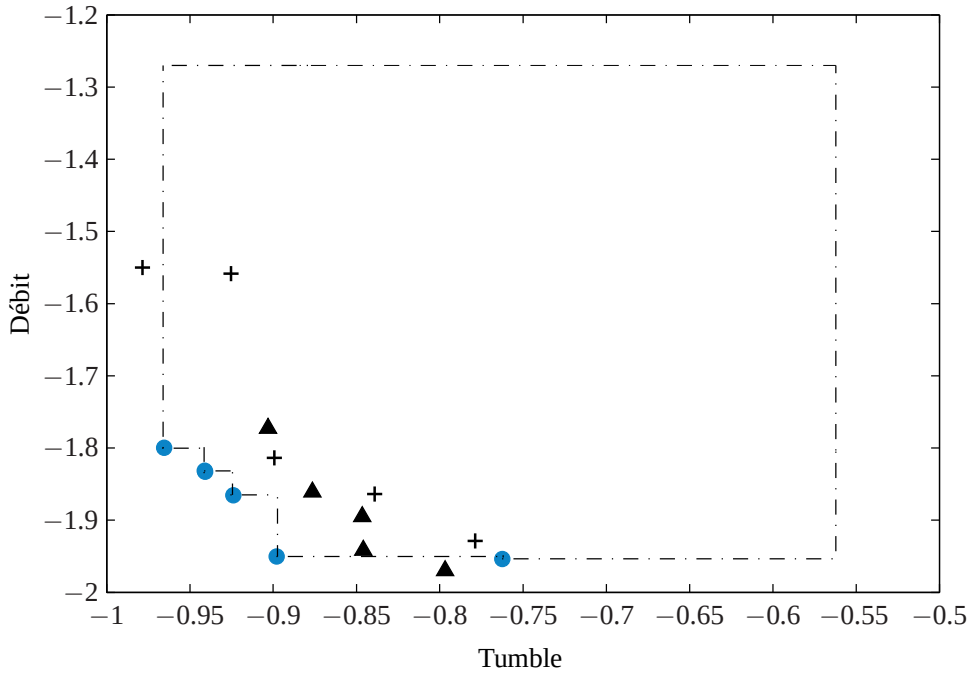


FIG. 5.4: *Représentation dans le plan tumble – débit des points optimaux au sens de Pareto parmi les vingt points choisis par un plan LHS (croix), par EGO (triangles) et par IAGO (disques). Le tumble et le débit ont été modifiés linéairement pour des raisons de confidentialité (les deux objectifs doivent maintenant être minimisés). Le trait mixte délimite l'ensemble des points dominés par les points Pareto optimaux obtenus avec IAGO et qui dominent le plus mauvais point obtenu pour chacun des deux objectifs. L'aire délimitée par ce trait est appelée* hypervolume *(Knowles et al., 2006) et est utilisée pour quantifier l'intérêt d'un ensemble de points Pareto optimaux.*

à-dire le point ayant pour coordonnées les meilleures valeurs obtenues pour chacun des objectifs, ici $[-0.97, -1.97]^{\mathsf{T}}$) est le résultat d'une simulation choisie par IAGO.

– Presque tous les points optimaux au sens de Pareto parmi ceux choisis par EGO sont dominés par des points choisis par IAGO. En d'autres termes, presque toutes les solutions trouvées grâce à EGO perdent tout intérêt face aux solutions trouvées par IAGO.
– L'aire de l'ensemble des points dominés par les points optimaux au sens de Pareto (cf. Knowles et al., 2006, pour des détails sur le mode de calcul de cette quantité) est de 0.31 pour IAGO (cette aire est représentée sur la figure 5.4), alors qu'elle n'est que de 0.26 pour EGO (la référence pour délimiter l'aire des points dominés est le point dont les coordonnées sont les pires valeurs obtenues pour chacun des objectifs).

Cette étude confirme la fois la possibilité d'automatiser la procédure de simulation et l'intérêt de l'optimisation globale bayésienne sur un problème multi-objectif.

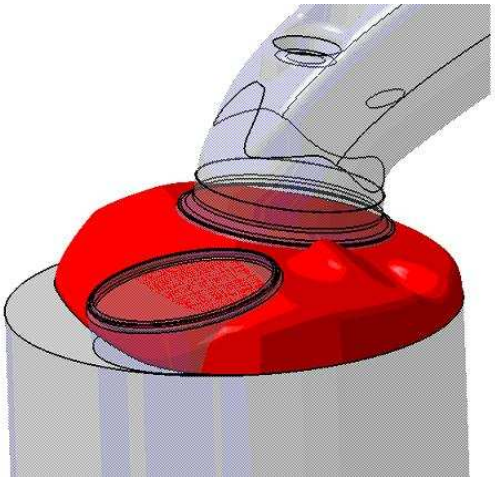


FIG. 5.5: *Représentation de la chasse combustion à optimiser.*

5.4 Optimisation d'une chasse combustion

La *chasse combustion* (cf. la figure 5.5) est la pièce qui relie le conduit d'admission à la chambre de combustion. Elle influence fortement la nature de la turbulence créée dans la chambre. L'objectif est ici de maximiser à la fois le *tumble* et le *swirl* (grandeur scalaire quantifiant le mouvement rotatif du mélange autour de l'axe du cylindre). L'optimisation des paramètres qui régissent la forme de la chasse combustion repose ici encore sur des simulations numériques (CAO, maillage et simulation de l'écoulement fluide par un solveur aux éléments finis). Dans cette étude, quatre paramètres de forme sont considérés.

5.4.1 Construction d'un cas test

Compte-tenu de la difficulté à automatiser la chaîne de simulation (voir les problèmes mentionnés à la section 5.1.2), l'optimisation réelle a été réalisée à l'aide d'un plan d'expériences orthogonal à 45 essais (Taguchi et Konishi, 1987). Pour beaucoup des simulations ainsi choisies, l'intervention humaine a été nécessaire, soit pour corriger à la main le maillage ou la CAO générés, soit pour augmenter le nombre d'itérations du solveur aux éléments finis. En effet, la convergence de l'estimation du *swirl*, et du *tumble*, et plus généralement la convergence de l'estimation de l'écoulement fluide dans la chambre de combustion, nécessite un nombre d'évaluations qui dépend fortement de la forme étudiée. Ainsi, dans certains cas, la convergence va être satisfaisante dès 1000 itérations du solveur, alors que dans d'autres, elle n'est toujours pas effective après 3000 itérations (cf. la figure 5.6).

Ces deux difficultés (simulations échouants ou incomplètes) ne vont pas disparaître lors des études futures. Il importe donc de démontrer que l'optimisation bayésienne s'applique bien dans ce contexte. Nous avons pour cela créé, à partir des 45 résultats de simulations à notre disposition,

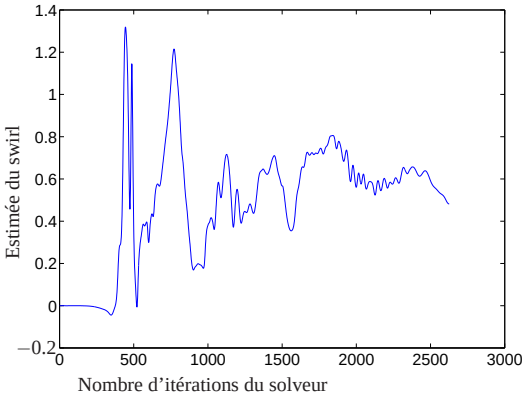


FIG. 5.6: *Évolution de l'estimée du swirl en fonction du nombre d'itérations du solveur aux éléments finis.*

un modèle de la relation des quatre paramètres de forme avec le *swirl* et le *tumble*. Ce modèle est la prédiction par krigeage du *swirl* et du *tumble* à partir des résultats de 45 évaluations. Nous avons ensuite modélisé l'échec potentiel de la simulation numérique par une probabilité, uniforme sur l'espace des facteurs, pour ce modèle de ne pas fournir de résultat à une évaluation.

À chacun des 45 résultats de simulations est associée une incertitude numérique. Pour en rendre compte dans notre cas test, nous l'avons modélisée par un bruit additif gaussien d'écart-type égal à l'écart-type de l'estimée du *swirl* ou du *tumble* sur les 500 dernières itérations du solveur. Cet écart-type, qui varie d'un résultat à l'autre, a ensuite été, lui aussi, modélisé par sa prédiction par krigeage. Nous disposons donc d'un modèle de la chasse combustion qui, pour chaque valeur du vecteur des paramètres, fournit une valeur de *swirl*, de *tumble* et les écarts-types associés. Ces écarts-types seront pris en compte, au cours de l'optimisation du cas test, par la prédiction par krigeage sur laquelle reposent les critères d'échantillonnage (cf. la figure 5.7 pour un exemple).

5.4.2 Résultats

Chaque simulation numérique requière plusieurs heures. Compte-tenu de cette durée et de la présence de bruit sur les résultats des évaluations, IAGO aurait été préféré à EGO si l'un ou l'autre avait pu être mis en œuvre sur le problème réel. Faute de mieux, nous allons appliquer IAGO à l'optimisation de notre cas test inspiré de cette étude. Considérons donc, dans un premier temps, les résultats obtenus par la version multi-objectif de IAGO (cf. la section 5.2.1) adaptée aux résultats d'évaluation bruités (cf. la section 2.4.1) en supposant nulle la probabilité d'erreur dans la chaîne de simulation. A titre de comparaison, un plan d'expériences orthogonal à cent points a aussi été réalisé.

De manière à disposer d'une référence pour les résultats, le front de Pareto du cas test a été

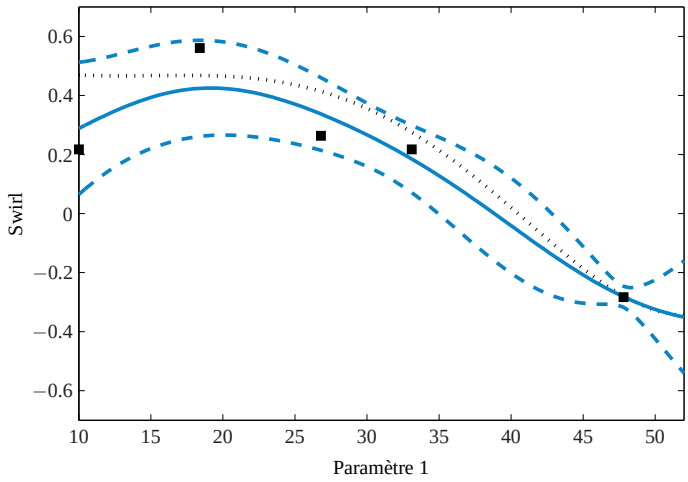


FIG. 5.7: *Coupe du modèle de swirl et de sa prédiction par krigeage à partir d'évaluations choisies arbitrairement (tous les paramètres sont fixés excepté le premier). Le modèle du swirl est représenté par des traits en pointillés courts, la prédiction par krigeage par le trait plein, la limite des intervalles de confiance à 95 % par les traits pointillés longs et les résultats bruités des évaluations par des carrés. Remarquons la diversité des écart-types associés à chaque résultat d'évaluation, qui se répercute sur les intervalles de confiance pour la prédiction (surtout remarquable au voisinage de l'évaluation la plus à droite).*

estimé par Monte-Carlo[11]. La figure 5.8 présente ainsi, pour seize itérations de IAGO comme pour le plan orthogonal, les résultats des évaluations obtenus et le front de Pareto du cas test. Cette figure présente en outre les estimées de ce front obtenues à l'issue des deux procédures. Ces estimées sont calculées par Monte-Carlo directement sur la prédiction par krigeage à partir des seize itérations de IAGO ou du plan orthogonal. A chacun des points obtenus est associée, pour attester de la qualité de l'estimation, une ellipse dont les demi-axes représentent l'écart-type de l'erreur de prédiction pour le *tumble* et pour le *swirl*. Au regard des résultats obtenus par le plan d'expériences, l'intérêt de IAGO apparaît nettement sur cette figure, puisque l'estimation du front de Pareto fournie après seize itérations est plus proche de la réalité et puisque l'incertitude qui y est associée est plus faible (conséquence directe de la proximité entre les résultats des évaluations choisies par IAGO et le front de Pareto réel).

Pour étudier l'influence des simulations qui ont échoué sur l'optimisation par IAGO, affectons maintenant au cas test une probabilité de 0.4 (uniforme sur l'espace des facteurs) de ne pas renvoyer de résultats. Cette probabilité est choisie supérieure aux probabilités observées en pratique pour vérifier le comportement de IAGO dans un cas défavorable. Il faut ainsi 28 itérations de IAGO pour obtenir les seize résultats d'évaluation présentés sur la figure 5.9. On y constate que IAGO continue à fonctionner dans ces conditions, et que la qualité de l'estimation du front de Pareto,

[11] La procédure est la suivante. Évaluation du cas test pour 100 000 points choisis à l'aide d'une loi uniforme, puis calcul des points optimaux au sens de Pareto parmi ces 100 000 points.

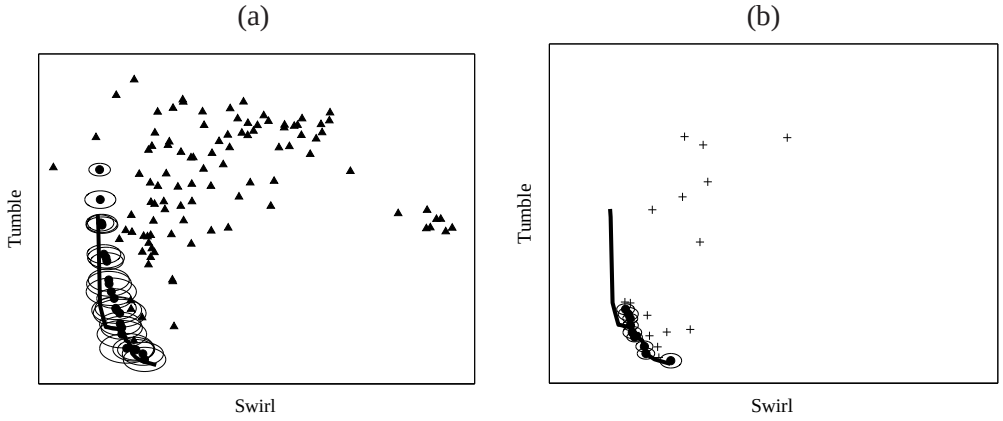


FIG. 5.8: *Estimation du front de Pareto dans le plan* swirl – tumble *à l'issue d'un plan orthogonal à cent essais (a) et de seize itérations de IAGO (b). Le trait plein représente le front de Pareto du cas test (estimé par Monte-Carlo). Les croix représentent les résultats des évaluations choisies par IAGO et les triangles les résultats des évaluations du plan orthogonal. Après application de chacune des deux approches, les prédictions par krigeage du* swirl *et du* tumble *sont calculées. Ces dernières sont ensuite utilisées pour estimer par Monte-Carlo un front de Pareto. Les disques représentent, pour chacune des figures, les points optimaux au sens de Pareto obtenus. Les demi-axes des ellipses associées à chacun de ces points représentent l'écart-type de l'erreur de prédiction pour le tumble et pour le swirl. Les ellipses représentent donc l'incertitude associée à l'estimation du front de Pareto. Les axes ne sont pas renseignés pour des raisons de confidentialité.*

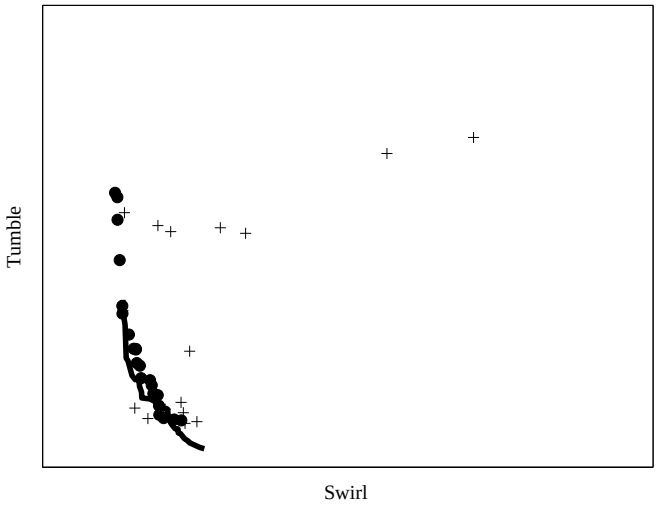


FIG. 5.9: *Estimation du front de Pareto dans le plan swirl – tumble à l'issue de seize itérations de IAGO. La différence avec la figure 5.8 est que le cas test a une probabilité de 0.4 de ne pas fournir de résultat. Il a ainsi fallu 28 calculs pour obtenir les 16 résultats présentés.*

pour un nombre d'évaluations *réussies* donné, ne semble pas affectée. Ces résultats sont néanmoins à nuancer, du fait de la régularité du cas test qui est lui même une prédiction par krigeage. Confrontée à des fonctions moins régulières, la prédiction par krigeage des résultats d'évaluation manquants serait plus éloignée de la réalité, et la convergence de IAGO serait sans doute ralentie.

5.5 Optimisation du contrôle de la direction assistée électrique

5.5.1 Présentation du problème

Cette section est consacrée aux lois de commande d'un modèle de *direction assistée électrique* (DAE). Cette technique d'aide à la direction est plus efficace (jusqu'à 0.2 litre d'économie aux 100km) que la direction assistée hydraulique, car elle consomme de l'énergie uniquement lorsque la direction est sollicitée, alors qu'une pompe hydraulique doit être actionnée en permanence.

La conception de la DAE implique le choix et le dimensionnement des organes, puis le réglage des lois de commande qui les gouvernent. Ce réglage est habituellement réalisé à l'aide d'essais sur banc ou sur circuit, au cours desquels, la direction est confrontée à plusieurs scénarios de conduite représentatifs de l'utilisation du véhicule. Par exemple, l'angle volant décrit un sinus d'amplitude 30° à 0,2Hz pendant 10 secondes à 90km/h (soit 2 périodes). La réponse de la direction est ensuite observée dans le plan angle volant – couple exercé sur la direction (cf. la figure 5.10 pour plusieurs exemples de ces réponses). La qualité de cette réponse, généralement estimée par le pilote ou par les expérimentateurs à partir de critères subjectifs (rappel élastique, pendulage, raideur de guidage...), peut néanmoins être quantifiée par une série de critères relevés directement

sur la courbe de réponse. Cette qualité peut ensuite être, par exemple, résumée par une simple grandeur scalaire, notée f, définie comme la somme, sur tous les scénarios de conduite, des écarts quadratiques entre les critères constatés et les valeurs de références fournies par les experts.

La conception de la DAE passe donc par l'optimisation de f qui dépend à la fois de la définition technique retenue, des lois de commande et du véhicule. L'objectif des campagnes d'essais est d'optimiser les lois de commande pour un véhicule et une définition technique donnés. La principale difficulté de ce problème est que la définition technique doit être choisie avant la campagne d'essais, et qu'il est difficile pour les concepteurs de prévoir les performances que pourra atteindre la définition technique retenue, une fois les lois de commande optimisées. Ils doivent aussi faire face à de nombreuses modifications du reste du véhicule dont certaines peuvent avoir d'importantes conséquences sur la direction assistée. Ainsi, les concepteurs ont à choisir, *a priori*, la définition technique qui offre la meilleure performance, une fois les lois de commande optimisées.

Pour palier cette difficulté, un modèle de DAE a été développé puis couplé avec un modèle de véhicule de manière à simuler numériquement les essais [12]. Cette simulation, bien qu'imparfaite, permet d'estimer les lois de commande optimales pour une définition technique donnée et par la même d'évaluer l'intérêt de cette définition. L'objectif est, à terme, de disposer d'un outil permettant aux concepteurs d'évaluer rapidement (une douzaine d'heures au maximum), à l'aide d'un ordinateur de bureau, l'intérêt de modifications de la définition technique et ce, de manière à réagir rapidement à tous les changements pouvant survenir au cours du développement.

La simulation des scénarios de conduite requiert environ 10 minutes. Le budget d'évaluation souhaité est donc limité à une centaine. Dans le cas qui nous a été soumis, les lois de commande peuvent être résumées par 32 paramètres.

Remarque 5.3. Pour cette application, les paramètres de la covariance ont été estimés à partir de données obtenues à l'aide d'un modèle du véhicule différent de celui utilisé pour l'optimisation. Ces données ne correspondent donc pas à la fonction à optimiser, mais il nous semble légitime de considérer qu'un même jeu de paramètres peut servir à leur représentation.

5.5.2 Résultats

Pour résoudre ce problème, la solution envisagée initialement consistait à réaliser un plan d'expériences LHS, à construire un modèle polynomial de f à partir des résultats du plan d'expériences, et enfin à évaluer f au minimiseur global du modèle polynomial (ce point est aussi appelé *point de validation*), avant de retenir pour solution finale le meilleur résultat d'évaluation.

Compte-tenu du temps de calcul relativement faible requis par chaque évaluation, c'est EGO que nous avons utilisé. 300 points d'évaluation ont été choisis à l'aide de ce critère et leur ré-

[12] Ces modèles sont réalisés en Matlab/Simulink.

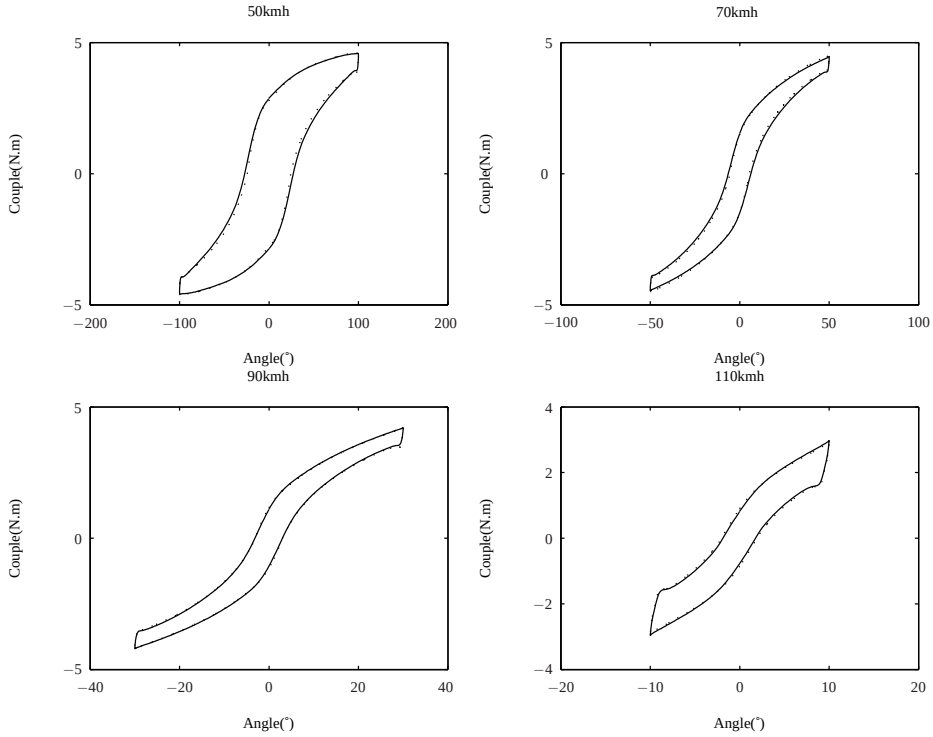


FIG. 5.10: *Réponses de la direction assistée à quatre scénarios de conduite (à différentes vitesses) dans le plan angle volant/couple exercé sur la direction. Les réponses en traits pointillés sont obtenus par simulation pour les lois de commande issues de l'optimisation par EGO (voir aussi la figure 5.11). Les traits continus représentent l'objectif à atteindre tel que défini par les experts.*

sultat comparés à ceux obtenus à l'aide d'un plan LHS à 300 essais et à l'aide de 300 itérations du simplexe de Nelder-Mead. Les résultats, présentés sur la figure 5.11, indiquent clairement la supériorité d'EGO qui trouve en 45 itérations un résultat meilleur que le résultat de validation du plan LHS à 300 essais, soit un gain de 85% .

Cette application confirme qu'en dimension élevée les critères d'échantillonage probabilistes restent utilisables et qu'ils fournissent de plus des résultats satisfaisants au regard de ce que peut offrir un plan d'expériences. Notons aussi le peu d'intérêt du simplexe de Nelder-Mead dans ce contexte (cf. la figure 5.11). Sur un plan pratique, le budget souhaité d'une centaine d'évaluations semble réaliste.

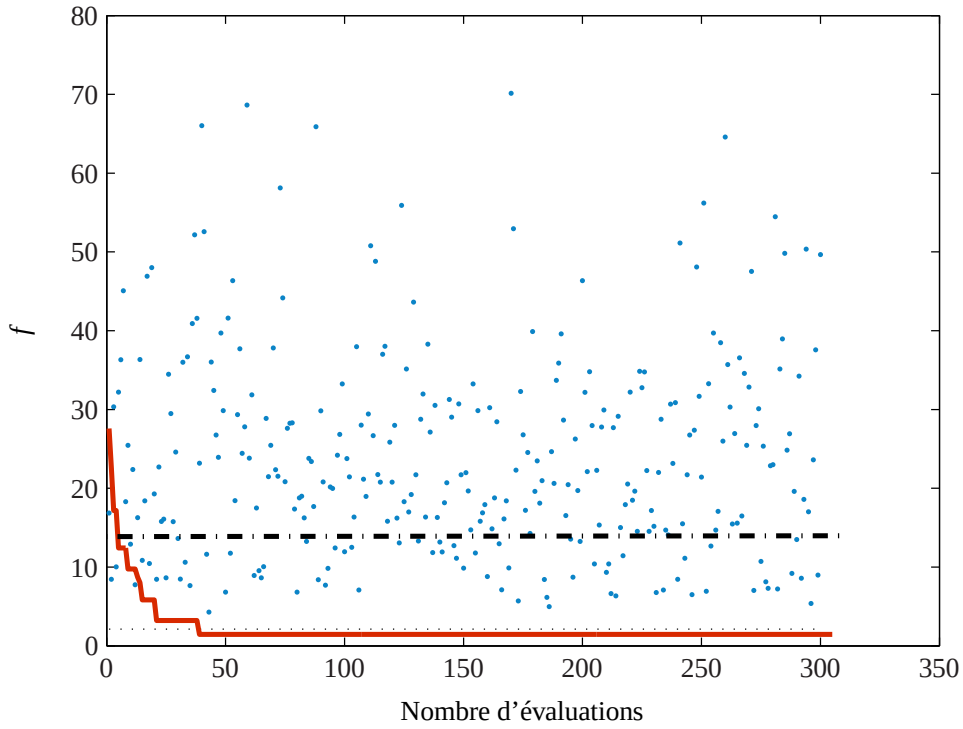


FIG. 5.11: *Application de EGO à l'optimisation des lois de commande de la DAE. Le trait en gras représente la meilleure solution obtenue par EGO en fonction du nombre des évaluations réalisées. Le trait mixte présente la solution obtenue à l'issue de 300 itérations du simplexe de Nelder-Mead. Les points représentent les résultats des évaluations d'un plan LHS à 300 essais. Enfin, les traits pointillés matérialisent le résultat de l'évaluation au point de validation.*

5.6 Optimisation de la masse d'un absorbeur de choc

5.6.1 Présentation du problème

L'application considérée ici a pour objectif de démontrer l'intérêt de l'optimisation globale bayésienne en dimension élevée en présence de contraintes, et traite de la minimisation sous contrainte de la masse d'un absorbeur de choc. Cette pièce, située derrière le pare choc arrière possède un cahier des charges portant, entre autres, sur l'intrusion maximale observée au cours d'un choc standard (nous ne donnerons pas ici les détails de l'expérience). Il s'agit donc de déterminer, parmi l'ensemble des formes permettant de satisfaire au cahier des charges, celle de poids minimal. Le problème d'optimisation obtenu présente 35 facteurs (des épaisseurs de nervures), une contrainte (intrusion plus faible que 80mm) et repose sur des simulations du choc qui requièrent chacune 25 minutes sur un serveur dédié.

Préalablement à notre étude, une première optimisation a été réalisée à l'aide d'un plan d'expériences de 150 simulations suivi d'une optimisation locale (évaluation au point qui minimise un modèle polynomial). Cette étude a permis de réaliser 500 g d'économie sur un absorbeur dont la masse était initialement de 3375 g. C'est ce résultat que nous utilisons comme référence. Notons que les économies de poids réalisées se transcrivent directement en économies de carburant et en diminution des émissions de polluants.

5.6.2 Résultats

Compte-tenu des temps de calculs faibles et de l'absence de bruit, nous avons réalisé l'optimisation avec EGO (l'utilisation de IAGO était programmée, mais n'a pu être réalisée dans les délais). La prise en compte de la contrainte s'est effectuée comme présenté à la section 2.4.3.

La figure 5.12 représente, dans le plan défini par l'intrusion et la masse, les résultats des 120 simulations choisies avec EGO. Il apparaît clairement qu'en comparaison de l'étude de référence, les évaluations choisies explorent efficacement la zone d'intérêt, c'est-à-dire l'intersection entre le front de Pareto (du problème bi-objectif de minimisation de la masse et de l'intrusion) et la contrainte sur l'intrusion. L'utilisation du budget d'évaluations est ainsi plus rationnelle et conduit à un meilleur résultat. En effet, la solution obtenue possède une masse de 2766 g face à 2886 g pour le résultat de l'étude de référence.

Sur le plan de la vitesse de convergence, là encore, EGO se comporte fort bien. On constate en effet sur la figure 5.13 que vingt itérations lui suffisent pour trouver un résultat meilleur que celui de l'étude de référence. Soit un gain en évaluation à iso-qualité de plus de 85%.

Cette étude constitue une preuve supplémentaire des capacités de l'optimisation globale bayésienne, qui se comporte bien en dimension élevée malgré la présence d'une contrainte.

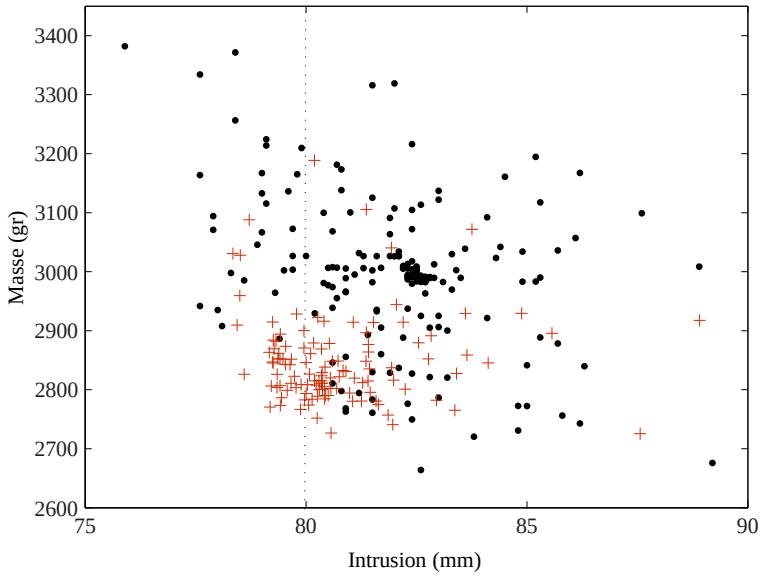


FIG. 5.12: *Représentation dans le plan masse/intrusion, des résultats des évaluations choisis par EGO (croix) et des évaluations réalisées au court de l'étude de référence (points). Les résultats à gauche du trait pointillé satisfont la contrainte d'intrusion.*

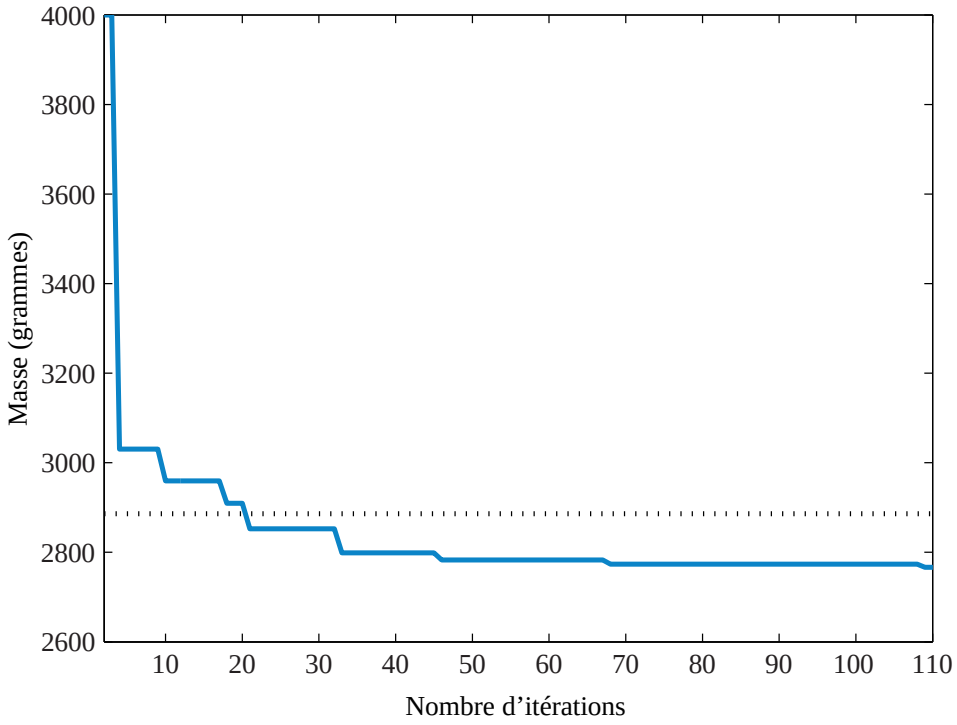


FIG. 5.13: *Meilleure solution (satisfaisant la contrainte) obtenue par EGO, en fonction du nombre de simulations réalisées (trait gras). Après seulement vingt itérations, la méthode trouve un résultat supérieur à celui issu de l'étude de référence qui avait nécessité 150 simulations (trait en pointillés).*

5.7 Conclusions

Les applications présentées dans ce chapitre confirment l'intérêt pratique de l'optimisation bayésienne. En effet, EGO et IAGO ont fait preuve d'une efficacité largement supérieure aux approches habituellement utilisées (surtout des plans d'expériences), dans des contextes d'utilisation réalistes et malgré un choix naïf des paramètres de la covariance. Il est certain que ces bons résultats proviennent de la relative simplicité des fonctions rencontrées (simplicité constatée, mais inconnue *a priori*), qui se sont avérées bien régulières. Cependant, il semble légitime de supposer, compte tenu de la variété des problèmes considérés, que cette simplicité demeure dans beaucoup de problèmes industriels[13]. Dans le cas contraire, les performances seront certes moindres, mais à l'image de celles des algorithmes concurrents.

Notons aussi la flexibilité offerte par IAGO puisque les difficultés supplémentaires telles la prise en compte de contraintes ou d'un bruit sur les résultats des évaluations, l'optimisation de plusieurs objectifs ou encore l'échec de simulations, ne posent pas de problème majeur.

Sur le plan de l'applicabilité en grande dimension, EGO démontre son efficacité en dimension 35. Reste cependant à appliquer IAGO à un problème similaire. Il importera aussi d'étudier l'intérêt pratique des méthodes d'optimisation robuste présentées au chapitre 3.

[13] La raison principale en est sans doute que les espaces de recherche sont généralement assez réduits. En effet, les concepteurs possèdent une idée assez précise du système optimal, et font appel à l'optimisation pour la décision finale.

Conclusions et perspectives

Dans ce mémoire, nous avons étudié l'intérêt d'une modélisation par processus gaussiens pour l'optimisation globale de fonctions coûteuses à évaluer. Cette approche est commune à de nombreux algorithmes conçus spécifiquement pour fonctionner en présence d'un budget réduit en évaluations. Cependant, ils ne tirent pas suffisamment parti du modèle et de la méthode de prédiction qu'on lui associe généralement, le krigeage. C'est pourquoi nous avons travaillé au développement de l'algorithme IAGO reposant sur le krigeage et sur le principe de *Stepwise Uncertainty Reduction*. Ces travaux ont été motivés en grande partie par les problèmes d'optimisation rencontrés chez Renault. Nous avons donc, tout au long de ce mémoire, insisté sur leur variété et proposé des extensions de IAGO pour y faire face.

Contributions

Notre objectif principal était de mettre en place une méthode capable à la fois de fonctionner correctement avec un budget d'évaluations réduit, mais aussi d'être suffisamment flexible pour faire face à la diversité des problèmes rencontrés chez Renault et plus généralement dans l'industrie.

La solution proposée dans ce mémoire est l'algorithme IAGO (et ses extensions), qui repose sur le critère d'échantillonage de minimisation de l'ECM. Son principe est de maximiser l'information apportée par l'évaluation sur les minimiseur de la fonction objectif, ou, plus formellement, de minimiser l'entropie conditionnelle des minimiseurs, c'est-à-dire l'entropie de la distribution de probabilité des minimiseurs conditionnellement au résultat de l'évaluation de la fonction au point candidat. En comparaison, les critères classiques (et en particulier, le critère de maximisation de l'EI) cherchent à améliorer l'estimation du minimum en échantillonnant la fonction objectif là où son apparition est la plus probable. Il nous semble plus raisonnable d'œuvrer à la diminution de l'incertitude associée à la position des minimiseurs globaux. Par exemple, il peut être très coûteux de raffiner cette estimation au voisinage d'un minimum potentiel, qui peut très bien être local, alors que quelques évaluations choisies avec l'ECM peuvent suggérer qu'une large part de l'espace de recherche a une probabilité très faible de contenir un minimiseur global.

Une bonne partie de notre travail (résumée principalement dans les chapitres 1 et 2) est consa-

crée à la mise au point du mode d'évaluation de l'ECM et de ses extensions. La plupart d'entre elles (prise en compte d'un bruit sur les résultats des évaluations, de résultats d'évaluation du gradient, de contraintes) repose sur le principe du cokrigeage, c'est-à-dire de la prédiction conjointe de phénomènes corrélés. Elles sont donc aisées à mettre en œuvre dans IAGO, qui tire, de par son principe même, davantage parti de la modélisation par processus gaussiens, que les autres algorithmes recensés au chapitre 1.

C'est finalement la flexibilité de la prédiction par krigeage qui nous permet de traiter toute la variété des problèmes rencontrés chez Renault. En particulier, nous avons développé au chapitre 3 une extension de IAGO aux problèmes d'optimisation en présence de bruit sur les facteurs. La robustesse de la solution retenue y est mesurée par la performance moyenne. Cette vision simple et pratique de l'optimisation robuste qu'est l'optimisation de la performance moyenne peut être utilisée pour choisir où dépenser le budget en évaluations ; le choix de la solution finalement retenue pouvant ensuite se faire à l'aide de mesures plus adaptées au problème considéré.

Pour donner de la substance à nos intuitions quant à l'efficacité de IAGO, nous avons appliqué chacun des algorithmes à des simulations du modèle gaussien qui leur est commun. Ainsi, nous avons pu estimer le comportement moyen de chacun des critères d'échantillonnage lorsque l'hypothèse gaussienne est satisfaite. Cette procédure de comparaison nous semble plus pertinente que le traitement d'un ensemble de fonctions-tests (elle aussi menée au chapitre 4). Elle permet en effet de tirer des conclusions plus fortes, par exemple que IAGO fonctionne mieux qu'EGO pour l'optimisation d'un processus gaussien de covariance connue lorsque les résultats des évaluations sont bruités. Elle constitue de plus une aide précieuse à la mise au point des algorithmes, notamment pour le choix du plan d'expériences initial. Cette méthode de comparaison nous a aussi permis de constater qu'une surestimation des paramètres d'une covariance de Matérn est beaucoup moins néfaste à la convergence de l'optimisation qu'une sous-estimation. Il faut néanmoins garder en mémoire que ces résultats ne valent en toute rigueur que pour la covariance du processus avec lequel ils sont obtenus. Nous estimons cependant avoir réalisé suffisamment de tests pour montrer que ces conclusions présentent une certaine généralité.

Fort des conclusions du chapitre 4, nous avons pu traiter quatre problèmes industriels et démontrer l'intérêt pratique de l'optimisation globale bayésienne.

Optimisation de fonctions coûteuses en pratique

Dans ce mémoire nous avons tenté de démontrer que, malgré la complexité des problèmes d'optimisation de fonctions coûteuses rencontrés dans l'industrie, il était possible de proposer des méthodes ayant une justification théorique et pratique.

Sur le plan théorique, l'algorithme IAGO et ses extensions nous apparaissent comme un outil très souple, que son coût calculatoire réserve à des problèmes au budget d'évaluation très réduit.

Quand ce coût devient prohibitif face au coût de l'évaluation, l'algorithme EGO peut être utilisé en remplacement de IAGO bien qu'il soit moins efficace en présence de bruit sur les résultats des évaluations.

Sur le plan pratique, nous avons tenté de présenter en détails, dans le chapitre 5, les difficultés rencontrées lors d'applications des algorithmes IAGO et EGO. Il ressort de ces applications que, mis à part les difficultés liées à l'automatisation de la chaîne de simulation numérique, les algorithmes d'optimisation bayésiens sont capables de fournir simplement une réponse satisfaisante à une grande variété de problèmes d'optimisation rencontrés. Cette réponse est sans doute éloignée de l'optimum, mais elle reste meilleure et est obtenue bien plus rapidement qu'à l'aide des d'algorithmes plus traditionnels. Ces résultats ont démontré la possibilité de traiter des problèmes d'optimisation aux caractéristiques variées, multi-objectif, de grande dimension (supérieure à 30), contraint, en présence de bruit sur les résultats des évaluations et enfin en présence d'erreurs dans le processus de simulation.

La qualité de ces résultats provient directement de la principale conclusion du chapitre 4, à savoir qu'une surestimation des paramètres de régularité et de portée de la covariance ne dégrade pas beaucoup les performances de l'optimisation. Ce constat nous a en effet permis de nous affranchir du problème de l'estimation des paramètres de la covariance, et de les choisir *a priori* suivant une règle empirique qui s'est avérée satisfaisante pour les problèmes que nous avons considérés.

Perspectives

Les travaux présentés dans ce mémoire ne sont que le point de départ de la construction d'un outil pour l'optimisation de fonctions coûteuses, utilisable pour différents types de problèmes par un utilisateur novice en matière d'optimisation. La mise au point d'un tel outil nécessite en effet encore de répondre à de nombreuses questions.

Tout d'abord, nous avons discuté de l'utilisation pratique de IAGO ou de EGO dans de nombreux contextes, mais pour la plupart de ceux-ci, des questions restent en suspens. En particulier, nous n'avons pas eu l'occasion de traiter un problème industriel d'optimisation robuste. Il importe aussi d'étudier avec davantage d'attention l'impact du passage à des problèmes multi-objectifs ou de l'apparition d'erreurs dans le processus de simulation.

Deuxièmement, le mode d'estimation de l'entropie conditionnelle des minimiseurs doit encore faire l'objet d'améliorations. Nous avons en effet proposé dans le chapitre 2, un estimateur simple reposant sur une estimation de la distribution des minimiseurs. L'utilisation d'un estimateur plus efficace permettrait peut-être de diminuer le nombre de simulations conditionnelles nécessaires.

Sur un plan plus technique, le paramétrage d'une méthode flexible et automatique nécessite des règles, nécessairement empiriques (mais reposant sur des considérations théoriques), notamment pour le choix de l'ensemble des points candidats à l'évaluation (cardinal et disposition) ou pour

les paramètres de la covariance (choix *a priori*, estimation, ou mélange des deux approches). Ces problèmes ont été abordés, mais ces questions restent ouvertes et les solutions proposées doivent être testées plus avant.

Sur le plan de l'utilisation pratique, insistons une fois de plus sur les difficultés inhérentes au couplage entre l'optimisation et la simulation numérique, en particulier pour l'optimisation de paramètres de forme à l'aide de solveurs aux éléments finis. Ce problème est aujourd'hui au centre des préoccupations des éditeurs de logiciels de CAO et devrait être résolu dans les prochaines années. Cependant, il est clair que cela reste un frein majeur à la mise en place dans l'industrie des méthodes proposées.

Enfin, concluons sur l'intérêt, pour la conception de systèmes complexes à simuler, de l'utilisation conjointe des deux idées sous-jacentes à IAGO : la prédiction par krigeage et l'approche SUR. En effet, le krigeage, de par sa simplicité, ses fondements théoriques et sa flexibilité (c'est-à-dire la possibilité de modéliser avec la même simplicité des systèmes observés avec ou sans bruit, des systèmes comportant des sorties corrélées ou encore d'estimer des dérivées) nous apparaît comme un outil de choix dans ce contexte. L'approche SUR nous semble, quant à elle, adaptée aux problèmes d'apprentissage avec un budget réduit en simulations (comme en témoignent les utilisations qu'en ont faites, dans des contextes très différents, Vergassola et al., 2007 ; Geman et Jedynak, 1995) et en particulier aux problèmes soulevés par la conception. Grâce à la prédiction par krigeage, nous l'avons mise en œuvre pour l'optimisation. Vazquez et Piera Martinez (2006) l'utilisent pour l'estimation de la probabilité de défaillance d'un système.

ANNEXE A

PRÉDICTION PAR KRIGEAGE

A.1 Choix d'une fonction de covariance

Choisir une fonction de covariance $k(\cdot,\cdot)$ adaptée à la modélisation d'une fonction donnée f est une question centrale à l'utilisation du krigeage. Cependant, comme nous l'avons vu au chapitre 4, les algorithmes proposés sont relativement robustes à une surestimation des paramètres de la covariance de Matèrn. Cette section sera donc assez brève, et nous renvoyons aux références qu'elle contient pour davantage de détails.

A.1.1 Classes de covariances

La théorie asymptotique du krigeage (Stein, 1999) montre l'importance du comportement de la covariance à l'origine. Ce comportement est en effet lié à la dérivabilité en moyenne quadratique du processus. Par exemple, si la fonction de covariance est continue à l'origine, alors le processus est continu en moyenne quadratique. Ainsi, les covariances classiquement utilisées (telles l'exponentielle $h \mapsto \sigma^2 \exp(-\theta|h|^\alpha)$, le produit d'exponentielles, ou la polynomiale), infiniment dérivables à l'origine ne permettent pas la description de processus dont les trajectoires ne sont pas analytiques. C'est pourquoi Stein (1999) conseille l'utilisation de la classe de covariance de Matérn, qui offre la possibilité de contrôler la régularité à l'origine à l'aide d'un unique paramètre. Cette classe de fonctions peut être paramétrée de la manière suivante :

$$k(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}h}{\rho} \right)^{\nu} \mathcal{K}_\nu \left(\frac{2\nu^{1/2}h}{\rho} \right), \tag{A.1}$$

avec ν le paramètre qui contrôle la régularité, ρ qui représente la portée et σ^2 la variance ($k(0) = \sigma^2$). La figure A.1 présente l'influence de ν sur la fonction de covariance et la figure A.2 son influence sur les trajectoires. Un choix soigneux des paramètres de la covariance semble donc indispensable pour une modélisation de qualité.

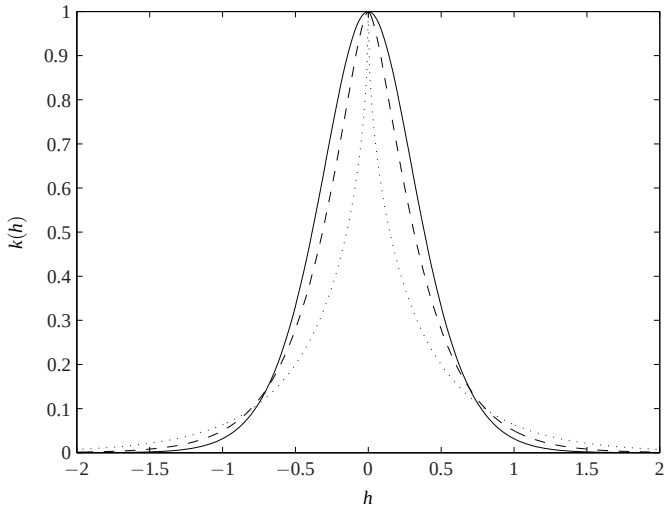


FIG. A.1: *Covariance de Matérn avec* $\rho = 0.5$, $\sigma^2 = 1$. *Le trait plein correspond à* $\nu = 4$, *le trait mixte à* $\nu = 1$ *et le trait en pointillés à* $\nu = 0.25$.

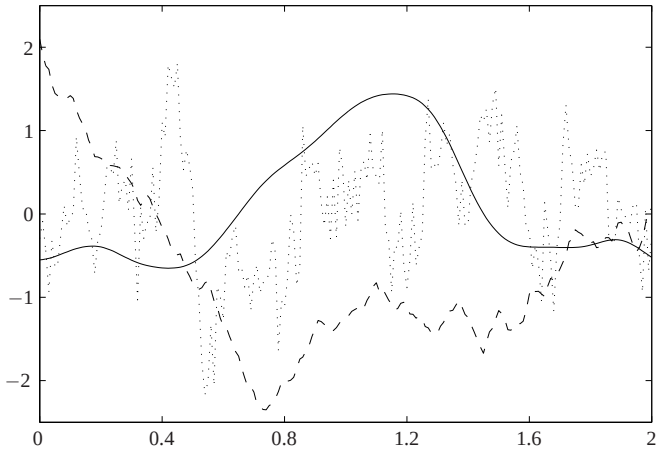


FIG. A.2: *Trois trajectoires d'un processus gaussien de moyenne nulle muni d'une covariance de Matérn. Les conventions graphiques sont identiques à celles de la figure A.1 :* $\nu = 4$ *pour le trait plein,* $\nu = 1$ *pour le trait mixte et* $\nu = 0.25$ *pour le trait en pointillés.*

A.1.2 Estimation des paramètres

Les paramètres, pour une classe de covariance donnée, peuvent être choisis *a priori* ou estimés à partir de données expérimentales. En géostatistique, cette estimation est réalisée en étudiant la correspondance entre la covariance empirique et sa modélisation (Chilès et Delfiner, 1999). Dans d'autres domaines, c'est la méthode de *validation croisée* (Wahba, 1998) ou la méthode du maximum de vraisemblance (Stein, 1999) qui sont les plus employées. Il n'existe pas à notre connaissance de preuve de la supériorité d'une approche sur les deux autres. Aussi utiliserons nous, pour sa simplicité et sa généralité, la méthode du maximum de vraisemblance.

Maximum de vraisemblance

Considérons la loi jointe du vecteur gaussien des observations $\mathbf{F}_n$ et supposons la moyenne $m(\mathbf{x})$ de F connue. L'estimateur $\hat{\boldsymbol{\theta}}$ par maximum de vraisemblance du vecteur $\boldsymbol{\theta}$ des paramètres de la covariance s'obtient alors en maximisant la log-vraisemblance

$$l(\mathbf{f_n}, \boldsymbol{\theta}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log\det \mathbf{K}(\boldsymbol{\theta}) - \frac{1}{2}\mathbf{f}_n^{\mathsf{T}}\mathbf{K}(\boldsymbol{\theta})^{-1}\mathbf{f}_n\,. \tag{A.2}$$

Lorsque la moyenne de $F(\mathbf{x})$ est inconnue, on peut par exemple utiliser, et c'est le cas pour les exemples de ce mémoire, le maximum de vraisemblance restreint (Stein, 1999).

Maximum de vraisemblance restreint

Supposons que la moyenne de F s'écrive $m(\mathbf{x}) = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{p}(\mathbf{x})$. Si le vecteur $\boldsymbol{\beta}$ n'est pas connu *a priori*, il est néanmoins possible d'estimer $\boldsymbol{\theta}$ en considérant la vraisemblance d'un vecteur de contrastes, c'est-à-dire d'une combinaison linéaire des observations dont la distribution jointe ne dépend pas de $\boldsymbol{\beta}$. De nombreuses simulations numériques (cf. par exemple Tunicliffe-Wilson, 1989) ont démontré la supériorité de cette approche, proposée initialement par Patterson et Thompson (1971), sur une estimation conjointe, par maximum de vraisemblance, de $\boldsymbol{\beta}$ et $\boldsymbol{\theta}$.

Pour calculer la vraisemblance des contrastes, considérons

$$\mathbf{Y}_n = (\mathbf{I}_n - \mathbf{P}(\mathbf{P}^{\mathsf{T}}\mathbf{P})^{-1}\mathbf{P}^{\mathsf{T}})\mathbf{F}_n.$$

$\mathbf{Y}_n$ est gaussien, et il est aisé de vérifier que sa moyenne est nulle. En revanche, obtenir sa covariance et la log-vraisemblance qui en découle se révèle plus fastidieux. McCullagh et Nelder (1989) l'écrivent sous la forme suivante :

$$\begin{aligned} l(\mathbf{y}_n, \boldsymbol{\theta}) = -\frac{n-l}{2}\log(2\pi) - \frac{1}{2}\log\det\{\mathbf{K}(\boldsymbol{\theta})\} - \frac{1}{2}\log\det\{\mathbf{W}(\boldsymbol{\theta})\} \\ - \frac{1}{2}\mathbf{y}_n^{\mathsf{T}}\{\mathbf{K}(\boldsymbol{\theta})^{-1} - \mathbf{K}(\boldsymbol{\theta})^{-1}\mathbf{P}\mathbf{W}(\boldsymbol{\theta})^{-1}\mathbf{P}^{\mathsf{T}}\mathbf{K}(\boldsymbol{\theta})^{-1}\}\mathbf{y}_n^{\mathsf{T}}, \end{aligned} \tag{A.3}$$

avec $\mathbf{W}(\boldsymbol{\theta}) = \mathbf{P}^{\mathsf{T}}\mathbf{K}(\boldsymbol{\theta})^{-1}\mathbf{P}$. L'estimation de $\boldsymbol{\theta}$ par maximum de vraisemblance restreint est ensuite obtenue par maximisation de (A.3).

A.2 Prédiction d'un processus convolué

Pour prédire $M(\mathbf{x}) = F * p_{\boldsymbol{\varepsilon}}(\mathbf{x})$, nous avons vu au chapitre 3 qu'il était nécessaire de calculer ou d'approcher

$$k_M(\mathbf{x},\mathbf{y}) = [k(\cdot,\cdot) * (p_{\boldsymbol{\varepsilon}}(\cdot)(p_{\boldsymbol{\varepsilon}}(\cdot))](\mathbf{x},\mathbf{y}),$$

et

$$k_{MF}(\mathbf{x},\mathbf{y}) = k(\cdot,\mathbf{y}) * p_{\boldsymbol{\varepsilon}}(\mathbf{x}).$$

Supposons le bruit $\boldsymbol{\varepsilon}$ gaussien de loi

$$p_{\boldsymbol{\varepsilon}}(\mathbf{u}) = \frac{1}{(2\pi)^{d/2}\sqrt{|\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}|}} \exp^{-\frac{1}{2}\mathbf{u}^\top \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \mathbf{u}}$$

avec $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ la matrice de covariance supposée diagonale (notons $\sigma^2_{\boldsymbol{\varepsilon},i}$ les éléments diagonaux). Dans cette section, nous décrivons un moyen d'approcher k_M et k_{MF} lorsque k est une covariance de Matérn.

Commençons par le cas plus simple où k est une covariance gaussienne ($k(h) = \sigma^2 \exp^{-\frac{h^2}{\rho^2}}$). k_{MF} est alors gaussienne, non isotrope, et s'écrit à l'issue de quelques manipulations simples

$$k_{MF}(\mathbf{x},\mathbf{y}) = \sigma^2_{MF} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{y})^\top \mathbf{R}^{-1}_{MF}(\mathbf{x}-\mathbf{y})\right),$$

avec

$$\sigma^2_{MF} = \frac{\sigma^2}{\prod_{i=1}^d \sqrt{\frac{2\sigma_{\boldsymbol{\varepsilon},i}}{\rho^2}+1}}, \tag{A.4}$$

et $\mathbf{R}^{-1}_{MF}$ une matrice diagonale d'éléments diagonaux

$$\frac{1}{\rho^2}\left(1 - \frac{1}{1+\frac{\rho^2}{2\sigma_{\boldsymbol{\varepsilon},i}}}\right) \quad \forall i \in [\![1:d]\!]. \tag{A.5}$$

$k_M(\mathbf{x},\mathbf{y}) = k_{MF}(\cdot,\mathbf{y}) * p_{\boldsymbol{\varepsilon}}(\mathbf{x})$ est elle aussi gaussienne, en tant que produit de convolution d'une covariance gaussienne, et ses paramètres s'obtiennent par des expressions similaires à (A.4) et (A.5).

Si l'on suppose désormais que k est une covariance de Matèrn de paramètres $\boldsymbol{\theta} = [\nu, \rho, \sigma]$, il n'est plus possible d'obtenir d'expression analytique pour k_M et k_{MF}. En revanche, il est légitime de tenter de les approcher par deux autres covariance de Matèrn de paramètres $\boldsymbol{\theta}_M = \{\nu_M, \mathbf{R}_M, \sigma_M\}$ et $\boldsymbol{\theta}_{MF} = \{\nu_{MF}, \mathbf{R}_{MF}, \sigma_{MF}\}$. Remarquons que pour faire face à l'anisotropie de ces covariances, la portée ρ a été remplacée par $\mathbf{R}$, une matrice diagonale rendant compte des disparités de portée suivant les directions. Ainsi, l'approximation de $k_{MF}(\mathbf{x},\mathbf{y})$ s'écrit

$$\frac{\sigma^2_{MF}}{2^{\nu_{MF}-1}\Gamma(\nu_{MF})}\left(2\sqrt{\nu_{MF}(\mathbf{x}-\mathbf{y})^\top \mathbf{R}^{-1}_{MF}(\mathbf{x}-\mathbf{y})}\right)^{\nu_{MF}} \mathcal{K}_\nu\left(2\sqrt{\nu_{MF}(\mathbf{x}-\mathbf{y})^\top \mathbf{R}^{-1}_{MF}(\mathbf{x}-\mathbf{y})}\right).$$

Reste à déterminer comment $\boldsymbol{\theta}_M$ et $\boldsymbol{\theta}_{MF}$ vont être estimés.

Tout d'abord, il faut remarquer que lorsque la régularité ν tend vers l'infini, une covariance de Matern de paramètres $\{\nu, \mathbf{R}, \sigma\}$ tend vers une covariance gaussienne $\sigma^2 \exp\{-(\mathbf{x}-\mathbf{y})^T \mathbf{R}^{-1}(\mathbf{x}-\mathbf{y})\}$. Ainsi, si l'on suppose que k_{MF} est effectivement une covariance de Matèrn, alors σ_{MF} et $\mathbf{R}_{MF}$ s'obtiennent grâce à (A.4) et (A.5). La régularité ν_{MF} est ensuite choisie pour que le hessien à l'origine de l'approximation de k_{MF} corresponde à celui estimé numériquement. En effet, le hessien d'une covariance de Matérn de paramètres $[\nu, \mathbf{R}, \sigma]$ s'écrit à l'origine comme

$$-2\sigma^2 \frac{\nu}{\nu - 1} \mathbf{R}^{-1}.$$

(Cette relation est obtenue, après calcul, grâce à deux propriétés des fonctions de Bessel

$$\dot{\mathcal{K}}_\nu(u) = -\mathcal{K}_{\nu-1}(u) - \frac{\nu}{u}\mathcal{K}_\nu(u) \ \forall u > 0, \tag{A.6}$$

et

$$u^{\nu-1}\mathcal{K}_{\nu-1}(u) \xrightarrow[u\to 0]{} 2^{\nu-1}\Gamma(\nu-1). \tag{A.7}$$

$k_M(\mathbf{x},\mathbf{y}) = k_{MF}(\cdot,\mathbf{y}) * p_{\boldsymbol{\varepsilon}}(\mathbf{x})$ est ensuite obtenue en itérant le processus. La figure A.3 présente l'approximation de k_{MF} ainsi obtenue.

Avec cette approche, il est donc possible de prédire M à l'aide d'une covariance de Matèrn, sans augmentation significative de la complexité, puisque k_M et k_{MF} peuvent être déterminées *a priori* moyennant une estimation numérique de leur hessien à l'origine.

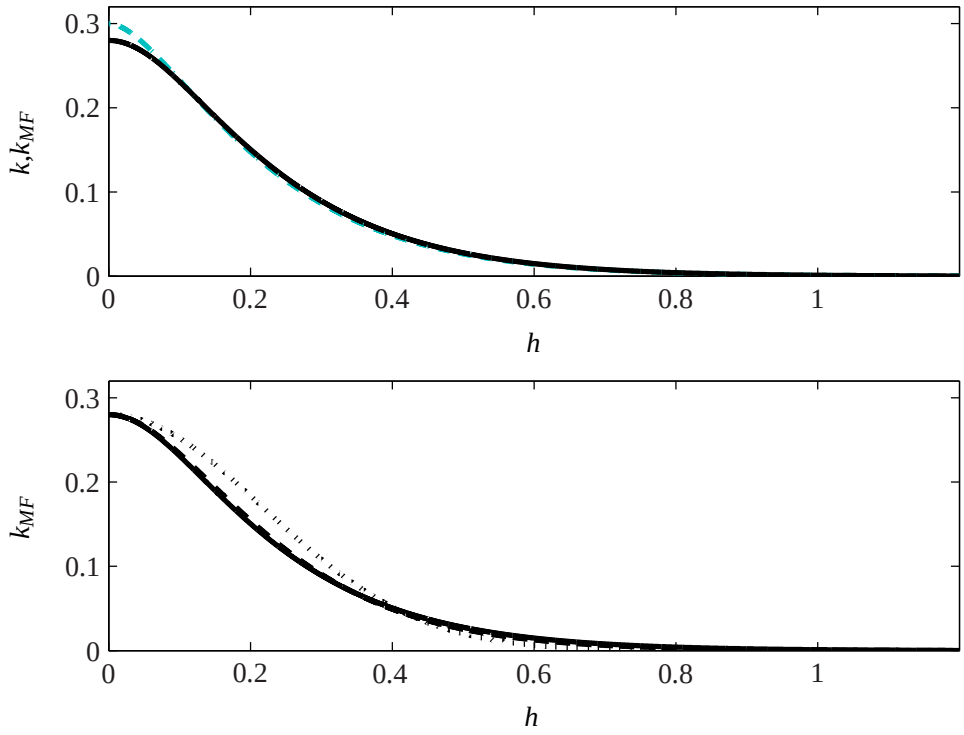


FIG. A.3: *(*partie supérieure*) k (trait pointillé), et* k_{MF} *(trait plein). (*partie inférieure*)* k_{MF} *(trait plein), approximation gaussienne de* k_{MF} *dont les paramètres sont donnés par* (A.4) *et* (A.5) *(trait en pointillés courts), et approximation par une covariance de Matèrn (trait en pointillés longs).*

ANNEXE B

EXTENTION DU KRIGEAGE À LA PRÉDICTION DE PLUSIEURS PHÉNOMÈNES CORRÉLÉS

B.1 Cokrigeage

La méthode appelée *cokrigeage* (voir par exemple Chilès et Delfiner, 1999) permet d'effectuer une prédiction lorsque l'on dispose d'observations de plusieurs phénomènes corrélés. Considérons ainsi q fonctions $f_1, ..., f_q$ définies sur $\mathbb{X}$. On peut par exemple vouloir prédire $f_\alpha(\mathbf{x})$ ($\mathbf{x} \in \mathbb{X}$) à partir de résultats d'évaluations $f_{\alpha_i}(\mathbf{x}_i)$, $i = 1, ..., n$ ($\alpha, \alpha_i, ..., \alpha_n \in [\![1 : q]\!]$). Le principe reste identique au cas où la fonction observée est la fonction prédite. Les fonctions $f_1, ..., f_q$ sont modélisées par q processus $F_1, ..., F_q$ gaussiens de moyennes nulles [1] et l'on cherche parmi les éléments de $\text{vect}\{F_{\alpha_1}(\mathbf{x}_1), ..., F_{\alpha_n}(\mathbf{x}_n)\}$ le prédicteur non biaisé à variance minimale. Ce dernier peut s'écrire

$$\hat{F}_\alpha(\mathbf{x}) = \sum_{i=1}^{n} \lambda_{\alpha,i}(\mathbf{x}) F_{\alpha_i}(\mathbf{x}_i),$$

et le vecteur $\boldsymbol{\lambda}_\alpha(\mathbf{x})$ des coefficients s'obtient par résolution du système linéaire

$$\begin{pmatrix} k_{\alpha_1,\alpha_1}(\mathbf{x}_1,\mathbf{x}_1) & k_{\alpha_1,\alpha_2}(\mathbf{x}_1,\mathbf{x}_2) & \dots & k_{\alpha_1,\alpha_n}(\mathbf{x}_1,\mathbf{x}_n) \\ k_{\alpha_2,\alpha_1}(\mathbf{x}_2,\mathbf{x}_1) & k_{\alpha_2,\alpha_2}(\mathbf{x}_2,\mathbf{x}_2) & \dots & k_{\alpha_2,\alpha_n}(\mathbf{x}_2,\mathbf{x}_n) \\ \vdots & \vdots & \ddots & \cdots \\ k_{\alpha_n,\alpha_1}(\mathbf{x}_n,\mathbf{x}_1) & k_{\alpha_n,\alpha_2}(\mathbf{x}_n,\mathbf{x}_2) & \dots & k_{\alpha_n,\alpha_n}(\mathbf{x}_n,\mathbf{x}_n) \end{pmatrix} \boldsymbol{\lambda}_\alpha(\mathbf{x}) = \begin{pmatrix} k_{\alpha,\alpha_1}(\mathbf{x},\mathbf{x}_1) \\ k_{\alpha,\alpha_2}(\mathbf{x},\mathbf{x}_2) \\ \vdots \\ k_{\alpha,\alpha_n}(\mathbf{x},\mathbf{x}_n) \end{pmatrix}, \tag{B.1}$$

où $k_{i,j}(\cdot,\cdot)$ ($(i,j) \in [\![1,q]\!]^2$) désigne la fonction de covariance entre F_i et F_j. On obtient ensuite la variance de l'erreur de prédiction comme

$$\mathbb{E}\left[F_\alpha(\mathbf{x}) - \hat{F}_\alpha(\mathbf{x})\right]^2 = k_{\alpha,\alpha}(\mathbf{x},\mathbf{x}) - \boldsymbol{\lambda}_\alpha(\mathbf{x})^\mathsf{T} \mathbf{k}_\alpha(\mathbf{x}), \tag{B.2}$$

[1] Cette hypothèse est faite pour simplifier les équations, il est là encore possible d'écrire ces moyennes comme combinaisons linéaires de fonctions connues.

avec

$$\mathbf{k}_\alpha(\mathbf{x}) = \begin{pmatrix} k_{\alpha,\alpha_1}(\mathbf{x},\mathbf{x}_1) \\ k_{\alpha,\alpha_2}(\mathbf{x},\mathbf{x}_2) \\ \vdots \\ k_{\alpha,\alpha_n}(\mathbf{x},\mathbf{x}_n) \end{pmatrix}.$$

Remarque B.1. Remarquons que l'on se ramène très simplement au cas du krigeage en interprétant l'indice α comme un paramètre supplémentaire. On s'intéresse alors au processus $F(\alpha,\mathbf{x})$ (défini sur $[\![1,q]\!]\times\mathbb{X}$) et à sa covariance $k([i,\mathbf{x}],[j,\mathbf{y}])$.

Cette approche est naturellement intéressante quand le système à modéliser comporte plusieurs sorties et que l'on souhaite utiliser d'éventuelles corrélations entre elles. On peut en particulier citer, dans le cadre de l'optimisation de fonctions coûteuses, l'utilisation simultanée de modèles physiques de qualités différentes. Par exemple, les simulations d'écoulement utilisées pour l'optimisation du conduit d'admission (section 5.3) ont été menées jusqu'à parfaite convergence. Mais si l'on avait choisi d'arrêter la simulation après quatre fois moins d'itérations du solveur, le résultat aurait été relativement proche de la solution après convergence, en tout cas suffisamment pour être utilisable pour l'optimisation. En allant plus loin, pourquoi ne pas utiliser ces simulations incomplètes pour quadrupler artificiellement le budget en évaluations ? L'appel aux simulations complètes peut alors servir uniquement à la fin de la recherche pour identifier avec précision un minimiseur global. L'utilisation du cokrigeage pour l'optimisation dans ce contexte, déjà proposée chez Jones et al. (1998), connaît actuellement un engouement important sous le nom d'optimisation multi-niveaux (ou *multi fidelity* en anglais, cf. Huang et Allen, 2005 ; Gano et al., 2005)).

Cependant, malgré des travaux prometteurs (Kennedy et O'Hagan, 2000), il nous apparaît aujourd'hui que ce type d'approche est inapplicable dans le contexte d'un budget d'évaluation réduit. En effet, le choix de la covariance, déjà complexe pour la prédiction d'un processus unique, devient réellement problématique puisqu'il faut maintenant modéliser les corrélations entre les processus, ce qui implique de nombreux paramètres supplémentaires à choisir ou à estimer. Le budget en évaluations a beau augmenter avec l'utilisation des modèles peu coûteux, le nombre de paramètres à estimer augmente lui aussi, et sans doute dans des proportions trop importantes. Ce constat est à nuancer si la connaissance *a priori* des relations entre les différents niveaux de simulation est importante (sans doute rare en pratique), ou si l'on ne s'intéresse qu'à deux ou trois de ces niveaux.

Dans la suite, nous nous intéresserons à d'autres applications du cokrigeage, qui, cette fois, ne nécessitent pas l'estimation de paramètres supplémentaires.

Remarque B.2. Si l'on souhaite générer des simulations conditionnelles de F_α ($\alpha\in[\![1:q]\!]$) sur $\mathbb{G}$, la procédure de conditionnement par krigeage décrite dans la section 2.2.1 s'applique encore, à la différence qu'il est désormais nécessaire de générer des simulations non conditionnelles de

F_α sur $\mathbb{G}$, mais aussi de les générer conjointement à des simulations de $F_{\alpha_i}(\mathbf{x}_i), i = 1,...,n$. Ainsi, si l'on considère $Z_1,...,Z_q$ des processus gaussiens de moyennes nulles de même loi jointe que $F_1,...,F_q$, il est possible de générer $t_\alpha(\mathbf{x})$ une simulation conditionnelle de $F_\alpha(\mathbf{x})$ ($\forall \mathbf{x} \in \mathbb{G}$), à partir d'une simulation non conditionnelle $z_\alpha(\mathbf{x})$ de F_α conjointe aux simulations $z(\alpha_i)(\mathbf{x}_i)$ de $Z(\alpha_i)(\mathbf{x}_i)$ ($i = 1,...,q$), en appliquant

$$t_\alpha(\mathbf{x}) = z_\alpha(\mathbf{x}) + \begin{pmatrix} \lambda_{\alpha,1}(\mathbf{x}) & \dots & \lambda_{\alpha,n}(\mathbf{x}) \end{pmatrix} \begin{pmatrix} z_{\alpha_1}(\mathbf{x}_1) - f_{\alpha_1}(\mathbf{x}_1) \\ \vdots \\ z_{\alpha_n}(\mathbf{x}_n) - f_{\alpha_n}(\mathbf{x}_n) \end{pmatrix}. \tag{B.3}$$

B.2 Prédiction à l'aide d'observations bruitées

L'application la plus courante du cokrigeage est la prise en compte d'un bruit additif sur les résultats des évaluations. Supposons en effet que l'évaluation de f en $\mathbf{x}$ produise une réalisation $f^{\text{obs}}(\mathbf{x})$ de la variable aléatoire $F^{\text{obs}}(\mathbf{x}) = F(\mathbf{x}) + b(\mathbf{x})$, où $b(\mathbf{x})$ est un bruit blanc, indépendant de F, de variance $\sigma_b^2(\mathbf{x})$ et de loi connue ou paramétrée (on peut alors estimer ses paramètres conjointement avec ceux de k). Alors la covariance entre F et F^{obs} est simplement $(\mathbf{x},\mathbf{y}) \to k(\mathbf{x},\mathbf{y}) + \delta_{\mathbf{x}=\mathbf{y}}\sigma_b^2(\mathbf{x})$, et les coefficients du prédicteur par krigeage $\hat{F}(\mathbf{x}) = \boldsymbol{\lambda}(\mathbf{x})^\mathsf{T}\mathbf{F}_n^{\text{obs}}$ de F à partir des résultats bruités d'évaluation $\mathbf{F}_n^{\text{obs}}$ s'obtiennent par résolution d'un système linéaire similaire au système (1.7) dans le cas non bruité

$$\begin{pmatrix} \mathbf{K} + \boldsymbol{\sigma}_b^2 & \mathbf{P} \\ \mathbf{P}^\mathsf{T} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}(\mathbf{x}) \\ \boldsymbol{\mu}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathbf{k}(\mathbf{x}) \\ \mathbf{p}(\mathbf{x}) \end{pmatrix}, \tag{B.4}$$

avec $\boldsymbol{\sigma}_b^2$ une matrice diagonale contenant $\sigma_b^2(\mathbf{x}_1),...,\sigma_b^2(\mathbf{x}_n)$. La variance de l'erreur de prédiction s'obtient toujours grâce à (1.8).

Remarque B.3. Cette méthode se généralise simplement au cas d'un bruit d'observation coloré. Il suffit en effet de connaître, ou d'estimer la covariance du bruit et d'appliquer les équations du cokrigeage (cf. par exemple Vazquez et Walter, 2005).

B.3 Prédiction jointe de f et de ses dérivées

Une application importante du cokrigeage est la possibilité de prédire conjointement une fonction et ses dérivées (cf. par exemple Vazquez et Walter, 2005). Pour décrire le principe de l'approche, supposons pour simplifier la présentation que $\mathbb{X} \subset \mathbb{R}$. L'extension au cas multidimensionnel ne pose pas de difficultés particulières et sera présenté dans le cas de la prédiction des composantes du gradient à partir d'observations de f.

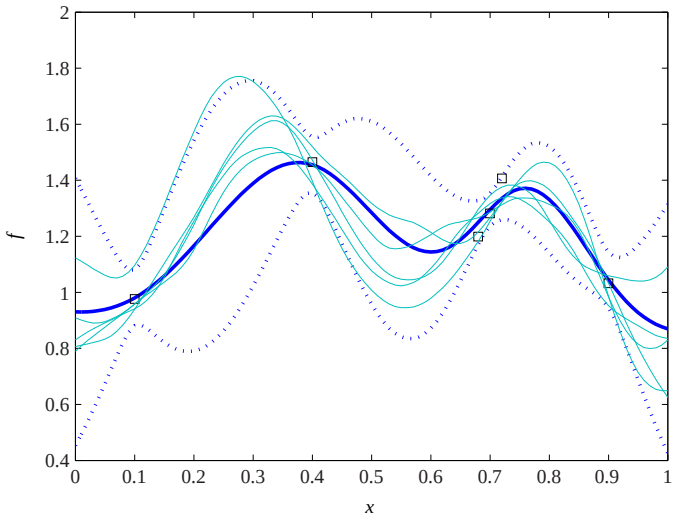


FIG. B.1: *Prédiction par krigeage en présence de bruit sur les résultats des évaluations. La prédiction (trait gras) n'interpole plus les données (carrés) et les intervalles de confiance à 95% (matérialisés par les traits pointillés) ne se réduisent plus à une valeur unique aux points d'évaluation. Les traits fins présentent quelques exemples de simulations conditionnelles de F.*

Notons, si elle existe, $F^{(j)}$ la dérivée j-ième de F en moyenne quadratique[2], il est immédiat de vérifier que

$$\forall (i,j) \in \mathbb{N}^2,\ \mathrm{cov}[F^{(i)}(\mathbf{x}), F^{(j)}(\mathbf{y})] = (-1)^r k^{i+j}(\mathbf{x}-\mathbf{y}). \tag{B.5}$$

Grâce à cette relation entre les dérivées de F, il est possible d'utiliser les équations du cokrigeage pour prédire une dérivée de f à partir d'évaluations simultanées de f et de ses dérivées[3].

En pratique, cette technique est surtout utile pour la prédiction du gradient et pour la prédiction à l'aide de résultats d'évaluation du gradient. Ce dernier cas de figure est utilisé en particulier pour prendre en compte des connaissances *a priori* sur la physique du système.

Remarque B.4. Notons que le cokrigeage permet d'estimer les dérivées de signaux bruités, ce qui est un problème délicat en traitement du signal.

Détail des équations pour la prédiction du gradient

Dans cette section, nous abandonnons l'hypothèse unidimensionnelle et détaillons les calculs nécessaires à la prédiction des composantes du gradient à l'aide de résultats d'évaluations de f lorsque la covariance utilisée est une covariance de Matèrn. Ces calculs peuvent être aisément

[2] La dérivabilité à l'ordre j en moyenne quadratique est assurée par une fonction de covariance $j+1$ fois dérivable sur la diagonale. Si F possède, et c'est le cas ici, une covariance de Matèrn, il suffit de choisir $\nu > j+1$.

[3] Dans le cas multidimensionnel, une relation de ce type demeure, mais sa forme générale est nettement plus lourde.

adaptés à la prédiction de f à partir d'évaluations simultanées de f et de son gradient (les équations qui suivent s'en trouvent cependant bien alourdies).

Soit $s \in [\![1:d]\!]$, pour toute fonction dérivable $g : \mathbb{X} \to \mathbb{R}$, notons $\nabla g_{[s]} := \frac{\partial g}{\partial x_{[s]}}$ la dérivée partielle de g par rapport à la s-ième composante de $\mathbf{x}$. De même, pour un processus G indexé sur $\mathbb{X}$ et dérivable en moyenne quadratique, notons $\nabla G_{[s]}$ la s-ième composante du gradient ∇G de G. La prédiction par krigeage de $\nabla F_{[s]}(\mathbf{x})$ à partir de $\mathbf{F}_n$ s'obtient alors par simple application de (B.1) comme

$$\nabla \hat{F}_{[s]}(\mathbf{x}) = \nabla \boldsymbol{\lambda}_{[s]}(\mathbf{x})^{\mathsf{T}} \mathbf{F}_n, \tag{B.6}$$

avec le vecteur $\nabla \boldsymbol{\lambda}_{[s]}(\mathbf{x})$ solution du système linéaire

$$\begin{pmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^{\mathsf{T}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \nabla \boldsymbol{\lambda}_{[s]}(\mathbf{x}) \\ \nabla \boldsymbol{\mu}_{[s]}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \nabla \mathbf{k}_{[s]}(\mathbf{x}) \\ \nabla \mathbf{p}_{[s]}(\mathbf{x}) \end{pmatrix}, \tag{B.7}$$

où $\nabla \boldsymbol{\mu}_{[s]}(\mathbf{x})$ est un vecteur de coefficients de Lagrange et

$$\nabla \mathbf{k}_{[s]}(\mathbf{x}) = [\mathrm{cov}(F(\mathbf{x}_1), \nabla F_{[s]}(\mathbf{x})), ..., \mathrm{cov}(F(\mathbf{x}_n), \nabla F_{[s]}(\mathbf{x}))]^{\mathsf{T}} \tag{B.8}$$

est le vecteur des covariances entre $\nabla F_{[s]}(\mathbf{x})$ et $\mathbf{F}_{\mathbb{S}}$.

La notation $\nabla \mathbf{k}_{[s]}(\mathbf{x})$ est cohérente avec notre notation des dérivées. En effet, il est aisé de vérifier que

$$\mathrm{cov}(F(\mathbf{x}), \nabla F_{[s]}(\mathbf{y})) = \frac{\partial k}{\partial y_{[s]}}(\mathbf{x}, \mathbf{y}) \ \forall (\mathbf{x}, \mathbf{y}) \in \mathbb{X}^2. \tag{B.9}$$

Ainsi, $\nabla \mathbf{k}_{[s]}(\mathbf{x})$ est effectivement la dérivée de $\mathbf{k}(\mathbf{x})$. Plus généralement, (B.7) peut être obtenue directement en dérivant les équations du krigeage (1.7), et $\nabla \boldsymbol{\lambda}_{[s]}(\mathbf{x})$ est effectivement la dérivée de $\boldsymbol{\lambda}(\mathbf{x})$ (cette correspondance fonctionne aussi pour $\nabla \boldsymbol{\mu}_{[s]}(\mathbf{x})$, $\nabla \mathbf{p}_{[s]}(\mathbf{x})$ et $\nabla \hat{F}_{[s]}$).

On peut aussi vérifier aisément que

$$\mathrm{cov}(\nabla F_{[s]}(\mathbf{x}), \nabla F_{[t]}(\mathbf{y})) = \frac{\partial^2 k}{\partial x_{[s]} \partial y_{[t]}}(\mathbf{x}, \mathbf{y}) \ \forall (\mathbf{x}, \mathbf{y}) \in \mathbb{X}^2 \text{ et } \forall 1 \leq s, t \leq d. \tag{B.10}$$

La variance de l'erreur de prédiction s'obtient donc par

$$\mathbb{E}\left[\nabla F_{[s]}(\mathbf{x}) - \nabla \hat{F}_{[s]}(\mathbf{x})\right]^2 = \frac{\partial^2 k}{\partial^2 x_{[s]}}(\mathbf{x}, \mathbf{x}) - \nabla \boldsymbol{\lambda}_{[s]}(\mathbf{x})^{\mathsf{T}} \nabla \mathbf{k}_{[s]}(\mathbf{x}) - \nabla \mathbf{p}_{[s]}(\mathbf{x})^{\mathsf{T}} \nabla \boldsymbol{\mu}_{[s]}(\mathbf{x}). \tag{B.11}$$

Le calcul des coefficients du krigeage par (B.7) et de la variance de l'erreur de prédiction par (B.9) (la figure B.2 présente un exemple en dimension un) requiert donc le calcul du gradient et du hessien de la covariance $k(\cdot,\cdot)$ de F, que nous avons supposée isotrope et appartenant à la classe des covariances de Matèrn. Rappelons que cette dernière peut s'écrire en fonction de $r = \frac{2\nu^{1/2}\|\mathbf{x}-\mathbf{y}\|}{\rho}$,

$$k(\mathbf{x}, \mathbf{y}) = k(r) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} r^{\nu} \mathcal{K}_{\nu}(r).$$

Ainsi, les composantes du gradient s'écrivent

$$\frac{\partial k}{\partial y_{[s]}}(\mathbf{x},\mathbf{y}) = \frac{\partial r}{\partial y_{[s]}}\dot{k}(r), \tag{B.12}$$

et celles du hessien

$$\frac{\partial^2 k}{\partial x_{[t]} y_{[s]}}(\mathbf{x},\mathbf{y}) = \frac{\partial r}{\partial y_{[s]}}\frac{\partial r}{\partial x_{[t]}}\ddot{k}(r) + \frac{\partial^2 r}{\partial x_{[t]} y_{[s]}}\dot{k}(r). \tag{B.13}$$

A l'aide de la propriété (A.6), on obtient

$$\dot{k}(r) = -\frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} r^{\nu}\mathcal{K}_{\nu-1}(r)$$

et

$$\ddot{k}(r) = -\frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left[r^{\nu}\mathcal{K}_{\nu-2}(r) - r^{\nu-1}\mathcal{K}_{\nu-1}(r)\right].$$

(B.12) devient alors

$$\mathrm{cov}[F(\mathbf{x}), \nabla F_{[s]}(\mathbf{y})] = \frac{\sigma^2 \nu^{1/2}}{2^{\nu-2}\Gamma(\nu)\rho}\frac{x_{[s]} - y_{[s]}}{\| \mathbf{x}-\mathbf{y} \|} r^{\nu}\mathcal{K}_{\nu-1}(r) \ \forall \mathbf{x} \neq \mathbf{y}, \tag{B.14}$$

et l'on définit par continuité $\mathrm{cov}[F(\mathbf{x}), \nabla F_{[s]}(\mathbf{x})] = 0$, puisque $r^{\nu-1}\mathcal{K}_{\nu-1}(r) \xrightarrow[r\to 0]{} 2^{\nu-1}\Gamma(\nu-1)$. De manière similaire, on obtient pour (B.13)

$$\mathrm{cov}[\nabla F_{[t]}(\mathbf{x}), \nabla F_{[s]}(\mathbf{y})] = \frac{(x_{[s]} - y_{[s]})(x_{[t]} - y_{[t]})}{\| \mathbf{x}-\mathbf{y} \|^2}\left[\frac{2\nu^{1/2}}{\rho}\dot{k}(r) - \frac{4\nu}{\rho^2}\ddot{k}(r)\right] \tag{B.15}$$

lorsque $s \neq t$, et

$$\mathrm{cov}[\nabla F_{[s]}(\mathbf{x}), \nabla F_{[s]}(\mathbf{y})] = -\frac{4\nu(x_{[s]} - y_{[s]})^2}{\rho^2 \| \mathbf{x}-\mathbf{y} \|^2}\ddot{k}(r) + \frac{2\nu^{1/2}\left[(x_{[s]} - y_{[s]})^2 - \| \mathbf{x}-\mathbf{y} \|^2\right]}{\rho \| \mathbf{x}-\mathbf{y} \|^3}\dot{k}(r) \tag{B.16}$$

lorsque $s = t$. On obtient ensuite par continuité

$$\mathrm{cov}[\nabla F_{[t]}(\mathbf{x}), \nabla F_{[s]}(\mathbf{x})] = \begin{cases} 0 & \text{si} \quad s \neq t \\ \frac{2\sigma^2\nu}{\rho^2(\nu-1)} & \text{sinon} \end{cases}.$$

Cette dernière expression est suffisante pour calculer la variance de l'erreur de prédiction du gradient.

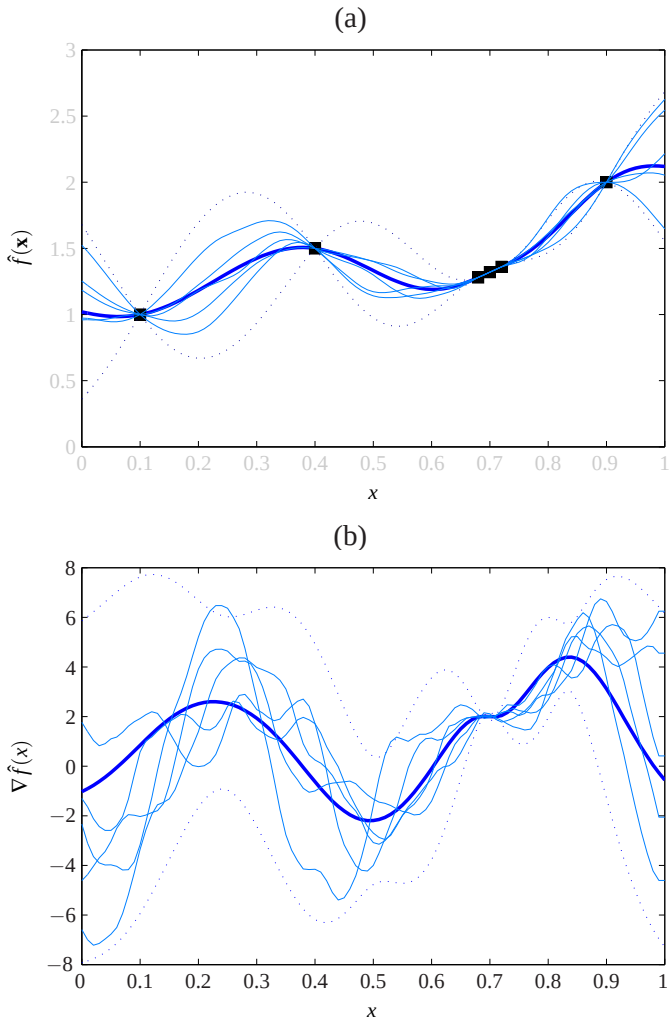


FIG. B.2: *Simulations (traits fins) de* F (a) *et de* $F^{(1)}$ (b) *conditionnellement aux résultats des évaluations de* f *(carrés). Les traits en gras sont les prédictions par krigeage de la fonction et de sa dérivée tandis que les traits pointillés représentent les intervalles de confiances à 95% pour ces prédictions.*

Table des figures

Liste des tableaux

Index

Références bibliographiques

P. Abrahamsen. « A review of Gaussian random fields and correlation functions ». Technical report, Norwegian Computing Center, 1997.

D.H. Ackley. *A Connectionist Machine for Genetic Hill-climbing*. Kluwer Academic Publishers, Norwell, 1987.

R.J. Adler. « On excursion sets, tubes formulas and maxima of random fields ». *Ann. Appl. Probab.*, 10(1) : 1–74, 2000.

D.W. Apley, J. Liu et W. Chen. « Understanding the effects of model uncertainty in robust design with computer experiments ». *J. Mech. Design*, 128 : 945–958, 2006.

R.R. Barton. « Minimization algorithms for functions with random noise ». *Am. J. Math. Manag. Sci.*, 4 : 109–138, 1984.

R. Bates et L. Pronzato. « Emulator-based global optimisation using lattices and Delaunay tesselation ». In P. Prado et R. Bolado (éditeurs), *Proc. 3rd Int. Symp. on Sensitivity Analysis of Model Output*, pp. 189–192, Madrid, June 2001.

J. Bect et E. Vazquez. « On the convergence of the expected improvement algorithm ». http ://fr.arXiv.org/abs/0712.3744, 2007.

J. Beirlant, E. Dudewicz, L. Gyorfi et E van der Meulen. « Nonparametric entropy estimation : An overview ». *Int. J. of Math. and Stat. Sci.*, 6 : 17–39, 1997.

R. Bettinger, P. Duchêne, L. Pronzato et E. Thierry. « Design of experiments for response diversity ». In *Proc. 6th International Conference on Inverse Problems in Engineering (ICIPE), Journal of Physics : Conference Series*, Dourdan (Paris), 2008.

G.E.P. Box et K.B. Wilson. « On the experimental attainment of optimum conditions (with discussion) ». *Journal of the Royal Statistical Society Series B*, 13(1) : 1–45, 1951.

F.H. Branin. « Widely convergent methods for finding multiple solutions of simultaneous nonlinear equations ». *IBM J. Res. Develop.*, 16 : 504–522, 1972.

J.M. Calvin. « A one-dimentional optimization algorithm and its convergence rate under the Wiener measure ». *J. Complexity*, 17 : 306–344, 2001.

J. P. Chilès. « How to adapt kriging to non-classical problems : three case studies ». In M. Guarascio, M. David et C. Huijbregts (éditeurs), *Advanced Geostatistics in the Mining Industry*, pp. 69–89, Dordrecht, Holland, 1976. Reidel.

J. P. Chilès. *Géostatistique des phénomènes non stationnaires (dans le plan)*. Mémoire de thèse, Université de Nancy-I, France, 1977.

J.P. Chilès et P. Delfiner. *Geostatistics, Modeling Spatial Uncertainty*. John Wiley & Sons, Inc, New York, 1999.

T. M. Cover et A. J. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc, New York, 1991.

D. Cox et S. John. « Sdo : a statistical method for global optimization ». In N. Alexandrov et M. Y Hussaini (éditeurs), *Multidisciplinary Design Optimization : State of the Art*, pp. 315–329, Philadelphia, 1997. SIAM.

I. Das. « Robustness optimization for constrained, nonlinear programming problems ». Technical Reports Dept. of Computational & applied Mathematics TR97-01, Rice University, Houston, 1997.

Z. S. Davies, R. J. Gilbert, R.J. Merry, D.B. Kell, M.K. Theodorou et G. W. Grifth. « Efficient improvement of silage additives by using genetic algorithms ». *Applied and Environmental Microbiology*, pp. 1435–1443, 2000.

P. Delfiner. *Shift Invariance Under Linear Models*. Mémoire de thèse, Princetown University, New Jersey, 1977.

T. Dvorak, R. Hoekstra et J. Pet-Armacost. « Improving exhaust header performance with multiple response surface methods ». SAE n˚2003-01-1389, 2003.

J.F. Elder IV. « Global R^d optimization when probes are expensive : the GROPE algorithm ». In *Proceedings of the 1992 IEEE International Conference on Systems, Man. and Cybernetics*, volume 1, pp. 577–582, Chicago, 1992.

D.E. Finkel et C.T. Kelley. « Convergence analysis of the DIRECT algorithm ». Technical Report CRSC-TR04-28, N. C. State University Center for Research in Scientific Computation, July 2004.

A. I. J. Forrester, A. Sóbester et A.J Keane. « Optimization with missing data ». *Proc. R. Soc. A*, 462 : 935–945, 2006.

S.E. Gano, J.E. Renaud, J.D. Martin et T.W. Simpson. « Update strategies for Kriging models for use in variable fidelity optimization ». 1st AIAA Multidisciplinary Design Optimization Specialist Conference, 18 – 21 April, Austin, Texas 2005.

D. Geman et B. Jedynak. « An active testing model for tracking roads in satellite images ». Technical Report 2757, Institut National de Recherche en Informatique et en Automatique (INRIA), December 1995.

K.C. Giannakoglou. « Design of optimal aerodynamic shapes using stochastic optimization methods and computational intelligence ». *International Review Journal Progress in Aerospace Sciences*, 38 : 43–76, 2002.

D. Ginsbourger, R. Le Riche et L. Carraro. « A multi-points criterion for deterministic parallel global optimization based on Gaussian processes ». In *Nonconvex Programming : Local and Global Aproaches (NCP)*, Rouen, France, 2007.

A. Girard. *Approximate Methods for Propagation of Uncertainty with Gaussian Process Models*. Mémoire de thèse, University of Glasgow, 2004.

H.M. Gutmann. « A radial basis function method for global optimization ». *J. Global Optim.*, 19 (3) : 201–227, 2001.

J.K. Hartman. « Some experiments in global optimization ». *Nav. Res. Logist. Q.*, 20 : 569–576, 1973.

R.G. Haylock et A. O'Hagan. « On inference for outputs of computationally expensive computer algorithms with uncertainty on the inputs ». In J.M. Bernardo, J.O. Berger, A.P. David et A.F.M. Smith (éditeurs), *Bayesian Statistics 5*, pp. 223–245, Oxford : University Press, 1996.

P.D.H. Hill. « A review of experimental design procedures for regression model discrimination ». *Technometrics*, 20 : 15–21, 1978.

D. Huang. *Experimental Planning and Sequential Kriging Optimization Using Variable Fidelity Data*. Mémoire de thèse, Ohio State University, 2005.

D. Huang et T. Allen. « Design and analysis of variable fidelity experimentation applied to engine valve heat treatment process design ». *Journal of Royal Statistics, Series C.*, 54, Part 2 : 443–463, 2005.

D. Huang, T. Allen, W. Notz et N. Zeng. « Global optimization of stochastic black-box systems via sequential Kriging meta-models ». *J. Global Optim.*, 34 : 441–466, 2006.

D.R. Jones. « A taxonomy of global optimization methods based on response surfaces ». *J. Global Optim.*, 21 : 345–383, 2001.

D.R. Jones, C.D. Perttunen et B.E. Stuckman. « Lipschitzian optimization without the Lipschitz constant ». *J. Opt. Theor. Appl.*, 79 : 157–181, 1993.

D.R. Jones, M. Schonlau et J. William. « Efficient global optimization of expensive black-box functions ». *J. Global Optim.*, 13 : 455–492, 1998.

A.G. Journel. « The indicator approach to estimation of spatial distributions ». In *Proceedings of the 17th APCOM (Application of Computers and Operations Research in the Mineral Industry) Symposium*. Soc. of Mining Engineers, 1982.

M. C. Kennedy et A. O'Hagan. « Predicting the output of a complex computer code when fast approximations are available ». *Biometrika*, 87 : 1–13, 2000.

J. Knowles. « Parego : A hybrid algorithm with on-line landscape approxiamtion for expensive multiobjective optimization problems ». *IEEE T. Evolut. Comput.*, 7(2) : 100–116, 2003.

J. Knowles, L. Thiele et E. Zitzler. « A tutorial on the performance assessment of stochastive multiobjective optimizers ». Technical Report 214, Computer Engineering and Networks Laboratory, ETH Zurich, February 2006.

D.G. Krige. « A Statistical Approach to Some Mine Valuations and Allied Problems at the Witwatersrand ». Master's thesis, University of Witwatersrand, 1951.

H.J. Kushner. « A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise ». *J. Basic Eng.*, 86 : 97–106, 1964.

J.S. Lehman, T.J. Santner et W.I. Notz. « Designing computer experiments to determine robust control variables ». *Stat. Sinica*, 14 : 571–590, 2004.

J.L. Lumley. *Engines : An Introduction*. Cambridge University Press, Cambridge, 1999.

G. Matheron. « Principles of geostatistics ». *Econ. Geol.*, 58 : 1246–1266, 1963.

G. Matheron. « Le krigeage universel ». In *Cahiers du Centre de Morphologie Mathématique de Fontainebleau*. Ecole des Mines de Paris, 1969.

P. McCullagh et J.A. Nelder. *Generalized Linear Models*. Chapman & Hall, London, 1989.

M.D. McKay, W.J. Conover et R.J. Beckman. « A comparison of three methods for selecting values of input variables in the analysis of output from a computer code ». *Technometrics*, 21 : 239–245, 1979.

K. Miettinen. *Nonlinear Multiobjective Optimization*. Kluwer Academic Publishers, 1999.

J. Mockus. *Bayesian Approach to Global Optimization*. Kluwer Academic publishers, Dordrecht, The Netherlands, 1989a.

J. Mockus. *Bayesian Approach to Global Optimization : Theory and Applications*. Kluwer Academic Publishers, Dordrecht, 1989b.

J. Mockus, V. Tiesis et A. Zilinskas. « The application of Bayesian methods for seeking the extremum ». In L.C.W. Dixon et G.P. Szego (éditeurs), *Towards Global Optimization 2*, pp. 117–129, North Holland, New York, 1978.

J. Oakley et A. O'Hagan. « Bayesian inference for the uncertainty distribution ». http ://www.citeseer.ist.psu.edu/oakley00bayesian.html, 1998.

S. O'Hagan, W.B. Dunn, M. Brown, J.D. Knowles et D.B. Kell. « Closed-loop, multiobjective optimization of analytical instrumentation : gas chromatography/time-of-ight mass spectrometry of the metabolomes of human serum and of yeast fermentations ». *Anal. Chem.*, 77(1) : 290–303, 2004.

E. Parzen. « A new approach to the synthesis of optimal smoothing and optimal systems ». In R. Bellman (éditeur), *Mathematical Optimization Techniques*, pp. 75–108. Univ. California Press, Berkeley, 1963.

H.D. Patterson et R. Thompson. « Recovery of inter-block information when block sizes are unequal ». *Biometrika*, 58 : 545–554, 1971.

C. Perttunen. « A computational geometric approach to feasible region division in constrained global optimization ». In *Proceedings of the 1991 IEEE Conference on Systems, Man. and Cybernetics*, volume 1, pp. 585–590, 1991.

L. Pronzato et E. Thierry. « Robust design with nonparametric models : prediction of second-order characteristics of process variability by kriging ». In *13th IFAC Symposium on System Identification*, pp. 560–565, Rotterdam, 2003.

L. Pronzato, H.P. Wynn et A.A. Zhigljavsky. « Using Renyi entropies to measure uncertainty in search problems ». In G. Yin et Q. Zhang (éditeurs), *Proc. 1996 AMS-SIAM Summer Seminar, Lectures in Applied Math, Mathematics of Stochastic Manufacturing Systems*, pp. 253–268. American Math. Soc., 1997.

J. Sacks, W.J. Welch et H.P. Mitchell, T.J. Wynn. « Design and analysis of computer experiments ». *Stat. Sci.*, 4(4) : 409–435, 1989.

M.J. Sasena, P. Papalambros et P. Goovaerts. « Exploration of metamodeling sampling criteria for constrained global optimization ». *Eng. Opt.*, 34 : 263–278, 2002.

M. Schonlau. *Computer Experiments and Global Optimization*. Mémoire de thèse, University of Waterloo, 1997.

E. Sjö. *Crossings and Maxima in Gaussian Fields and Seas*. Mémoire de thèse, Lund Institute of Technology, 2000.

A.J. Smola. *Learning with Kernels*. Mémoire de thèse, Technische Universität Berlin, 1998.

M.L. Stein. *Interpolation of Spatial Data : Some Theory for Kriging*. Springer, New-York, 1999.

G. Taguchi et S. Konishi. *Orthogonal Arrays and Linear Graphs*. American Supplier Institute, Inc, Dearborn, 1987.

G. Tunicliffe-Wilson. « On the use of marginal likelihood in time series model estimation ». *J. Roy. Statist. Soc. B*, 50 : 297–312, 1989.

S. Vaidyanathan, D. Kell et R. Goodacre. « Selective detection of proteins in mixtures using electrospray ionization mass spectrometry : influence of instrumental settings and implications for proteomics ». *Anal. Chem.*, 76 : 5024–5032, 2004.

E. Vazquez et M. Piera Martinez. « Estimation of the volume of an excursion set of a Gaussian process using intrinsic kriging ». http ://www.citebase.org/abstract ?id=oai :arXiv.org :math/0611273, 2006.

E. Vazquez, J. Villemonteix, M. Sidorkiewicz et É. Walter. « Global optimization based on noisy evaluation : an empirical study of two statistical approaches ». In *6th International Conference on Inverse Problems in Engineering : Theory and Practice*, Dourdan (Paris), France, June 15–19, 2008.

E. Vazquez et É. Walter. « Estimating derivatives and integrals with kriging ». In *Proceedings of the joint 44th IEEE Conference on Decision and European Control Conference*, pp. 8156–8161, Seville, 2005.

Emmanuel Vazquez. *Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et application*. Mémoire de thèse, Université Paris-Sud, UFR Scientifique d'Orsay, 2005.

M. Vergassola, E. Villermaux et B. I. Shraiman. « Infotaxis as a strategy for searching without gradients ». *Nature*, 445 : 406–409, 2007.

J. Villemonteix, E. Vazquez, M. Sidorkiewicz et É. Walter. « Gradient-based IAGO strategy for the global optimization of expensive-to-evaluate functions and application to intake-port design ». Advances in Global Optimization : Methods and Applications, Myconos (Greece), June 13–17, 2007a.

J. Villemonteix, E. Vazquez, M. Sidorkiewicz et É. Walter. « Global optimization of expensive-to-evaluate functions : an empirical comparison of two sampling criteria ». *To appear in the Mykonos special issue of the J. Global Optim.*, 2008a.

J. Villemonteix, E. Vazquez et É. Walter. « Identification of expensive-to-simulate parametric models using kriging and stepwise uncertainty reduction ». In *46th IEEE Conference on Decision and Control*, pp. 5505–5510, New Orleans (USA), December 12-14, 2007b.

J. Villemonteix, E. Vazquez et É. Walter. « An informational approach to the global optimization of expensive-to-evaluate functions ». *To appear in J. Global Optim.*, 2008b.

G. Wahba. « Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV ». In B. Schölkopf, C.J.C. Burges et A.J. Smola (éditeurs), *Advances in Kernel Methods - Support Vector Learning*, volume 6, pp. 69–87, Boston, 1998. MIT Press.

É. Walter et L. Pronzato. *Identification of Parametric Models from Experimental Data*. Springer, London, 1997.

D. Weuster-Botz et Wandrey C. « Medium optimization by genetic algorithm for continuous production of formate dehydrogenase ». *Process Biochem.*, 30 : 563–371, 1995.

D. Wiesmann, U. Hammel et T. Bäck. « Robust design of multilayer optical coatings by means of evolutionary algorithms ». *IEEE T. Evolt. Comput.*, 2(4) : 162–167, 1998.

B.J. Williams, T.J. Santner et W.I. Notz. « Sequential design of computer experiments to minimize integrated response functions ». *Stat. Sinica*, 10 : 1133–1152, 2000.

C.K.I Williams et C.E. Rasmussen. « Gaussian processes for regression ». In D.S. Touretzky, M.C. Mayer et M.E. Hasselmo (éditeurs), *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1996.

Z. Zhu et H. Zhang. « Spatial sampling design under the infill asymptotic framework ». *Environmetrics*, 17 : 323–337, 2006.

A. Zilinskas. « A review of statistical models for global optimization ». *J. Global Optim.*, 2 : 145–153, 1992.

Résumé — Cette thèse traite d'une question centrale dans de nombreux problèmes d'optimisation, en particulier en ingénierie. Comment optimiser une fonction lorsque le nombre d'évaluations autorisé est très limité au regard de la dimension et de la complexité du problème ? Par exemple, lorsque le budget d'évaluations est limité par la durée des simulations numériques du système à optimiser, il n'est pas rare de devoir optimiser trente paramètres avec moins de cent évaluations. Ce travail traite d'algorithmes d'optimisation spécifiques à ce contexte pour lequel la plupart des méthodes classiques sont inadaptées.

Le principe commun aux méthodes proposées est d'exploiter les propriétés des processus gaussiens et du krigeage pour construire une approximation peu coûteuse de la fonction à optimiser. Cette approximation est ensuite utilisée pour choisir itérativement les évaluations à réaliser. Ce choix est dicté par un critère d'échantillonnage qui combine recherche locale, à proximité des résultats prometteurs, et recherche globale, dans les zones non explorées. La plupart des critères proposés dans la littérature, tel celui de l'algorithme EGO (pour *Efficient Global Optimization*), cherchent à échantillonner la fonction là où l'apparition d'un optimum est jugée la plus probable. En comparaison, l'algorithme IAGO (pour *Informational Approach to Global Optimization*), principale contribution de nos travaux, cherche à maximiser la quantité d'information apportée, sur la position de l'optimum, par l'évaluation réalisée.

Des problématiques industrielles ont guidé l'organisation de ce mémoire, qui se destine à la communauté de l'optimisation tout comme aux praticiens confrontés à des fonctions à l'évaluation coûteuse. Aussi les applications industrielles y tiennent-elles une place importante tout comme la mise en place de l'algorithme IAGO. Nous détaillons non seulement le cas standard de l'optimisation d'une fonction réelle, mais aussi la prise en compte de contraintes, de bruit sur les résultats des évaluations, de résultats d'évaluation du gradient, de problèmes multi-objectifs, ou encore d'incertitudes de fabrication significatives.

Abstract — This dissertation is driven by a question central to many industrial optimization problems : how to optimize a function when the budget for its evaluation is severely limited by either time or cost ? For example, when optimization relies on computer simulations, each taking several hours, the dimension and complexity of the optimization problem may seem irreconcilable with the evaluation budget (typically thirty parameters to be optimized with less than one hundred evaluations). This work is devoted to optimization algorithms dedicated to this context, which is out of range for most classical methods.

The common principle of the methods discussed is to use Gaussian processes and Kriging to build a cheap proxy for the function to be optimized. This approximation is then used iteratively to choose the evaluation points. This choice is guided by a sampling criterion which combines local search, near promising evaluation results, and global search, in unexplored areas. Most of the criteria proposed over the years, such as the one underlying the classical EGO (for Efficient Global Optimization) algorithm, sample where the optimum is most likely to appear. By contrast, we propose an algorithm, named IAGO for Informational Approach to Global Optimization, which samples where the information gain on the optimizer location is deemed to be highest.

The organisation of this dissertation is a direct consequence of the industrial concerns which drove this work. We hope it can be of use to the optimization community, but most of all to practitioners confronted with expensive-to-evaluate functions. This is why we insist on the practical use of IAGO for the optimization of functions encountered in actual industrial problems. We also discuss how to handle constraints, noisy evaluation results, multi-objective problems, derivative evaluation results, or significant manufacturing uncertainties.