Sequential search strategies based on kriging

Emmanuel Vazquez

Emmanuel Vazquez

Ancien élève de l'École Normale Supérieure de Cachan
Ancien Professeur Agrégé
Docteur en Sciences de l'Université Paris XI Orsay
Enseignant–Chercheur à CentraleSupélec

Sequential search strategies based on kriging

Mémoire présenté en vue d'obtenir l'Habilitation à Diriger des Recherches

Le 15 juillet 2015

Devant la commission d'examen consituée de

M.	IOOSS	Bertrand	Ingénieur-Chercheur EDF R&D
M.	GAMBOA	Fabrice	Pr. Inst. Math. Toulouse
M.	GARNIER	Josselin	Pr. Univ. Paris Diderot
M.	PRONZATO	Luc	DR CNRS
M^{me}	SEBAG	Michèle	DR CNRS

Preface

This manuscript has been written to obtain the French *Habilitation à Diriger des Recherches*. It is not intended to provide new academic results nor should it be considered as a reference textbook. Instead, this manuscript is a brief (and incomplete) summary of my teaching and research activities. You will find in this manuscript a compilation of some articles in which I had a significant contribution, together with some introductory paragraphs about sequential search strategies based on kriging.

Gif-sur-Yvette, March 2015 *Emmanuel Vazquez*

Acknowledgements

I would like to thank Miguel, Julien, Sylvain, Aimad, Sándor, Lily, Romain, Benoit and Paul for their hard work and commitment during the preparation of their PhD thesis. I hope I have been able to provide them a good training and a useful experience for their career. I can safely say that I could not have done it without them. I would also like to thank Éric Walter for having guided me toward a very interesting research domain. Eventually, I would like to thank Albert Cohen, Ron DeVore and above all my colleague Julien Bect, who provided me invaluable help to formalize and develop new ideas.

Introduction

I was admitted at the ENS Cachan in 1997. After my *Agrégation de Physique* in 2000, I decided to undertake a Master of Research degree in applied mathematics. This drove me toward statistics, machine learning and signal processing.

I started the preparation of my PhD under the supervision of Éric Walter who suggested that I study the application of kriging for modeling computer experiments, in the path of [4, 13]. I can safely say that the application of kriging/Gaussian processes to modeling computer experiments was a quite confidential area when I started, at least in France. My PhD dissertation [20] was focused on making bridges between the literature of kriging [2, 3, 7, 8, 19, ...], the literature of reproducing kernel methods [15, 28, 29, ...], and the literature of Support Vector methods in machine learning [16–18, ...]. Today, I think that the textbooks [12, 14] cover much of what I have written in my dissertation. Modeling computer experiments using Gaussian processes has become a popular topic, as evidenced, for example, by its strong presence in the French Research Group MASCOT-NUM created in 2007.

In 2004, I was offered a position as an Assistant Researcher at Supélec, and been asked to co-supervise the PhD thesis of Miguel Piera-Martinez who studied the problem of computing probabilities of failure of a system [9]. At first, we concentrated our efforts on using the extreme value theory to obtain statistical models of the tail distribution of the output of a computer model [10, 11]. In the second part of his thesis, I suggested using a sequential approach for the estimation of a probability of failure [25]. This was the first instance of a SUR (*stepwise uncertain reduction*) algorithm.

In 2005, I began the supervision of Julien Villemonteix who had obtained a PhD grant from Renault S.A. to work on the reduction of pollutants of combustion engines [26]—in a context of increasingly stringent European emission standards. To reduce emissions, the idea was to optimize, using computer simulations, the shape of the intake ports of an engine. Since the simulation of the flow of the mixture of air and fuel in the intake ports was time-consuming, it was important to consider optimization algorithms which could provide a good approximation of the optimum with a limited budget of computer simulations. My idea was to adopt a Bayesian approach, by modeling the model output by a Gaussian process, and to devise an

algorithm that would reduce the entropy of the distribution of the optimizer by making new simulations sequentially. It turned out that this new idea worked very well in practice. Our algorithm, which Julien called IAGO (*Informational Approach to Global Optimization*), could be compared very favorably to the best algorithms in the literature, and in particular to the *expected improvement* optimization algorithm [27]. IAGO was again a SUR algorithm.

In 2009, I began a collaboration with my colleague Julien Bect who was newly recruited at Supélec. Julien provided me invaluable help to formalize a number of ideas I had during the supervision of my former PhD students [21–24]. We also decided to undertake the supervision of several PhD theses: a first one on unconstrained single-objective Bayesian optimization (Romain Benassi, 2009–2013), a second one on the problem of estimating probabilities of failure (Ling Li, 2009–2012), a third on constrained multi-objective optimization (Paul Féliot, since 2014), and a fourth on the design and analysis of computer experiments with several levels of predictive precision (Rémi Stroh, since 2015). Supervising these theses was very fruitful since it led us to the elaboration of several new algorithms: a fully Bayesian optimization algorithm using an expected-improvement sampling criterion and sequential Monte Carlo techniques [1], a Bayesian Subset Simulation (BSS) algorithm for the estimation of a probability of failure [6], a new Bayesian algorithm for constrained multi-objective optimization [5]. . . In this manuscript, I have chosen not to focus on these algorithms. Instead, I will present in the foregoing paragraphs the framework of my research activities on sequential search strategies based on kriging, on the period 2007-2014.

To conclude this introduction, I would like to add that my main activity has been devoted to teaching, so far. I also spend a significant part of my time working on contracts with industrial partners. I try to set aside about 20% of my time for research work.

The first part of this report consists of a presentation of my resume, an overview of my teaching activities and an assessment of my research activities. In the second part, I will focus on my research activities about sequential search strategies. In the third and last part, I will discuss future orientations.

References

1. R. Benassi, J. Bect, and E. Vazquez, *Bayesian optimization using sequential Monte Carlo*, Learning and Intelligent Optimization – 6th International Conference, LION 6 (Paris, France) (Y. Hamadi and M. Schoenauer, eds.), Springer, January 16-20 2012, pp. 339–342.
2. J.-P. Chilès and P. Delfiner, *Geostatistics: Modeling spatial uncertainty*, Wiley, New York, 1999.
3. N. Cressie, *Statistics for spatial data*, Wiley, New York, 1993.
4. C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker, *Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments*, J. Amer. Statist. Assoc. **86** (1991), no. 416, 953–963.

5. P. Féliot, J. Bect, and E. Vazquez, *A Bayesian approach to constrained multi-objective optimization*, Learning and Intelligent Optimization – 9th International Conference, LION 9 (Lille, France), Springer, January 12-15 2015.
6. L. Li, J. Bect, and E. Vazquez, *Bayesian Subset Simulation : a kriging-based subset simulation algorithm for the estimation of small probabilities of failure*, Proceedings of PSAM 11 & ESREL 2012 (Helsinki, Finland), June 25-29 2012.
7. G. Matheron, *Principles of geostatistics*, Economic Geology **58** (1963), 1246–1266.
8. _____, *The intrinsic random functions, and their applications*, Adv. Appl. Prob. **5** (1973), 439–468.
9. M. Piera-Martinez, *Modélisation des comportements extrêmes en ingénierie*, Ph.D. thesis, Université Paris Sud - Paris XI, Orsay, France, 2008.
10. M. Piera-Martinez, E. Vazquez, E. Walter, and G. Fleury, *RKHS classification for multivariate extreme-value analysis*, Statistics for Data Mining, Learning and Knowledge Extraction, IASC07 (Aveiro, Portugal), August 30-September 1 2007.
11. M. Piera-Martinez, E. Vazquez, E. Walter, G. Fleury, and R. Kielbasa, *Estimation of extreme values with application to uncertain systems*, 14th IFAC Symposium on System Identification, SYSID 2006 (Newcastle, Australia), March 29-31 2006, pp. 1027–1032.
12. C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, The MIT Press, Cambridge, 2006.
13. J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and analysis of computer experiments*, Statist. Sci. **4** (1989), no. 4, 409–435.
14. T. J. Santner, B. J. Williams, and W. Notz, *The Design and Analysis of Computer Experiments*, Springer Verlag, 2003.
15. R. Schaback, *Native Hilbert spaces for radial basis functions*, 1998.
16. B. Schölkopf, R. Herbrich, and A. J. Smola, *A generalized representer theorem*, Proceedings of the Annual Conference on Computational Learning Theory, 2001, pp. 416–426.
17. A. J. Smola, *Learning with kernels*, Ph.D. thesis, Technische Universität Berlin, 1998.
18. A. J. Smola, T. Friess, and B. Schölkopf, *Semiparametric support vector and linear programming machines*, Advances in Neural Information Processing Systems (M. Kearns, S. Solla, and D. Cohn, eds.), 11, MIT Press, Cambridge, 1999, pp. 585 – 591.
19. M. L. Stein, *Interpolation of spatial data: Some theory for Kriging*, Springer, New York, 1999.
20. E. Vazquez, *Modélisation comportementale des systèmes non linéaires multivariables par méthodes à noyaux et applications*, Ph.D. thesis, Université Paris-Sud XI, Orsay, France, May 2005.
21. E. Vazquez and J. Bect, *A sequential Bayesian algorithm to estimate a probability of failure*, Proceedings of the 15th IFAC Symposium on System Identification, SYSID 2009 15th IFAC Symposium on System Identification, SYSID 2009 (Saint-Malo France), 2009.
22. _____, *Convergence properties of the expected improvement algorithm with fixed mean and covariance functions*, Journal of Statistical Planning and Inference **140** (2010), no. 11, 3088–3095.
23. _____, *Pointwise consistency of the kriging predictor with known mean and covariance functions*, mODa 9 — Advances in Model-Oriented Design and Analysis, (Proc. of the 9th Int. Workshop in Model-Oriented Design and Analysis (Bertinoro, Italy), Physica-Verlag, Contributions to Statistics, Springer, June 14-18 2010, pp. 221–228.
24. _____, *Sequential search based on kriging: convergence analysis of some algorithms*, ISI & 58th World Statistics Congress of the International Statistical Institute (ISI'11) (Dublin, Ireland), August 21-26 2011.
25. E. Vazquez and M. Piera-Martinez, *Estimation of the volume of an excursion set of a Gaussian process using intrinsic kriging*, Tech. Report arXiv:math/0611273, arXiv.org, 2006.
26. J. Villemonteix, *Optimisation de fonctions coûteuses*, Ph.D. thesis, Université Paris-Sud XI, Faculté des Sciences d'Orsay, 2008.
27. J. Villemonteix, E. Vazquez, and E. Walter, *An informational approach to the global optimization of expensive-to-evaluate functions*, 2006.

28. G. Wahba, *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59, SIAM, Philadelphia, 1990.
29. Z. Wu and R. Schaback, *Local error estimates for radial basis function interpolation of scattered data*, IMA J. Numer. Anal. **13** (1993), 13–27.

Contents

Part III Conclusions & Future work

Part I
Curriculum Vitae

Chapter 1
Resume

EMMANUEL VAZQUEZ
SUPELEC, 3 rue Joliot Curie
91192 Gif-sur-Yvette, France
Telephone: +33(0)169851416
E-mail: emmanuel.vazquez@centralesupelec.fr

CURRENT POSITION
Associate Professor at CentraleSupélec

PERSONAL DETAILS
Age & DOB: 39, Feb. 12, 1976
Nationality: French

Education

2001–2005 **Ph.D. Université Paris-Sud**

Title: *Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications*
Supervision: Eric Walter
Defense: May 12, 2005 at Supélec, Gif-sur-Yvette
Jury: Georges Bastin, Pascal Bondon, Luc Pronzato, Robert Schaback, Hans Wackernagel, Éric Walter
Host laboratory: Laboratoire des Signaux et Systèmes (UMR8506)
Funding: *Allocation Couplée* (from ENS)

2000–2001 **DEA Mathématiques, Vision et Apprentissage**

At: Centre de Mathématiques et de Leurs Applications (UMR8536), ENS Cachan
Honors: *Mention Très Bien*

1997–2000 **École Normale Supérieure de Cachan**

- *Agrégation Externe de Sciences Physiques, option physique et électricité appliquée*, rank: 7th
- *Magistère de Physique Appliquée*
- *Maîtrise et Licence EEA*, at Université Paris-Sud Orsay, Honors: *mention Bien*

1994–1997 **Classes préparatoires aux Grandes Écoles**

At Lycée Marcelin Berthelot, Val-de-Marne, Option P'

1994 **Baccalauréat C**

At: Lycée Camille Jullian, Bordeaux
Honors: *Mention Bien*

Professional experience

2004 – today **Associate Professor at Supélec/CentraleSupélec**

Teaching: about 220h/year (in *équivalent TD*)
Research: statistics, design of experiments, Gaussian processes for modeling and optimization of systems
Contracts with industrial partners
Supervision: Ph.D. and postdoctoral students

2001 – 2004 **Monitorat**

Description: *Travaux Dirigés* on Probability and *Travaux Pratiques* on Signal Processing
At: Supélec, Service des Mesures
Duration: 64-80h/year (in *équivalent TD*)

2002 **Continuing Education Teaching**

Description: Time Series Analysis
At: Supélec
Duration: 3h30

2000 **Teaching**

Description: *Travaux Pratiques* on Linear Control Theory (40h)
At: IUT de Cachan
Duration: 40h

1999 **Industrial-research training formation**

Description: *Positionnement sous-marin par effet Döppler, Codes large spectre*
At: Thomson-Marconi-Sonar, Sophia-Antipolis
Duration: 3 months

1999 **Teaching**

Description: Lecture and tutorial sessions in Physics
At: Lycée Marcelin Berthelot
Duration: 6h

Chapter 2
Teaching experience & industrial contracts

2.1 Teaching

The number of hours of my teaching service at Supélec is about 220–260 hours/year on the period 2004–2014. Since Sept. 2014, I have reduced my teaching service to about 210 hours. A large part of this service is dedicated to the supervision of student projects.

Table 2.1: Teaching service

Title	Type	Period	Volume (in *équivalent TD* / year)	Place	Level
Present teaching activities					
Probability	lectures + small classes	since 2009	33	Supélec	1st year
Markov Chains Monte Carlo	lectures	since 2004	13.5	Supélec	3rd year
Bayesian Optimization	lectures	since 2011	3	Supélec	lifelong learning
Statistics	small classes	since 2004	6	Supélec	1st year
Introduction to stochastic processes	practical classes	since 2001	108 (2001–2009), 54 (since 2009)	Supélec	2nd year
Project Supervision	projects	since 2004	≈ 100	Supélec	1rd to 3rd year

Past teaching activities

Hilbert Spaces and Multiresolution Analysis	lectures	2004–2014	27	Supélec	2nd year
Introduction to numerical analysis	small classes	2004–2014	6	Supélec	2nd year
Introduction to kriging	lectures	2012	9	ATSI – Univ. Paris-Sud 11	MSc
Modeling extreme events from computer simulations	lectures + small classes	2011	26	Summer School CEA-EDF-INRIA	
Sequential Bayesian decision theory for designing systems from expensive computer simulations	lectures	2012	9	PhD week – Risk and Uncertainty ECP-Polimi-Supélec	
Spectral representation of stationary processes	lectures	2004–2007	13.5	ATSI – Univ. Paris-Sud 11	MSc
Signal Processing	practical classes	2008–2009	54	Supélec	1st year
Introduction to stochastic processes	small classes	2004–2009	6	Supélec	2nd year
Non-stationary time series analysis	lectures	2002	5.25	Supélec	lifelong learning
Linear control theory	Practical classes	2000	40	IUT Cachan	2nd year

2.2 Industrial contracts

I spend a significant part of my time working on contracts with industrial partners. This activity has three parts. The first part consists of the supervision of third-year students during their industrial projects (which represents a volume of about 250 hours). As a supervisor, it is often necessary to participate in the project's progress and to assume the role of interface with the industrial partner. The second part consists of taking part in collaborative projects funded by public agencies. The third part consists of bilateral contracts.

Table 2.2: Industrial contracts

Title of the study	Project/Industrial partner	Period & Effort
Supervision of student industrial projects		
Prévision de hauteur de la nappe phréatique sous un site industriel	EDF R&D STEP	2004–2005

Étude statistique de données de mesures de champ électromagnétique effectuées par l'ANFR à Paris	BOUYGUES TELECOM	2004-2005
Modélisation et prévision des jeux de calage des conduites du circuit primaire d'une centrale nucléaire	EDF R&D STEP	2005–206
Étude statistique de l'émission électromagnétique d'un télephone portable en environnement réel	BOUYGUES TELECOM	2005-2006
Estimation du gradient hydraulique à partir de mesures de hauteur de nappe phréatique effectuées sur des piézomètres non uniformément répartis spatialement	EDF R&D STEP	2006-2007
Étude et optimisation d'une stratégie d'arbitrage	ABC ARBITRAGE	2006-2007
Étude statistique de données de vibration d'un moteur d'avion, détection de valeurs aberantes	HISPANO SUIZA	2006-2007
Modélisation par krigeage d'un code de calcul aéronautique	EADS	2006–2007
Analyse des performances d'un logiciel d'estimation de composition isotopique	CEA LIST	2007–2008
Modélisation statistique des flux sur une grille de calcul	LAB. DE RECHERCHE EN INFORMATIQUE	2007-2008
Récupération de texte dactylographié à partir de l'analyse acoustique d'un clavier	THALES	2008–2009
Méthode d'estimation d'une composition isotopique à partir de mesure spectrale X/γ	CEA LIST	2008–2009
Méta-modèles pour l'estimation d'une probabilité de défaillance et l'évaluation d'un risque de crue	EDF R&D MRI	2009–2010
Utilisation de code adjoint pour l'optimisation fondée sur des méta-modèles	RENAULT	2009–2010
Étude et comparaison de méthodes de modélisation pour la simulation de programmes de maintenance	EDF R&D MRI	2010-2011
Estimation d'incertitudes de mesures en régime transitoire	LNE	2010–2011
Étude comparative de méthodes de modélisation de charges non linéaires sur le réseau électrique BT	EDF R&D MIRE	2011-2012
Analyse statistique spatiale de données de fissuration sur aéroréfrigérants	EDF R&D MRI	2011–2012
Analyse statistique d'une procédure de calage d'un modèle de circuit de refroidissement	EDF R&D LNHE	2012–2013
Stratégies de planification d'expériences en vue de modéliser des charges non linéaires	EDF R&D MIRE	2012–2013
Méthodes d'analyse et de visualisation de données fonctionnelles	EDF R&D MRI	2012–2013
Optimisation de l'équilibrage d'un rotor	SNECMA	2013–2014
Optimisation des techniques d'équilibrage d'un rotor	SNECMA	2014–2015
Calage d'un modèle numérique de performance énergétique d'un bâtiment	EDF R&D MRI	2014–2015

Collaborative projects

Étude statistique de l'inclinaison de tranches via les capteurs de température des modules de cuisson dans le processus de fabrication de circuits intégrés	Usine Numérique Sytem@tic FUI 0	2006–2007 2 months
Method for Forward Collision Warning based on Extreme Value Theory	ISI-PADAS European project	2008–2011
Évaluation d'un risque d'inondation fluviale par planification séquentielle d'expériences – Modélisation boîte noire de systèmes couplés	OPUS ANR project	2008-2011
Sequential design of computer experiments for the estimation of a probability of failure	CSDL Sytem@tic FUI 7	2009–2012
Optimisation d'un système électrique en présence d'énergies renouvelables	APOTEOSE ANR project	2012–2016
Optimisation de fonctions coûteuses	ROM – Outils de conception et de simulation IRT SystemX	2013–2016 2.2 m/y

Bilateral projects

Estimation du gradient d'une nappe phréatique	EDF R&D STEP	2006 2 months
Estimation du gradient de pression d'un écoulement biphasique	SCHLUMBERGER	2008 2 months
Calage du logiciel CooliSS	EDF R&D LNHE	2009 0.5 months
Étude de performances d'une procédure de calage du logiciel CooliSS	EDF R&D LNHE	2012 1 months

Chapter 3
Research experience

3.1 Summary of research work

3.1.1 Modeling systems using kriging

I started the preparation of my PhD thesis in 2001. The thesis was focused on the problem of modeling a system from computer experiments, using reproducing kernel regression and kriging techniques [20]. In computer-aided design, the term *computer experiments* refers to the idea of experimenting with a computer model [8, 18, 35]. More precisely, assume a knowledge-based model of a system in the form of a computer program—for example, a finite-element program simulating the physical behavior of some component in a device. For a number of reasons, the computer program used for this purpose must often be seen as a black box, only known by the numerical values that it produces (the program outputs) for given numerical values of its input arguments.

To study the influence of these inputs on the values taken by the outputs, for instance to find the combination of input arguments that leads to the most satisfactory values of the output arguments, one has to call the same program again and again. When the computational cost of the program is high, the number of runs is limited. One can use a second level of modeling, often called a *meta-model*, to summarize what has been learned from the simulations and to infer the response of the original computer model without actually running it again. The problem of constructing a meta-model from simulation results is in fact that of *function approximation* from pointwise evaluations. Kriging and kernel regression (which includes splines, radial basis functions, and support vector regression...) have proven effective in this context.

Mathematical links between kriging and kernel-based regularized regression have been noticed very soon in the literature. My PhD dissertation presents a synthesis of these links, drawn from the domains of function approximation, machine learning, time series prediction and geostatistics. These links are essential, for instance in order to understand how regularized regression via kernel-based methods

should be formulated in the context of the approximation of vector-valued functions for the approximation of MIMO systems [28], or to approximate the derivative of a noisy function [30], or to take into account prior knowledge via semi-regularized regression (or intrinsic kriging) [31] ... The question of how kernel-based regression methods can be applied in engineering was addressed via the consideration of real problems [20, 25, 29].

The specific problem of choosing the experiments, that is, which computer simulations should be carried out to obtain relevant information about the black box model, is not addressed in the dissertation, which at that time was left for future work.

After my thesis defense in 2005, I continued working on the specific problem of modeling systems using kriging. In particular, from 2006 to 2009, I worked on the problem of constructing black box approximations of continuous-time dynamical systems [34].

3.1.2 Modeling extreme events from computer simulations

Uncertainty may appear in a system due to external perturbations or dispersion of the design parameters. In this case, a deterministic approach to design systems, which assumes a perfect knowledge of its environment, becomes questionable. The need of reliable systems leads us to elaborate statistical models that are able to deal with this randomness. In this context, I undertook the supervision of Miguel Piera-Martinez during his PhD thesis. We focused our work on the problem of modeling the occurrence of extreme values in the output of a computer program, following the idea that these extreme values may correspond to abnormal or dangerous operating conditions. A simple Monte Carlo analysis of extreme values requires many simulations of the system, which are often very expensive. It is thus desirable to analyze extreme events with as few system simulations as possible.

Our first idea was to use the statistical theory of *extreme values* to model the tail distribution of the output of a computer model. We could obtain satisfactory results by applying Extreme Value Theory (EVT) to various problems in the domain of electronic design and electromagnetic compatibility engineering [15, 17]. We were also interested in using EVT to estimate multivariate quantiles. In this context, we proposed a method combining EVT and one-class SVM, using the following principle [16]: one-class SVM constructs a function, which when thresholded, yields an empirical multivariate quantile, that is, a set (in the space of the observations) that contains a given proportion of the observations; by using EVT, we could adjust the threshold in such a way that the quantile remains accurate even when the quantile probability is close to one.

In the second part of Miguel's thesis, I suggested using a sequential approach for the estimation of a probability of failure [26] using a kriging approach. This was the first instance of a SUR (*stepwise uncertain reduction*) algorithm. This work was improved and published during the period 2009–2013 [3, 7, 21].

In 2007, I was presented the opportunity to co-supervise the PhD thesis of Aimad El Habachi, which was about the safety evaluation of electromagnetic emissions from wireless communications devices on human health. More precisely, the objective was to assess using computer models the probability that the SAR (Specific Absorption Rate) in human body exceeds a given level given the variability of human anatomy [10]. Aimad and his co-supervisors from Orange Labs were able to obtain interesting practical results using the sequential techniques that I had developed with Miguel.

3.1.3 Informational approach to global optimization and Bayesian optimization

Julien Villemonteix's PhD thesis (2005–2009) was driven by a question central to many industrial optimization problems: how to optimize a function when its evaluation is time-consuming?

I suggested that we could construct a new optimization algorithm using the Bayesian viewpoint of kriging and by focusing on the posterior probability distribution of the optimizer points. More precisely, the Informational Approach to Global Optimization (IAGO) [33] consists in sequentially choosing new evaluation points of the function to be optimized in order to decrease the entropy of the optimizer points.

This idea yields a global optimization algorithm, which combines local search, near promising evaluation results, and global search, in unexplored areas. This instance of *Bayesian optimization* algorithms has proven effective in comparison to other algorithms in this class, and in particular the classical EGO (Efficient Global Optimization) algorithm [12].

The work of Julien was above all guided by industrial concerns. Julien focused on the use of IAGO for solving actual industrial problems. In his dissertation [32], he also discusses a couple of important practical problems: constrained optimization, noisy evaluation results, multi-objective problems, robust optimization in presence of manufacturing uncertainties...

By 2008, I also had the opportunity to help Sándor Bilicz whose PhD thesis was on using Bayesian optimization for dealing with inverse problems in Eddy-current testing [5].

More recently, I participated in the writing of an article with Héloïse Dutrieux, PhD student (co-)supervised by my colleague Julien Bect, on the application of IAGO to optimize very noisy functions [9], where we proposed a new sampling criterion where to circumvent the problem of the estimation of the entropy when evaluation noise becomes large.

3.1.4 Study of kriging-based sequential strategies

By 2006, I was interested in the asymptotic performance of the SUR procedure to estimate a probability of failure and wanted to know if using approximations could yield better estimators than the simple Monte Carlo estimator [26, 27].

Later, Julien Bect and I worked on the problem of the consistency of kriging-based sequential strategies [22–24]; more precisely, the problem was to understand whether kriging-based sequential strategies could provide consistent estimators of a given quantity of interest (for instance, when seeking the global maximum of a function using the Expected Improvement [14, 19] criterion). This topic has been left out since 2011 by lack of time and dedication, but is among my favorite research activities.

3.1.5 Efficient algorithms for optimization and estimation of probabilities of failure

Since September 2009, Julien Bect and I began the supervision of PhD students on the problem of the *implementation* of kriging-based sequential strategies.

1. Implementation of SUR strategies to estimate a probability of failure

   The objective was to work on the limitation of the SUR strategies to estimate a probability of failure that we had developed before 2009. More precisely, the strategies we were using until that time consisted in constructing a (Bayesian) estimator of the (simple) Monte Carlo estimator of the volume of excursion of a function of interest above a given threshold. The applicability of this idea, that we called *meta-estimation* [1, 3], is limited by the cost of computing the Bayesian posterior distribution of the function of interest. Julien and I suggested the idea of a new algorithm, called *Bayesian Subset Simulation*, that combines a SUR strategy and the classical Subset Simulation algorithm [2, 6]. By using this new algorithm, we were able to obtain very good results for the problem of estimating very small probabilities of failure [13].

2. Fully Bayesian optimization

   The objective was to work on an efficient implementation of Bayesian optimization based on the EI criterion. When working with EI-based algorithms, two difficulties can be identified. The first problem is the choice of a Gaussian process prior, and more precisely, the classical use of a maximum likelihood approach to estimate the parameters of the covariance function of the Gaussian process model of the EI-based approach. Using maximum likelihood turns out to lack robustness and to yield poor optimizations in some configurations. The second problem is the maximization of the multi-modal EI criterion after each new evaluation. We suggested combining a fully Bayesian approach and a Sequential Monte Carlo approach to deal with both problems [4]: at each step of the optimization algorithm, we deal with a set of so-called *particles*, where

each particle corresponds to a pair of a likely position for the maximizer of the EI criterion and a parameter value of the covariance function. We could obtain very interesting numerical results that showed that our algorithm compares very favorably to other methods of the literature.

3. Constrained multi-objective optimization

   Since January 2014, we work on the problem of dealing efficiently with expensive-to-evaluate constraints in multi-objective optimization problems. By combining several principles from our previous work, we were able to obtain promising results [11]

References

1. A. Arnaud, J. Bect, M. Couplet, A. Pasanisi, and E. Vazquez, *Évaluation d'un risque d'inondation fluviale par planification séquentielle d'expériences*, 42es Journées de Statistique, JdS 2010 (Marseille, France), 2010.
2. S. K. Au and J. Beck, *Estimation of small failure probabilities in high dimensions by subset simulation*, Probab. Engrg. Mechan. **16** (2001), no. 4, 263–277.
3. J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez, *Sequential design of computer experiments for the estimation of a probability of failure*, Statist. Comput. **22** (2012), no. 3, 773–793.
4. R. Benassi, J. Bect, and E. Vazquez, *Bayesian optimization using sequential Monte Carlo*, Learning and Intelligent Optimization – 6th International Conference, LION 6 (Paris, France) (Y. Hamadi and M. Schoenauer, eds.), Springer, January 16-20 2012, pp. 339–342.
5. S. Bilicz, E. Vazquez, S. Gyimóthy, J. Pávó, and M. Lambert, *Kriging for eddy-current testing problems*, IEEE Transactions on Magnetics **46** (2010), no. 8, 3165–3168.
6. F. Cérou, P. Del Moral, T. Furon, and A. Guyader, *Sequential monte carlo for rare event estimation*, Statist. Comput. (2011), 1–14.
7. C. Chevalier, J. Bect, D. Ginsbourger, Y. Richet, V. Picheny, and E. Vazquez, *Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set*, Technometrics **56** (2014), no. 4, 455–465.
8. C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker, *Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments*, J. Amer. Statist. Assoc. **86** (1991), no. 416, 953–963.
9. H. Dutrieux, I. Aleksovska, J. Bect, E. Vazquez, G. Delille, and B. François, *The informational approach to global optimization in presence of very noisy evaluation results. Application to the optimization of renewable energy integration strategies*, 47e journées de Statistique, JdS 2015 (Lille, France), June 1-5 2015.
10. A. El Habachi, E. Conil, A. Hadjem, E. Vazquez, M. F. Wong, A. Gati, G. Fleury, and J. Wiart, *Statistical analysis of whole-body absorption depending on anatomical human characteristics at a frequency of 2.1 GHz*, Physics in Medicine and Biology **55** (2010), no. 7, 1875.
11. P. Féliot, J. Bect, and E. Vazquez, *A Bayesian approach to constrained multi-objective optimization*, Learning and Intelligent Optimization – 9th International Conference, LION 9 (Lille, France), Springer, January 12-15 2015.
12. D. R. Jones, M. Schonlau, and W. J. Welch, *Efficient global optimization of expensive black-box functions*, J. Global Optim. **13** (1998), 455–492.
13. L. Li, J. Bect, and E. Vazquez, *Bayesian Subset Simulation : a kriging-based subset simulation algorithm for the estimation of small probabilities of failure*, Proceedings of PSAM 11 & ESREL 2012 (Helsinki, Finland), June 25-29 2012.
14. J. Mockus, V. Tiesis, and A. Zilinskas, *The application of Bayesian methods for seeking the extremum*, Towards Global Optimization (North Holland, New York) (L.C.W. Dixon and G.P. Szego, eds.), vol. 2, 1978, pp. 117–129.

15. M. Piera-Martinez, *Modélisation des comportements extrêmes en ingénierie*, Ph.D. thesis, Université Paris Sud - Paris XI, Orsay, France, 2008.
16. M. Piera-Martinez, E. Vazquez, E. Walter, and G. Fleury, *RKHS classification for multivariate extreme-value analysis*, Statistics for Data Mining, Learning and Knowledge Extraction, IASC07 (Aveiro, Portugal), August 30-September 1 2007.
17. M. Piera-Martinez, E. Vazquez, E. Walter, G. Fleury, and R. Kielbasa, *Estimation of extreme values with application to uncertain systems*, 14th IFAC Symposium on System Identification, SYSID 2006 (Newcastle, Australia), March 29-31 2006, pp. 1027–1032.
18. J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and analysis of computer experiments*, Statist. Sci. **4** (1989), no. 4, 409–435.
19. M. Schonlau, *Computer experiments and global optimization*, Ph.D. thesis, University of Waterloo, Waterloo, Ontario, Canada, 1997.
20. E. Vazquez, *Modélisation comportementale des systèmes non linéaires multivariables par méthodes à noyaux et applications*, Ph.D. thesis, Université Paris-Sud XI, Orsay, France, May 2005.
21. E. Vazquez and J. Bect, *Sequential Bayesian algorithm to estimate a probability of failure*, 15th IFAC Symposium on System Identification, SYSID09 (Saint-Malo, France), July 6-8 2009, pp. 546–550.
22. E. Vazquez and J. Bect, *Convergence properties of the expected improvement algorithm with fixed mean and covariance functions*, Journal of Statistical Planning and Inference **140** (2010), no. 11, 3088–3095.
23. _____, *Pointwise consistency of the kriging predictor with known mean and covariance functions*, mODa 9 — Advances in Model-Oriented Design and Analysis, (Proc. of the 9th Int. Workshop in Model-Oriented Design and Analysis (Bertinoro, Italy), Physica-Verlag, Contributions to Statistics, Springer, June 14-18 2010, pp. 221–228.
24. _____, *Sequential search based on kriging: convergence analysis of some algorithms*, ISI & 58th World Statistics Congress of the International Statistical Institute (ISI'11) (Dublin, Ireland), August 21-26 2011.
25. E. Vazquez, G. Fleury, and E. Walter, *Kriging for indirect measurement, with application to flow measurement*, IEEE Trans. Instrumentation and Measurement **55** (2006), no. 1, 343–349.
26. E. Vazquez and M. Piera-Martinez, *Estimation of the volume of an excursion set of a Gaussian process using intrinsic kriging*, Tech. Report arXiv:math/0611273, arXiv.org, 2006.
27. _____, *Estimation du volume des ensembles d'excursion d'un processus aléatoire par krigeage intrinséque*, 39èmes Journées de Statistiques, JDS 2007 (Angers, France), June 11-15 2007.
28. E. Vazquez and E. Walter, *Multi-output vector regression*, 13th IFAC Symposium on System Identification, SYSID 2003 (Rotterdam, Netherlands), 2003, pp. 1820–1825.
29. E. Vazquez and E. Walter, *Choix d'une covariance pour la prediction par krigeage de series chronologiques échantillonnées irrégulièrement*, 20e Colloque Gretsi sur le Traitement du Signal et des Images, GRETSI 05 (Louvain-La-Neuve, Belgique), September 06-09 2005.
30. E. Vazquez and E. Walter, *Estimating derivatives and integrals with kriging*, IEEE 44th Conference on Decision and Control and European Control Conference (Seville, Spain), December 12-15 2005, pp. 8156–8161.
31. E. Vazquez, E. Walter, and G. Fleury, *Intrinsic kriging and prior information*, Appl. Stoch. Models Bus. Ind. **21** (2005), no. 2, 215–226.
32. J. Villemonteix, *Optimisation de fonctions coûteuses*, Ph.D. thesis, Université Paris-Sud XI, Faculté des Sciences d'Orsay, 2008.
33. J. Villemonteix, E. Vazquez, and E. Walter, *An informational approach to the global optimization of expensive-to-evaluate functions*, J. Global Optim. **44** (2009), no. 4, 509–534, doi: 10.1007/s10898-008-9354-2.
34. S. Vinet and E. Vazquez, *Black-box identification and simulation of continuous-time nonlinear systems with random processes*, 17th World Congress The International Federation of Automatic Control, IFAC 2008 (Seoul, Korea), July 6-11 2008, pp. 14391–14396.

35. W. J. Welch, R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, and M. D. Morris, *Screening, predicting, and computer experiments*, Technometrics **34** (1992), no. 1, 15–25.

3.2 List of publications

I have co-authored 14 journal articles, 31 conference articles and 22 conference talks, which are listed below. I also list a number of technical reports and invited talks.

A. Journal articles

1. C. Chevalier, J. Bect, D. Ginsbourger, Y. Richet, V. Picheny, and E. Vazquez. Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4):455–465, 2014.
2. J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Stat. Comput.*, 22(3):773–793, 2012.
3. E. Vazquez and J. Bect. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *J. Statist. Plann. Inference*, 140(11):3088–3095, 2010.
4. A. El Habachi, E. Conil, A. Hadjem, E. Vazquez, M. F. Wong, A. Gati, G. Fleury, and J. Wiart. Statistical analysis of whole-body absorption depending on anatomical human characteristics at a frequency of 2.1 GHz. *Physics in Medicine and Biology*, 55(7):1875, 2010.
5. S. Bilicz, M. Lambert, E. Vazquez, and S. Gyimóthy. Combination of maximin and kriging prediction methods for eddy-current testing database generation. *J. Phys.: Conf. Ser.*, 255(1):012003, 2010.
6. S. Bilicz, M. Lambert, E. Vazquez, and S. Gyimóthy. A new database generation method combining maximin method and kriging prediction for eddy-current testing. In J. Knopp, M. Blodgett, B. Wincheski, and N. Bowler, editors, *Studies in Applied Electromagnetics and Mechanics: Electromagnetic Nondestructive Evaluation (XIII)*, pages 199–206, Amsterdam, 2010. IOS Press.
7. S. Bilicz, E. Vazquez, S. Gyimóthy, J. Pávó, and M. Lambert. Kriging for eddy-current testing problems. *IEEE Trans. Magnetics*, 46(8):3165–3168, August 2010.
8. S. Bilicz, E. Vazquez, M. Lambert, S. Gyimóthy, and J. Pávó. Characterization of a 3d defect using the expected improvement algorithm. *COMPEL -The International Journal For Computation And Mathematics In Electrical And Electronic Engineering*, 28(4):851–864, 2009.
9. J. Villemonteix, E. Vazquez, M. Sidorkiewicz, and E. Walter. Global optimization of expensive-to-evaluate functions: an empirical comparison of two sampling criteria. *J. Global Optim.*, 43(2-3):373–389, 2009.
10. J. A. Egea, E. Vazquez, J. R. Banga, and R. Marti. Improved scatter search for the global optimization of computationally expensive dynamic models. *J. Global Optim.*, 43(2-3):175–190, 2009.
11. E. Vazquez, J. Villemonteix, M. Sidorkiewicz, and E. Walter. Global optimization based on noisy evaluations: An empirical study of two statistical approaches. *J. Phys.: Conf. Ser.*, 135(012100):8pp, 2008. doi: 10.1088/1742-6596/135/1/012100.
12. J. Villemonteix, E. Vazquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *J. Global Optim.*, 44(4):509–534, 2009. doi: 10.1007/s10898-008-9354-2.
13. E. Vazquez, G. Fleury, and E. Walter. Kriging for indirect measurement, with application to flow measurement. *IEEE Trans. Instrumentation and Measurement*, 55(1):343–349, 2006.
14. E. Vazquez, E. Walter, and G. Fleury. Intrinsic kriging and prior information. *Appl. Stoch. Models Bus. Ind.*, 21(2):215–226, 2005.

B. Conference articles

1. H. Dutrieux, I. Aleksovska, J. Bect, F. Bruno, G. Delille, and E. Vazquez. The informational approach to global optimization in presence of very noisy evaluation results. application to the optimization of renewable energy integration strategies. In *47e Journées de Statistique, JdS 2015*, Lille, France, June 1-5 2015.
2. J. Bect, R. Sueur, A. Gérossier, L. Mongellaz, S. Petit, and E. Vazquez. Echantillonnage préférentiel et méta-modèles : méthodes bayésiennes optimale et défensive. In *47e Journées de Statistique, JdS 2015*, Lille, France, June 1-5 2015.
3. P. Féliot, J. Bect, and E. Vazquez. A Bayesian approach to constrained multi-objective optimization. In *Learning and Intelligent Optimization – 9th International Conference, LION 9*, Lille, France, January 12-15 2015. Springer.
4. J. Bect, N. Bousquet, B. Iooss, S. Liu, A. Mabille, A.-L. Popelin, T. Rivière, R. Stroh, R. Sueur, and E. Vazquez. Quantification et réduction de l'incertitude concernant les propriétés de monotonie d'un code de calcul coûteux à évaluer. In *46e Journées de Statistique, JdS 2014*, Rennes, France, June 2-6 2014.
5. B. Jan, J. Bect, E. Vazquez, and P. Lefranc. Approche bayésienne pour l'estimation d'indices de sobol. In *45e Journées de Statistique, JdS 2013*, Toulouse, France, May 27-31 2013.
6. R. Benassi, J. Bect, and E. Vazquez. Bayesian optimization using sequential Monte Carlo. In Y. Hamadi and M. Schoenauer, editors, *Learning and Intelligent Optimization – 6th International Conference, LION 6*, pages 339–342, Paris, France, January 16-20 2012. Springer.
7. L. Li, J. Bect, and E. Vazquez. Bayesian Subset Simulation : a kriging-based subset simulation algorithm for the estimation of small probabilities of failure. In *Proceedings of PSAM 11 & ESREL 2012*, Helsinki, Finland, June 25-29 2012.
8. R. Benassi, J. Bect, and E. Vazquez. Optimisation bayésienne par méthodes SMC. In *44e Journées de Statistique, JdS 2012*, Bruxelles, Belgium, May 21-25 2012.
9. L. Li, J. Bect, and E. Vazquez. A numerical comparison of kriging-based sequential strategies for estimating a probability of failure. In *11th International Conference on Applications of Statistics and Probability Civil Engeneering, ICASP'11*, Zurich, Switzerland, August 1-4 2011.
10. E. Vazquez and J. Bect. Sequential search based on kriging: convergence analysis of some algorithms. In *ISI & 58th World Statistics Congress of the International Statistical Institute (ISI'11)*, Dublin, Ireland, August 21-26 2011.
11. R. Benassi, J. Bect, and E. Vazquez. Robust gaussian process-based global optimization using a fully bayesian expected improvement criterion. In Coello-Coello and A. Carlos, editors, *Learning and Intelligent Optimization, 5th International Conference, LION 5*, pages 176–190, Rome, Italy, January 17-21 2011. Springer.
12. A. Arnaud, J. Bect, M. Couplet, A. Pasanisi, and E. Vazquez. Évaluation d'un risque d'inondation fluviale par planification séquentielle d'expériences. In *42e Journées de Statistique, JdS 2010*, Marseille, France, 2010.
13. E. Vazquez and J. Bect. Pointwise consistency of the kriging predictor with known mean and covariance functions. In *mODa 9 — Advances in Model-Oriented Design and Analysis, Proc. of the 9th Int. Workshop in Model-Oriented Design and Analysis*, Physica-Verlag, Contributions to Statistics, pages 221–228, Bertinoro, Italy, June 14-18 2010. Springer.
14. E. Vazquez and J. Bect. Sequential bayesian algorithm to estimate a probability of failure. In *15th IFAC Symposium on System Identification, SYSID09*, pages 546–550, Saint-Malo, France, July 6-8 2009.
15. J. Villemonteix, E. Vazquez, and E. Walter. Bayesian optimization for parameter identification on a small simulation budget. In *15th IFAC Symposium on System Identification, SYSID09*, pages 1603–1608, Saint-Malo, France, July 6-8 2009.
16. S. Bilicz, E. Vazquez, J. Pávó, M. Lambert, and S. Gyimóthy. Eddy-current testing with the expected improvement optimization algorithm. In *15th IFAC Symposium on System Identification, SYSID09*, pages 1750–1755, Saint-Malo, France, July 6-8 2009.

17. S. Bilicz, E. Vazquez, M. Lambert, J. Pávó, and S. Gyimóthy. Characterization of a 3d defect using the expected improvement algorithm. In *13th International IGTE Symposium on Numerical Field Calculation in Electrical Engineering*, pages 157–162, Graz, Austria, September 21-24 2008.
18. S. Vinet and E. Vazquez. Identification boîte noire et simulation de systèmes non-linéaires à temps continu par prédiction linéaire de processus aléatoires. In *Actes de la Conférence Internationale Francophone d'Automatique 2008, CIFA 2008*, page 6pp, Bucarest, Roumanie, September 3-5 2008.
19. S. Vinet and E. Vazquez. Black-box identification and simulation of continuous-time nonlinear systems with random processes. In *17th World Congress The International Federation of Automatic Control, IFAC 2008*, pages 14391–14396, Seoul, Korea, July 6-11 2008.
20. J. Villemonteix, E. Vazquez, and E. Walter. Identification of expensive-to-simulate parametric models using kriging and stepwise uncertainty reduction. In *46th IEEE Conference on Decision and Control, CDC 2007*, pages 5505–5510, New Orleans, LA, USA, December 12-14 2007.
21. E. Vazquez and M. Piera-Martinez. Estimation du volume des ensembles d'excursion d'un processus aléatoire par krigeage intrinséque. In *39e Journées de Statistiques, JdS 2007*, Angers, France, June 11-15 2007.
22. M. Piera-Martinez, E. Vazquez, E. Walter, and G. Fleury. RKHS classification for multivariate extreme-value analysis. In *Statistics for Data Mining, Learning and Knowledge Extraction, IASC 07*, Aveiro, Portugal, August 30-September 1 2007.
23. Piera-Martinez M., E. Vazquez, E. Walter, Fleury G., and Kielbasa R. Estimation of extreme values with application to uncertain systems. In *14th IFAC Symposium on System Identification, SYSID 2006*, pages 1027–1032, Newcastle, Australia, March 29-31 2006.
24. M. Piera-Martinez, E. Vazquez, G. Fleury, and E. Walter. Application de la classification à vecteurs de support pour l'estimation de quantiles multidimensionnels extrêmes. In *38e Journées de Statistiques, JdS 2006*, Clamart, France, May 29-June 02 2006.
25. E. Vazquez and E. Walter. Estimating derivatives and integrals with kriging. In *IEEE 44th Conference on Decision and Control and European Control Conference*, pages 8156–8161, Seville, Spain, December 12-15 2005.
26. E. Vazquez and E. Walter. Choix d'une covariance pour la prediction par krigeage de series chronologiques échantillonnées irrégulièrement. In *20e Colloque Gretsi sur le Traitement du Signal et des Images, GRETSI 05*, Louvain-La-Neuve, Belgique, September 06-09 2005.
27. M. Braci, S. Diop, E. Vazquez, and E. Walter. On higher order numerical differentiation schemes for nonlinear estimation. In P. Borne, M. Benerjeb, N. Dangoumau, and L. Lorimier, editors, *IMACS World Congress on Scientific Computation, Applied Mathematics and Simulation*, Lille, France, July 11-15 2005.
28. E. Vazquez and E. Walter. Régression régularisée multivariable, méthodes à noyaux et krigeage. In *IEEE Conférence internationale francophone d'automatique, CIFA 2004*, Douz, Tunisie, 2004.
29. E. Vazquez and E. Walter. Choix d'un noyau pour la régression à vecteurs de support par analyse structurelle : application à la régression multivariable. In *19e Colloque Gretsi sur le Traitement du Signal et des Images, GRETSI 2003*, Paris, France, 2003.
30. E. Vazquez and E. Walter. Multi-output vector regression. In *13th IFAC Symposium on System Identification, SYSID 2003*, pages 1820–1825, Rotterdam, Netherlands, 2003.
31. E. Vazquez and E. Walter. Intrinsic kriging and prior information. In *Conference on Statistical Learning, Theory and Applications*, pages 93–97, CNAM, Paris, France, November 2002.

C. Communications

1. P. Féliot, J. Bect, and E. Vazquez. A Bayesian approach to constrained multi-objective optimization of expensive-to-evaluate functions. In *Journées annuelles du MASCOT-NUM 2015*, Saint-Étienne, France, April 8-10 2015.
2. P. Féliot, J. Bect, and E. Vazquez. A Bayesian subset simulation approach to constrained global optimization of expensive-to-evaluate black-box functions. In *Conference on Optimization and Practices in Industry : PGMO-COPI'14*, Palaiseau, France, Oct. 2014.
3. J. Bect, N. Bousquet, B. Iooss, S. Liu, A. Mabille, A.-L. Popelin, T. Rivière, R. Stroh, R. Sueur, and E. Vazquez. Uncertainty quantification and reduction for the monotonicity properties of expensive-to-evaluate computer models. In *Uncertainty in Computer Models 2014 Conference, UCM2014*, Sheffield, England, July 28-30 2013.
4. J. Bect and E. Vazquez. Bayes-optimal importance sampling for computer experiments. In *7th International Workshop on Simulation, IWS 2013*, Rimini, Italy, May 21-25 2013.
5. B. Jan, J. Bect, and E. Vazquez. Fully bayesian approach for the estimation of (first-order) Sobol indices. In *7th International Conference on Sensitivity Analysis of Model Output, MASCOT-SAMO 2013*, Nice, France, July 1-4 2013.
6. R. Benassi, J. Bect, and E. Vazquez. Optimisation bayésienne par méthodes SMC. In *Journées annuelles du GdR MASCOT-NUM 2012*, Bruyères-le-Châtel, France, March 21-23 2012.
7. N. Fischer, E. Georgin, A. Ismail, E. Vazquez, and Le Brusquet L. A nonparametric model for sensors used in a dynamical context. In *7th International Workshop on Analysis of Dynamical Measurements*, Paris, France, 2012.
8. R. Benassi, J. Bect, and E. Vazquez. étude d'un nouveau critère d'optimisation bayésienne. In *Journées annuelles du GdR MASCOT-NUM 2010*, Avignon, France, March 17-19 2010.
9. L. Li, J. Bect, and E. Vazquez. A numerical comparison of two sequential kriging-based algorithms to estimate a probability of failure. In *Uncertainty in Computer Model Conference, UCM 2010*, Sheffield, UK, July, 12-14 2010.
10. J. Bect, L. Li, and E. Vazquez. Planification séquentielle pour l'estimation de probabilities de défaillance. In *Atelier Événements rares du GdR MASCOT-NUM*, Paris, France, May 4 2010.
11. A. El Habachi, E. Conil, J. Carette, A. Hadjem, E. Vazquez, A. Gati, M. F. Wong, G. Fleury, and J. Wiart. Multidimensional collocation stochastic method to evaluate the whole specific absorption rate for a given population. In *32nd Annual Meeting of the Bioelectromagnetics Society 2010, BioEM'10*, pages 453–454, Seoul, South Korea, June 13-18 2010.
12. A. El Habachi, E. Conil, A. Hadjem, E. Vazquez, A. Gati, M. F. Wong, G. Fleury, and J. Wiart. Bayesian experiment planning applied to numerical dosimetry. In *32nd Annual Meeting of the Bioelectromagnetics Society 2010, BioEM'10*, pages 455–456, Seoul, South Korea, June 13-18 2010.
13. S. Bilicz, E. Vazquez, S. Gyimóthy, J. Pávó, and M. Lambert. Kriging for eddy-current testing problems. In *COMPUMAG—17th Conference on the Computation of Electromagnetic Fields*, Florianópolis, Brasil, November 22-26 2009.
14. S. Bilicz, E. Vazquez, M. Lambert, S. Gyimóthy, and J. Pávó. The expected improvement global optimization algorithm for the solution of eddy-current testing inverse problems. In *14th International Workshop on Electromagnetic Nondestructive Evaluation, ENDE 2009*, Dayton OH, USA, July 21-23 2009.
15. S. Bilicz, M. Lambert, E. Vazquez, and S. Gyimóthy. A new database generation method combining maximin method and kriging prediction for eddy-curent testing. In *14th International Workshop on Electromagnetic Nondestructive Evaluation, ENDE 2009*, Dayton OH, USA, July 21-23 2009.
16. S. Bilicz, M. Lambert, E. Vazquez, and S. Gyimóthy. Combination of maximin and kriging prediction methods for eddy-current testing database generation. In *Workshop on Electromagnetic Inverse Problems*, Manchester, England, June 15 2009.
17. A. El Habachi, E. Conil, A. Hadjem, E. Vazquez, G. Fleury, and J. Wiart. Identification of factors influencing the Whole Body Absorption Rate using statistical analysis. In *Joint meet-*

ing of the Biolecetromagnetics Society and the European Bioelectromagnetics Association, BioEM'09, Davos, Switzerland, June 14-19 2009.
18. A. El Habachi, E. Conil, A. Hadjem, E. Vazquez, G. Fleury, and J. Wiart. Identification des facteurs morphologiques impactant le Débit d'Absorption Spécifique du Corps Entier. In *Aremif'09*, Paris, France, April 30 2009.
19. A. El Habachi, E. Conil, G. Fleury, E. Vazquez, A. Hadjem, M.F. Wong, and J. Wiart. Analyse statistique de la puissance absorbée par le corps entier en radiofréquence. In *16e Journées Nationales Microondes, JNM2009*, Grenoble, France, May, 27-29 2009.
20. E. Vazquez, J. Villemonteix, M. Sidorkiewicz, and E. Walter. Global optimization based on noisy evaluations: an empirical study of two statistical approaches. In *6th International Conference on Inverse Problems in Engineering: Theory and Practice*, Dourdan, France, June 15-19 2008.
21. C. Germain Renaud, E. Vazquez, and D. Colling. Towards a statistical model of egee load. In *3rd EGEE User Forum*, Clermont-Ferrand, France, 2008.
22. J. Villemonteix, E. Vazquez, M. Sidorkiewicz, and E. Walter. Gradient-based IAGO strategy for the global optimization of expensive-to-evaluate functions and application to intake-port design. In *Advances in Global Optimization: Method and Applications*, Myconos, Greece, June 13-17 2007.
23. J. A. Egea, E. Vazquez, J. R. Banga, and R. Marti. Improved scatter search for the global optimization of computationally expensive dynamic models. In *Advances in Global Optimization: Method and Applications*, Myconos, Greece, June 13-17 2007.

D. Technical reports

1. E. Vazquez and J Bect. A new integral loss function for Bayesian optimization. Technical Report arXiv:1408.4622, arXiv.org, 2014.
2. C. Chevalier, J. Bect, D. Ginsbourger, V. Picheny, Y. Richet, and Vazquez E. Fast kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. Technical Report HAL-00641108, version 1, hal.archives-ouvertes.fr, 2011.
3. E. Vazquez and M. Piera-Martinez. Estimation of the volume of an excursion set of a Gaussian process using intrinsic kriging. Technical Report arXiv:math/0611273, arXiv.org, 2006.

E. Talks

1. E. Vazquez. New loss functions for Bayesian optimization. In *Horizon Maths 2014 – Fondation Sciences Mathématiques de Paris*, IFPEN, Reuil-Malmaison, France, December 2014.
2. E. Vazquez. Bayesian optimization with noisy evaluation results. In *LRI/TAO Seminar*, Gif-sur-Yvette, France, June 2014. LRI.
3. E. Vazquez. Conception de systèmes à partir d'expériences numériques coûteuses. In *Journée Calcul et Simulation*, Orsay, France, June 2014.
4. E. Vazquez. Conception de systèmes à partir d'expériences numériques coûteuses. In *École ASPEN 2014 – Analyse de Sensibilité, Propagation d'incertitudes et exploration numérique de modèles en sciences de l'environnement*, Les Houches, France, May 2014.
5. E. Vazquez. Conception de systèmes à partir d'expériences numériques coûteuses. In *Séminaire Éric Walter*, Gif-sur-Yvette, France, March 2014.
6. E. Vazquez. Conception de systèmes à partir d'expériences numériques coûteuses. In *Séminaire LSCE*, Gif-sur-Yvette, France, February 2014.
7. E. Vazquez. Conception de systèmes à partir d'expériences numériques coûteuses. In *WorkStat 2013–EDF R&D*, Clamart, France, November 21-22 2013.
8. J Bect and E. Vazquez. Convergence rate of a greedy learning strategy with application to nearly MMSE-optimal designs for computer experiments. In *Seminar at the Departement of Mathematics and Statistics*, Bern, Switzerland, April 2013.
9. E. Vazquez, J. Bect, and L. Li. Bayesian Subset Sampling, a kriging-based Subset Simulation algorithm for the estimation of small probabilities of failure. In *Workhop Évenements Rares, ONERA*, Palaiseau, France, November 2012.
10. E. Vazquez. Stratégies bayésiennes pour l'optimisation et l'estimation de probabilités de défaillance. In *Computer experiments and uncertainty analysis, OPUS Final workshop*, Institut Henri Poincaré, Paris, France, October 2011.
11. E. Vazquez. Bayesian optimization. In *Seminar EDF R&D and CSDL*, Clarmart, France, July 2011.
12. E. Vazquez. Sequential bayesian algorithms in the domain of computer experiments. In *High Dimensional Problems and Solutions, SMAI-AFA and the Fondation Sciences Mathematiques de Paris*, Paris, France, June 21-22 2010.
13. E. Vazquez. Bayesian optimization. how to optimize an expensive-to-evaluate function? In *Seminar of the Department of Information Management, Tilburg School of Economics and Management*, Tilburg, Netherlands, April 2010.
14. E. Vazquez. Estimation d'une probabilité de défaillance par krigeage. In *Revue annuelle du projet ANR OPUS*, Paris, France, April 2010.
15. E. Vazquez. Approximation de modèles numériques par méthodes à noyaux et applications. In *Séminaire Institut de Mathématiques de Toulouse*, Toulouse, France, March 2010.
16. E. Vazquez. Approximation de modèles numériques par méthodes à noyaux et applications. In *Séminaire Laboratoire Jacques-Louis Lions, Chevaleret*, Paris, France, March 2009.
17. E. Vazquez. Modèles non-stationnaires, fonctions aléatoires intrinsèques. In *Atelier GDR MASCOT-NUM, Institut Henri Poincaré*, Paris, France, April 2008.
18. E. Vazquez. Kriging and sequential search algorithms. In *Séminaire Groupe Spatial, AgroParisTech*, Paris, France, October 2008.
19. E. Vazquez. Kriging and sequential search algorithms. In *Workshop OPUS 2008, EDF R&D*, Clamart, France, October 2008.
20. E. Vazquez. Autour des méthodes de régression dans les RKHS. In *Rencontres du GDR MASCOT-NUM, Institut Henri Poincaré*, Paris, France, December 2008.
21. E. Vazquez. Construction de modèles de code numérique par méthodes à noyaux. In *Assemblée Générale du Département Mathématique et Informatique Appliquées, INRA*, Roquebrune-Cap-Martin, France, March 17-20 2008.
22. E. Vazquez. Construction de modèles de code numérique par méthodes à noyaux. In *Journées Incertitudes et Simulation 2007, CEA*, Saclay, France, October 3-4 2007.

3.3 Supervision

This section reviews the students I have supervised.

3.3.1 PhD theses

1. MIGUEL PIERA-MARTINEZ

Period:	December 2004–September 2008
Subject:	Modeling extreme events in systems
Defense:	September 2008, Supélec, Gif-sur-Yvette
Jury:	Maurice Lemaire, Igor Nikiforov, Michel Broniatowski, Fabien Mangeant, Emmanuel Vazquez, Eric Walter
Funding:	Fondation EADS
Supervision:	Eric Walter (thesis director), Gilles Fleury, Emmanuel Vazquez (estimation of supervision effort: 80%)
Current position:	Aerospace industry (European Spatial Agency)

2. JULIEN VILLEMONTEIX

Period:	September 2005 – December 2008
Subject:	Optimisation des performances d'un moteur à explosion à l'aide de méta-modèles
Defense:	December 2008, Supélec, Gif-sur-Yvette
Jury:	Donald Jones, Luc Pronzato, Régis Langelle, Olivier Teytaud, Maryan Sidorkiewicz, Emmanuel Vazquez, Éric Walter
Funding:	Renault SA
Supervision:	Eric Walter (thesis director), Marian Sidorkiewicz, Emmanuel Vazquez (estimation of supervision effort: 60%)
Current position:	Software design and development

3. AIMAD EL HABACHI

Period:	September 2007 – January 2011
Subject:	Characterization of the absorption of electromagnetic emissions in the human body as a function of morphological parameters.
Defense:	January 2011, Supélec, Gif-sur-Yvette
Jury:	Laurent Carraro, Gilles Fleury, Marc Helier, David Lautru, Éric Walter, Joe Wiart
Funding:	Orange Labs
Supervision:	Gilles Fleury (thesis director), Joe Wiart, Emmanuelle Conil, Emmanuel Vazquez (estimation of supervision effort: 10%)
Current position:	Postdoctoral researcher at IFSTTAR & UCBL

4. SÁNDOR BILICZ

Period:	September 2008 – May 2011
Subject:	Kriging-based optimization method for electromagnetic nondestructive control
Defense:	May 2011, Supélec, Gif-sur-Yvette
Jury:	Patrick Dular, Harsányi Gábor, Jérôme Idier, József Bíró, Dominique Lesselier
Funding:	French RTRA Digiteo & PhD scholarship grant of Hungary
Supervision:	Marc Lambert (thesis co-director), Szabolcs Gyimóthy (thesis co-director), Pierre Calmon, Emmanuel Vazquez (estimation of supervision effort: 5%)
Current position:	Assistant Professor at Budapest University of Technology and Economics

5. LING LI

Period:	September 2009 – May 2012
Subject:	Sequential design of experiments to estimate a probability of failure
Defense:	May 2012, Supélec, Gif-sur-Yvette
Jury:	Luc Pronzato, Bruno Sudret, Bertrand Iooss, Gilles Fleury, Éric Walter Julien Bect, Emmanuel Vazquez
Funding:	*Fond Unique Interministériel*
Supervision:	Gille Fleury (thesis director), Julien Bect, Emmanuel Vazquez (supervision effort: 50%)
Current position:	Research engineer in UK

6. ROMAIN BENASSI

Period:	September 2009 – June 2013
Subject:	Construction of a Bayesian optimization algorithm that maximizes the Expected Improvement (EI) criterion using a Sequential Monte Carlo approach
Defense:	June 2013, Supélec, Gif-sur-Yvette
Jury:	Olivier Cappé, Luc Pronzato, Rodolphe Le Riche, Gilles Duc, Éric Walter
Funding:	French Ministry of National Education
Supervision:	Julien Bect, Emmanuel Vazquez (thesis director, supervision effort: 50%)
Current position:	Software engineering

3.3.2 Ongoing theses

1. PAUL FÉLIOT

Period:	January 2014 – *today*
Subject:	Constrained multi-objective Bayesian optimization
Funding:	Institut de Recherche Technologique SystemX
Supervision:	Julien Bect, Emmanuel Vazquez (supervision effort: 50%), Jérôme Juillard (thesis director)

2. RÉMI STROH

Period:	January 2015 – *today*
Subject:	Design and analysis of multi-fidelity computer experiments applied to fire safety
Funding:	LNE, Laboratoire national de métrologie et d'essais
Supervision:	Julien Bect (thesis director), Sandrine Demeyer, Nicolas Fischer, Emmanuel Vazquez

3.3.3 Discontinued theses

1. SYLVAIN VINET

Period:	September 2006 – June 2008
Subject:	Black box modeling of nonlinear continuous-time systems by kriging
Funding:	French Ministry of National Education
Supervision:	Jacques Oksman, Emmanuel Vazquez (supervision effort: 100%)
Current position:	*Professeur Agrégé* in mathematics

2. BENOIT JAN

Period:	September 2012 – June 2013
Subject:	Modeling high-dimensional systems by kriging
Funding:	French Ministry of National Education
Supervision:	Jerôme Juillard (thesis director), Julien Bect, Emmanuel Vazquez (supervision effort: 50%)
Current position:	Research engineer

3.3.4 Master of Science students

1. SYLVAIN VINET

   Period: April 2006 – September 2006

Subject: Modeling of nonlinear continuous-time systems by kriging

2. VIET-HUNG TRAN

   Period: February 2009 – June 2009
   Subject: Multi-dimensional quantile estimation for vehicle collision warning

3. ROMAIN BENASSI

   Period: April 2009 – September 2009
   Subject: Bandit algorithms for continuous optimization

4. BENOIT JAN

   Period: April 2012 – September 2012
   Subject: Modeling high-dimensional systems by kriging

5. IVANA ALEKSOVSKA

   Period: April 2014 – September 2014
   Subject: Informational approach to global optimization

3.3.5 Postdoctoral fellows

1. CLAIRE CANNAMELA

   Period: September 2007 – September 2008
   Subject: Estimation of failure probabilities
   Funding: French *Agence Nationale de la Recherche*
   Supervision: Emmanuel Vazquez
   Current position: CEA

2. RÉGIS BETTINGER

   Period: October 2009 – December 2010
   Subject: Modeling of infinite-dimensional systems by kriging
   Funding: French *Agence Nationale de la Recherche*
   Supervision: Emmanuel Vazquez
   Current position: Risk Analyst Engineer

3.4 Various activities

3.4.1 Invitations to participate in PhD Jurys

1. Defense of Jose Egea

Date: May 2008
Subject: Bayesian and stochastic optimization
Place: Vigo, Spain

2. Defense of Régis Bettinger

Date: October 2009
Subject: Kriging-based sequential search strategies
Place: Sophia Antipolis, France

3. Defense of François Bachoc

Date: October 2013
Subject: Parametric estimation of covariance function in Gaussian-process models
Place: Univ. Paris 7, Paris, France

3.4.2 Project supervision

1. OPUS — OPEN PLATEFORM FOR UNCERTAINTY IN SIMULATIONS

Period: April 2008 – September 2011
Supervision: From January 2011, organization and supervision of work-package 2 (dealing with the scientific developments of the project)
Funding: French *Agence Nationale de la Recherche*
Project partners: CEA, Dassault Aviation, EADS, EDF, Softia, École Centrale Paris, INRIA, SUPELEC, Université Joseph Fourier, Université Paris 7

2. CSDL — COMPLEX SYSTEMS DESIGN LAB

Period: September 2009 – September 2012
Supervision: From January 2011, organization and supervision of work-package 2.1 (dealing with the problem of the estimation of a probability of failure)
Funding: French *Fond Universel Interministériel 7*
Project partners: Ansys, Armines, Bull, CS, Dassault Aviation, Digiteo, Distene, École Centrale Paris, EDF, Enginsoft, ENS Cachan, Esi Group, Eurodecision, MBDA, Onera, Oxalya, Renault, Samtech, Supelec

3. IRT SYSTEMX — PROGRAMME OCS (Outils de Conception et de Simulation)

Period: February 2012 – January 2013
Supervision: From February 2012 to January 2013, supervision and organization of the Programme *Outils de Conception et de Sim-*

ulation (elaboration of the specifications and the technical contents of two three-year research projects)
Project partners: IRT SystemX, Renault, EADS, ANSYS, CEA, École Centrale Paris, ESI Group, SNECMA

3.4.3 Committees

1. DIGICOSME@UPSAY, LABEX of the Foundation for Scientific Cooperation (FSC) Paris-Saclay

   Period: Since 2015
   Role: Scientific Committee

2. Groupement de Recherche MASCOT-NUM – CNRS

   Period: Since 2015
   Role: Steering Committee

3. CDS@UPSAY, Center For Data Science, Foundation for Scientific Cooperation (FSC) Paris-Saclay

   Period: Since March 2014
   Role: Steering Committee

4. IRT SYSTEMX, OCS Programme, ROM Project

   Period: Since March 2014
   Role: Steering Committee

5. NIPS Workshop on Bayesian Optimization, Experimental Design and Bandits, Sierra Nevada, Spain

   Period: December 2011
   Role: Program Committee

6. DIGITEO French RTRA

   Period: From September 2011 to December 2014
   Role: Program Committee

7. Groupement de Recherche MASCOT-NUM – CNRS

   Period: From September 2009 to December 2014
   Role: Scientific Committee

3.4.4 Workshop organization

1. *Reproducing Kernels in Black-Box Modeling* — Invited Session at the 13th IFAC Symposium on System Identification (SYSID-2003)

   Place: Rotterdam, Netherlands
   Date: August 2003

2. *Rare events and estimation of a probability of failure* — Workshop of the ANR project OPUS

   Organizers: Alberto Pasanisi, Emmanuel Vazquez
   Place: Institut Henri Poincaré, Paris, France
   Date: June 2010

3. Closing Workshop of the ANR project OPUS

   Organizers: Alberto Pasanisi, Emmanuel Vazquez
   Place: Institut Henri Poincaré, Paris, France
   Date: October 2011

Part II
Sequential search strategies based on kriging

Chapter 1
General framework

1.1 Introduction

The basic framework which underlines most kriging-related studies is that of decision theory. In its simplest form, the problem is as follows. Consider a set $\mathscr{F}$ of real-valued functions defined on a set $\mathbb{X} \subseteq \mathbb{R}^d$ and possessing some regularity property, and let $\phi : \mathscr{F} \to \mathscr{G}$ be a given mapping. Our objective is to make inference about $\phi(f)$, with $f \in \mathscr{F}$ unknown, from a finite set of (possibly noisy) evaluations of f.

Under the framework of decision theory, the act of evaluating f at a given point is an *experiment*. The experiments can be prescribed in advance or chosen adaptively, in which case the experimenter takes observations in sequence and decides at each stage which new experiment should be performed on the basis of the information thus far collected. The usual mathematical structure to describe such experiments consists of a sequence of random variables $Z_1, Z_2, \ldots \in \mathbb{R}$ that model the outcomes of the evaluations of f at points $X_1, X_2, \ldots \in \mathbb{X}$, which are also random variables, such that for all $n = 1, 2, \ldots$, X_{n+1} depends on the information $I_n = \{(X_1, Z_1), \ldots, (X_n, Z_n)\}$ collected after n experiments. The initial point X_1 may be chosen arbitrarily. We simply have $Z_n = f(X_n)$ when evaluations are exact (noiseless). Note that we use random models for the X_ns and Z_ns because the X_ns may be chosen using a stochastic procedure, or because we may assume a random model for f. The sequence of decision rules $\underline{X} = (X_1, X_2, \ldots)$ is commonly referred to as the *sampling strategy*.

To describe an estimation problem, one needs two additional objects: an estimator $\widehat{\phi}_n \in \mathscr{G}$ of $\phi(f)$ that depends measurably on I_n, and a positive function $L : \mathscr{G} \times \mathscr{G} \to \mathbb{R}_+$ such that $L(\phi(f), \widehat{\phi}_n)$ quantifies the loss incurred by choosing the estimator $\widehat{\phi}_n$ instead of $\phi(f)$.

In our work, we consider the following four classes of problems, which originate from problems of system design in the industry.

a) Approximation: find a function $\widehat{\phi}_n$ such that it is close to f in some sense.

b) Optimization: find $\widehat{\phi}_n = (x_n^\star, M_n) \in \mathbb{X} \times \mathbb{R}$ such that M_n is close to the maximum value of f over $\mathbb{X}$ and $x_n^\star$ is close to (one of) the corresponding global maximizer.
c) Estimation of the probability of failure of a system: given a probability distribution $\mathsf{P}_{\mathbb{X}}$ on $(\mathbb{X}, \mathscr{B}(\mathbb{X}))$, which models the fact that the input factors of a system are uncertain, and a threshold $u \in \mathbb{R}$ which corresponds to a critical value, find an estimate $\widehat{\phi}_n$ of the probability α of f being above u, that is,

$$\alpha = \mathsf{P}_{\mathbb{X}}\{x \in \mathbb{X} : f(x) > u\}\,.$$

d) Estimation of a quantile: given a probability distribution $\mathsf{P}_{\mathbb{X}}$ on $(\mathbb{X}, \mathscr{B}(\mathbb{X}))$ and a real number $0 \leq \alpha \leq 1$, find an estimate $\widehat{\phi}_n$ of

$$q_\alpha = \inf\{u \in \mathbb{R} : \mathsf{P}_{\mathbb{X}}\{f \leq u\} \geq \alpha\}\,.$$

Note however that a given real industrial problem can seldom be simply expressed under the form of one of the four classes stated above. In practice, several objectives may be sought at the same time. For instance, when doing simulation-based car-crash testing, one not only seeks to minimize the probability of injury of the passengers, but also, to minimize the mass of the vehicle, the cost of fabrication, etc.

The performance of the estimation procedure after n experiments depends on the terminal decision, that is, the choice of $\widehat{\phi}_n$, and of the quantity of information collected by the sampling strategy. The choice of a sampling strategy for the problems mentioned above is discussed further in the next section.

1.2 SUR sampling strategies

In our work, we focus on *stepwise uncertainty reduction* (SUR) sampling strategies [7, 66, 68, 69, ...]. Let ξ be a real-valued random process defined on some probability space $(\Omega, \mathscr{B}, \mathsf{P}_0)$ with parameter in $\mathbb{X}$, and assume that f is a sample path of ξ. From a Bayesian decision-theoretic point of view, ξ represents prior knowledge about f and makes it possible to infer a quantity of interest $\phi(\xi)$ by considering its posterior distribution after the results of the evaluations of ξ at $X_1, X_2, \ldots$ This point of view, which has been introduced in the 60s in different domains [33–35, 39, 40, 54], has been widely explored in the domain of optimization and computer experiments (see, e.g., [17, 48, 61, 62, 71]). Then, $\underline{X}$ is a random sequence in $\mathbb{X}$, with the property that X_{n+1} is measurable with respect to the σ-algebra $\mathscr{F}_n$ generated by the random variables $X_1, Z_1 = \xi(X_1), \ldots, X_n, Z_n = \xi(X_n)$. (We exclude here strategies where the X_ns would be chosen using a stochastic mechanism. Moreover, for the sake of clarity, we assume here that there is no noise.)

Under this setting, the performance of a given strategy $\underline{X}$ at step n can be assessed using the risk

$$\upsilon_n(\underline{X}) := \mathsf{E}_0\, L(\phi(\xi), \widehat{\phi}_n(\xi))\,,$$

where E_0 denotes expectation with respect to P_0.

When the total number of experiments, say N, is known in advance, it is well-known (see, e.g., [12, 13]) that there exists a strategy $\underline{X}^\star$ such that

$$\upsilon_N(\underline{X}^\star) = \inf_{\underline{X}\in\mathscr{A}} \upsilon_N(\underline{X}),$$

where $\mathscr{A}$ denotes the class of all sampling strategies, which can be formally obtained by dynamic programming. Let $r_N = \mathsf{E}_N\left(L(\phi(\xi),\widehat{\phi}_N(\xi))\right)$, where E_n, $n = 1, 2, \ldots$ denotes the conditional expectation with respect to $\mathscr{F}_n$. The quantity r_N is the terminal risk, or in other words, the actual loss incurred at the end of the estimation procedure. Define by backward induction

$$r_n = \mathsf{E}_n\big(r_{n+1} \mid X_{n+1}\big), \quad n = N-1,\ldots,0. \tag{1.1}$$

Notice that $r_n = \mathsf{E}_n\big(r_{n+1} \mid X_{n+1}\big)$ is an $\mathscr{F}_n$-measurable random variable, since X_{n+1} is measurable with respect to $\mathscr{F}_n$. Moreover, we have $r_0 = \upsilon_N(\underline{X})$. The optimal strategy $\underline{X}^\star$ is obtained by choosing, at each iteration n, the next evaluation point

$$X_{n+1}^\star = \operatorname*{arg\,min}_{x\in\mathbb{X}} \mathsf{E}_n\big(r_{n+1} \mid X_{n+1} = x\big), \quad n = 1,\ldots,N-1. \tag{1.2}$$

Unfortunately, solving (1.1)–(1.2) numerically is intractable when the horizon N is greater than a few steps.

A very general approach to construct a sub-optimal strategy consists in replacing the risk r_n with $\tilde{r}_n = E_n\big(L(\phi(\xi),\widehat{\phi}_n(\xi))\big)$ in (1.2), and to choose the next evaluation point using the following one-step look-ahead decision rule:

$$\begin{aligned} X_{n+1} &= \operatorname*{arg\,min}_{x\in\mathbb{X}} \mathsf{E}_n\big(\tilde{r}_{n+1} \,\big|\, X_{n+1} = x\big), \\ &= \operatorname*{arg\,min}_{x\in\mathbb{X}} \mathsf{E}_n\big\{\mathsf{E}_{n+1}\big(L(\phi(\xi),\widehat{\phi}_{n+1}(\xi))\big) \,\big|\, X_{n+1} = x\big\} \\ &= \operatorname*{arg\,min}_{x\in\mathbb{X}} \mathsf{E}_n\big(L(\phi(\xi),\widehat{\phi}_{n+1}(\xi)) \,\big|\, X_{n+1} = x\big). \end{aligned} \tag{1.3}$$

The $\mathscr{F}_n$-measurable random variable $\tilde{r}_n$ can often be viewed as a measure of the residual uncertainty about $\phi(\xi)$. Then, we say that the strategy (1.3) is a *stepwise uncertainty reduction* strategy. In (1.3), the function

$$\gamma_n : x \mapsto \mathsf{E}_n\big(L(\phi(\xi),\widehat{\phi}_{n+1}(\xi)) \,\big|\, X_{n+1} = x\big) \tag{1.4}$$

is viewed as a *sampling criterion*—the minimum of which indicates the next evaluation to be performed.

A fundamental example — Let $f : \mathbb{X} \to \mathbb{R}$ be a real-valued continuous function defined on a compact subset $\mathbb{X}$ of $\mathbb{R}^d$, $d \geq 1$ and consider the problem of finding an approximation of the maximum of f,

$$M = \max_{x \in \mathbb{X}} f(x)\,.$$

The unknown function f is considered as a sample path of a random process ξ defined on some probability space $(\Omega, \mathscr{B}, \mathsf{P}_0)$, with parameter $x \in \mathbb{X}$. For a given f, the efficiency of an optimization strategy $\underline{X}$ can be measured using the loss function

$$L(M, M_n) = M - M_n\,, \tag{1.5}$$

with $M_n = f(X_1) \vee \cdots \vee f(X_n)$. Notice that $M \in [M_n, M_n + m]$ with ($\mathscr{F}_n$-conditional) probability at least $1 - \mathsf{E}_n(M - M_n)/m$ by Markov's inequality. Thus, $\mathsf{E}_n(M - M_n)$ can be viewed as a measure of residual uncertainty about M. At iteration n, a SUR strategy for choosing X_{n+1} consists in minimizing the risk of $\underline{X}$ after $n+1$ evaluation results:

$$\begin{aligned} X_{n+1} &= \operatorname*{arg\,min}_{x \in \mathbb{X}} \mathsf{E}_n\left(M - M_{n+1} \mid X_{n+1} = x\right) \\ &= \operatorname*{arg\,max}_{x \in \mathbb{X}} \mathsf{E}_n\left(M_{n+1} \mid X_{n+1} = x\right) \\ &= \operatorname*{arg\,max}_{x \in \mathbb{X}} \rho_n(x) := \mathsf{E}_n\big((\xi(x) - M_n)_+\big)\,, \end{aligned} \tag{1.6}$$

where, for $z \in \mathbb{R}$, $z_+ := z \vee 0$. This Bayesian decision-theoretic point of view for optimization has been explored between 1970 and 1990 by J. Mockus, A. Žilinskas and their coauthors (see [41, 42, 64, 79] and references therein). The sampling criterion ρ_n, introduced by [42] and popularized through the EGO algorithm [31], is known as the *expected improvement* (EI). □

To derive a numerical approximation of the conditional expectation E_n that appear in (1.3), we restrict ξ to be a *Gaussian process*. When ξ is Gaussian, it is often possible (see, for instance, [7, 14]) to derive a closed-form expression of $\tilde{r}_n$, or sometimes of an upper-bound of $\tilde{r}_n$, that depends only on the posterior mean $\widehat{\xi}_n$ and the posterior standard deviation s_n of ξ given $\xi(X_1), \ldots, \xi(X_n)$. The functions $\widehat{\xi}_n$ and s_n can be computed with a moderate computational effort using the framework of kriging (see next section). Then, the conditional expectation E_n in (1.3) is a one-dimensional integral with respect to the Gaussian posterior distribution of Z_{n+1}.
Example — When ξ is a Gaussian process, the case of optimization mentioned above is a particular case of a SUR strategy where the sampling criterion itself can be obtained as a closed-form expression that depends on M_n, $\widehat{\xi}_n$ and s_n:

$$\rho_n(x) = \begin{cases} s_n(x)\,\Phi'\left(\frac{\widehat{\xi}_n(x) - M_n}{s_n(x)}\right) + (\widehat{\xi}_n(x) - M_n)\,\Phi\left(\frac{\widehat{\xi}_n(x) - M_n}{s_n(x)}\right) & \text{if } s_n(x) > 0, \\ \left(\widehat{\xi}_n(x) - M_n\right)_+ & \text{if } s_n(x) = 0, \end{cases} \tag{1.7}$$

with Φ standing for the Gaussian cumulative distribution function. □

In the next section, we recall the framework of kriging that is used to obtain the posterior distribution of a Gaussian process.

1.3 Kriging

As mentioned above, we use the framework of kriging to obtain the posterior distribution of a Gaussian process from observations. Hereafter, the notation $\xi \sim \mathrm{GP}(m,k)$ means that ξ is a Gaussian process with mean function $m : x \in \mathbb{X} \mapsto \mathsf{E}(\xi(x))$ and covariance function $k : (x,y) \in \mathbb{X}^2 \mapsto \mathrm{cov}(\xi(x),\xi(y))$, with respect to a probability space $(\Omega, \mathscr{B}, \mathsf{P}_0)$.

Consider the model ξ defined by

$$\begin{cases} \xi \sim \mathrm{GP}(m,k)\,, \\ m(\cdot) = \sum_{i=1}^{q} \beta_i p_i(\cdot)\,, \\ \beta_1, \ldots, \beta_q \in \mathbb{R}\,, \end{cases} \tag{1.8}$$

where the β_is are unknown parameters, the p_is form a basis of d-variate polynomials, and k is a continuous, strictly positive-definite function.

Notations. Let $\tilde{\Lambda}$ be the vector space of all finite-support measures on $\mathbb{X} \subseteq \mathbb{R}^d$. In other words, $\tilde{\Lambda}$ is the space of linear combinations $\sum_{i=1}^{n} c_i \delta_{x^i}$, where for all i, $c_i \in \mathbb{R}$, and δ_{x^i} stands for the Dirac measure at $x^i \in \mathbb{X}$. (In what follows, we shall use the notation $\langle \lambda, \varphi \rangle := \int \varphi \, \mathrm{d}\lambda$, where φ is a measurable function and λ is a signed measure.)

The linear map

$$\begin{array}{rccl} \xi : & \tilde{\Lambda} & \to & \mathscr{H} = \mathrm{span}\{\xi(x)\,;\, x \in \mathbb{X}\} \subset L^2(\Omega, \mathscr{B}, \mathscr{P}_0) \\ & \lambda = \sum_{i=1}^{n} c_i \delta_{x^i} & \mapsto & \xi(\lambda) := \sum_{i=1}^{n} c_i \xi(x^i)\,, \end{array} \tag{1.9}$$

extends ξ on $\tilde{\Lambda}$. We can also define the linear maps

$$m : \lambda \in \tilde{\Lambda} \mapsto \mathsf{E}(\xi(\lambda)) = \langle \lambda, m \rangle$$

and

$$k : (\lambda,\mu) \in \tilde{\Lambda}^2 \mapsto \mathrm{cov}[\xi(\lambda),\xi(\mu)] = \iint k(x,y)\,\mathrm{d}\lambda(x)\mathrm{d}\mu(y)\,.$$

The bilinear form $(\lambda,\mu)_{\tilde{\Lambda}} := k(\lambda,\mu) + \sum_{i=1}^{q} \langle \lambda, p_i \rangle \langle \mu, p_i \rangle$ defines an inner product on $\tilde{\Lambda}$ (since k is strictly positive-definite). Denote by Λ the completion of $\tilde{\Lambda}$ under this inner product. Since ξ and m are bounded under $(\cdot,\cdot)_{\Lambda}$, they can be extended on Λ by continuity. Similarly, k can be extended by continuity on Λ^2.

Let $\mathscr{P}$ be the q-dimensional vector space spanned by the functions p_i and denote by $\Lambda_{\mathscr{P}^\perp}$ the subset of elements of Λ that vanish on $\mathscr{P}$:

$$\lambda \in \Lambda_{\mathscr{P}^\perp} \implies \forall \varphi \in \mathscr{P},\ \langle \lambda, \varphi \rangle = 0\,.$$

Notice that for all $\lambda \in \Lambda_{\mathscr{P}^\perp}$, $\xi(\lambda)$ is a zero-mean random variable.

We recall now the notion of kriging prediction.

Definition 1.1 (Kriging predictor). Suppose we observe random variables $\xi^{i,\mathrm{obs}} = \xi^i + \varepsilon^i$, $i = 1, \dots, n$, where $\xi^1, \dots, \xi^n \in \mathscr{H}$, and where the ε^is are zero-mean Gaussian random variables independent of $\mathscr{H}$. Denote by λ^i the (unique) element of Λ such that $\xi^i = \xi(\lambda^i)$. For a given $\lambda \in \Lambda$, the *kriging predictor* of $\xi(\lambda) \in \mathscr{H}$ based on the observations is the linear projection

$$\widehat{\xi}_n(\lambda) := \sum_{i=1}^n a_\lambda^i \xi^{i,\mathrm{obs}} = \xi(\widehat{\lambda}_n) + \sum_{i=1}^n a_\lambda^i \varepsilon^i \quad \left(\text{with } \widehat{\lambda}_n = \sum_{i=1}^n a_\lambda^i \lambda^i\right) \tag{1.10}$$

of $\xi(\lambda)$ onto $\mathrm{span}\{\xi^{i,\mathrm{obs}}, i = 1, \dots, n\}$, such that the variance of the prediction error $\xi(\lambda) - \widehat{\xi}_n(\lambda)$ is minimized under the constraint

$$\lambda - \widehat{\lambda}_n = \lambda - \sum_{i=1}^n a_\lambda^i \lambda^i \in \Lambda_{\mathscr{P}^\perp}. \tag{1.11}$$

Proposition 1.1. *The kriging coefficients a_λ^i, $i = 1, \dots, n$, are solutions of a system of linear equations, which can be written in matrix form as*

$$\begin{pmatrix} \mathrm{K} + \mathrm{K}_\varepsilon & \mathrm{P}^\mathsf{T} \\ \mathrm{P} & 0 \end{pmatrix} \begin{pmatrix} \mathrm{a}_\lambda \\ \mathrm{b}_\lambda \end{pmatrix} = \begin{pmatrix} \mathrm{k}_\lambda \\ \mathrm{p}_\lambda \end{pmatrix}, \tag{1.12}$$

where K *is the $n \times n$ matrix with entries $k(\lambda^i, \lambda^j)$, $i, j = 1, \dots, n$,* K_ε *is the covariance matrix of the ε^is,* P *is a $q \times n$ matrix with entries $\langle \lambda^j, p_i \rangle$ for $j = 1, \dots, n$ and $i = 1, \dots, q$,* b_λ *is a vector of Lagrange multipliers,* k_λ *is a vector of size n with entries $k(\lambda^i, \lambda)$ and* p_λ *is a vector of size q with entries $\langle \lambda, p_i \rangle$, $i = 1, \dots, q$.*

Proposition 1.2. *The kriging predictor is unbiased.*

Definition 1.2 (Kriging covariance). The covariance function k_n of the prediction error is called the *kriging covariance function*. For all $\lambda, \mu \in \Lambda$,

$$\begin{aligned} k_n(\lambda, \mu) &:= \mathrm{cov}\left(\xi(\lambda) - \widehat{\xi}_n(\lambda), \xi(\mu) - \widehat{\xi}_n(\mu)\right) \\ &= k(\lambda, \mu) - \mathrm{a}_\lambda^\mathsf{T} \mathrm{k}_\mu - \mathrm{a}_\mu^\mathsf{T} \mathrm{k}_\lambda + \mathrm{a}_\lambda^\mathsf{T} (\mathrm{K} + \mathrm{K}_\varepsilon) \mathrm{a}_\mu \,, \\ &= k(\lambda, \mu) - \mathrm{a}_\mu^\mathsf{T} \mathrm{k}_\lambda - \mathrm{b}_\mu^\mathsf{T} \mathrm{p}_\lambda \,. \end{aligned} \tag{1.13}$$

Also define the *kriging standard deviation* s_n as

$$s_n(\lambda) := \sqrt{k_n(\lambda, \lambda)}, \quad \lambda \in \Lambda \,. \tag{1.14}$$

When $\lambda = \delta_x$, for some $x \in \mathbb{X}$, we shall use the simplified notations $\widehat{\xi}_n(x) := \widehat{\xi}_n(\delta_x)$ and $s_n(x) := s_n(\delta_x)$.

One fundamental property of a Gaussian process is the following proposition.

Proposition 1.3. *Let*

$$m = \sum_{i=1}^q \beta_i p_i \in \mathscr{P} \,,$$

with $\beta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, s^2)$, $s > 0$, *and* k *be a strictly positive-definite covariance function. Let* ξ^s *be a Gaussian random process such that, for all* $\lambda \in \Lambda$,

$$\xi^s(\lambda) = \eta(\lambda) + m(\lambda)\,, \tag{1.15}$$

where $\eta \sim \mathrm{GP}(0, k)$ *and, for all* λ *and all* i, $\mathsf{E}(\eta(\lambda)\beta_i) = 0$.

Given $\lambda^1, \ldots, \lambda^n \in \Lambda$, *denote by* $\xi^s \mid \mathscr{F}_n^s$ *the random process* ξ^s *conditioned on the* σ*-algebra* $\mathscr{F}_n^s$ *generated by* $\xi^{s,i} = \xi^s(\lambda^i)$, $i = 1, \ldots, n$. *Then, as* s *becomes large,*

$$\xi^s \mid \mathscr{F}_n^s \xrightarrow{\text{d}} \mathrm{GP}\left(\tilde{\xi}_n^s, k_n\right),$$

where $\tilde{\xi}_n^s(\lambda) := \sum_{i=1}^n a_\lambda^i\, \xi^{s,i}$ *and* k_n *are respectively the kriging predictor and the kriging covariance function using the coefficients* a_λ^i *from (1.12)—obtained under the model (1.8).*

Under the model (1.8), where the parameters β_i in m are deterministic but unknown, the kriging predictor $\widehat{\xi}_n$ does not correspond to the posterior mean $\mathsf{E}_n(\xi(\lambda))$ of ξ in general (because $\mathsf{E}_n(\xi(\lambda))$ depends explicitly on m whereas $\widehat{\xi}_n$ does not). Similarly, k_n does not correspond to the posterior covariance of ξ. This is not desirable under a Bayesian framework. However, Proposition 1.3 shows that $\widehat{\xi}_n$ and k_n do correspond to the posterior mean and the posterior covariance of the model (1.15) when $s \to \infty$. Note that (1.15) can be rewritten under the form of a hierarchical Bayesian model:

$$\xi^s : \quad \begin{cases} \xi^s \mid \beta_1, \ldots, \beta_q \sim \mathrm{GP}\Big(\displaystyle\sum_{i=1}^q \beta_i p_i, k\Big)\,, \\ \\ \beta_1, \ldots, \beta_q \sim \mathrm{N}(0, s^2)\,. \end{cases} \tag{1.16}$$

As $s \to \infty$, the prior (1.16) becomes improper. Thus, we can say that the kriging predictor and the kriging covariance receive a Bayesian interpretation as conditional mean and covariance functions under an improper prior, which we shall denote again by ξ [34, 47].

1.4 Consistency of the kriging predictor

The properties of the kriging estimator are partially understood. Most results come from the literature of approximation in *reproducing kernel Hilbert spaces* (RKHS).

Throughout this section, assume that the distribution of ξ is given by (1.16) with $s = \infty$, and that there is no evaluation noise.

Denote by $\mathscr{R}$ the topological dual space of $(\Lambda, \|\cdot\|_\Lambda)$, which can be identified (see [65], p. 65) to the RKHS of functions generated by the kernel

$$k' : (x,y) \mapsto k(x,y) + \sum_{i=1}^{q} p_i(x)p_i(y)\,.$$

Note that k' is the reproducing kernel of the linear space of functions $f = f_0 + f_1 + \cdots + f_q$, where f_0 is in the RKHS $(\mathscr{R}_0, \|\cdot\|_0)$ generated by the kernel k, and the f_is, $i = 1, \ldots, q$, are in the RKHSs $(\mathscr{R}_i, \|\cdot\|_i)$ generated by the kernels $x, y \mapsto p_i(x)p_i(y)$, endowed with a norm $\|\cdot\|_{\mathscr{R}}$ defined by

$$\|f\|_{\mathscr{R}}^2 = \min\left[\|f\|_0^2 + \|f\|_1^2 + \cdots + \|f\|_q^2\right], \tag{1.17}$$

the minimum taken for all the decompositions of $f = f_0 + f_1 + \cdots f_q$ with $f_i \in \mathscr{R}_i$ (see [4], p. 353).

Proposition 1.4. *Let $\lambda \in \Lambda$, $(\lambda^n)_{n\geq 1} \in \Lambda^{\mathbb{N}}$, and consider the sequence $(\widehat{\lambda}_n)_{n\geq 1}$ of kriging predictors of λ from the λ^is defined by (1.10).*

Then,

$$\lim_{n\to\infty} s_n(\lambda) = 0 \implies \forall f \in \mathscr{R}, \quad \lim_{n\to\infty} \langle \widehat{\lambda}_n, f\rangle = \langle \lambda, f\rangle\,.$$

If we take $\lambda = \delta_x$, for some $x \in \mathbb{X}$, and $\lambda^i = \delta_{x^i}$, $i = 1, 2, \ldots$, for some sequence of evaluation points $\{x^i; i \geq 1\}$ such that x is a point of adherence of this sequence, then $s_n(\delta_x)$ converges to zero (since k is continuous and for all i, $s_n(\delta_x)^2 \leq \mathrm{var}\big(\xi(x) - \xi(x^i)\big)$). Thus, *pointwise consistency* of the kriging predictor holds for all sample paths $f \in \mathscr{R}$. However, it is well-known that $\mathsf{P}_0\{\xi \in \mathscr{R}\} = 0$ (see for instance Driscoll's theorem in [37]), which means that, under the prior ξ, observations are not generated from a function coming from $\mathscr{R}$. Fortunately, $\mathscr{R}$ is in fact smaller than the space of all sample paths for which consistency holds, which is of probability one, as shown in the following proposition.

Proposition 1.5. *Let $\lambda \in \Lambda$ and $(\lambda^n)_{n\geq 1} \in \Lambda^{\mathbb{N}}$ such that*

$$\lim_{n\to\infty} s_n(\lambda) = 0.$$

Define

$$\mathscr{G} = \{f \in \mathbb{R}^{\mathbb{X}} : \lim_{n\to\infty} \langle \widehat{\lambda}_n, f\rangle = \langle \lambda, f\rangle)\}$$

Then $\{\xi \notin \mathscr{G}\}$ is P_0-negligible.

Determining interesting classes of functions included in $\mathscr{G}$ is related to the underlying properties of regularized approximation in reproducing kernel Hilbert spaces and the notion of *Lebesgue constant*. More precisely, recall that the total variation of any $\lambda = \sum_i c_i \delta_{x^i} \in \tilde{\Lambda}$ is the positive measure $|\lambda| := \sum_i |c_i| \delta_{x^i}$. For given $\lambda, \lambda^1, \ldots, \lambda^n \in \tilde{\Lambda}$, the Lebesgue constant is defined as the operator norm $|\widehat{\lambda}_n|(\mathbb{X})$ of the linear form $f \in (C(\mathbb{X}), \|\cdot\|_\infty) \mapsto \langle \widehat{\lambda}_n, f\rangle$, that maps continuous functions to their kriging interpolation at λ. This constant provides a measure of stability of the interpolation process. The following proposition specifies the role of the Lebesgue constant.

Proposition 1.6. *Assume there exists $\Phi \in L^2(\mathbb{R}^d)$ such that $k(x,y) = \Phi(x-y)$. Denote by $\widehat{\Phi} : u \mapsto \int_{\mathbb{R}^d} \Phi(x) e^{-\imath(x,u)}\,\mathrm{d}u$ the Fourier transform of Φ, and assume that Φ^{-1} has at most polynomial growth.*

Let $(x_n)_{n\geq 1} \in \mathbb{X}^{\mathbb{N}}$ be a bounded sequence and denote by $\mathbb{X}_0$ its compact closure. Set $\lambda^i = \delta_{x^i}$ for all $i \geq 1$. Then, for $x \in \mathbb{X}_0$,

$$\sup_{n\geq 1} |\widehat{\lambda}_n|(\mathbb{X}) < \infty \implies \forall f \in C(\mathbb{R}^d), \quad \lim_{n\to\infty} \langle \widehat{\lambda}_n, f\rangle = f(x)\,.$$

Conversely, if the Lebesgue constant $|\widehat{\lambda}_n|(\mathbb{X})$ is not bounded, there exists a dense subset V of $\left(C(\mathbb{R}^d), \|\cdot\|_\infty\right)$ such that, for all $f \in V$, $\sup_{n\geq 1} |\langle \widehat{\lambda}_n, f\rangle| = +\infty$

Note that Proposition 1.6 is a correction of a false claim in [76] (see [67]). Results about the Lebesgue constant for RKHS interpolation were obtained only recently by [28]. In the case Φ satisfies

$$c_1(1+\|u\|_2^2)^{-s} \leq \widehat{\Phi}(u) \leq c_2(1+\|u\|_2^2)^{-s} \quad (u \in \mathbb{R}^d) \tag{1.18}$$

with $s > d/2$ and constants $0 < c_1 \leq c_2$, [28] proves that the Lebesgue constant is bounded for *quasi-uniform* designs; that is, when there exists $c > 0$ such that for all $n \geq 1$,

$$q_n \leq h_n \leq c q_n,$$

where

$$q_n = \frac{1}{2} \min_{1\leq i<j\leq n} |x_i - x_j|$$

is the *separation distance* of (x_n) and

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

is the *fill distance* of (x_n).

1.5 Choice of a covariance function

1.5.1 Matérn Gaussian processes

As mentioned above, the rationale for restricting ξ to be a Gaussian process, or in other words, for choosing a Gaussian prior for f, is that the posterior distribution of ξ given observational data can be computed with a moderate computational effort. However, experience suggests that a Gaussian process prior carries a high amount of information about f, and it is often difficult in practice to elicit such a prior before any evaluation is made.

For numerical studies, we most often use a Gaussian process with a low-degree unknown polynomial mean and a *Matérn* covariance function whose parameters are

also assumed to be unknown. The Matérn covariance is a general purpose positive-definite function on $\mathbb{R}^d$, for every d. Recall that the Matérn correlation structure may be written as

$$\kappa_\nu(h) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(2\nu^{1/2}h\right)^\nu \mathscr{K}_\nu \left(2\nu^{1/2}h\right) \qquad (h \in \mathbb{R}_+) \tag{1.19}$$

where Γ is the Gamma function and $\mathscr{K}_\nu$ is the modified Bessel function of the third kind [26, 27, 63]. A corresponding anisotropic covariance function may be written as

$$k_{\{\theta=(\sigma,\nu,\rho_1,\ldots,\rho_d)\}}(x,y) = \sigma^2 \kappa_\nu \left(\sqrt{\sum_{i=1}^d \frac{(x_{[i]} - y_{[i]})^2}{\rho_i^2}} \right) \qquad (x,y \in \mathbb{R}^d). \tag{1.20}$$

The parameter $\nu > 0$ controls regularity of the covariance, since for scale parameters $\rho_1 = \cdots = \rho_d = 1$, the Fourier transform $\widehat{g}$ of $g : h \in \mathbb{R}^d \mapsto k_\theta(0,h)$ may be written up to a multiplicative constant as (Theorem 6.13 in [72])

$$\widehat{g}(u) \propto (4\nu + |u|^2)^{-\nu-d/2} \qquad (u \in \mathbb{R}^d).$$

Of course, there are other interesting choices for correlations structures: the inverse multiquadrics $x \in \mathbb{R}^d \mapsto (c^2 + \|x\|_2^2)^{-\beta}$ (with $c > 0$, $\beta > 0$), Wendland's compactly supported kernels [72]... We recommend against considering the Gaussian covariance function in the context of sequential search algorithms because it may yield inconsistent estimators; see [67, 77].

To deal with the unknown parameters of the covariance function, we generally choose a prior distribution for those parameters (see, e.g., [26, 32, 47, 56]) and we follow one of the two approaches described below.

1.5.2 Standard fully Bayesian approach

To deal with the parameters of the covariance function, we consider the model

$$\xi : \quad \begin{cases} \xi \mid \beta_1, \ldots, \beta_q, \theta \sim \mathrm{GP}\Big(\sum_{i=1}^q \beta_i p_i, k_\theta\Big), \\ \beta_1, \ldots, \beta_q \sim \mathrm{N}(0,\infty), \\ \theta \sim \pi_0. \end{cases} \tag{1.21}$$

In geostatistics and spatial statistics, the introduction of a prior for the parameters of the covariance function is frequently called *Bayesian kriging* [26, 49, 58]. The predictive distribution of ξ is generally obtained using conjugate priors (see, e.g., [11, 26, 36, 56]) or Monte Carlo sampling techniques (see, e.g., [18, 30, 45, 46, 74]).

Methods for approximating the posterior distribution of ξ have also been proposed [32, 44].

Under the model (1.21), the SUR sampling criterion (1.4) may be written as

$$\begin{aligned}\gamma_n(x) &= \mathsf{E}_n\big(L(\phi(\xi),\widehat{\phi}_{n+1}(\xi)) \,\big|\, X_{n+1}=x\big)\\ &= \mathsf{E}_n\Big\{\mathsf{E}_n\big[L(\phi(\xi),\widehat{\phi}_{n+1}(\xi)) \,\big|\, \theta, X_{n+1}=x\big] \,\Big|\, X_{n+1}=x\Big\}\\ &= \int \bar{\gamma}_n(x;\theta)\,\mathrm{d}\pi_n(\theta) \end{aligned} \tag{1.22}$$

where $\bar{\gamma}_n(x;\theta) = \mathsf{E}_n\big(L(\phi(\xi),\widehat{\phi}_{n+1}(\xi)) \,\big|\, \theta, X_{n+1}=x\big)$ and π_n stands for the posterior distribution of θ. Sampling techniques (Monte Carlo Markov Chains, Sequential Monte Carlo…) are generally used to approximate the integral with respect to π_n in (1.22) (see, e.g., [10, 25, 50–52] and references therein). The corresponding computational is often significant. It may be useful to obtain a closed-form expression of the integral with respect to components of θ when possible, using conjugate priors, as illustrated in the following example.

Example — Let r be a correlation function and assume that m and σ^2 are independent, with $m \sim \mathrm{N}(0,\infty)$ and σ^2 following an inverse-gamma distribution $\mathrm{IG}(a_0,b_0)$ with shape parameter a_0 and scale parameter b_0. Consider the case of a process ξ with conditional distribution $\xi \mid m,\sigma^2 \sim \mathrm{GP}(m,\sigma^2 r)$. The prior for σ^2 is conjugate [8, 9, 21, 36, 58, 73]: the conditional distribution of σ^2 given $\mathscr{F}_n$ is $\mathrm{IG}(a_n,b_n)$, with

$$\begin{cases} a_n &= a_0 + \frac{n-1}{2}\,,\\ b_n &= b_0 + \frac{1}{2}\left(\underline{\xi}_n - \widehat{m}_n \mathbb{1}_n\right)^{\mathsf{T}} \mathrm{R}_n^{-1}\left(\underline{\xi}_n - \widehat{m}_n \mathbb{1}_n\right)\,, \end{cases}$$

where $\underline{\xi}_n = (\xi(X_1),\ldots,\xi(X_n))^{\mathsf{T}}$, $\mathbb{1}_n = (1,\ldots,1)^{\mathsf{T}} \in \mathbb{R}^n$, R_n is the correlation matrix of $\underline{\xi}_n$, and $\widehat{m}_n = \frac{\mathbb{1}_n^{\mathsf{T}}\mathrm{R}_n^{-1}\underline{\xi}_n}{\mathbb{1}_n^{\mathsf{T}}\mathrm{R}_n^{-1}\mathbb{1}_n}$ is the weighted least squares estimate of m. Let

$$\varsigma_n^2(x) = \frac{b_n}{a_n}\left(1 - \mathrm{r}_n(x)^{\mathsf{T}}\mathrm{R}_n^{-1}\mathrm{r}_n(x) + \frac{(1-\mathrm{r}_n(x)^{\mathsf{T}}\mathrm{R}_n^{-1}\mathbb{1}_n)^2}{\mathbb{1}_n^{\mathsf{T}}\mathrm{R}_n^{-1}\mathbb{1}_n}\right),$$

where $\mathrm{r}_n(x)$ is the correlation vector between $\xi(x)$ and $\underline{\xi}_n$.

Then, for all $x \in \mathbb{X}$, the posterior distribution of $\xi(x)$ may be written as

$$\frac{\xi(x) - \widehat{\xi}_n(x)}{\varsigma_n(x)} \mid \mathscr{F}_n \sim t_{\eta_n},$$

with t_η the Student distribution with $\eta > 0$ degrees of freedom and $\eta_n = 2a_n$. In other words, the predictive distribution at x is a location-scale Student distribution with η_n degrees of freedom, location parameter $\widehat{\xi}_n(x)$ and scale parameter $\varsigma_n(x)$.

Using ξ as a model for a function f to be optimized, the EI criterion (1.6) has a closed-form expression, which is a generalization of (1.7) [9]:

$$\mathsf{E}_n\big((\xi(x)-M_n)_+\big) \;=\; \varsigma_n(x)\left(\frac{\eta_n+u^2}{\eta_n-1}F'_{\eta_n}(u)+u\,F_{\eta_n}(u)\right), \tag{1.23}$$

with $u=\big(\widehat{\xi}_n(x)-M_n\big)/\varsigma_n(x)$ and F_η being the cumulative distribution function of t_η. □

1.5.3 Empirical Bayes approach

Very often however, we use the popular *empirical Bayes* approach, in which the prior is estimated from the data inside a family of prior processes $\{\xi_\theta;\theta\in\Theta\}$ of the type (1.16) with $s=\infty$, in contrast to the standard fully Bayesian approach where the prior is fixed before any data are observed. Regarding SUR-based sequential search strategies, this means that we are considering a *family* of Bayesian procedures indexed by θ, corresponding to sampling criteria $\gamma_{\theta,n}:x\mapsto\mathsf{E}_n\big(L(\phi(\xi_\theta),\widehat{\phi}_{n+1}(\xi_\theta))\;\big|\;X_{n+1}=x\big)$. Then, we choose at each iteration of the sequential strategy a sampling criterion $\gamma_{\widehat{\theta}_n,n}$ corresponding to a member $\xi_{\widehat{\theta}_n}$ in the family of priors that we think models observed data well. This is often called a *plug-in* approach.

Note that we can also think of the plug-in approach as an *approximate* fully Bayesian method for dealing with covariance parameters: when the posterior distribution π_n of θ is concentrated enough around a point estimate $\widehat{\theta}_n$ of θ,

$$\gamma_n(x)=\int\bar{\gamma}_n(x;\theta)\,\mathrm{d}\pi_n(\theta)\;\approx\;\bar{\gamma}_n(x;\widehat{\theta}_n)=\gamma_{\widehat{\theta}_n,n}(x)\,.$$

Most often in practice, we use restricted maximum likelihood (REML) for estimating θ [15, 29, 55, 63]. We prefer using REML estimation over maximum likelihood estimation because REML estimation does not require to integrate out the mean of ξ_θ to carry out the estimation of θ. Recall that estimating the parameters of the covariance by REML involves writing the likelihood of some zero-mean differences of the observed data $\underline{\xi}_{n,\theta}=(\xi_\theta(\lambda^1),\ldots,\xi_\theta(\lambda^n))^\mathsf{T}$, also called contrasts. Contrasts are obtained by constructing a $n\times(n-q)$ matrix W of rank $n-q$ such that

$$\mathrm{PW}=0\,,$$

with P as in (1.12). (The columns of W are in the null space of P.) Then, for all $i\in\{1,\ldots,n\}$, $\sum_{j=1}^n \mathrm{W}_{[i,j]}\lambda^j\in\tilde{\Lambda}_{\mathscr{P}^\perp}$. The contrast vector $\mathrm{Z}=\mathrm{W}^\mathsf{T}\underline{\xi}_{\theta,n}\in\mathbb{R}^{n-l-1}$ is a zero-mean Gaussian random vector with covariance matrix $\mathrm{W}^\mathsf{T}\mathrm{K}_\theta\mathrm{W}$, where K_θ is the covariance matrix of $\underline{\xi}_{\theta,n}$ with entries $k_\theta(\lambda^i,\lambda^j)$. The log-likelihood of contrasts can be written as

$$l(\mathrm{z};\theta)=-\frac{n-q}{2}\log 2\pi-\frac{1}{2}\log\det(\mathrm{W}^\mathsf{T}\mathrm{K}_\theta\mathrm{W})-\frac{1}{2}\mathrm{z}^\mathsf{T}(\mathrm{W}^\mathsf{T}\mathrm{K}_\theta\mathrm{W})^{-1}\mathrm{z}. \tag{1.24}$$

To the best of our knowledge, determining the properties of the REML estimator of the parameters of a covariance is still an open problem—in the setting of *infill* asymptotics. We refer the reader to [5] for a comprehensive review of current results. One important result is that only *microergodic parameters* ([63], p. 162) of the covariance function may be estimated consistently [2, 78]. In particular, parameters ν, σ and $\rho_1, \ldots, \rho_d$ of the Matérn covariance interplay, and literature suggests that one should not expect to be able to identify all parameters from data. To deal with this issue, we suggest using a penalized version of REML:

$$\tilde{l}(\mathrm{z};\theta) = -\frac{n-q}{2}\log 2\pi - \frac{1}{2}\log\det(\mathrm{W}^\mathsf{T}\mathrm{K}_\theta\mathrm{W}) - \frac{1}{2}\mathrm{z}^\mathsf{T}(\mathrm{W}^\mathsf{T}\mathrm{K}_\theta\mathrm{W})^{-1}\mathrm{z} - v(\theta)\,,$$

where $v(\theta) = (\theta-\theta^\star)^\mathsf{T}\mathrm{V}^{-1}(\theta-\theta^\star)$ is a quadratic penalization that constrains θ in a neighborhood of a parameter $\theta^\star$ that is deemed plausible by the user.

Eventually, note that using a stationary Matérn Gaussian process as above has the merit of simplicity although some theoretical questions remain open. This framework has been extended in several directions in the literature: non-Gaussian random processes [18, 24, ...], non-stationary and local models [3, 6, 16, 20, 22, 53, 57, 59, ...], modeling based on ANOVA decompositions and sensitivity analysis [19, 23, 38, 43, 70, ...].

1.6 Proofs

1.6.1 Proof of Proposition 1.1

The kriging coefficients a_λ^i, $i = 1,\ldots,n$, are solutions of a quadratic problem with linear constraints, which can be written under matrix form as:

$$\begin{cases} \min_{\mathrm{a}_\lambda\in\mathbb{R}^n} \quad k(\lambda,\lambda) - 2\mathrm{a}_\lambda^\mathsf{T}\mathrm{k}_\lambda + \mathrm{a}_\lambda^\mathsf{T}(\mathrm{K}+\mathrm{K}_\varepsilon)\mathrm{a}_\lambda\,, \\ \mathrm{P}\mathrm{a}_\lambda = p_\lambda\,. \end{cases}$$

It is well-known that such a problem can be rewritten under the form (1.12) using the method of Lagrange multipliers which express collinearity of the gradients of the linear form $\mathrm{P}\mathrm{a}_\lambda$ and the quadratic form $k(\lambda,\lambda) - 2\mathrm{a}_\lambda^\mathsf{T}\mathrm{k}_\lambda + \mathrm{a}_\lambda^\mathsf{T}(\mathrm{K}+\mathrm{K}_\varepsilon)\mathrm{a}_\lambda$. □

1.6.2 Proof of Proposition 1.2

Since $\lambda - \widehat{\lambda}_n \in \Lambda_{\mathscr{P}^\perp}$. □

1.6.3 Proof of Proposition 1.3

The solution $(\mathrm{a}_\lambda, b_\lambda)$ of (1.12) under the model $\xi \sim \mathrm{GP}(m,k)$, where m is assumed deterministic, can be written as

$$\begin{cases} \mathrm{a}_\lambda &= \mathrm{K}^{-1}k_\lambda - \mathrm{K}^{-1}\mathrm{P}^\mathsf{T}(\mathrm{PK}^{-1}\mathrm{P}^\mathsf{T})^{-1}(\mathrm{PK}^{-1}k_\lambda - p_\lambda)\,, \\ b_\lambda &= (\mathrm{PK}^{-1}\mathrm{P}^\mathsf{T})^{-1}(\mathrm{PK}^{-1}k_\lambda - p_\lambda)\,. \end{cases}$$

For $\beta_i \overset{\text{i.i.d.}}{\sim} \mathscr{N}(0,s^2)$, m is a Gaussian process with distribution $\mathrm{GP}(0,k')$ where, for all λ and μ in Λ,

$$k'(\lambda,\mu) = p_\lambda^\mathsf{T}\Sigma p_\mu\,,$$

with $\Sigma = s^2\mathrm{I}_q$ standing for the covariance matrix of the random vector $(\beta_1,\ldots,\beta_q)$.

Since the mean of ξ^s is zero, $\xi^s \mid \mathscr{F}_n$ is a Gaussian process with distribution $\mathrm{GP}(\widehat{\xi}_n^s, k_n^s)$, where $\widehat{\xi}_n^s$ and k_n^s stand for the kriging predictor and the kriging covariance function from observations $\xi^{s,i} = \xi^s(\lambda^i)$, $i = 1,\ldots,n$, such that

$$\widehat{\xi}_n^s(\lambda) = \mathsf{E}_n(\xi^s(\lambda)) \qquad (\lambda \in \Lambda)$$

and

$$k_n^s(\lambda,\mu) = \mathrm{cov}\left(\xi^s(\lambda) - \widehat{\xi}_n^s(\lambda), \xi^s(\mu) - \widehat{\xi}_n^s(\mu)\right) \qquad (\lambda,\mu \in \Lambda)$$

Note that $\widehat{\xi}_n^s(\lambda)$ is also the orthogonal projection of $\xi^s(\lambda)$ onto $\mathrm{span}\{\xi^{s,i}, i = 1,\ldots,n\}$. Thus,

$$\widehat{\xi}_n^s(\lambda) = \sum_{i=1}^n \mathrm{a}_\lambda^{s,i}\xi^{s,i},$$

where the $\mathrm{a}_\lambda^{s,i}$s satisfy

$$\mathrm{a}_\lambda^s = (\mathrm{a}_\lambda^{s,1} \;\ldots\; \mathrm{a}_\lambda^{s,n})^\mathsf{T} = (\mathrm{K}+\mathrm{P}^\mathsf{T}\Sigma\mathrm{P})^{-1}(k_\lambda + \mathrm{P}^\mathsf{T}\Sigma p_\lambda)\,.$$

Using the Sherman-Morrison-Woodburry formula applied to the term $(\mathrm{K}+\mathrm{P}^\mathsf{T}\Sigma\mathrm{P})^{-1}$, rewrite a_λ^s as

$$\begin{aligned} \mathrm{a}_\lambda^s =& \mathrm{K}^{-1}k_\lambda - \mathrm{K}^{-1}\mathrm{P}^\mathsf{T} \\ &\cdot\left[(\Sigma^{-1}+\mathrm{PK}^{-1}\mathrm{P}^\mathsf{T})^{-1}\mathrm{PK}^{-1}k_\lambda - \left(\mathrm{I}_q - (\Sigma^{-1}+\mathrm{PK}^{-1}\mathrm{P}^\mathsf{T})^{-1}\mathrm{PK}^{-1}\mathrm{P}^\mathsf{T}\right)\Sigma p_\lambda\right]. \end{aligned}$$

Then, for s large enough,

$$\begin{aligned} &\left(\mathrm{I}_q - (\Sigma^{-1} + \mathrm{PK}^{-1}\mathrm{P}^\mathsf{T})^{-1}\mathrm{PK}^{-1}\mathrm{P}^\mathsf{T}\right)\Sigma p_\lambda \\ &\qquad = \left(\mathrm{I}_q - \left(\mathrm{I}_q + (\mathrm{PK}^{-1}\mathrm{P}^\mathsf{T})^{-1}\Sigma^{-1}\right)^{-1}\right)\Sigma p_\lambda \\ &\qquad = (\mathrm{PK}^{-1}\mathrm{P}^\mathsf{T})^{-1}p_\lambda - (\mathrm{PK}^{-1}\mathrm{P}^\mathsf{T})^{-1}\Sigma^{-1}(\mathrm{PK}^{-1}\mathrm{P}^\mathsf{T})^{-1}p_\lambda + o(s^{-1})\,, \end{aligned}$$

since

$$\begin{aligned}&\left(\mathrm{I}_q + (\mathrm{P}\mathrm{K}^{-1}\mathrm{P}^\mathsf{T})^{-1}\Sigma^{-1}\right)^{-1}\\ &\quad = \mathrm{I}_q - (\mathrm{P}\mathrm{K}^{-1}\mathrm{P}^\mathsf{T})^{-1}\Sigma^{-1} + (\mathrm{P}\mathrm{K}^{-1}\mathrm{P}^\mathsf{T})^{-1}\Sigma^{-1}(\mathrm{P}\mathrm{K}^{-1}\mathrm{P}^\mathsf{T})^{-1}\Sigma^{-1} + o(s^{-2})\end{aligned}$$

(Neumann series).

Therefore, $\mathrm{a}_\lambda = (a_\lambda^1, \ldots, a_\lambda^n)^\mathsf{T} = \lim_{s\to\infty} \mathrm{a}_\lambda^s$, and $\widehat{\xi}_n^s(\lambda) \xrightarrow{\text{p.s.}} \tilde{\xi}_n^s(\lambda) = \sum_i a_\lambda^i \xi^{i,s}$. Moreover, it can be easily shown that

$$\mathrm{a}_\lambda^{s\,\mathsf{T}}\mathrm{P}^\mathsf{T} = p_\lambda^\mathsf{T} + \left((\mathrm{P}\mathrm{K}^{-1}k_\lambda)^\mathsf{T} - p_\lambda^\mathsf{T}\right)(\mathrm{P}\mathrm{K}^{-1}\mathrm{P}^\mathsf{T})^{-1}\Sigma^{-1} + o(s^{-1})\,.$$

Thus, as $s \to \infty$, the kriging covariance function k_n^s can be written as

$$\begin{aligned}k_n^s(\lambda,\mu) &= k(\lambda,\mu) - \mathrm{a}_\lambda^{s\,\mathsf{T}} k_\mu + (p_\lambda^\mathsf{T} - \mathrm{a}_\lambda^{s\,\mathsf{T}}\mathrm{P}^\mathsf{T})\Sigma p_\mu\\ &= k(\lambda,\mu) - \mathrm{a}_\lambda^\mathsf{T} k_\mu - b_\lambda^\mathsf{T} p_\mu + o(1)\\ &= k_n(\lambda,\mu) + o(1)\,.\end{aligned}$$

□

1.6.4 Proof of Proposition 1.4

Since $\lambda - \widehat{\lambda}_n \in \Lambda_{\mathscr{P}^\perp}$,

$$\begin{aligned}\|\lambda - \widehat{\lambda}_n\|_\Lambda^2 &= k(\lambda - \widehat{\lambda}_n, \lambda - \widehat{\lambda}_n) = \|\xi(\lambda) - \xi(\widehat{\lambda}_n)\|_{L^2}^2 = \|\xi(\lambda) - \widehat{\xi}_n(\lambda)\|_{L^2}^2\\ &= s_n(\lambda)^2\,.\end{aligned}$$

Therefore, the convergence $\widehat{\lambda}_n \to \lambda$ holds strongly in Λ if and only if the kriging predictor is $L^2(\Omega, \mathscr{A}, \mathsf{P})$-consistent at λ; that is, if the kriging variance $s_n(\lambda)^2$ converges to zero. Since strong convergence implies weak convergence, if $\lim_{n\to\infty} s_n(\lambda)^2 = 0$, then for all functions in the dual $\mathscr{R}$ of Λ

$$\lim_{n\to\infty} \langle \widehat{\lambda}_n, f\rangle = \langle \lambda, f\rangle.$$

□

1.6.5 Proof of Proposition 1.5

Since ξ^s defined by (1.16) is a zero-mean process, $\widehat{\xi}_n^s(\lambda) = \mathsf{E}(\xi^s(\lambda) \mid \mathscr{F}_n)$ a.s. Thus, $\left(\mathsf{E}(\xi^s(\lambda) \mid \mathscr{F}_n)\right)_n$ is an L^2-bounded martingale sequence, and we know that

$\big(\widehat{\xi}_n^s(\lambda)\big)_n$ converges a.s. and in L^2-norm to the same limit (see, e.g., [75]). Therefore, $k_n^s(\lambda,\lambda) \to 0 \implies \langle \widehat{\lambda}_n^s, \xi^s \rangle \overset{L^2,\,\text{a.s.}}{\longrightarrow} \langle \lambda, \xi^s \rangle$. □

1.6.6 Proof of Proposition 1.6

Recall that the Sobolev space $H^s(\mathbb{R}^d)$ can be defined as the class of elements $f \in L^2(\mathbb{R}^d)$ such that $u \mapsto u^\alpha \widehat{f}(u) \in L^2(\mathbb{R}^d)$ for all multi-indices α satisfying $|\alpha| \le s$, where $\widehat{f}$ stands for the Fourier transform of f. $H^s(\mathbb{R}^d)$ is endowed with the inner product

$$(f,g)_s = \int (1+\|u\|_2^2)^s \widehat{f}(u)\overline{\widehat{g}(u)}\,\mathrm{d}u\,.$$

Recall also that the space $C_0^\infty(\mathbb{R}^d)$ of C^∞-functions with compact support on $\mathbb{R}^d$ is dense in H^s for every s (see, e.g., [1], Theorem 3.18). $C_0^\infty(\mathbb{R}^d)$ is also dense in the class $C(\mathbb{R}^d)$ of continuous functions for the topology of uniform convergence on compact sets (see, e.g., [1], Lemma 2.18).

We have the following properties:

(i) For all $f \in \mathscr{R}_0$, we have (see, e.g., [65])

$$\|f\|_{\mathscr{R}_0}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \big|\tilde{f}(u)\big|^2\, \widehat{\Phi}(u)^{-1}\,\mathrm{d}u\,.$$

(ii) Since $\widehat{\Phi}^{-1}$ has at most polynomial growth, there exists s_0 such that for all $f \in \mathscr{R}_0$,

$$\|f\|_{\mathscr{R}_0}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \big|\tilde{f}(u)\big|^2\, \widehat{\Phi}(u)^{-1} du \le \frac{C}{(2\pi)^d} \int_{\mathbb{R}^d} \big|\tilde{f}(u)\big|^2 \left(1+\|u\|_2^2\right)^{s_0} du < +\infty\,.$$

Thus, $H^{s_0}(\mathbb{R}^d) \hookrightarrow \mathscr{R}_0$.

(iii) $\mathscr{R}_0$ is continuously embedded in $\mathscr{R}$, due to (1.17).

Using (ii)-(iii) gives $C_0^\infty(\mathbb{R}^d) \subset \mathscr{R}$. Thus,

$$x \in \mathbb{X}_0 \implies \forall f \in C_0^\infty(\mathbb{R}^d), \quad \lim_{n\to\infty} \langle \widehat{\lambda}_n, f \rangle = f(x)\,. \tag{1.25}$$

Let $f \in C(\mathbb{R}^d)$, and let (ϕ_k) be a sequence of $C_0^\infty(\mathbb{R}^d)$-functions that converges to f uniformly on $\mathbb{X}_0$. Then,

$$\begin{aligned} \big|\langle \widehat{\lambda}_n, f\rangle - f(x)\big| &\le \big|\langle \widehat{\lambda}_n, f-\phi_k\rangle\big| + \big|\langle \widehat{\lambda}_n - \delta_x, \phi_k\rangle\big| + \big|\phi_k(x) - f(x)\big| \\ &\le \big(1+|\widehat{\lambda}_n|(\mathbb{X}_0)\big) \sup_{\mathbb{X}_0} \big|f-\phi_k\big| + \big|\langle \widehat{\lambda}_n - \delta_x, \phi_k\rangle\big|\,. \end{aligned}$$

If we assume that the Lebesgue constant is upper-bounded by $K > 0$, then

$$\limsup_{n\to\infty} \left|\langle \widehat{\lambda}_n, f\rangle - f(x)\right| \;\leq\; \left(1+K\right) \sup_{\mathbb{X}_0} \left|f-\phi_k\right| \;\xrightarrow[k\to\infty]{}\; 0\,.$$

Conversely, if the Lebesgue constant is not bounded, there exists a dense subset V of $\left(C(\mathbb{R}^d), \|\cdot\|_\infty\right)$ such that, for all $f \in V$, $\sup_{n\geq 1} |\langle \widehat{\lambda}_n, f\rangle| = +\infty$ (Banach-Steinhaus theorem; see, e.g., Section 5.8 of [60]).

References

1. R. A. Adams, *Sobolev spaces*, Academic Press, 1975.
2. E. B. Anderes, *On the consistent separation of scale and variance for Gaussian random fields*, Ann. Statist. (2010), 870–893.
3. E. B. Anderes and M. L. Stein, *Local likelihood estimation for nonstationary random fields*, J. Multivariate Anal. **102** (2011), no. 3, 506–520.
4. N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc. **68** (1950), 337–404.
5. F. Bachoc, *Parametric estimation of covariance function in Gaussian-process based kriging models. application to uncertainty quantification for computer experiments*, Ph.D. thesis, Université Paris-Diderot-Paris VII, 2013.
6. S. Banerjee, B. P. Carlin, and A. E. Gelfand, *Hierarchical modeling and analysis for spatial data*, 2nd ed., Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Crc Press, 2014.
7. J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez, *Sequential design of computer experiments for the estimation of a probability of failure*, Statist. Comput. **22** (2012), no. 3, 773–793.
8. R. Benassi, *Nouvel algorithme d'optimisation bayésien utilisant une approche monte-carlo séquentielle.*, Ph.D. thesis, Supelec, 2013.
9. R. Benassi, J. Bect, and E. Vazquez, *Robust Gaussian process-based global optimization using a fully Bayesian expected improvement criterion*, Learning and Intelligent Optimization, 5th International Conference, LION 5 (Rome, Italy) (Coello-Coello C. A., ed.), Springer, January 17-21 2011, pp. 176–190.
10. ———, *Bayesian optimization using sequential Monte Carlo*, Learning and Intelligent Optimization – 6th International Conference, LION 6 (Paris, France) (Y. Hamadi and M. Schoenauer, eds.), Springer, January 16-20 2012, pp. 339–342.
11. J. O. Berger, V. De Oliveira, and B. Sansó, *Objective Bayesian analysis of spatially correlated data*, J. Amer. Statist. Assoc. **96** (2001), no. 456, 1361–1374.
12. D. A. Berry and B. Fristedt, *Bandit problems: sequential allocation of experiments*, Chapman & Hall, 1985.
13. D. P. Bertsekas, *Dynamic programming and optimal control vol. 1*, Athena Scientific, 1995.
14. C. Chevalier, J. Bect, D. Ginsbourger, Y. Richet, V. Picheny, and E. Vazquez, *Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set*, Technometrics **56** (2014), no. 4, 455–465.
15. N. Cressie, *Statistics for spatial data*, Wiley, New York, 1993.
16. N. Cressie and G. Johannesson, *Fixed rank kriging for very large spatial data sets*, J. R. Stat. Soc. Ser. B Stat. Method. **70** (2008), no. 1, 209–226.
17. C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker, *Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments*, J. Amer. Statist. Assoc. **86** (1991), no. 416, 953–963.
18. P. J. Diggle, J. A. Tawn, and R. A. Moyeed, *Model-based geostatistics*, Journal of the Royal Statistical Society: Series C (Applied Statistics) **47** (1998), no. 3, 299–350.

19. N. Durrande, D. Ginsbourger, O. Roustant, and L. Carraro, *ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis*, J. Multivariate Anal **115** (2013), 57–67.
20. M. Fuentes, *A high frequency kriging approach for non-stationary environmental processes*, Environmetrics **12** (2001), no. 5, 469–483.
21. M. Gaudard, M. Karson, E. Linder, and D. Sinha, *Bayesian spatial prediction*, Environmental and Ecological Statistics **6** (1999), no. 2, 147–171.
22. A. E. Gelfand, A. M. Schmidt, S. Banerjee, and C. F. Sirmans, *Nonstationary multivariate process modeling through spatially varying coregionalization*, TEST **13** (2004), no. 2, 263–312.
23. D. Ginsbourger, O. Roustant, D. Schuhmacher, N. Durrande, and N. Lenz, *On ANOVA decompositions of kernels and Gaussian random field paths*, Tech. Report arXiv:1409.6008, arXiv.org, 2014.
24. R. B. Gramacy and H. Lee, *Bayesian treed Gaussian process models with an application to computer modeling*, J. Amer. Statist. Assoc. **103** (2008), no. 483, 1119–1130.
25. R. B. Gramacy and N. G. Polson, *Particle learning of Gaussian process models for sequential design and optimization*, Journal of Computational and Graphical Statistics **20** (2011), no. 1, 102–118.
26. M. S. Handcock and M. L. Stein, *A Bayesian analysis of kriging*, Technometrics **35** (1993), no. 4, 403–410.
27. M. S. Handcock and J. R. Wallis, *An approach to statistical spatial-temporal modeling of meteorological fields*, J. Amer. Statist. Assoc. **89** (1994), no. 426, 368–378.
28. T. Hangelbroek, F. J. Narcowich, and J. D. Ward, *Kernel approximation on manifolds i. bounding the lebesgue constant*, SIAM J. Math. Anal. **42** (2010), 1732–1760.
29. D. A. Harville, *Bayesian inference for variance components using only the error contrasts*, Biometrika **61** (1974), 383–385.
30. D. Higdon, J. Gattiker, B. Williams, and M. Rightley, *Computer model calibration using high-dimensional output*, J. Amer. Statist. Assoc. **103** (2008), no. 482, 570–583.
31. D. R. Jones, M. Schonlau, and W. J. Welch, *Efficient global optimization of expensive black-box functions*, J. Global Optim. **13** (1998), 455–492.
32. M. C. Kennedy and A. O'Hagan, *Bayesian calibration of computer models*, Journal of the Royal Statistical Society. Series B (Statistical Methodology) **63** (2001), no. 3, 425–464.
33. G. Kimeldorf and G. Wahba, *Spline functions and stochastic processes*, Sankhyä: the Indian Journal of Statistics: Series A **32** (1970), no. 2, 173–180.
34. G. S. Kimeldorf and G. Wahba, *A correspondence between Bayesian estimation on stochastic processes and smoothing by splines*, Ann. Math. Statist. **41** (1970), no. 2, 495–502.
35. H. J. Kushner, *A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise*, J. Basic Engineering **86** (1964), 97–106.
36. M. Locatelli and F. Schoen, *An adaptive stochastic global optimization algorithm for one-dimensional functions*, Symposium on Applied Mathematical Programming and Modeling, APMOD93 (Budapest, Hungary), January 1993, pp. 374–381.
37. M. N. Lukic and J. H. Beder, *Stochastic processes with sample paths in reproducing kernel Hilbert spaces*, Trans. Amer. Math. Soc. **353** (2001), no. 10, 3945–3969.
38. A. Marrel, B. Iooss, B. Laurent, and O. Roustant, *Calculations of Sobol indices for the Gaussian process metamodel*, Reliab. Eng. & System Safety **94** (2009), no. 3, 742–751.
39. G. Matheron, *La théorie des variables régionalisées et ses applications*, Les cahiers du Centre de Morphologie Mathématique, no. 5, Ecole Nationale Supérieure des Mines De Paris, 1971, 212 p.
40. B. Matérn, *Spatial variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations*, Tech. Report Band 49, Nr 5, 144 pp, Meddelanden Fran Statens Skogsforskningsinstitut, Stockholm, Sweden, 1960.
41. J. Mockus, *Bayesian approach to global optimization: Theory and applications*, Kluwer Acad. Publ., Dordrecht-Boston-London, 1989.

42. J. Mockus, V. Tiesis, and A. Zilinskas, *The application of Bayesian methods for seeking the extremum*, Towards Global Optimization (North Holland, New York) (L.C.W. Dixon and G.P. Szego, eds.), vol. 2, 1978, pp. 117–129.
43. T. Muehlenstaedt, O. Roustant, L. Carraro, and S. Kuhnt, *Data-driven kriging models based on FANOVA-decomposition*, Statist. Comput. **22** (2012), no. 3, 723–738.
44. B. Nagy, J. L. Loeppky, and W. J. Welch, *Fast Bayesian inference for Gaussian process models*, Tech. Report 230, The University of British Columbia, Department of Statistics, 2007.
45. R. M. Neal, *Monte carlo implementation of Gaussian process models for Bayesian regression and classification*, Tech. Report arXiv:physics/9701026, arXiv.org, 1997.
46. J. E. Oakley, *Decision-theoretic sensitivity analysis for complex computer models*, Technometrics **51** (2009), no. 2, 121–129.
47. A. O'Hagan, *Curve fitting and optimal design for prediction*, Journal of the Royal Statistical Society. Series B (Methodological) **40** (1978), no. 1, 1–42.
48. _____, *Some Bayesian numerical analysis*, Bayesian statistics 4: proceedings of the Fourth Valencia International Meeting, April 15-20, 1991, Oxford University Press, 1992.
49. H. Omre, *Bayesian kriging—merging observations and qualified guesses in kriging*, Mathematical Geology **19** (1987), no. 1, 25–39.
50. M. A. Osborne, *Bayesian Gaussian processes for sequential prediction optimisation and quadrature*, Ph.D. thesis, University of Oxford, 2010.
51. M. A. Osborne, R. Garnett, and S. J. Roberts, *Gaussian processes for global optimization*, 3rd International Conference on Learning and Intelligent Optimization (LION3), online proceedings (Trento, Italy), 2009.
52. M. A. Osborne, S.J. Roberts, A. Rogers, S.D. Ramchurn, and N.R. Jennings, *Towards real-time information processing of sensor network data using computationally efficient multi-output Gaussian processes*, Proceedings of the 7th International Conference on Information Processing in Sensor Networks, IEE Computer Society, 2008, pp. 109–120.
53. C. J. Paciorek and M. J. Schervish, *Spatial modelling using a new class of nonstationary covariance functions*, Environmetrics **17** (2006), no. 5, 483–506.
54. E. Parzen, *An approach to time series analysis*, Ann. Math. Stat. **32** (1962), 951–989.
55. H. D. Patterson and R. Thompson, *Recovery of inter-block information when block sizes are unequal*, Biometrika **58** (1971), no. 3, 545–554.
56. R. Paulo, *Default priors for Gaussian processes*, Ann. Statist. **33** (2005), no. 2, 556–582.
57. V. Picheny and D. Ginsbourger, *A nonstationary space-time Gaussian process model for partially converged simulations*, SIAM/ASA J. Uncertain. Quantification **1** (2013), no. 1, 57–78.
58. J. Pilz and G. Spöck, *Why do we need and how should we implement Bayesian kriging methods*, Stochastic Environmental Research and Risk Assessment **22** (2008), no. 5, 621–632.
59. L. Pronzato and J. Rendas, *Bayesian local kriging*, Tech. Report HAL Id: hal-01093466, HAL, 2014.
60. W. Rudin, *Real and complex analysis*, 3rd ed., McGraw-Hill, New York, 1987.
61. J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and analysis of computer experiments*, Statistical Science **4** (1989), no. 4, 409–435.
62. T. J. Santner, B. J. Williams, and W. Notz, *The Design and Analysis of Computer Experiments*, Springer Verlag, 2003.
63. M. L. Stein, *Interpolation of spatial data: Some theory for Kriging*, Springer, New York, 1999.
64. A. Törn and A. Zilinskas, *Global optimization*, Springer, Berlin, 1989.
65. E. Vazquez, *Modélisation comportementale des systèmes non linéaires multivariables par méthodes à noyaux et applications*, Ph.D. thesis, Université Paris-Sud XI, Orsay, France, May 2005.
66. E. Vazquez and J. Bect, *Sequential Bayesian algorithm to estimate a probability of failure*, 15th IFAC Symposium on System Identification, SYSID09 (Saint-Malo, France), July 6-8 2009, pp. 546–550.
67. E. Vazquez and J. Bect, *Pointwise consistency of the kriging predictor with known mean and covariance functions*, mODa 9 — Advances in Model-Oriented Design and Analysis, (Proc.

of the 9th Int. Workshop in Model-Oriented Design and Analysis (Bertinoro, Italy), Physica-Verlag, Contributions to Statistics, Springer, June 14-18 2010, pp. 221–228.
68. E. Vazquez and M. Piera-Martinez, *Estimation of the volume of an excursion set of a Gaussian process using intrinsic kriging*, Tech. Report arXiv:math/0611273, arXiv.org, 2006.
69. J. Villemonteix, E. Vazquez, and E. Walter, *An informational approach to the global optimization of expensive-to-evaluate functions*, J. Global Optim. **44** (2009), no. 4, 509–534.
70. G. Wahba, *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59, SIAM, Philadelphia, 1990.
71. W. J. Welch, R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, and M. D. Morris, *Screening, predicting and computer experiments*, Technometrics **34** (1992), 15–25.
72. H. Wendland, *Scattered data approximation*, Monographs on Applied and Computational Mathematics, Cambridge Univ. Press, Cambridge, 2005.
73. B. J. Williams, T. J. Santner, and W. I. Notz, *Sequential design of computer experiments to minimize integrated response functions*, Statistica Sinica **10** (2000), no. 4, 1133–1152.
74. C. K. I. Williams, *Regression with Gaussian processes*, Mathematics of Neural Networks: Models, Algorithms and Applications (S.W. Ellacott, J. C. Mason, and I. J. Anderson, eds.), Kluwer, 1997, Presented at the Mathematics of Neural Networks and Applications Conference, Oxford, 1995.
75. D. Williams, *Probability with martingales*, Cambridge University Press, Cambridge, 1991.
76. S. J. Yakowitz and F. Szidarovszky, *A comparison of kriging with nonparametric regression methods*, J. Multivariate Analysis **16** (1985), 21–53.
77. D. Yarotsky, *Examples of inconsistency in optimization by expected improvement*, Journal of Global Optimization **56** (2013), no. 4, 1773–1790.
78. H. Zhang, *Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics*, J. Amer. Statist. Assoc. **99** (2004), no. 465, 250–261.
79. A. Zilinskas, *A review of statistical models for global optimization*, J. Global Optim. **2** (1992), 145–153.

Chapter 2
Sequential search strategies

2.1 Overview

This chapter presents a selection of articles dealing with SUR/one-step look-ahead strategies that I consider important and for which I had a significant contribution.

Contents

2.2 Sequential search based on kriging: convergence analysis of some algorithms

This contribution was presented to the ISI & 58th World Statistics Congress of the International Statistical Institute (ISI11), in Dublin Ireland, in 2011.

Sequential search based on kriging: convergence analysis of some algorithms

Vazquez, Emmanuel

Bect, Julien

SUPELEC, Gif-sur-Yvette, France
e-mail: emmanuel.vazquez@supelec.fr, julien.bect@supelec.fr

1 Introduction

Let $\mathcal{F}$ be a set of real-valued functions on a set $\mathbb{X}$ and let $S : \mathcal{F} \to \mathcal{G}$ be an arbitrary mapping. We consider the problem of making inference about $S(f)$, with $f \in \mathcal{F}$ unknown, from a finite set of pointwise evaluations of f. We are mainly interested in the problems of approximation and optimization. Formally, a deterministic algorithm to infer a quantity of interest $S(f)$ from a set of n evaluations of f is a pair $\left(\underline{X}_n, \widehat{S}_n\right)$ consisting of a deterministic *search strategy*

$$\underline{X}_n : f \mapsto \underline{X}_n(f) = (X_1(f), X_2(f), \ldots, X_n(f)) \in \mathbb{X}^n \,,$$

and a mapping $\widehat{S}_n : \mathcal{F} \to \mathcal{G}$, such that:

a) $X_1(f) = x_1$, for some arbitrary $x_1 \in \mathbb{X}$

b) For all $1 \leq i < n$, $X_{i+1}(f)$ depends measurably on $\mathcal{I}_i(f)$, where $\mathcal{I}_i = ((X_1, Z_1), \ldots, (X_i, Z_i))$, and $Z_i(f) = f(X_i(f))$, $1 \leq i \leq n$.

c) There exists a measurable function ϕ_n such that $\widehat{S}_n = \phi_n \circ \mathcal{I}_n$.

The algorithm $\left(\underline{X}_n, \widehat{S}_n\right)$ describes a sequence of decisions, made from an increasing amount of information: for each $i = 1, \ldots, n-1$, the algorithm uses information $\mathcal{I}_i(f)$ to choose the next evaluation point $X_{i+1}(f)$. The estimator $\widehat{S}_n(f)$ of $S(f)$ is the terminal decision. We shall denote by $\mathcal{A}_n$ the class of all strategies $\underline{X}_n$ that query sequentially n evaluations of f and also define the subclass $\mathcal{A}_n^0 \subset \mathcal{A}_n$ of non-adaptive strategies, that is, the class of all strategies such that the X_is do not depend on f.

A classical approach to study the performance of a sequential strategy is to consider the worst error of estimation on some class of functions $\mathcal{F}$

$$\epsilon_{\text{worstcase}}(\underline{X}_n) := \sup_{f \in \mathcal{F}} L(S(f), \widehat{S}_n(f)) \,,$$

where L is a loss function. There are many results dealing with the problems of function approximation and optimization in the worst case setting. Two noticeable results concern convex and symmetric classes of bounded functions. For such classes, from a worst-case point of view, any strategy will behave similarly for the problem of global optimization and that of function approximation. Moreover the use of adaptive methods can not be justified by a worst case analysis (see, e.g., Novak, 1988, Propositions 1.3.2 and 1.3.3). These results, combined with the fact that most optimization algorithms are adaptive, lead to think that the worst-case setting may not be the most appropriate framework to assess the performance of a search algorithm in practice. Indeed, it would be also important, in practice, to know whether the loss $L(S(f), \widehat{S}_n(f))$ is close to, or on the contrary much smaller than $\epsilon_{\text{worstcase}}$, for "typical" functions $f \in \mathcal{F}$ not corresponding to worst cases. To address this question, a classical approach is to adopt a Bayesian point of view.

In this paper, we consider methods where f is seen as a sample path of a real-valued random process ξ defined on some probability space $(\Omega, \mathcal{B}, \mathsf{P}_0)$ with parameter in $\mathbb{X}$. Then, $\underline{X}_n(\xi)$ is a random

sequence in $\mathbb{X}$, with the property that $X_{n+1}(\xi)$ is measurable with respect to the σ-algebra generated by $\xi(X_1(\xi))$, ..., $\xi(X_n(\xi))$. From a Bayesian decision-theoretic point of view, the random process represents prior knowledge about f and makes it possible to infer a quantity of interest before evaluating the function. This point of view has been widely explored in the domain of optimization and computer experiments. Under this setting, the performance of a given strategy $\underline{X}_n$ can be assessed by studying the average loss

$$\epsilon_{\text{average}}(\underline{X}_n) := \mathsf{E}\, L(S(\xi), \widehat{S}_n(\xi))\,.$$

How much does adaption help on the average, and is it possible to derive rates of decay for errors in average? In this article, we shall make a brief review of results concerning average error bounds of Bayesian search methods based on a random process prior.

This article has three parts. The precise assumptions about ξ are given in Section 2. Section 3 deals with the problem of function approximation, while Section 4 deals with the problem of optimization.

2 Framework

Let ξ be a random process defined on a probability space $(\Omega, \mathcal{B}, \mathsf{P}_0)$, with parameter $x \in \mathbb{R}^d$. Assume moreover that ξ has a zero mean and a continuous covariance function. The kriging predictor of $\xi(x)$, based on the observations $\xi(X_i(\xi))$, $i = 1, \ldots, n$, is the orthogonal projection

$$\widehat{\xi}_n(x) \;:=\; \sum_{i=1}^{n} \lambda^i(x; \underline{X}_n(\xi))\, \xi(X_i(\xi)) \tag{1}$$

of $\xi(x)$ onto $\operatorname{span}\{\xi(X_i(\xi)), i = 1, \ldots, n\}$ in $L^2(\Omega, \mathcal{B}, \mathsf{P}_0)$. At step $n \geq 1$, given evaluation points $\underline{X}_n(\xi)$, the kriging coefficients $\lambda^i(x; \underline{X}_n(\xi))$ can be obtained by solving a system of linear equations (see, e.g., Chilès and Delfiner, 1999). Note that for any sample path $f = \xi(\omega, \cdot)$, $\omega \in \Omega$, the value $\widehat{\xi}_n(\omega, x)$ is a function of $\mathcal{I}_n(f)$ only.

The mean-square error (MSE) of estimation at a fixed point $x \in \mathbb{R}^d$ will be denoted by

$$\sigma_n^2(x) := \mathsf{E}\{(\xi(x) - \widehat{\xi}(x; \underline{X}_n(\xi)))^2\}\,.$$

It is generally not possible to compute $\sigma_n^2(x)$ when $\underline{X}_n$ is an adaptive strategy.

Regularity assumptions. Assume that there exists $\Phi : \mathbb{R}^d \to \mathbb{R}$ such that $k(x, y) = \Phi(x - y)$, which is in $L^2(\mathbb{R}^d)$ and has a Fourier transform

$$\tilde{\Phi}(u) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \Phi(x) e^{i(x,u)} dx$$

that satisfies

$$c_1(1 + \|u\|_2^2)^{-s} \leq \tilde{\Phi}(u) \leq c_2(1 + \|u\|_2^2)^{-s}\,, \quad u \in \mathbb{R}^d\,, \tag{2}$$

with $s > d/2$ and constants $0 < c_1 \leq c_2$. Note that the Matérn covariance with regularity parameter ν (see, e.g., Stein, 1999) satisfies such a regularity assumption, with $s = \nu + d/2$. Tensor-product covariance functions, however, never satisfy such a condition (see Ritter, 2000, chapter 7, for some results in this case).

Let $\mathcal{H}$ be the RKHS of functions generated by k. Denote by $(\cdot\,, \cdot\,)_{\mathcal{H}}$ the inner product of $\mathcal{H}$, and by $\|\cdot\|_{\mathcal{H}}$ the corresponding norm. It is well known (see, e.g. Wendland, 2005) that $\mathcal{H}$ is the Sobolev space

$$W_2^s(\mathbb{R}^d) = \left\{ f \in L^2(\mathbb{R}^d);\ \tilde{f}(\cdot)(1 + \|\cdot\|_2^2)^{s/2} \in L^2(\mathbb{R}^d) \right\}$$

due to the following result.

Proposition 1. *$\mathcal{H} \subset L^2(\mathbb{R}^d)$ and $\forall f \in \mathcal{H}$,*

$$\|f\|_{\mathcal{H}}^2 = \int_{\mathbb{R}^d} |\tilde{f}(u)|^2 \, \tilde{\Phi}(u)^{-1} \, du \, .$$

$\|f\|_{\mathcal{H}}^2$ is equivalent to the Sobolev norm

$$\|f\|_{W_2^s(\mathbb{R}^d)}^2 = \|\tilde{f}(\,\cdot\,) \left(1 + \|\cdot\|_2^2\right)^{s/2}\|_{L^2(\mathbb{R}^d)}$$

3 Approximation

We first consider the problem of approximation, with the point of view exposed in Section 2. Using the notations introduced above, the problem of approximation corresponds to considering operators S and $\widehat{S}_n$ defined by $S(\xi) := \xi_{|\mathbb{X}}$ and $\widehat{S}_n(\xi) := \widehat{\xi}_{n\,|\mathbb{X}}$, with $\mathbb{X} \subset \mathbb{R}^d$ a compact domain with non-empty interior. For the design of computer experiments, classical criteria for assessing the quality of a strategy $\underline{X}_n \in \mathcal{A}_n$ for the approximation problem are the maximum mean-square error (MMSE)

$$\epsilon_{\text{MMSE}}(\underline{X}_n) \;:=\; \sup_{x \in \mathbb{X}} \mathsf{E}\left(\left(\xi(x) - \widehat{\xi}_n(x)\right)^2\right) \;=\; \sup_{x \in \mathbb{X}} \sigma_n^2(x)$$

and the integrated mean-square error (IMSE)

$$\epsilon_{\text{IMSE}}(\underline{X}_n) \;:=\; \mathsf{E}\left(\|\xi - \widehat{\xi}_n\|_{L^2(\mathbb{X},\mu)}^2\right) \;=\; \int_{\mathbb{X}} \sigma_n(x)^2 \, \mu(\mathrm{d}x)$$

(see, e.g., Sacks et al., 1989; Currin et al., 1991; Welch et al., 1992; Santner et al., 2003). These criteria correspond to G-optimality and I-optimality in the theory of (parametric) optimal design.

As mentioned earlier, computing $\sigma_n^2(x)$ is usually not possible in the case of adaptive sampling strategies, even for a Gaussian process. From a theoretical point of view, however, it is important to know if adaptive strategies can improve upon non-adaptive strategies for the approximation problem.

Proposition 2. *Assume that ξ is a Gaussian process. Then adaptivity does not help for the approximation problem, with respect to either the MMSE or the IMSE criterion.*

Proof. For any adaptive strategy $\underline{X}_n$, it can be proved by induction (using the fact that X_{i+1} only depends on $\mathcal{I}_i$) that, for each $x \in \mathbb{X}$,

$$\sigma_n^2(x) \;=\; \mathsf{E}\left(\sigma^2(x; X_1(\xi), \ldots, X_n(\xi))\right) \,, \tag{3}$$

where $\sigma^2(x; x_1, \ldots, x_n)$, $x_1, \ldots, x_n \in \mathbb{X}$, denotes the MSE at x of the non-adaptive strategy that selects the points $x_1, \ldots, x_n$. Therefore, for each $x \in \mathbb{X}$,

$$\sigma_n^2(x) \;\geq\; \min_{x_1, \ldots, x_n \in \mathbb{X}} \sigma^2(x; x_1, \ldots, x_n) \,,$$

which proves the claim in the case of the MMSE criterion. Similarly, integrating (3) yields

$$\begin{aligned} \int_{\mathbb{X}} \sigma_n^2 \, d\mu \;&=\; \mathsf{E}\left\{\int_{\mathbb{X}} \sigma^2(x; \underline{X}_n(\xi)) \, \mu(\mathrm{d}x)\right\} \\ &\geq\; \min_{x_1, \ldots, x_n \in \mathbb{X}} \int_{\mathbb{X}} \sigma^2(x; x_1, \ldots, x_n) \, \mu(\mathrm{d}x) \,, \end{aligned}$$

which proves the claim in the case of the IMSE criterion. □

In the case of the IMSE criterion, Proposition 2 can be seen as a special case of a general result about linear problems (see, e.g., Ritter, 2000, Chapter 7). The following proposition establishes a connection between the MMSE criterion and the worst-case L^∞-error of approximation in the unit ball of $\mathcal{H}$, which will be useful to establish the optimal rate for IMSE- and MMSE-optimal designs.

Proposition 3. *Let $\mathcal{H}_1$ denote the unit ball of $\mathcal{H}$. For any non-adaptive strategy $\underline{X}_n \in \mathcal{A}_n^0$, the MMSE criterion equals the squared worst-case L^∞-error of approximation in $\mathcal{H}_1$ using $\widehat{S}_n$:*

$$\epsilon_{\text{MMSE}}(\underline{X}_n) = \left(\sup_{f \in \mathcal{H}_1} \|S(f) - \widehat{S}_n(f)\|_{L^\infty(\mathbb{X})} \right)^2 .$$

Proof. Let $\underline{X}_n \in \mathcal{A}_n^0$ be a non-adaptive strategy such that $X_i(\xi) = x_i$, $i = 1, \ldots, n$, for some arbitrary x_is in $\mathbb{X}$. Denote by $\lambda_i(x) = \lambda_i(x; \underline{X}_n(\xi))$ the corresponding kriging coefficients (which do not depend on ξ). Using the fact that the mapping $\xi(x) \mapsto k(x, \cdot)$ extends linearly to an isometry from $\overline{\text{span}}\{\xi(y),\, y \in \mathbb{R}^d\}$ to $\mathcal{H}$, we have for all $x \in \mathbb{X}$

$$\begin{aligned}
\sigma_n(x) &= \left\| \xi(x) - \widehat{\xi}_n(x) \right\|_{L^2(\Omega, \mathcal{B}, \mathrm{P}_0)} \\
&= \left\| k(x, \cdot) - \textstyle\sum_i \lambda^i(x)\, k(x_i, \cdot) \right\|_{\mathcal{H}} \\
&= \sup_{f \in \mathcal{H}_1} \left(f\,,\, k(x, \cdot) - \textstyle\sum_i \lambda^i(x)\, k(x_i, \cdot) \right)_{\mathcal{H}} . \\
&= \sup_{f \in \mathcal{H}_1} (f - \widehat{S}_n f)(x) .
\end{aligned}$$

Thus,

$$\sup_{x \in \mathbb{X}} \sigma_n(x) \;=\; \sup_{f \in \mathcal{H}_1} \sup_{x \in \mathbb{X}} (f - \widehat{S}_n f)(x) \;=\; \sup_{f \in \mathcal{H}_1} \left\| f - \widehat{S}_n f \right\|_{L^\infty(\mathbb{X})} .$$

□

The following proposition summarizes known results concerning the optimal rate of decay in the class of non-adaptive strategies for both the IMSE criterion and the MMSE criterion. Note that, by Proposition 2, this rate is also the optimal rate of decay in the class of all adaptive strategies if ξ is a Gaussian process.

Proposition 4. *Assume that ξ has a continuous covariance function satisfying the regularity assumptions of Section 2, and let $\nu = s - d/2 > 0$. Then there exists $C_1 > 0$ such that, for any $\underline{X}_n \in \mathcal{A}_n^0$,*

$$C_1\, n^{-2\nu/d} \;\le\; \epsilon_{\text{IMSE}}(\underline{X}_n) \;\le\; \mu(\mathbb{X})\, \epsilon_{\text{MMSE}}(\underline{X}_n) \tag{4}$$

Moreover, if $\mathbb{X}$ has a Lipschitz boundary and satisfies an interior cone condition, then there exists $C_2 > 0$ such that

$$\inf_{\underline{X}_n \in \mathcal{A}_n^0} \epsilon_{\text{IMSE}}(\underline{X}_n) \;\le\; \mu(\mathbb{X}) \inf_{\underline{X}_n \in \mathcal{A}_n^0} \epsilon_{\text{MMSE}}(\underline{X}_n) \;\le\; C_2\, n^{-2\nu/d} . \tag{5}$$

The optimal rate of decay is therefore $n^{-2\nu/d}$ for both criteria.

Proof. It is proved in (Ritter, 2000, Chapter 7, Proposition 8) that there exists $C_1 > 0$ such that $\epsilon_{\text{IMSE}}(\underline{X}_n) \;\ge\; C_1\, n^{-2\nu/d}$ in the case where $\mathbb{X} = [0; 1]^d$. This readily proves the lower bound (4) since any $\mathbb{X}$ with non-empty interior contains an hypercube on which Ritter's result holds.

If $\mathbb{X}$ is a bounded Lipschitz domain satisfying an interior cone condition, then (Narcowich et al., 2005, Proposition 3.2) there exists $c_1 > 0$ such that $\|S(f) - \widehat{S}_n(f)\|_{L^\infty(\mathbb{X})} \le c_1 h_n^{s-d/2} \|S(f)\|_{W_2^s(\mathbb{X})}$ for

all $f \in \mathcal{H}$, where $h_n = \sup_{x\in\mathbb{X}} \min_{i\in\{1,\ldots,n\}} \|x - X_i(f)\|_2$ is the fill distance of the non-adaptive strategy $\underline{X}_n$ in $\mathbb{X}$. Therefore

$$\|S(f) - \widehat{S}_n(f)\|_{L^\infty(\mathbb{X})} \;\le\; c_1 h_n^\nu \, \|S(f)\|_{W_2^s(\mathbb{X})} \;\le\; c_1 h_n^\nu \, \|f\|_{W_2^s(\mathbb{R}^d)} \;\le\; c_2 h_n^\nu \, \|f\|_{\mathcal{H}}$$

for some $c_2 > 0$, using the equivalence of the Sobolev $W_2^s(\mathbb{R}^d)$ norm with the RKHS norm (see Section 2). Considering any non-adaptive space-filling strategy $\underline{X}_n$ with a fill distance $h_n = O(n^{-1/d})$ yields

$$\inf_{\underline{X}_n \in \mathcal{A}_n^0} \sup_{f\in\mathcal{H}_1} \left\| f - \widehat{S}_n f \right\|_{L^\infty(\mathbb{X})} \;\le\; c_3 \, n^{-\nu/d}$$

for some $c_3 > 0$ and the upper-bound (5) then follows from Proposition 3. □

Finding a non-adaptive MMSE-optimal design is a difficult non-convex optimization problem in nd dimensions. Instead of addressing directly such a high-dimensional global optimization problem, we can use the classical sequential non-adaptive greedy strategy $\underline{X}_n(\,\cdot\,) = (x_1, \ldots, x_n) \in \mathbb{X}^n$ defined by

$$x_{i+1} \;=\; \operatorname*{argmax}_{x\in\mathbb{X}} \, \sigma^2\,(x; x_1, \ldots, x_i)\;, \quad 1 \le i < n\,. \tag{6}$$

Of course, the strategy is suboptimal but it only involves simpler optimization problems in d dimensions and has the advantage that it can be stopped at any time. Following Binev et al. (2010), it can be established that this greedy strategy is rate optimal.

Proposition 5. *Assume that ξ has a continuous covariance function satisfying the regularity assumptions of Section 2, and let $\nu = s - d/2 > 0$. Let $\underline{X}_n$ be the sequential strategy defined by (6). Then,*

$$\epsilon_{\text{MMSE}}(\underline{X}_n) = \mathrm{O}(n^{2\nu/d})\,.$$

Proof. Theorem 3.1 in Binev et al. (2010), applied to the compact subset $\{\xi(x), x \in \mathbb{X}\}$ in $L^2(\Omega, \mathcal{B}, \mathsf{P}_0)$, states that the greedy algorithm (6) preserves polynomial rates of decay. The result follows from Proposition 4. □

4 Optimization

In this section, we consider the problem of global optimization on a compact domain $\mathbb{X} \subset \mathbb{R}^d$, which corresponds formally to operators S and $\widehat{S}_n$ defined by $S(\xi) = \sup_{x\in\mathbb{X}} \xi(x)$ and $\widehat{S}_n(\xi) = \max_{i\in 1,\ldots,n} \xi(X_i(\xi))$.

In a Bayesian setting, a classical criterion to assess the performance of an optimization procedure is the average error

$$\epsilon_{\overline{\text{OPT}}}(\underline{X}_n) := \mathsf{E}(S(\xi) - \widehat{S}_n(\xi))\,.$$

Although it may be not possible in the context of this article to make a comprehensive review of known results concerning the average case in the Gaussian case, it can be safely said however that such results are scarce and specific.

In fact, most available results about the average-case error concern the one-dimensional Wiener process ξ on the interval $[0, 1]$. Under this setting, Ritter (1990) shows that the average error of the best non-adaptive optimization procedure decreases at rate $n^{-1/2}$ (extensions of this result for

non-adaptive algorithms and the r-fold Wiener measure can be found in Wasilkowski, 1992). Under the same assumptions for ξ, Calvin (1997) derives the exact limiting distribution of the error of a particular adaptive algorithm, which suggests that adaptivity does yield a better average error for the optimization problem—the result is that, for any $0 < \delta < 1$, it is possible to find an adaptive strategy such that $n^{(1-\delta)}(S(\xi) - \widehat{S}_n(\xi))$ converges in distribution.

A theoretical result concerning the optimal average-error criterion for less restrictive Gaussian priors is also available. If the covariance of a Gaussian process ξ is α-Hölder continuous, then Grünewälder et al. (2010) show that a space filling strategy $\underline{X}_n$ achieves

$$\epsilon_{\overline{\mathrm{OPT}}}(\underline{X}_n) = \mathrm{O}(n^{-\alpha/(2d)}(\log n)^{1/2})\,. \tag{7}$$

Thus, under the assumptions of Section 2, for a Matérn covariance with regularity parameter ν, the rate of the optimal average error of estimation of the optimum is less than $n^{-\nu/d}(\log n)^{1/2}$ (since a Matérn covariance is α-Hölder continuous with $\alpha = 2\nu$). Note that this bound is not sharp in general since the optimal non-adaptive rate is $n^{-1/2}$ for the Brownian motion on $[0; 1]$, the covariance function of which is α-Hölder continuous with $\alpha = 1$.

In view of these results, we can safely say that characterizing the average behavior of adaptive sequential optimization algorithms is still an open (and apparently difficult) problem. At present, the only way to draw useful conclusions about the interest of a particular optimization algorithm is to resort to numerical simulations.

In the following paragraphs, we shall illustrate the kind of results that can be expected from such empirical studies. Benassi et al. (2011) provide an empirical comparison between four optimization algorithms. The first algorithm is a non-adaptive space-filling strategy. The second algorithm assumes a Gaussian prior about the objective function and use the expected improvement (EI) sampling criterion (Mockus et al., 1978) for choosing the evaluation points. In practice however, it is often difficult to choose a Gaussian prior before any evaluation is made. As a result, the covariance function of ξ is usually chosen in some parametric class of positive definite functions, the value of the parameters assumed to be unknown. The third algorithm compared in Benassi et al. (2011) is a fully Bayesian algorithm (FBA), which is used to deal with uncertain parameters of the covariance of ξ. The fourth strategy is the popular efficient global optimization (EGO) algorithm introduced by Jones et al. (1998), which assumes a Gaussian process prior and takes a plug-in approach to deal with the uncertain parameters of the covariance.

In order to compare the four optimization strategies, Benassi et al. (2011) build several testbeds $\mathcal{T}_k$, $k = 1, 2, \ldots$, of functions $f_{k,l}$, $l = 1, \ldots, L$, corresponding to sample paths of a Gaussian process, with zero-mean and a Matérn covariance function, simulated on a set of $q = 600$ points in $[0, 1]^d$ generated using a Latin hypercube sampling (LHS), with different values for d and for the parameters of the covariance. Here, we present the results obtained for two testbeds in dimension 1 and 4 (the actual parameters are provided in Table 1).

Figures 1 and 2 show the average errors and also the distributions of the error of estimation of the global optimum. These empirical results show that the EI strategy performs much better in average than the space-filling strategy. Large errors are also less frequent with the EI strategy. Moreover, we can also assess the cost of estimating the parameters of the covariance. EGO and FBA have very similar average performances. In fact, both of them perform almost as well, in this experiments, as the EI strategy, where the true parameters are assumed to be known. Comparing the tails of complementary cumulative distribution function of the error $Sf - \widehat{S}_n f$ shows, however, that using a fully Bayesian approach brings a reduction of the occurrence of large errors with respect to the EGO algorithm. In other words, the fully Bayesian approach appears to be statistically more robust than the plug-in approach, while retaining the same average performance. Empirical studies such as the

Parameter \ Testbed	$\mathcal{T}_1$	$\mathcal{T}_2$
Dimension d	1	4
Number of sample paths L	20000	20000
Variance σ^2	1.0	1.0
Regularity ν	2.5	2.5
Scale $\beta = (\beta_1, \ldots, \beta_d)$	0.1	(0.7, 0.7, 0.7, 0.7)

Table 1: Parameters used for building the testbeds of Gaussian-process sample-paths. The Gaussian process has a zero-mean and a isotropic Matérn covariance function $k_{[\nu,\sigma^2,\rho]} : (x, y) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto \sigma^2 \kappa_\nu(\|x - y\|/\rho)$ with $\kappa_\nu(h) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(2\nu^{1/2}h\right)^\nu \mathcal{K}_\nu \left(2\nu^{1/2}h\right)$, $h \in \mathbb{R}$, where Γ is the Gamma function, $\mathcal{K}_\nu$ is the modified Bessel function of the second kind, and ν, σ^2 ρ are strictly positive scalar parameters (see Stein, 1999).

one presented here are therefore very useful from a practical point of view, since they make it possible to obtain fine and sound performance assessments of any strategy with a reasonable computational cost.

References

R. Benassi, J. Bect, and E. Vazquez. Robust gaussian process-based global optimization using a fully bayesian expected improvement criterion. In *Proceedings of fifth Learning and Intelligent Optimization Conference (LION 5)*, Rome, 2011.

P. Binev, A. Cohen, W. Dahmen, R. DeVore, G; Petrova, and P. Wojtaszczyk. *Convergence Rates for Greedy Algorithms in Reduced Basis Methods*, volume IGPM Report 310. RWTH Aachen, 2010.

J.M. Calvin. Average performance of a class of adaptive algorithms for global optimization. *The Annals of Applied Probability*, 7(3):711–730, 1997.

J.-P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York, 1999.

C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Amer. Statist. Assoc.*, pages 953–963, 1991.

S. Grünewälder, J.Y. Audibert, M. Opper, and J. Shawe-Taylor. Regret bounds for gaussian process bandit problems. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *J. Global Optim.*, 13:455–492, 1998.

J. Mockus, V. Tiesis, and A. Zilinskas. The application of Bayesian methods for seeking the extremum. In L.C.W. Dixon and G.P. Szego, editors, *Towards Global Optimization*, volume 2, pages 117–129, North Holland, New York, 1978.

F. J. Narcowich, J. D. Ward, and H. Wendland. Sololev bounds on functions with scattered zeros, with applications to radial basis function surface fitting. *Math. Comp.*, 74:743–763, 2005.

E. Novak. *Deterministic and stochastic error bounds in numerical analysis*, volume 1349 of *Lecture Notes in Mathematics*. Springer-Verlag, 1988.

K. Ritter. Approximation and optimization on the wiener space. *Journal of Complexity*, 6(4):337–364, 1990.

K. Ritter. *Average-case analysis of numerical problems*, volume 1733 of *Lecture Notes in Mathematics*. Springer Verlag, 2000.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statist. Sci.*, 4(4):409–435, 1989.

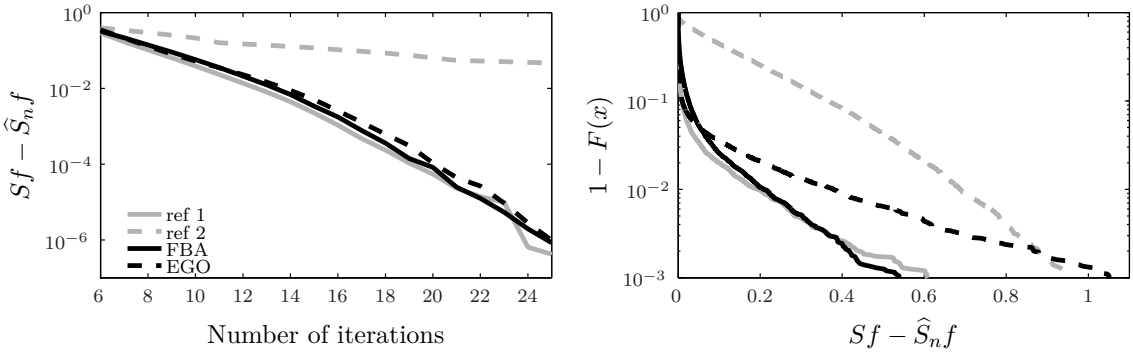
T. J. Santner, B. J. Williams, and W. I. Notz. *The Design and Analysis of Computer Experiments*. Springer, 2003.

M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, 1999.

G.W. Wasilkowski. On average complexity of global optimization problems. *Mathematical programming*, 57(1):313–324, 1992.

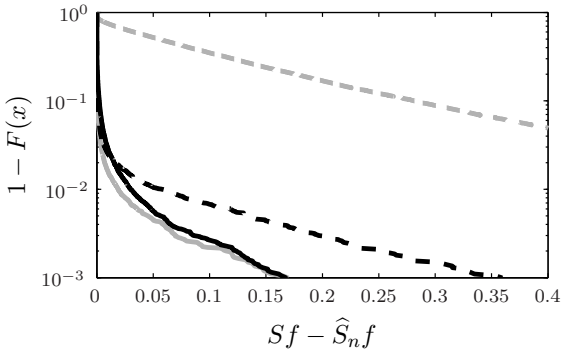
W. J. Welch, R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, and M. D. Morris. Screening, predicting, and computer experiments. *Technometrics*, 34(1):15–25, 1992.

H. Wendland. *Scattered Data Approximation*. Monographs on Applied and Computational Mathematics. Cambridge Univ. Press, Cambridge, 2005.



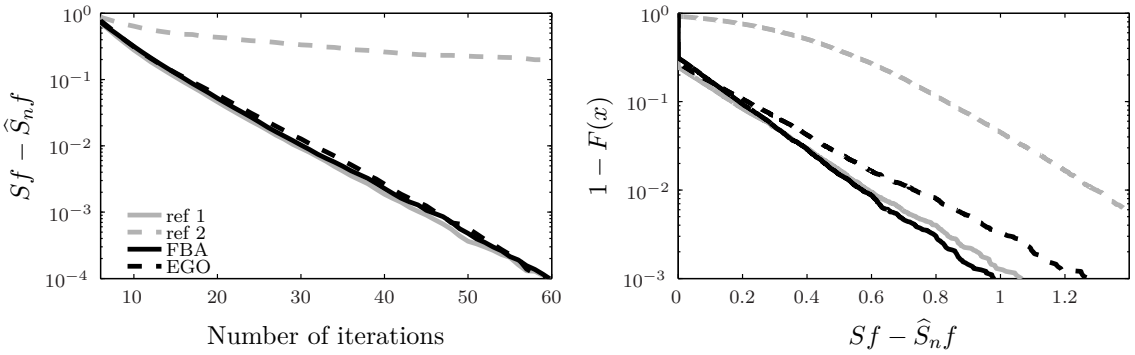
(a) Average error to the maximum

(b) Distribution of errors at iteration 13



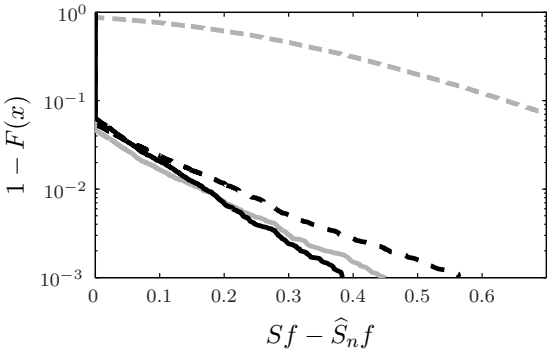
(c) Distribution of errors at iteration 16

Figure 1: Average results and error distributions for testbed $\mathcal{T}_1$, for FBA (solid black line), EGO (dashed black line), the EI with the parameters used to generate sample paths (solid gray line), the space-filling strategy (dashed gray line). More precisely, (a) represents the average approximation error as a function of the number of evaluation points. In (b) and (c), $F(x)$ stands for the cumulative distribution function of the approximation error. We plot $1-F(x)$ in logarithmic scale in order to analyze the behavior of the tail of the distribution (big errors with small probabilities of occurrence). Small values for $1 - F(x)$ mean better results.



(a) Average error to the maximum

(b) Distribution of errors at iteration 20



(c) Distribution of errors at iteration 34

Figure 2: Average results and distribution of errors for testbed $\mathcal{T}_2$. See Figure 1 for details.

2.3 Sequential design of computer experiments for the estimation of a probability of failure

This contribution was published in *Statistics and Computing* (Springer) in 2012. It is an extension of a work initiated with Miguel Piera-Martinez (PhD student, 2004–2008), and continued with my colleague Julien Bect with whom I have supervised the PhD preparation of Ling Li (2009–2012). For writing the article, we also collaborated with David Ginsbourger and Victor Picheny.

Stat Comput (2012) 22:773–793
DOI 10.1007/s11222-011-9241-4

Sequential design of computer experiments for the estimation of a probability of failure

Julien Bect · David Ginsbourger · Ling Li · Victor Picheny · Emmanuel Vazquez

Received: 8 September 2010 / Accepted: 22 February 2011 / Published online: 21 April 2011

Abstract This paper deals with the problem of estimating the volume of the excursion set of a function $f : \mathbb{R}^d \to \mathbb{R}$ above a given threshold, under a probability measure on $\mathbb{R}^d$ that is assumed to be known. In the industrial world, this corresponds to the problem of estimating a probability of failure of a system. When only an expensive-to-simulate model of the system is available, the budget for simulations is usually severely limited and therefore classical Monte Carlo methods ought to be avoided. One of the main contributions of this article is to derive *SUR* (*stepwise uncertainty reduction*) strategies from a Bayesian formulation of the problem of estimating a probability of failure. These sequential strategies use a Gaussian process model of f and aim at performing evaluations of f as efficiently as possible to infer the value of the probability of failure. We compare these strategies to other strategies also based on a Gaussian process model for estimating a probability of failure.

J. Bect (✉) · L. Li · E. Vazquez
SUPELEC, Gif-sur-Yvette, France
e-mail: julien.bect@supelec.fr

L. Li
e-mail: ling.li@supelec.fr

E. Vazquez (✉)
e-mail: emmanuel.vazquez@supelec.fr

V. Picheny
Ecole Centrale Paris, Chatenay-Malabry, France
e-mail: victor.picheny@ecp.fr

D. Ginsbourger
Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland
e-mail: david.ginsbourger@stat.unibe.ch

Keywords Computer experiments · Sequential design · Gaussian processes · Probability of failure · Stepwise uncertainty reduction

1 Introduction

The design of a system or a technological product has to take into account the fact that some design parameters are subject to unknown variations that may affect the reliability of the system. In particular, it is important to estimate the probability of the system to work under abnormal or dangerous operating conditions due to random dispersions of its characteristic parameters. The *probability of failure* of a system is usually expressed as the probability of the excursion set of a function above a fixed threshold. More precisely, let f be a measurable real function defined over a probability space $(\mathbb{X}, \mathcal{B}(\mathbb{X}), \mathsf{P}_{\mathbb{X}})$, with $\mathbb{X} \subseteq \mathbb{R}^d$, and let $u \in \mathbb{R}$ be a threshold. The problem to be considered in this paper is the estimation of the volume, under $\mathsf{P}_{\mathbb{X}}$, of the excursion set

$$\Gamma := \{x \in \mathbb{X} : f(x) > u\} \tag{1}$$

of the function f above the level u. In the context of robust design, the volume $\alpha := \mathsf{P}_{\mathbb{X}}(\Gamma)$ can be viewed as the probability of failure of a system: the probability $\mathsf{P}_{\mathbb{X}}$ models the uncertainty on the input vector $x \in \mathbb{X}$ of the system—the components of which are sometimes called *design variables* or *factors*—and f is some deterministic performance function derived from the outputs of a deterministic model of the system.[1] The evaluation of the outputs of the model

[1] Stochastic simulators are also of considerable practical interest, but raise specific modeling and computational issues that will not be considered in this paper.

for a given set of input factors may involve complex and time-consuming computer simulations, which turns f into an expensive-to-evaluate function. Therefore, the estimation of α must be carried out with a *restricted number of evaluations* of f, generally excluding the estimation of the probability of excursion by a Monte Carlo approach. Indeed, consider the empirical estimator

$$\alpha_m := \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{f(X_i)>u\}}, \tag{2}$$

where the X_is are independent random variables with distribution $\mathsf{P}_{\mathbb{X}}$. According to the strong law of large numbers, the estimator α_m converges to α almost surely when m increases. Moreover, it is an unbiased estimator of α, i.e. $\mathsf{E}(\alpha_m) = \alpha$. Its mean square error is

$$\mathsf{E}\big((\alpha_m - \alpha)^2\big) = \frac{1}{m}\alpha\big(1-\alpha\big).$$

If the probability of failure α is small, then the standard deviation of α_m is approximately $\sqrt{\alpha/m}$. To achieve a given standard deviation $\delta\alpha$ thus requires approximately $1/(\delta^2\alpha)$ evaluations, which can be prohibitively high if α is small. By way of illustration, if $\alpha = 2\times 10^{-3}$ and $\delta = 0.1$, we obtain $m = 50000$. If one evaluation of f takes, say, one minute, then the entire estimation procedure will take about 35 days to complete. Of course, a host of refined random sampling methods have been proposed to improve over the basic Monte Carlo convergence rate; for instance, methods based on importance sampling with cross-entropy (Rubinstein and Kroese 2004), subset sampling (Au and Beck 2001) or line sampling (Pradlwarter et al. 2007). They will not be considered here for the sake of brevity and because the required number of function evaluations is still very high.

Until recently, all the methods that do not require a large number of evaluations of f were based on the use of parametric approximations for either the function f itself or the boundary $\partial\Gamma$ of Γ. The so-called response surface method falls in the first category (see, e.g., Bucher and Bourgund 1990; Rajashekhar and Ellingwood 1993, and references therein). The most popular approaches in the second category are the first- and second-order reliability method (FORM and SORM), which are based on a linear or quadratic approximation of $\partial\Gamma$ around the *most probable failure point* (see, e.g., Bjerager 1990). In all these methods, the accuracy of the estimator depends on the actual shape of either f or $\partial\Gamma$ and its resemblance to the approximant: they do not provide statistically consistent estimators of the probability of failure.

This paper focuses on sequential sampling strategies based on Gaussian processes and kriging, which can be seen as a *nonparametric* approximation method. Several strategies of this kind have been proposed recently by Ranjan et al. (2008), Bichon et al. (2008), Picheny et al. (2010) and Echard et al. (2010a, 2010b). The idea is that the Gaussian process model, which captures prior knowledge about the unknown function f, makes it possible to assess the uncertainty about the position of Γ given a set of evaluation results. This line of research has its roots in the field of design and analysis of computer experiments (see, e.g., Sacks et al. 1989; Currin et al. 1991; Welch et al. 1992; Oakley and O'Hagan, 2001, 2004; Oakley 2004; Bayarri et al. 2007). More specifically, kriging-based sequential strategies for the estimation of a probability of failure are closely related to the field of Bayesian global optimization (Mockus et al. 1978; Mockus 1989; Jones et al. 1998; Villemonteix 2008; Villemonteix et al. 2009; Ginsbourger 2009).

The contribution of this paper is twofold. First, we introduce a Bayesian decision-theoretic framework from which the theoretical form of an optimal strategy for the estimation of a probability of failure can be derived. One-step lookahead sub-optimal strategies are then proposed,[2] which are suitable for numerical evaluation and implementation on computers. These strategies will be called SUR (stepwise uncertainty reduction) strategies in reference to the work of D. Geman and its collaborators (see, e.g. Fleuret and Geman 1999). Second, we provide a review in a unified framework of all the kriging-based strategies and compare them numerically with the SUR strategies proposed in this paper.

The outline of the paper is as follows. Section 2 introduces the Bayesian framework and recalls the basics of dynamic programming and Gaussian processes. Section 3 introduces SUR strategies, from the decision-theoretic underpinnings, down to the implementation level. Section 4 provides a review of other kriging-based strategies proposed in the literature. Section 5 provides some illustrations and reports an empirical comparison of these sampling criteria. Finally, Section 6 presents conclusions and offers perspectives for future work.

2 Bayesian decision-theoretic framework

2.1 Bayes risk and sequential strategies

Let f be a continuous function. We shall assume that f corresponds to a computer program whose output is not a closed-form expression of the inputs. Our objective is to obtain a numerical approximation of the probability of failure

$$\begin{aligned}\alpha(f) &= \mathsf{P}_{\mathbb{X}}\{x\in\mathbb{X} : f(x) > u\}\\ &= \int_{\mathbb{X}} \mathbb{1}_{f>u}\,\mathrm{d}\mathsf{P}_{\mathbb{X}},\end{aligned} \tag{3}$$

[2]Preliminary accounts of this work have been presented in Vazquez and Piera-Martinez (2007) and Vazquez and Bect (2009).

where $\mathbb{1}_{f>u}$ stands for the indicator function of the excursion set Γ, such that for any $x \in \mathbb{X}$, $\mathbb{1}_{f>u}(x)$ equals one if $x \in \Gamma$ and zero otherwise. The approximation of $\alpha(f)$ has to be built from a set of computer experiments, where an experiment simply consists in choosing an $x \in \mathbb{X}$ and computing the value of f at x. The result of a pointwise evaluation of f carries information about f and quantities depending on f and, in particular, about $\mathbb{1}_{f>u}$ and $\alpha(f)$. In the context of expensive computer experiments, we shall also suppose that the number of evaluations is limited. Thus, the estimation of $\alpha(f)$ must be carried out using a fixed number, say N, of evaluations of f.

A sequential non-randomized algorithm to estimate $\alpha(f)$ with a budget of N evaluations is a pair $(\underline{X}_N, \widehat{\alpha}_N)$,

$$\underline{X}_N : f \mapsto \underline{X}_N(f) = (X_1(f), X_2(f), \ldots, X_N(f)) \in \mathbb{X}^N,$$
$$\widehat{\alpha}_N : f \mapsto \widehat{\alpha}_N(f) \in \mathbb{R}_+,$$

with the following properties:

(a) There exists $x_1 \in \mathbb{X}$ such that $X_1(f) = x_1$, i.e. X_1 does not depend on f.
(b) Let $Z_n(f) = f(X_n(f))$, $1 \le n \le N$. For all $1 \le n < N$, $X_{n+1}(f)$ depends measurably[3] on $\mathcal{I}_n(f)$, where $\mathcal{I}_n = ((X_1, Z_1), \ldots, (X_n, Z_n))$.
(c) $\widehat{\alpha}_N(f)$ depends measurably on $\mathcal{I}_N(f)$.

The mapping $\underline{X}_N$ will be referred to as a strategy, or policy, or design of experiments, and $\widehat{\alpha}_N$ will be called an estimator. The algorithm $(\underline{X}_N, \widehat{\alpha}_N)$ describes a sequence of decisions, made from an increasing amount of information: $X_1(f) = x_1$ is chosen prior to any evaluation; for each $n = 1, \ldots, N-1$, the algorithm uses information $\mathcal{I}_n(f)$ to choose the next evaluation point $X_{n+1}(f)$; the estimation $\widehat{\alpha}_N(f)$ of $\alpha(f)$ is the terminal decision. In some applications, the class of sequential algorithms must be further restricted: for instance, when K computer simulations can be run in parallel, algorithms that query batches of K evaluations at a time may be preferred (see, e.g. Ginsbourger et al. 2010). In this paper no such restriction is imposed.

The choice of the estimator $\widehat{\alpha}_N$ will be addressed in Sect. 2.4: for now, we simply assume that an estimator has been chosen, and focus on the problem of finding a good strategy $\underline{X}_N$; that is, one that will produce a good final approximation $\widehat{\alpha}_N(f)$ of $\alpha(f)$. Let $\mathcal{A}_N$ be the class of all strategies $\underline{X}_N$ that query sequentially N evaluations of f. Given a loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we define the error of approximation of a strategy $\underline{X}_N \in \mathcal{A}_N$ on f as $\epsilon(\underline{X}_N, f) = L(\widehat{\alpha}_N(f), \alpha(f))$. In this paper, we shall consider the quadratic loss function, so that $\epsilon(\underline{X}_N, f) = (\widehat{\alpha}_N(f) - \alpha(f))^2$.

We adopt a Bayesian approach to this decision problem: the unknown function f is considered as a sample path of a real-valued random process ξ defined on some probability space $(\Omega, \mathcal{B}, \mathsf{P}_0)$ with parameter in $x \in \mathbb{X}$, and a good strategy is a strategy that achieves, or gets close to, the *Bayes risk*

$$r_{\mathrm{B}} := \inf_{\underline{X}_N \in \mathcal{A}_N} \mathsf{E}_0(\epsilon(\underline{X}_N, \xi)),$$

where E_0 denotes the expectation with respect to P_0. From a subjective Bayesian point of view, the stochastic model ξ is a representation of our uncertain initial knowledge about f. From a more pragmatic perspective, the prior distribution can be seen as a tool to define a notion of a good strategy in an average sense. Another interesting route, not followed in this paper, would have been to consider the minimax risk $\inf_{\underline{X}_N \in \mathcal{A}_N} \max_f \epsilon(\underline{X}_N, f)$ over some class of functions.

Notation From now on, we shall consider the stochastic model ξ instead of the deterministic function f and, for abbreviation, the explicit dependence on ξ will be dropped when there is no risk of confusion; e.g., $\widehat{\alpha}_N$ will denote the random variable $\widehat{\alpha}_N(\xi)$, X_n will denote the random variable $X_n(\xi)$, etc. We will use the notations $\mathcal{F}_n$, P_n and E_n to denote respectively the σ-algebra generated by $\mathcal{I}_n$, the conditional distribution $\mathsf{P}_0(\cdot \mid \mathcal{F}_n)$ and the conditional expectation $\mathsf{E}_0(\cdot \mid \mathcal{F}_n)$. Note that the dependence of X_{n+1} on $\mathcal{I}_n$ can be rephrased by saying that X_{n+1} is $\mathcal{F}_n$-measurable. Recall that $\mathsf{E}_n(Z)$ is $\mathcal{F}_n$-measurable, and thus can be seen as a measurable function of $\mathcal{I}_n$, for any random variable Z.

2.2 Optimal and k-step lookahead strategies

It is well-known (see, e.g., Berry and Fristedt 1985; Mockus 1989; Bertsekas 1995) that an optimal strategy for such a finite horizon problem[4], i.e. a strategy $\underline{X}_N^\star \in \mathcal{A}_N$ such that $\mathsf{E}_0(\epsilon(\underline{X}_N^\star, \xi)) = r_{\mathrm{B}}$, can be formally obtained by *dynamic programming*: let $R_N = \mathsf{E}_N(\epsilon(\underline{X}_N, \xi)) = \mathsf{E}_N((\widehat{\alpha}_N - \alpha)^2)$ denote the terminal risk and define by backward induction

$$R_n = \min_{x \in \mathbb{X}} \mathsf{E}_n\big(R_{n+1} \mid X_{n+1} = x\big), \quad n = N-1, \ldots, 0. \tag{4}$$

To get an insight into (4), notice that R_{n+1}, $n = 0, \ldots, N-1$, depends measurably on $\mathcal{I}_{n+1} = (\mathcal{I}_n, X_{n+1}, Z_{n+1})$,

[3] I.e., there is a measurable map $\varphi_n : (\mathbb{X} \times \mathbb{R})^n \to \mathbb{X}$ such that $X_{n+1} = \varphi_n \circ \mathcal{I}_n$.

[4] In other words, a sequential decision problem where the total number of steps to be performed is known from the start.

so that $\mathsf{E}_n(R_{n+1} \mid X_{n+1} = x)$ is in fact an expectation with respect to Z_{n+1}, and R_n is an $\mathcal{F}_n$-measurable random variable. Then, we have $R_0 = r_{\mathrm{B}}$, and the strategy $\underline{X}_N^\star$ defined by

$$X_{n+1}^\star = \operatorname*{argmin}_{x\in\mathbb{X}} \mathsf{E}_n\big(R_{n+1} \mid X_{n+1} = x\big), \quad n = 1, \ldots, N-1, \tag{5}$$

is optimal.[5] It is crucial to observe here that, for this dynamic programming problem, both the space of possible actions and the space of possible outcomes at each step are continuous, and the state space $(\mathbb{X} \times \mathbb{R})^n$ at step n is of dimension $n(d+1)$. Any direct attempt at solving (4)–(5) numerically, over an horizon N of more than a few steps, will suffer from the curse of dimensionality.

Using (4), the optimal strategy can be expanded as

$$X_{n+1}^\star = \operatorname*{argmin}_{x\in\mathbb{X}} \mathsf{E}_n\Big(\min_{X_{n+2}} \mathsf{E}_{n+1} \cdots \min_{X_N} \mathsf{E}_{N-1} R_N \;\Big|\; X_{n+1} = x\Big).$$

A very general approach to construct sub-optimal—but hopefully good—strategies is to truncate this expansion after k terms, replacing the exact risk R_{n+k} by any available surrogate $\widetilde{R}_{n+k}$. Examples of such surrogates will be given in Sects. 3 and 4. The resulting strategy,

$$X_{n+1} = \operatorname*{argmin}_{x\in\mathbb{X}} \mathsf{E}_n\Big(\min_{X_{n+2}} \mathsf{E}_{n+1} \cdots \min_{X_{n+k}} \mathsf{E}_{n+k-1} \widetilde{R}_{n+k} \;\Big|\; X_{n+1} = x\Big) \tag{6}$$

is called a *k-step lookahead strategy* (see, e.g., Bertsekas 1995, Sect. 6.3). Note that both the optimal strategy (5) and the k-step lookahead strategy implicitly define a *sampling criterion* $J_n(x)$, $\mathcal{F}_n$-measurable, the minimum of which indicates the next evaluation to be performed. For instance, in the case of the k-step lookahead strategy, the sampling criterion is

$$J_n(x) = \mathsf{E}_n\left(\min_{X_{n+2}} \mathsf{E}_{n+1} \cdots \min_{X_{n+k}} \mathsf{E}_{n+k-1} \widetilde{R}_{n+k} \;\middle|\; X_{n+1} = x\right).$$

In the rest of the paper, we restrict our attention to the class of one-step lookahead strategies, which is, as we shall see in Sect. 3, large enough to provide very efficient algorithms. We leave aside the interesting question of whether more complex k-step lookahead strategies (with $k \geq 2$) could provide a significant improvement over the strategies examined in this paper.

Remark 1 In practice, the analysis of a computer code usually begins with an exploratory phase, during which the output of the code is computed on a *space-filling design* of size $n_0 < N$ (see, e.g., Santner et al. 2003). Such an exploratory phase will be colloquially referred to as the *initial design*. Sequential strategies such as (5) and (6) are meant to be used after this initial design, at steps $n_0 + 1, \ldots, N$. An important (and largely open) question is the choice of the size n_0 of the initial design, for a given global budget N. As a rule of thumb, some authors recommend to start with a sample size proportional to the dimension d of the input space $\mathbb{X}$, for instance $n_0 = 10\,d$; see Loeppky et al. (2009) and the references therein.

2.3 Gaussian process priors

Restricting ξ to be a Gaussian process makes it possible to deal with the conditional distributions P_n and conditional expectations E_n that appear in the strategies above. The idea of modeling an unknown function f by a Gaussian process has originally been introduced approximately in 1960 in time series analysis (Parzen 1962), optimization theory (Kushner 1964) and geostatistics (see, e.g., Chilès and Delfiner 1999, and the references therein). Today, the Gaussian process model plays a central role in the design and analysis of computer experiments (see, e.g., Sacks et al. 1989; Currin et al. 1991; Welch et al. 1992; Santner et al. 2003). Recall that the distribution of a Gaussian process ξ is uniquely determined by its mean function $m(x) := \mathsf{E}_0(\xi(x))$, $x \in \mathbb{X}$, and its covariance function $k(x, y) := \mathsf{E}_0((\xi(x) - m(x))(\xi(y) - m(y)))$, $x, y \in \mathbb{X}$. Hereafter, we shall use the notation $\xi \sim \mathrm{GP}(m, k)$ to say that ξ is a Gaussian process with mean function m and covariance function k.

Let $\xi \sim \mathrm{GP}(0, k)$ be a zero-mean Gaussian process. The best linear unbiased predictor (BLUP) of $\xi(x)$ from observations $\xi(x_i)$, $i = 1, \ldots, n$, also called the *kriging predictor* of $\xi(x)$, is the orthogonal projection

$$\widehat{\xi}(x; \underline{x}_n) := \sum_{i=1}^{n} \lambda_i(x; \underline{x}_n)\xi(x_i) \tag{7}$$

of $\xi(x)$ onto $\mathrm{span}\{\xi(x_i), i = 1, \ldots, n\}$. Here, the notation $\underline{x}_n$ stands for the set of points $\underline{x}_n = \{x_1, \ldots, x_n\}$. The weights $\lambda_i(x; \underline{x}_n)$ are the solutions of a system of linear equations

$$k(\underline{x}_n, \underline{x}_n)\lambda(x; \underline{x}_n) = k(x, \underline{x}_n) \tag{8}$$

where $k(\underline{x}_n, \underline{x}_n)$ stands for the $n \times n$ covariance matrix of the observation vector, $\lambda(x; \underline{x}_n) = (\lambda_1(x; \underline{x}_n), \ldots, \lambda_n(x; \underline{x}_n))^{\mathsf{T}}$,

[5] Proving rigorously that, for a given P_0 and $\widehat{\alpha}_N$, (4) and (5) actually define a (measurable!) strategy $\underline{X}_N^\star \in \mathcal{A}_N$ is a technical problem that is not of primary interest in this paper. This can be done for instance, in the case of a Gaussian process with continuous covariance function (as considered later), by proving that $x \mapsto \mathsf{E}_n(R_{n+1} \mid X_{n+1}(\xi) = x)$ is a continuous function on $\mathbb{X}$ and then using a measurable selection theorem.

and $k(x, \underline{x}_n)$ is a vector with entries $k(x, x_i)$. The function $x \mapsto \widehat{\xi}(x; \underline{x}_n)$ conditioned on $\xi(x_1) = f(x_1), \ldots, \xi(x_n) = f(x_n)$, is deterministic, and provides a cheap *surrogate model* for the true function f (see, e.g., Santner et al. 2003). The covariance function of the error of prediction, also called *kriging covariance* is given by

$$k(x, y; \underline{x}_n) := \operatorname{cov}\left(\xi(x) - \widehat{\xi}(x; \underline{x}_n), \xi(y) - \widehat{\xi}(y; \underline{x}_n)\right) = k(x, y) - \sum_i \lambda_i(x; \underline{x}_n)\, k(y, x_i). \tag{9}$$

The variance of the prediction error, also called the *kriging variance*, is defined as $\sigma^2(x; \underline{x}_n) = k(x, x; \underline{x}_n)$. One fundamental property of a zero-mean Gaussian process is the following (see, e.g., Chilès and Delfiner 1999, Chap. 3):

Proposition 1 *If $\xi \sim \mathrm{GP}(0, k)$, then the random process ξ conditioned on the σ-algebra $\mathcal{F}_n$ generated by $\xi(x_1), \ldots, \xi(x_n)$, which we shall denote by $\xi \mid \mathcal{F}_n$, is a Gaussian process with mean $\widehat{\xi}(\cdot; \underline{x}_n)$ given by (7)–(8) and covariance $k(\cdot, \cdot; \underline{x}_n)$ given by (9). In particular, $\widehat{\xi}(x; \underline{x}_n) = \mathsf{E}_0(\xi(x) \mid \mathcal{F}_n)$ is the best $\mathcal{F}_n$-measurable predictor of $\xi(x)$, for all $x \in \mathbb{X}$.*

In the domain of computer experiments, the mean of a Gaussian process is generally written as a linear parametric function

$$m(\cdot) = \beta^{\mathsf{T}} h(\cdot), \tag{10}$$

where β is an l-dimensional vector of unknown parameters, and $h = (h_1, \ldots, h_l)^{\mathsf{T}}$ is an l-dimensional vector of functions (in practice, polynomials). The simplest case is when the mean function is assumed to be an unknown constant m, in which case we can take $\beta = m$ and $h : x \in \mathbb{X} \mapsto 1$. The covariance function is generally chosen to be a translation-invariant function:

$$k : (x, y) \in \mathbb{X}^2 \mapsto \sigma^2 \rho_\theta(x - y),$$

where σ^2 is the variance of the (stationary) Gaussian process and ρ_θ is the correlation function, which generally depends on a parameter vector θ. When the mean is written under the form (10), the kriging predictor is again a linear combination of the observations, as in (7), and the weights $\lambda_i(x; \underline{x}_n)$ are again solutions of a system of linear equations (see, e.g., Chilès and Delfiner 1999), which can be written under a matrix form as

$$\begin{pmatrix} k(\underline{x}_n, \underline{x}_n) & h(\underline{x}_n)^{\mathsf{T}} \\ h(\underline{x}_n) & 0 \end{pmatrix} \begin{pmatrix} \lambda(x; \underline{x}_n) \\ \mu(x) \end{pmatrix} = \begin{pmatrix} k(x, \underline{x}_n) \\ h(x) \end{pmatrix}, \tag{11}$$

where $h(\underline{x}_n)$ is an $l \times n$ matrix with entries $h_i(x_j)$, $i = 1, \ldots, l$, $j = 1, \ldots, n$, μ is an l-dimensional vector of Lagrange coefficients ($k(\underline{x}_n, \underline{x}_n)$, $\lambda(x; \underline{x}_n)$, $k(x, \underline{x}_n)$ as above).

The kriging covariance function is given in this case by

$$k(x, y; \underline{x}_n) := \operatorname{cov}\left(\xi(x) - \widehat{\xi}(x; \underline{x}_n), \xi(y) - \widehat{\xi}(y; \underline{x}_n)\right) = k(x, y) - \lambda(x; \underline{x}_n)^{\mathsf{T}} k(y, \underline{x}_n) - \mu(x)^{\mathsf{T}} h(y). \tag{12}$$

The following result holds (Kimeldorf and Wahba 1970; O'Hagan 1978):

Proposition 2 *Let k be a covariance function.*

$$\textit{If} \quad \begin{cases} \xi \mid m \sim \mathrm{GP}(m, k) \\ m : x \mapsto \beta^{\mathsf{T}} h(x), \beta \sim \mathcal{U}_{\mathbb{R}^l} \end{cases}$$

$$\textit{then} \quad \xi \mid \mathcal{F}_n \sim \mathrm{GP}\left(\widehat{\xi}(\cdot; \underline{x}_n),\, k(\cdot, \cdot; \underline{x}_n)\right),$$

where $\mathcal{U}_{\mathbb{R}^l}$ stands for the (improper) uniform distribution over $\mathbb{R}^l$, and where $\widehat{\xi}(\cdot; \underline{x}_n)$ and $k(\cdot, \cdot; \underline{x}_n)$ are given by (7), (11) and (12).

Proposition 2 justifies the use of kriging in a Bayesian framework provided that the covariance function of ξ is known. However, the covariance function is rarely assumed to be known in applications. Instead, the covariance function is generally taken in some parametric class (in this paper, we use the so-called Matérn covariance function, see Appendix A). A *fully Bayesian* approach also requires to choose a prior distribution for the unknown parameters of the covariance (see, e.g., Handcock and Stein 1993; Kennedy and O'Hagan 2001; Paulo 2005). Sampling techniques (Monte Carlo Markov Chains, Sequential Monte Carlo…) are then generally used to approximate the posterior distribution of the unknown covariance parameters. Very often, the popular *empirical Bayes* approach is used instead, which consists in plugging-in the maximum likelihood (ML) estimate to approximate the posterior distribution of ξ. This approach has been used in previous papers about contour estimation or probability of failure estimation (Picheny et al. 2010; Ranjan et al. 2008; Bichon et al. 2008). In Sect. 5.2 we will adopt a plug-in approach as well.

Simplified notation In the rest of the paper, we shall use the following simplified notations when there is no risk of confusion: $\widehat{\xi}_n(x) := \widehat{\xi}(x; \underline{X}_n)$, $\sigma_n^2(x) := \sigma^2(x; \underline{X}_n)$.

2.4 Estimators of the probability of failure

Given a random process ξ and a strategy $\underline{X}_N$, the optimal estimator that minimizes $\mathsf{E}_0((\alpha - \widehat{\alpha}_n)^2)$ among all $\mathcal{F}_n$-measurable estimators $\widehat{\alpha}_n$, $1 \le n \le N$, is

$$\widehat{\alpha}_n = \mathsf{E}_n(\alpha) = \mathsf{E}_n\left(\int_{\mathbb{X}} \mathbb{1}_{\xi > u}\, \mathrm{d}\mathsf{P}_{\mathbb{X}}\right) = \int_{\mathbb{X}} p_n\, \mathrm{d}\mathsf{P}_{\mathbb{X}}, \tag{13}$$

where

$$p_n : x \in \mathbb{X} \mapsto \mathsf{P}_n\{\xi(x) > u\}. \tag{14}$$

When ξ is a Gaussian process, the probability $p_n(x)$ of exceeding u at $x \in \mathbb{X}$ given $\mathcal{I}_n$ has a simple closed-form expression:

$$p_n(x) = 1 - \Phi\left(\frac{u - \widehat{\xi}_n(x)}{\sigma_n(x)}\right) = \Phi\left(\frac{\widehat{\xi}_n(x) - u}{\sigma_n(x)}\right), \tag{15}$$

where Φ is the cumulative distribution function of the normal distribution. Thus, in the Gaussian case, the estimator (13) is amenable to a numerical approximation, by integrating the excess probability p_n over $\mathbb{X}$ (for instance using Monte Carlo sampling, see Sect. 3.3).

Another natural way to obtain an estimator of α given $\mathcal{I}_n$ is to approximate the excess indicator $\mathbb{1}_{\xi>u}$ by a hard classifier $\eta_n : \mathbb{X} \to \{0, 1\}$, where "hard" refers to the fact that η_n takes its values in $\{0, 1\}$. If η_n is close in some sense to $\mathbb{1}_{\xi>u}$, the estimator

$$\widehat{\alpha}_n = \int_{\mathbb{X}} \eta_n \mathrm{d}\mathsf{P}_{\mathbb{X}} \tag{16}$$

should be close to α. More precisely,

$$\begin{aligned}\mathsf{E}_n\left((\widehat{\alpha}_n - \alpha)^2\right) &= \mathsf{E}_n\left[\left(\int (\eta_n - \mathbb{1}_{\xi>u})\mathrm{d}\mathsf{P}_{\mathbb{X}}\right)^2\right] \\ &\le \int \mathsf{E}_n\left((\eta_n - \mathbb{1}_{\xi>u})^2\right)\mathrm{d}\mathsf{P}_{\mathbb{X}}. \end{aligned}\tag{17}$$

Let $\tau_n(x) = \mathsf{P}_n\{\eta_n(x) \neq \mathbb{1}_{\xi(x)>u}\} = \mathsf{E}_n((\eta_n(x) - \mathbb{1}_{\xi(x)>u})^2)$ be the probability of misclassification; that is, the probability to predict a point above (resp. under) the threshold when the true value is under (resp. above) the threshold. Thus, (17) shows that it is desirable to use a classifier η_n such that τ_n is small for all $x \in \mathbb{X}$. For instance, the method called SMART (Deheeger and Lemaire 2007) uses a support vector machine to build η_n. Note that

$$\tau_n(x) = p_n(x) + (1 - 2p_n(x))\,\eta_n(x).$$

Therefore, the right-hand side of (17) is minimized if we set

$$\eta_n(x) = \mathbb{1}_{p_n(x)>1/2} = \mathbb{1}_{\bar{\xi}_n(x)>u}, \tag{18}$$

where $\bar{\xi}_n(x)$ denotes the posterior median of $\xi(x)$. Then, we have

$$\tau_n(x) = \min(p_n(x), 1 - p_n(x)).$$

In the case of a Gaussian process, the posterior median and the posterior mean are equal. Then, the classifier that minimizes $\tau_n(x)$ for each $x \in \mathbb{X}$ is $\eta_n = \mathbb{1}_{\widehat{\xi}_n>u}$, in which case

$$\begin{aligned}\tau_n(x) &= \mathsf{P}_n\left((\xi(x) - u)(\widehat{\xi}_n(x) - u) < 0\right) \\ &= 1 - \Phi\left(\frac{|\widehat{\xi}_n(x) - u|}{\sigma_n(x)}\right). \end{aligned}\tag{19}$$

Notice that for $\eta_n = \mathbb{1}_{\widehat{\xi}_n>u}$, we have $\widehat{\alpha}_n = \alpha(\widehat{\xi}_n)$. Therefore, this approach to obtain an estimator of α can be seen as a type of plug-in estimation.

Standing assumption It will assumed in the rest of the paper that ξ is a Gaussian process, or more generally that $\xi \mid \mathcal{F}_n \sim \mathrm{GP}(\widehat{\xi}_n, k(\cdot, \cdot; \underline{x}_n))$ for all $n \ge 1$ as in Proposition 2.

3 Stepwise uncertainty reduction

3.1 Principle

A very natural and straightforward way of building a one-step lookahead strategy is to select *greedily* each evaluation as if it were the last one. This kind of strategy, sometimes called *myopic*, has been successfully applied in the field of Bayesian global optimization (Mockus et al. 1978; Mockus 1989), yielding the famous *expected improvement* criterion later popularized in the Efficient Global Optimization (EGO) algorithm of Jones et al. (1998).

When the Bayesian risk provides a measure of the estimation error or uncertainty (as in the present case), we call such a strategy a *stepwise uncertainty reduction* (SUR) strategy. In the field of global optimization, the Informational Approach to Global Optimization (IAGO) of Villemonteix et al. (2009) is an example of a SUR strategy, where the Shannon entropy of the minimizer is used instead of the quadratic cost. When considered in terms of utility rather than cost, such strategies have also been called *knowledge gradient policies* by Frazier et al. (2008).

Given a sequence of estimators $(\widehat{\alpha}_n)_{n\ge1}$, a direct application of the above principle using the quadratic loss function yields the sampling criterion (to be minimized)

$$J_n(x) = \mathsf{E}_n\left((\alpha - \widehat{\alpha}_{n+1})^2 \mid X_{n+1} = x\right). \tag{20}$$

Having found no closed-form expression for this criterion, and no efficient numerical procedure for its approximation, we will proceed by upper-bounding and discretizing (20) in order to get an expression that will lend itself to a numerically tractable approximation. By doing so, several SUR strategies will be derived, depending on the choice of estimator (the posterior mean (13) or the plug-in estimator (16) with (18)) and bounding technique.

3.2 Upper bounds of the SUR sampling criterion

Recall that $\tau_n(x) = \min(p_n(x), 1 - p_n(x))$ is the probability of misclassification at x using the optimal classifier $\mathbb{1}_{\widehat{\xi}_n(x)>u}$. Let us further denote by $v_n(x) := p_n(x)(1 - p_n(x))$ the variance of the excess indicator $\mathbb{1}_{\xi(x)\geq u}$.

Proposition 3 *Assume that either* $\widehat{\alpha}_n = \mathsf{E}_n(\alpha)$ *or* $\widehat{\alpha}_n = \int \mathbb{1}_{\widehat{\xi}_n\geq u} \mathrm{dP}_{\mathbb{X}}$. *Define* $G_n := \int_{\mathbb{X}} \sqrt{\gamma_n(y)} \mathrm{dP}_{\mathbb{X}}$ *for all* $n \in \{0, \ldots, N-1\}$, *with*

$$\gamma_n := \begin{cases} v_n = p_n(1-p_n) = \tau_n(1-\tau_n), \\ \quad \textit{if } \widehat{\alpha}_n = \mathsf{E}_n(\alpha), \\ \tau_n = \min(p_n, 1-p_n), \\ \quad \textit{if } \widehat{\alpha}_n = \int \mathbb{1}_{\widehat{\xi}_n\geq u} \mathrm{dP}_{\mathbb{X}}. \end{cases}$$

Then, for all $x \in \mathbb{X}$ *and all* $n \in \{0, \ldots, N-1\}$,

$$J_n(x) \leq \widetilde{J}_n(x) := \mathsf{E}_n\left(G_{n+1}^2 \mid X_{n+1} = x\right).$$

Note that $\gamma_n(x)$ is a function of $p_n(x)$ that vanishes at 0 and 1, and reaches its maximum at $1/2$; that is, when the uncertainty on $\mathbb{1}_{\widehat{\xi}_n(x)>u}$ is maximal (see Fig. 1).

Proof First, observe that, for all $n \geq 0$, $\alpha - \widehat{\alpha}_n = \int U_n \, \mathrm{dP}_{\mathbb{X}}$, with

$$U_n : x \in \mathbb{X} \mapsto U_n(x) = \begin{cases} \mathbb{1}_{\xi(x)>u} - p_n(x) \\ \quad \textit{if } \widehat{\alpha}_n = \mathsf{E}_n(\alpha), \\ \mathbb{1}_{\xi(x)>u} - \mathbb{1}_{\widehat{\xi}_n(x)>u} \\ \quad \textit{if } \widehat{\alpha}_n = \int \mathbb{1}_{\widehat{\xi}_n\geq u} \mathrm{dP}_{\mathbb{X}}. \end{cases} \tag{21}$$

Moreover, note that $\gamma_n = \|U_n\|_n^2$ in both cases, where $\|\cdot\|_n : L^2(\Omega, \mathcal{B}, \mathsf{P}) \to L^2(\Omega, \mathcal{F}_n, \mathsf{P})$, $W \mapsto \mathsf{E}_n(W^2)^{1/2}$. Then, using the generalized Minkowski inequality (see, e.g., Vestrup 2003, Sect. 10.7) we get that

$$\begin{aligned} \left\| \int U_n \, \mathrm{dP}_{\mathbb{X}} \right\|_n &\leq \int \|U_n\|_n \, \mathrm{dP}_{\mathbb{X}} \\ &= \int \sqrt{\gamma_n} \, \mathrm{dP}_{\mathbb{X}} = G_n. \end{aligned} \tag{22}$$

Finally, it follows from the tower property of conditional expectations and (22) that, for all $n \geq 0$,

$$\begin{aligned} J_n(x) &= \mathsf{E}_n\left(\|\alpha - \widehat{\alpha}_{n+1}\|_{n+1}^2 \mid X_{n+1} = x\right) \\ &= \mathsf{E}_n\left(\left\| \int U_{n+1} \, \mathrm{dP}_{\mathbb{X}} \right\|_{n+1}^2 \,\middle|\, X_{n+1} = x\right) \\ &\leq \mathsf{E}_n\left(G_{n+1}^2 \mid X_{n+1} = x\right). \end{aligned}$$

□

Note that two other upper-bounding sampling criteria readily follow from those of Proposition 3, by using the Cauchy-Schwarz inequality in $L^2(\mathbb{X}, \mathcal{B}(\mathbb{X}), \mathsf{P}_{\mathbb{X}})$:

$$\widetilde{J}_n(x) \leq \mathsf{E}_n\left(\int \gamma_{n+1} \, \mathrm{dP}_{\mathbb{X}} \,\middle|\, X_{n+1} = x\right). \tag{23}$$

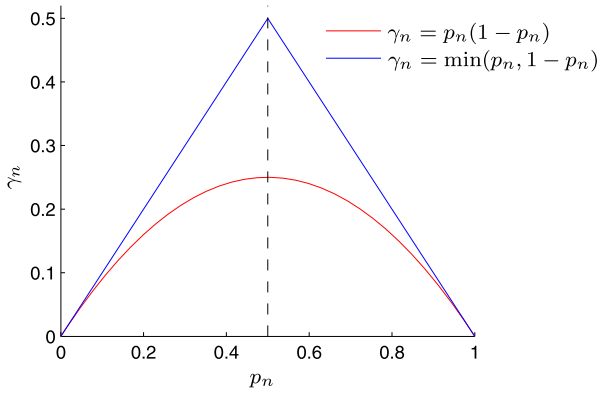


Fig. 1 γ_n as a function of p_n (see Proposition 3). In both cases, γ_n is maximum at $p_n = 1/2$

As a result, we can write four SUR criteria, whose expressions are summarized in Table 1. Criterion $J_{1,n}^{\mathrm{SUR}}$ has been proposed in the PhD thesis of Piera-Martinez (2008) and in conference papers (Vazquez and Piera-Martinez 2007; Vazquez and Bect 2009); the other ones, to the best of our knowledge, are new. Each criterion is expressed as the conditional expectation of some (possibly squared) $\mathcal{F}_{n+1}$-measurable integral criterion, with an integrand that can be expressed as a function of the probability of misclassification τ_{n+1}. It is interesting to note that the integral in J_4^{SUR} is the integrated mean square error (IMSE)[6] for the process $\mathbb{1}_{\xi>u}$.

Remark 2 The conclusions of Proposition 3 still hold in the general case when ξ is not assumed to be a Gaussian process, provided that the posterior median $\widetilde{\xi}_n$ is substituted to posterior the mean $\widehat{\xi}_n$.

3.3 Discretizations

In this section, we proceed with the necessary integral discretizations of the SUR criteria to make them suitable for numerical evaluation and implementation on computers. Assume that n steps of the algorithm have already been performed and consider, for instance, the criterion

$$J_{3,n}^{\mathrm{SUR}}(x) = \mathsf{E}_n\left(\int \tau_{n+1}(y) \, \mathsf{P}_{\mathbb{X}}(\mathrm{d}y) \,\middle|\, X_{n+1} = x\right). \tag{24}$$

[6] The IMSE criterion is usually applied to the response surface ξ itself (see, e.g., Box and Draper 1987; Sacks et al. 1989). The originality here is to consider the IMSE of the process $\mathbb{1}_{\xi>u}$ instead. Another way of adapting the IMSE criterion for the estimation of a probability of failure, proposed by Picheny et al. (2010), is recalled in Sect. 4.2.

Table 1 Expressions of four SUR-type criteria

SUR-type sampling criterion	How it is obtained
$J_{1,n}^{\mathrm{SUR}}(x) = \mathsf{E}_n((\int \sqrt{\tau_{n+1}}\,\mathrm{d}\mathsf{P}_{\mathbb{X}})^2 \mid X_{n+1} = x)$	Proposition 3 with $\widehat{\alpha}_n = \int \mathbb{1}_{\widehat{\xi}_n > u}\,\mathrm{d}\mathsf{P}_{\mathbb{X}}$
$J_{2,n}^{\mathrm{SUR}}(x) = \mathsf{E}_n((\int \sqrt{v_{n+1}}\,\mathrm{d}\mathsf{P}_{\mathbb{X}})^2 \mid X_{n+1} = x)$	Proposition 3 with $\widehat{\alpha}_n = \mathsf{E}_n(\alpha)$
$J_{3,n}^{\mathrm{SUR}}(x) = \mathsf{E}_n(\int \tau_{n+1}\,\mathrm{d}\mathsf{P}_{\mathbb{X}} \mid X_{n+1} = x)$	Equation (23) with $\widehat{\alpha}_n = \int \mathbb{1}_{\widehat{\xi}_n > u}\,\mathrm{d}\mathsf{P}_{\mathbb{X}}$
$J_{4,n}^{\mathrm{SUR}}(x) = \mathsf{E}_n(\int v_{n+1} \mid X_{n+1} = x)$	Equation (23) with $\widehat{\alpha}_n = \mathsf{E}_n(\alpha)$

Remember that, for each $y \in \mathbb{X}$, the probability of misclassification $\tau_{n+1}(y)$ is $\mathcal{F}_{n+1}$-measurable and, therefore, is a function of $\mathcal{I}_{n+1} = (\mathcal{I}_n, X_{n+1}, Z_{n+1})$. Since $\mathcal{I}_n$ is known at this point, we introduce the notation $v_{n+1}(y; X_{n+1}, Z_{n+1}) = \tau_{n+1}(y)$ to emphasize the fact that, when a new evaluation point must be chosen at step $(n+1)$, $\tau_{n+1}(y)$ depends on the choice of X_{n+1} and the random outcome Z_{n+1}. Let us further denote by $\mathsf{Q}_{n,x}$ the probability distribution of $\xi(x)$ under P_n. Then, (24) can be rewritten as

$$J_{3,n}^{\mathrm{SUR}}(x) = \iint_{\mathbb{R}\times\mathbb{X}} v_{n+1}(y; x, z)\mathsf{Q}_{n,x}(\mathrm{d}z)\,\mathsf{P}_{\mathbb{X}}(\mathrm{d}y),$$

and the corresponding strategy is:

$$X_{n+1} = \operatorname*{argmin}_{x\in\mathbb{X}} \iint_{\mathbb{R}\times\mathbb{X}} v_{n+1}(y; x, z) \times \mathsf{Q}_{n,x}(\mathrm{d}z)\,\mathsf{P}_{\mathbb{X}}(\mathrm{d}y). \tag{25}$$

Given $\mathcal{I}_n$ and a triple (x, y, z), $v_{n+1}(y; x, z)$ can be computed efficiently using the equations provided in Sects. 2.3 and 2.4.

At this point, we need to address: (1) the computation of the integral on $\mathbb{X}$ with respect to $\mathsf{P}_{\mathbb{X}}$; (2) the computation of the integral on $\mathbb{R}$ with respect to $\mathsf{Q}_{n,x}$; (3) the minimization of the resulting criterion with respect to $x \in \mathbb{X}$.

To solve the first problem, we draw an i.i.d. sequence $Y_1, \ldots, Y_m \sim \mathsf{P}_{\mathbb{X}}$ and use the Monte Carlo approximation:

$$\int_{\mathbb{X}} v_{n+1}(y; x, z)\mathsf{P}_{\mathbb{X}}(\mathrm{d}y) \approx \frac{1}{m}\sum_{j=1}^{m} v_{n+1}(Y_j; x, z).$$

An increasing sample size $n \mapsto m_n$ should be used to build a convergent algorithm for the estimation of α (possibly with a different sequence $Y_{n,1}, \ldots, Y_{n,m_n}$ at each step). In this paper we adopt a different approach instead, which is to take a fixed sample size $m > 0$ and keep the same sample $Y_1, \ldots, Y_m$ throughout the iterations. Equivalently, it means that we choose to work from the start on a discretized version of the problem: we replace $\mathsf{P}_{\mathbb{X}}$ by the empirical distribution $\widehat{\mathsf{P}}_{\mathbb{X},n} = \frac{1}{m}\sum_{j=1}^{m}\delta_{Y_j}$, and our goal is now to *estimate the Monte Carlo estimator* $\alpha_m = \int \mathbb{1}_{\xi>u}\mathrm{d}\widehat{\mathsf{P}}_{\mathbb{X},n} = \frac{1}{m}\sum_{j=1}^{m}\mathbb{1}_{\xi(Y_j)>u}$, using either the posterior mean $\mathsf{E}_n(\alpha_m) = \frac{1}{m}\sum_j p_n(Y_j)$ or the plug-in estimate $\frac{1}{m}\sum_j \mathbb{1}_{\widehat{\xi}(Y_j;\underline{X}_n)>u}$.

This kind of approach has be coined *meta-estimation* by Arnaud et al. (2010): the objective is to estimate the value of a precise Monte Carlo estimator of $\alpha(f)$ (m being large), using prior information on f to alleviate the computational burden of running m times the computer code f. This point of view also underlies the work in structural reliability of Hurtado (2004, 2007), Deheeger and Lemaire (2007), Deheeger (2008), and more recently Echard et al. (2010a, 2010b).

This new point of view suggests a natural solution for the third problem, which is to replace the continuous search for a minimizer $x \in \mathbb{X}$ by a discrete search over the set $\mathbb{X}_m := \{Y_1, \ldots, Y_m\}$. This is obviously sub-optimal, even in the meta-estimation framework introduced above, since picking $x \in \mathbb{X}\setminus\mathbb{X}_m$ can sometimes bring more information about $\xi(Y_1), \ldots, \xi(Y_m)$ than the best possible choice in $\mathbb{X}_m$. Global optimization algorithms may of course be used to tackle directly the continuous search problem: for instance, (Ranjan et al. 2008) use a combination of a genetic algorithm and local search technique, (Bichon et al. 2008) use the DIRECT algorithm and (Picheny et al. 2010) use a covariance-matrix-adaptation evolution strategy. In this paper we will stick to the discrete search approach, since it is much simpler to implement (we shall present in Sect. 3.4 a method to handle the case of large m) and provides satisfactory results (see Sect. 5).

Finally, remark that the second problem boils down to the computation of a one-dimensional integral with respect to Lebesgue's measure. Indeed, since ξ is a Gaussian process, $\mathsf{Q}_{n,x}$ is a Gaussian probability distribution with mean $\widehat{\xi}_n(x)$ and variance $\sigma_n^2(x)$ as explained in Sect. 2.3. The integral can be computed using a standard Gauss-Hermite quadrature with Q points (see, e.g., Press et al. 1992, Chap. 4):

$$\int v_{n+1}(y; x, z)\mathsf{Q}_{n,x}(\mathrm{d}z) \approx \frac{1}{\sqrt{\pi}}\sum_{q=1}^{Q} w_q\, v_{n+1}\big(y; x, \widehat{\xi}_n(x) + \sigma_n(x)u_q\sqrt{2}\big),$$

where $u_1, \ldots, u_Q$ denote the quadrature points and $w_1, \ldots, w_Q$ the corresponding weights. Note that this is equivalent to replacing under P_n the random variable $\xi(x)$ by a quantized random variable with probability distribution

$\sum_{q=1}^{Q} w'_q \delta_{z_{n+1,q}(x)}$, where $w'_q = w_q/\sqrt{\pi}$ and $z_{n+1,q}(x) = \widehat{\xi}_n(x) + \sigma_n(x) u_q \sqrt{2}$.

Taking all three discretizations into account, the proposed strategy is:

$$X_{n+1} = \operatorname*{argmin}_{1\le k\le m} \sum_{j=1}^{m} \sum_{q=1}^{Q} w'_q v_{n+1}\left(Y_j;\, Y_k, z_{n+1,q}(Y_k)\right). \quad (26)$$

3.4 Implementation

This section gives implementation guidelines for the SUR strategies described in Sect. 3. As said in Sect. 3.3, the strategy (26) can, in principle, be translated directly into a computer program. In practice however, we feel that there is still room for different implementations. In particular, it is important to keep the computational complexity of the strategies at a reasonable level. We shall explain in this section some simplifications we have made to achieve this goal.

A straight implementation of (26) for the choice of an additional evaluation point is described in Table 2. This procedure is meant to be called iteratively in a sequential algorithm, such as that described for instance in Table 3. Note that the only parameter to be specified in the SUR strategy (26) is Q, which tunes the precision of the approximation of the integral on $\mathbb{R}$ with respect to $\mathsf{Q}_{n,x}$. In our numerical experiments, it was observed that taking $Q = 12$ achieves a good compromise between precision and numerical complexity.

To assess the complexity of a SUR sampling strategy, recall that kriging takes $O(mn^2)$ operations to predict the value of f at m locations from n evaluation results of f (we suppose that $m > n$ and no approximation is carried out). In the procedure to select an evaluation, a first kriging prediction is performed at Step 1 and then, m different predictions have to performed at Step 2.1. This cost becomes rapidly burdensome for large values of n and m, and we must further simplify (26) to be able to work on applications where m must be large. A natural idea to alleviate the computational cost of the strategy is to avoid dealing with candidate points that have a very low probability of misclassification, since they are probably far from the frontier of the domain

Table 2 Procedure to select a new evaluation point $X_{n+1} \in \mathbb{X}$ using a SUR strategy

Require computer representations of

(a) A set $\mathcal{I}_n = \{(X_1, f(X_1)), \ldots, (X_n, f(X_n))\}$ of evaluation results;
(b) A Gaussian process prior ξ with a (possibly unknown linear parametric) mean function and a covariance function k_θ, with parameter θ;
(c) A (pseudo-)random sample $\mathbb{X}_m = \{Y_1, \ldots, Y_m\}$ of size m drawn from the distribution $\mathsf{P}_{\mathbb{X}}$;
(d) Quadrature points $u_1, \ldots, u_Q$ and corresponding weights $w'_1, \ldots, w'_Q$;
(e) A threshold u.

1. Compute the kriging approximation $\widehat{f_n}$ and kriging variance σ_n^2 on $\mathbb{X}_m$ from $\mathcal{I}_n$
2. For each candidate point Y_j, $j \in \{1, \ldots, m\}$,
   2.1 For each point Y_k, $k \in \{1, \ldots, m\}$, compute the kriging weights $\lambda_i(Y_k; \{\underline{X}_n, Y_j\})$, $i \in \{1, \ldots, (n+1)\}$, and the kriging variances $\sigma^2(Y_k; \{\underline{X}_n, Y_j\})$
   2.2 Compute $z_{n+1,q}(Y_j) = \widehat{f_n}(Y_j) + \sigma_n(Y_j) u_q \sqrt{2}$, for $q = 1, \ldots, Q$
   2.3 For each $z_{n+1,q}(Y_j)$, $q \in \{1, \ldots, Q\}$,
      2.3.1 Compute the kriging approximation $\tilde{f}_{n+1,j,q}$ on $\mathbb{X}_m$ from $\mathcal{I}_n \cup (Y_j, f(Y_j) = z_{n+1,q}(Y_j))$, using the weights $\lambda_i(Y_k; \{\underline{X}_n, Y_j\})$, $i = 1, \ldots, (n+1)$, $k = 1, \ldots, m$, obtained at Step 2.1.
      2.3.2 For each $k \in \{1, \ldots, m\}$, compute $v_{n+1}(Y_k;\, Y_j, z_{n+1,q}(Y_j))$, using u, $\tilde{f}_{n+1,j,q}$ obtained in 2.3.1, and $\sigma^2(Y_k; \{\underline{X}_n, Y_j\})$ obtained in 2.1
   2.4 Compute $J_n(Y_j) = \sum_{k=1}^{m} \sum_{q=1}^{Q} w'_q\, v_{n+1}(Y_k;\, Y_j, z_{n+1,q}(Y_j))$.
3. Find $j^\star = \operatorname{argmin}_j J_n(Y_j)$ and set $X_{n+1} = Y_{j^\star}$

Table 3 Sequential estimation of a probability of failure

1. Construct an initial design of size $n_0 < N$ and evaluate f at the points of the initial design.
2. Choose a Gaussian process ξ (in practice, this amounts to choosing a parametric form for the mean of ξ and a parametric covariance function k_θ)
3. Generate a Monte Carlo sample $\mathbb{X}_m = \{Y_1, \ldots, Y_m\}$ of size m from $\mathsf{P}_{\mathbb{X}}$
4. While the evaluation budget N is not exhausted,
   4.1 Optional step: estimate the parameters of the covariance function (case of a plug-in approach);
   4.2 Select a new evaluation point, using past evaluation results, the prior ξ and $\mathbb{X}_m$;
   4.3 Perform the new evaluation.
5. Estimate the probability of failure obtained from the N evaluations of f (for instance, by using $\mathsf{E}_N(\alpha_m) = \frac{1}{m}\sum_j p_N(Y_j)$).

of failure. It is also likely that those points with a low probability of misclassification will have a very small contribution in the variance of the error of estimation $\widehat{\alpha}_n - \alpha_m$.

Therefore, the idea is to rewrite the sampling strategy described by (26), in such a way that the first summation (over m) and the search set for the minimizer is restricted to a subset of points Y_j corresponding to the m_0 largest values of $\tau_n(Y_j)$. The corresponding algorithm is not described here for the sake of brevity but can easily be adapted from that of Table 2. Sections 5.2 and 5.3 will show that this *pruning* scheme has almost no consequence on the performances of the SUR strategies, even when one considers small values for m_0.

4 Other strategies

4.1 Estimation of a probability of failure and closely related objectives

Given a real function f defined over $\mathbb{X} \subseteq \mathbb{R}^d$, and a threshold $u \in \mathbb{R}$, consider the following possible goals:

1. estimate a region $\Gamma \subset \mathbb{X}$ of the form $\Gamma = \{x \in \mathbb{X} \mid f(x) > u\}$;
2. estimate the level set $\partial\Gamma = \{x \in \mathbb{X} \mid f(x) = u\}$;
3. estimate f precisely in a neighborhood of $\partial\Gamma$;
4. estimate the probability of failure $\alpha = \mathsf{P}_{\mathbb{X}}(\Gamma)$ for a given probability measure $\mathsf{P}_{\mathbb{X}}$.

These different goals are, in fact, closely related: indeed, they all require, more or less explicitly, to select sampling points in order to get a fine knowledge of the function f in a neighborhood of the level set $\partial\Gamma$ (the location of which is unknown before the first evaluation). Any strategy proposed for one of the first three objectives is therefore expected to perform reasonably well on the fourth one, which is the topic of this paper.

Several strategies recently introduced are presented in Sects. 4.2 and 4.3, and will be compared numerically to the SUR strategy in Sect. 5. Each of these strategies has been initially proposed by its authors to address one or several of the above objectives, but they will only be discussed from the point of view of their performance on the fourth one. Of course, a comparison focused on any other objective would probably be based on different performance metrics, and thus could yield a different performance ranking of the strategies.

4.2 The targeted IMSE criterion

The *targeted IMSE* proposed in Picheny et al. (2010) is a modification of the IMSE (Integrated Mean Square Error) sampling criterion (Sacks et al. 1989). While the IMSE sampling criterion computes the average of the kriging variance (over a compact domain $\mathbb{X}$) in order to achieve a space-filling design, the targeted IMSE computes a weighted average of the kriging variance for a better exploration of the regions near the frontier of the domain of failure, as in Oakley (2004). The idea is to put a large weight in regions where the kriging prediction is close to the threshold u, and a small one otherwise. Given $\mathcal{I}_n$, the targeted IMSE sampling criterion, hereafter abbreviated as tIMSE, can be written as

$$J_n^{\mathrm{tIMSE}}(x) = \mathsf{E}_n\left(\int_{\mathbb{X}} \left(\xi - \widehat{\xi}_{n+1}\right)^2 W_n \, \mathrm{d}\mathsf{P}_{\mathbb{X}} \,\middle|\, X_{n+1} = x \right) \quad (27)$$

$$= \int_{\mathbb{X}} \sigma^2\left(y; X_1, \ldots, X_n, x\right) \, W_n(y) \, \mathsf{P}_{\mathbb{X}}(\mathrm{d}y), \quad (28)$$

where W_n is a weight function based on $\mathcal{I}_n$. The weight function suggested by Picheny et al. (2010) is

$$W_n(x) = \frac{1}{s_n(x)\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\widehat{\xi}_n(x) - u}{s_n(x)} \right)^2 \right), \quad (29)$$

where $s_n^2(x) = \sigma_\varepsilon^2 + \sigma_n^2(x)$. Note that $W_n(x)$ is large when $\widehat{\xi}_n(x) \approx u$ and $\sigma_n^2(x) \approx 0$, i.e., when the function is known to be close to u.

The tIMSE criterion operates a trade-off between global uncertainty reduction (high kriging variance σ_n^2) and exploration of target regions (high weight function W_n). The weight function depends on a parameter $\sigma_\varepsilon > 0$, which allows to tune the width of the "window of interest" around the threshold. For large values of σ_ε, J^{tIMSE} behaves approximately like the IMSE sampling criterion. The choice of an appropriate value for σ_ε, when the goal is to estimate a probability of failure, will be discussed on the basis of numerical experiments in Sect. 5.3.

The tIMSE strategy requires a computation of the expectation with respect to $\xi(x)$ in (27), which can be done analytically, yielding (28). The computation of the integral with respect to $\mathsf{P}_{\mathbb{X}}$ on $\mathbb{X}$ can be carried out with a Monte Carlo approach, as explained in Sect. 3.3. Finally, the optimization of the criterion is replaced by a discrete search in our implementation.

4.3 Criteria based on the marginal distributions

Other sampling criteria proposed by Ranjan et al. (2008), Bichon et al. (2008) and Echard et al. (2010a, 2010b) are briefly reviewed in this section.[7] A common feature of these three criteria is that, unlike the SUR and tIMSE criteria

[7] Note that the paper of Ranjan et al. (2008) is the only one in this category that does not address the problem of estimating a probability of failure (i.e., Objective 4 of Sect. 4.1).

discussed so far, they only depend on the *marginal posterior distribution* at the considered candidate point $x \in \mathbb{X}$, which is a Gaussian $\mathcal{N}(\widehat{\xi}_n(x), \sigma_n^2(x))$ distribution. As a consequence, they are of course much cheaper to compute than integral criteria like SUR and tIMSE.

A natural idea, in order to sequentially improve the estimation of the probability of failure, is to visit the point $x \in \mathbb{X}$ where the event $\{\xi(x) \geq u\}$ is the most uncertain. This idea, which has been explored by Echard et al. (2010a, 2010b), corresponds formally to the sampling criterion

$$J_n^{\mathrm{EGL}}(x) = \tau_n(x) = 1 - \Phi\left(\frac{|u - \widehat{\xi}_n(x)|}{\sigma_n(x)}\right). \tag{30}$$

As in the case of the tIMSE criterion and also, less explicitly, in SUR criteria, a trade-off is realized between global uncertainty reduction (choosing points with a high $\sigma_n^2(x)$) and exploration of the neighborhood of the estimated contour (where $|u - \widehat{\xi}_n(x)|$ is small).

The same leading principle motivates the criteria proposed by Ranjan et al. (2008) and Bichon et al. (2008), which can be seen as special cases of the following sampling criterion:

$$J_n^{\mathrm{RB}}(x) := \mathsf{E}_n\left(\max\left(0, \epsilon(x)^\delta - |u - \xi(x)|^\delta\right)\right), \tag{31}$$

where $\epsilon(x) = \kappa\,\sigma_n(x)$, $\kappa, \delta > 0$. The following proposition provides some insights into this sampling criterion:

Proposition 4 *Define* $G_{\kappa,\delta} :]0,1[\to \mathbb{R}_+$ *by*

$$G_{\kappa,\delta}(p) := \mathsf{E}\left(\max\left(0, \kappa^\delta - \left|\Phi^{-1}(p) + U\right|\right)\right),$$

where U *is a Gaussian* $\mathcal{N}(0,1)$ *random variable. Let* φ *and* Φ *denote respectively the probability density function and the cumulative distribution function of* U.

(a) $G_{\kappa,\delta}(p) = G_{\kappa,\delta}(1-p)$ *for all* $p \in]0,1[$.

(b) $G_{\kappa,\delta}$ *is strictly increasing on* $]0,1/2]$ *and vanishes at* 0. *Therefore,* $G_{\kappa,\delta}$ *is also strictly decreasing on* $[1/2,1[$, *vanishes at* 1, *and has a unique maximum at* $p = 1/2$.

(c) *Criterion* (31) *can be rewritten as*

$$J_n^{\mathrm{RB}}(x) = \sigma_n(x)^\delta\, G_{\kappa,\delta}\big(p_n(x)\big). \tag{32}$$

(d) $G_{\kappa,1}$ *has the following closed-form expression*:

$$\begin{aligned} G_{\kappa,1}(p) = {} & \kappa\big(\Phi(t^+) - \Phi(t^-)\big) \\ & - t\left(2\Phi(t) - \Phi(t^+) - \Phi(t^-)\right) \\ & - \big(2\varphi(t) - \varphi(t^+) - \varphi(t^-)\big), \end{aligned} \tag{33}$$

where $t = \Phi^{-1}(1-p)$, $t^+ = t + \kappa$ *and* $t^- = t - \kappa$.

(e) $G_{\kappa,2}$ *has the following closed-form expression*:

$$\begin{aligned} G_{\kappa,2}(p) = {} & \big(\kappa^2 - 1 - t^2\big)\big(\Phi(t^+) - \Phi(t^-)\big) \\ & - 2t\big(\varphi(t^+) - \varphi(t^-)\big) \\ & + t^+\varphi(t^+) - t^-\varphi(t^-), \end{aligned} \tag{34}$$

with the same notations.

It follows from (a) and (b) that $J_n^{\mathrm{RB}}(x)$ can also be seen as a function of the kriging variance $\sigma_n^2(x)$ and the probability of misclassification $\tau_n(x) = \min(p_n(x), 1 - p_n(x))$. Note that, in the computation of $G_{\kappa,\delta}(p_n(x))$, the quantity denoted by t in (33) and (34) is equal to $(u - \widehat{\xi}_n(x))/\sigma_n(x)$, i.e., equal to the normalized distance between the predicted value and the threshold.

Bichon et al.'s *expected feasibility* function corresponds to (32) with $\delta = 1$, and can be computed efficiently using (33). Similarly, Ranjan et al.'s *expected improvement*[8] function corresponds to (32) with $\delta = 2$, and can be computed efficiently using (34). The proof of Proposition 4 is provided in Appendix B.

Remark 3 In the case $\delta = 1$, our result coincides with the expression given by Bichon et al. (2008, (17)). In the case $\delta = 2$, we have found and corrected a mistake in the computations of Ranjan et al. (2008, (8) and Appendix B).

5 Numerical experiments

5.1 A one-dimensional illustration of a SUR strategy

The objective of this section is to illustrate a SUR strategy in a simple one-dimensional case. We wish to estimate $\alpha = \mathsf{P}_{\mathbb{X}}\{f > 1\}$, where $f : \mathbb{X} = \mathbb{R} \to \mathbb{R}$ is such that $\forall x \in \mathbb{R}$,

$$\begin{aligned} f(x) = {} & (0.4x - 0.3)^2 + \exp\big(-11.534|x|^{1.95}\big) \\ & + \exp\big(-5(x - 0.8)^2\big), \end{aligned}$$

and where $\mathbb{X}$ is endowed with the probability distribution $\mathsf{P}_{\mathbb{X}} = \mathcal{N}(0, \sigma_{\mathbb{X}}^2)$, $\sigma_{\mathbb{X}} = 0.4$, as depicted in Fig. 2. We know in advance that $\alpha \approx 0.2$. Thus, a Monte Carlo sample of size $m = 1500$ will give a good estimate of α.

[8] Despite its name and some similarity between the formulas, this criterion should not be confused with the well-known EI criterion in the field of optimization (Mockus et al. 1978; Jones et al. 1998).

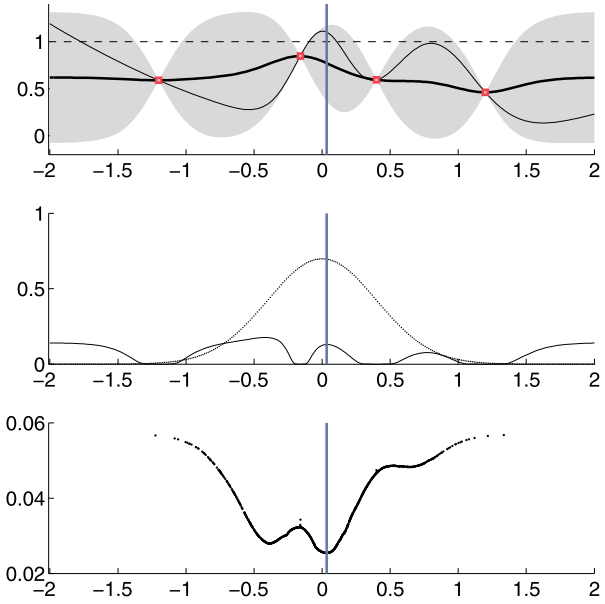


Fig. 2 Illustration of a SUR strategy. This figure shows the initial design. *Top*: threshold $u = 1$ (*horizontal dashed line*); function f (*thin line*); $n = 4$ initial evaluations (*squares*); kriging approximation f_n (*thick line*); 95% confidence intervals computed from the kriging variance (*shaded area*). *Middle*: probability of excursion (*solid line*); probability density of $\mathsf{P}_{\mathbb{X}}$ (*dotted line*). *Bottom*: graph of $J^{\mathrm{SUR}}_{1,n=4}(Y_i)$, $i = 1, \ldots, m = 1500$, the minimum of which indicates where the next evaluation of f should be done (i.e., near the origin)

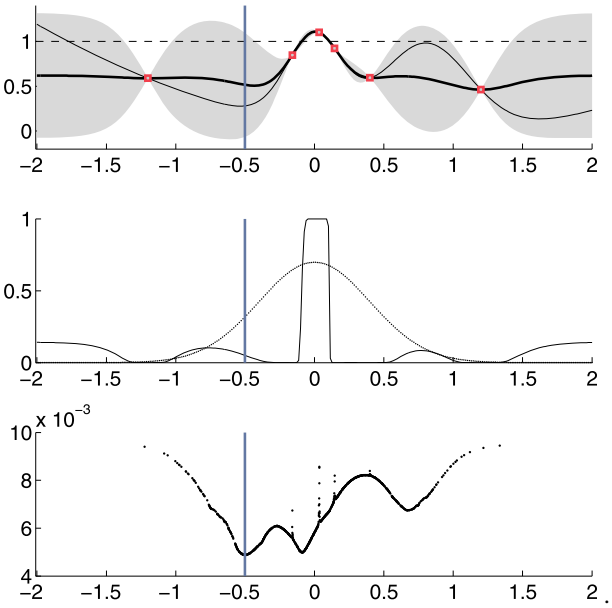


Fig. 3 Illustration of a SUR strategy (see also Figs. 2 and 4). This figure shows the progress of the SUR strategy after two iterations—a total of $n = 6$ evaluations (*squares*) have been performed. The next evaluation point will be approximately at $x = -0.5$

In this illustration, ξ is a Gaussian process with constant but unknown mean and a Matérn covariance function, whose parameters are kept *fixed*, for the sake of simplicity. Figure 2 shows an initial design of four points and the sampling criterion $J^{\mathrm{SUR}}_{1,n=4}$. Notice that the sampling criterion is only computed at the points of the Monte Carlo sample. Figures 3 and 4 show the progress of the SUR strategy after a few iterations. Observe that the probability of excursion p_n is very close to either zero or one in the region where the density of $\mathsf{P}_{\mathbb{X}}$ is high.

5.2 An example in structural reliability

In this section, we evaluate all criteria discussed in Sects. 3 and 4 through a classical benchmark example in structural reliability (see, e.g., Borri and Speranzini 1997; Waarts 2000; Schueremans 2001; Deheeger 2008). Echard et al. (2010a, 2010b) used this benchmark to make a comparison among several methods proposed in Schueremans and Gemert (2005), some of which are based on the construction of a response surface. The objective of the benchmark is to estimate the probability of failure of a so-called *four-branch*

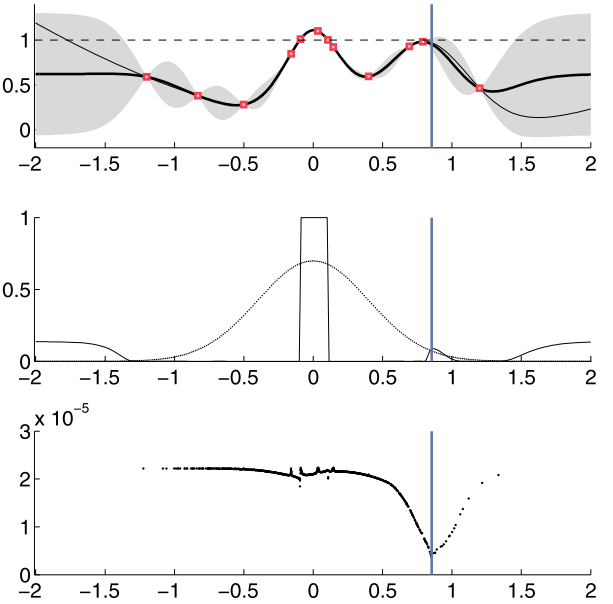


Fig. 4 Illustration of a SUR strategy (see also Figs. 2 and 3). This figure shows the progress of the SUR strategy after eight iterations—a total of $n = 12$ evaluations (*squares*) have been performed. At this stage, the probability of excursion p_n almost equals 0 or 1 in the region where the density of $\mathsf{P}_{\mathbb{X}}$ is high

series system. A failure happens when the system is working under the threshold $u = 0$. The performance function f for

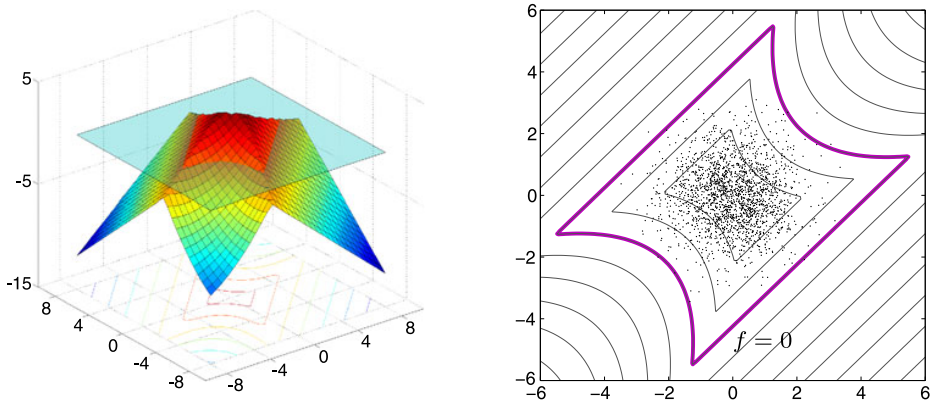


Fig. 5 *Left*: mesh plot of the performance function f corresponding to the four-branch series system; a failure happens when f is below the transparent plane; *Right*: contour plot of f; limit state $f = 0$ (*thick line*); sample of size $m = 3 \times 10^3$ from $\mathsf{P}_{\mathbb{X}}$ (*dots*)

this system is defined as

$$f : (x_1, x_2) \in \mathbb{R}^2 \mapsto$$

$$f(x_1, x_2) = \min \begin{Bmatrix} 3 + 0.1(x_1 - x_2)^2 - (x_1 + x_2)/\sqrt{2}; \\ 3 + 0.1(x_1 - x_2)^2 + (x_1 + x_2)/\sqrt{2}; \\ (x_1 - x_2) + 6/\sqrt{2}; \\ (x_2 - x_1) + 6/\sqrt{2} \end{Bmatrix}.$$

The uncertain input factors are supposed to be independent and have standard normal distribution. Figure 5 shows the performance function, the failure domain and the input distribution. Observe that f has a first-derivative discontinuity along four straight lines originating from the point $(0, 0)$.

For each sequential method, we will follow the procedure described in Table 3. We generate an initial design of $n_0 = 10$ points (five times the dimension of the factor space) using a maximin LHS (Latin Hypercube Sampling)[9] on $[-6; 6] \times [-6; 6]$. We choose a Monte Carlo sample of size $m = 30000$. Since the true probability of failure is approximately $\alpha = 0.4\%$ in this example, the coefficient of variation for α_m is $1/\sqrt{m\,\alpha} \approx 9\%$. The same initial design and Monte Carlo sample are used for all methods.

A Gaussian process with constant unknown mean and a Matérn covariance function is used as our prior information about f. The parameters of the Matérn covariance functions are estimated on the initial design by REML (see, e.g. Stein 1999). In this experiment, we follow the common practice of re-estimating the parameters of the covariance function during the sequential strategy, but only once every ten iterations to save some computation time.

The probability of failure is estimated by (13). To evaluate the rate of convergence, we compute the number n_γ of iterations that must be performed using a given strategy to observe a stabilization of the relative error of estimation within an interval of length 2γ:

$$n_\gamma = \min \left\{ n \geq 0; \forall k \geq n, \frac{|\widehat{\alpha}_{n_0+k} - \alpha_m|}{\alpha_m} < \gamma \right\}.$$

All the available sequential strategies are run 100 times, with different initial designs and Monte Carlo samples. The results for $\gamma = 0.10$, $\gamma = 0.03$ and $\gamma = 0.01$ are summarized in Table 4. We shall consider that $n_{0.1}$ provides a measure of the performance of the strategy in the "initial phase", where a rough estimate of α is to be found, whereas $n_{0.03}$ and $n_{0.01}$ measure the performance in the "refinement phase".

The four variants of the SUR strategy (see Table 1) have been run with $Q = 12$ and either $m_0 = 10$ or $m_0 = 500$. The performance are similar for all four variants and for both values of m_0. It appears, however, that the criterions J_1^{SUR} and $J_2^{\mathrm{SUR}}2$ (i.e., the criterions given directly by Proposition 3) are slightly better than J_3^{SUR} and J_4^{SUR}; this will be confirmed by the simulations of Sect. 5.3. It also seems that the SUR algorithm is slightly slower to obtain a rough estimate of the probability of failure when m_0 is very small, but performs very well in the refinement phase. (Note that $m_0 = 10$ is a drastic pruning for a sample of size $m = 30000$.)

The tIMSE strategy has been run for three different values of its tuning parameter σ_ε^2, using the pruning scheme with $m_0 = 500$. The best performance is obtained for $\sigma_\varepsilon^2 \approx 0$, and is almost as good as the performance of SUR strategies with the same value of m_0 (a small loss of performance, of about one evaluation on average, can be noticed in the refinement phase). Note that the required accuracy was not reached after 200 iterations in 17% of the runs for $\sigma_\varepsilon^2 = 1$. In fact, the tIMSE strategy tends to behave like a space-filling strategy in this case. Figure 6 shows the points that have been evaluated in three cases: the evaluations are less concentrated on the boundary between the safe and the failure region when $\sigma_\varepsilon^2 = 1$.

Finally, the results obtained for J^{RB} and J^{EGL} indicate that the corresponding strategies are clearly less efficient in

[9] More precisely, we use Matlab's `lhsdesign()` function to select the best design according to the maximin criterion among 10^4 randomly generated LHS designs.

Table 4 Comparison of the convergence to α_m in the benchmark example Sect. 5.2 for different sampling strategies. The first number (bold text) is the average value of n_γ over 100 runs. The numbers between brackets indicate the 10th and 90th percentile

Criterion	Parameters	$\gamma = 0.10$	$\gamma = 0.03$	$\gamma = 0.01$
J_1^{SUR}	$m_0 = 500$	**16.1** [10–22]	**25.7** [17–35]	**36.0** [26–48]
	$m_0 = 10$	**19.4** [11–28]	**28.1** [19–38]	**35.4** [26–44]
J_2^{SUR}	$m_0 = 500$	**16.4** [10–24]	**25.7** [19–33]	**35.5** [25–45]
	$m_0 = 10$	**20.0** [11–30]	**28.3** [20–39]	**35.3** [26–44]
J_3^{SUR}	$m_0 = 500$	**18.2** [10–27]	**26.9** [18–37]	**35.9** [27–46]
	$m_0 = 10$	**20.1** [11–30]	**28.0** [20–36]	**35.2** [25–44]
J_4^{SUR}	$m_0 = 500$	**17.2** [10–28]	**26.5** [20–36]	**35.2** [25–45]
	$m_0 = 10$	**21.4** [13–30]	**28.9** [20–38]	**35.5** [27–44]
J^{tIMSE}	$\sigma_\varepsilon^2 = 10^{-6}$	**16.6** [10–23]	**26.5** [19–36]	**37.3** [28–49]
	$\sigma_\varepsilon^2 = 0.1$	**15.9** [10–22]	**29.1** [19–43]	**50.5** [30–79]
	$\sigma_\varepsilon^2 = 1$	**21.7** [11–31]	**52.4** [31–85]	**79.5** [42–133][a]
J^{EGL}	–	**21.0** [11–31]	**29.2** [21–39]	**36.4** [28–44]
J^{RB}	$\delta = 1, \kappa = 0.5$	**18.7** [10–27]	**27.5** [20–35]	**36.6** [27–44]
	$\delta = 1, \kappa = 2.0$	**18.9** [11–28]	**28.3** [21–35]	**37.7** [30–45]
	$\delta = 2, \kappa = 0.5$	**17.6** [10–24]	**27.6** [20–34]	**37.1** [29–45]
	$\delta = 2, \kappa = 2.0$	**17.0** [10–21]	**27.1** [20–34]	**36.8** [29–44]

[a]The required accuracy was not reached after 200 iterations in 17% of the runs

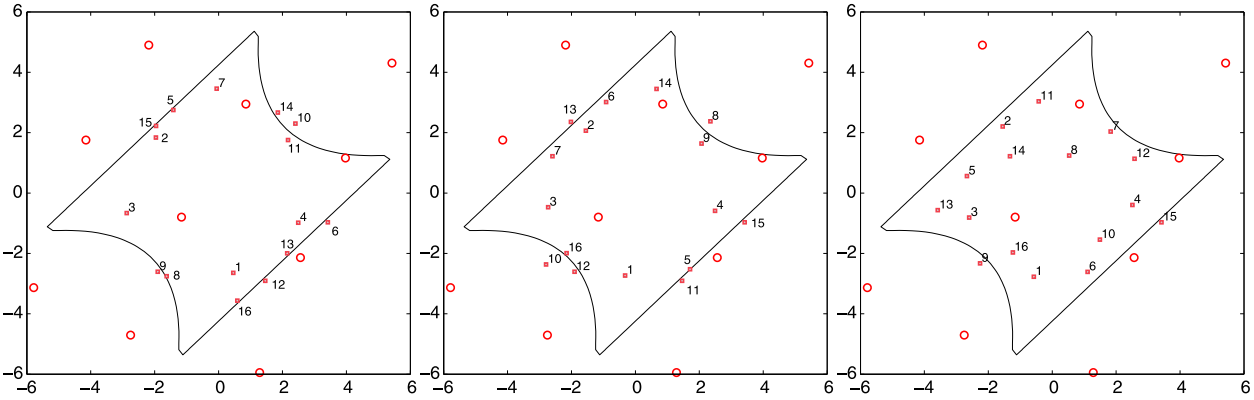


Fig. 6 The first 16 points (*squares*) evaluated using sampling criterion J_1^{SUR} (*left*), J^{tIMSE} with $\sigma_\varepsilon^2 = 0.1$ (*middle*), J^{tIMSE} with $\sigma_\varepsilon^2 = 1$ (*right*). Numbers near squares indicate the order of evaluation. The location of the $n_0 = 10$ points of the initial design are indicated by *circles*

the "initial phase" than strategies based on J_1^{SUR} or J_2^{SUR}. For $\gamma = 0.1$, the average loss with respect to J_1^{SUR} is between approximately 0.9 evaluations for the best case (criterion J^{RB} with $\delta = 2$, $\kappa = 2$) and 3.9 evaluations for the worst case. For $\gamma = 0.03$, the loss is between 1.4 evaluations (also for (criterion J^{RB} with $\delta = 2$, $\kappa = 2$) and 3.5 evaluations. This loss of efficiency can also be observed very clearly on the 90th percentile in the initial phase. Criterion J^{RB} seems to perform best with $\delta = 2$ and $\kappa = 2$ in this experiment, but this will not be confirmed by the simulations of Sect. 5.3. Tuning the parameters of this criterion for the estimation of a probability of failure does not seem to be an easy task.

5.3 Average performance on sample paths of a Gaussian process

This section provides a comparison of all the criteria introduced or recalled in this paper, on the basis of their average performance on the sample paths of a zero-mean Gaussian process defined on $\mathbb{X} = [0, 1]^d$, for $d \in \{1, 2, 3\}$. In all experiments, the same covariance function is used for the gen-

Table 5 Size of the initial design and covariance parameters for the experiments of Sect. 5.3. The parametrization of the Matérn covariance function used here is defined in Appendix A

d	n_0	σ^2	ν	ρ
1	3	1.0	2.0	0.100
2	10	1.0	2.0	0.252
3	15	1.0	2.0	0.363

eration of the sample paths and for the computation of the sampling criteria. We have considered isotropic Matérn covariance functions, whose parameters are given in Table 5. An initial maximin LHS design of size n_0 (also given in the table) is used: note that the value of n reported on the x-axis of Figs. 7–11 is the total number of evaluations, including the initial design.

The d input variables are assumed to be independent and uniformly distributed on $[0, 1]$, i.e., $\mathsf{P}_{\mathbb{X}}$ is the uniform distribution on $\mathbb{X}$. An m-sample $Y_1, \ldots, Y_m$ from $\mathsf{P}_{\mathbb{X}}$ is drawn one and for all, and used both for the approximation of integrals (in SUR and tIMSE criteria) and for the discrete search of the next sampling point (for all criteria). We take $m = 500$ and use the same MC sample for all criteria in a given dimension d.

We adopt the meta-estimation framework as described in Sect. 3.3; in other words, our goal is to estimate the MC estimator α_m. We choose to adjust the threshold u in order to have $\alpha_m = 0.02$ for all sample paths (note that, as a consequence, there are exactly $m\alpha_m = 10$ points in the failure region) and we measure the performance of a strategy after n evaluations by its relative mean-square error (MSE) expressed in decibels (dB):

$$\mathrm{rMSE} := 10 \log_{10} \left(\frac{1}{L} \sum_{l=1}^{L} \frac{(\widehat{\alpha}_{m,n}^{(l)} - \alpha_m)^2}{\alpha_m^2} \right),$$

where $\widehat{\alpha}_{m,n}^{(l)} = \frac{1}{m} \sum_{j=1}^{m} p_n^{(l)}(Y_j)$ is the posterior mean of the MC estimator α_m after n evaluations on the lth simulated sample path ($L = 4000$).

We use a sequential maximin strategy as a reference in all of our experiments. This simple space-filling strategy is defined by $X_{n+1} = \operatorname{argmax}_j \min_{1 \le i \le n} |Y_j - X_i|$, where the argmax runs over all indices j such that $Y_j \notin \{X_1, \ldots, X_n\}$. Note that this strategy does not depend on the choice of a Gaussian process model.

Our first experiment (Fig. 7) provides a comparison of the four SUR strategies proposed in Sect. 3.2. It appears that all of them perform roughly the same when compared to the reference strategy. A closer look, however, reveals that the strategies J_1^{SUR} and J_2^{SUR} provided by Proposition 3 perform slightly better than the other two (noticeably so in the case $d = 3$).

The performance of the tIMSE strategy is shown on Fig. 8 for several values of its tuning parameter σ_ε^2 (other values, not shown here, have been tried as well). It is clear that the performance of this strategy improves when σ_ε^2 goes to zero, whatever the dimension.

The performance of the strategy based on $J_{\kappa,\delta}^{\mathrm{RB}}$ is shown on Fig. 9 for several values of its parameters. It appears that the criterion proposed by Bichon et al. (2008), which corresponds to $\delta = 1$, performs better than the one proposed by Ranjan et al. (2008), which corresponds to $\delta = 2$, for the same value of κ. Moreover, the value $\kappa = 0.5$ has been found in our experiments to produce the best results.

Figure 10 illustrates that the loss of performance associated to the "pruning trick" introduced in Sect. 3.4 can be negligible if the size m_0 of the pruned MC sample is large enough (here, m_0 has been taken equal to 50). In practice, the value of m_0 should be chosen small enough to keep the overhead of the sequential strategy reasonable—in other words, large values of m_0 should only be used for very complex computer codes.

Finally, a comparison involving the best strategy obtained in each category is presented on Fig. 11. The best result is consistently obtained with the SUR strategy based on $J_{1,n}^{\mathrm{SUR}}$. The tIMSE strategy with $\sigma_\varepsilon^2 \approx 0$ provides results which are almost as good. Note that both strategies are one-step lookahead strategies based on the approximation of the risk by an integral criterion, which makes them rather expensive to compute. Simpler strategies based on the marginal distribution (criteria J_n^{RB} and J_n^{EGL}) provide interesting alternatives for moderately expensive computer codes: their performances, although not as good as those of one-step lookahead criterions, are still much better than that of the reference space-filling strategy.

6 Concluding remarks

One of the main objectives of this paper was to present a synthetic viewpoint on sequential strategies based on a Gaussian process model and kriging for the estimation of a probability of failure. The starting point of this presentation is a Bayesian decision-theoretic framework from which the theoretical form of an optimal strategy for the estimation of a probability of failure can be derived. Unfortunately, the dynamic programming problem corresponding to this strategy is not numerically tractable. It is nonetheless possible to derive from there the ingredients of a sub-optimal strategy: the idea is to focus on one-step lookahead suboptimal strategies, where the exact risk is replaced by a substitute risk that accounts for the information gain about α expected from a new evaluation. We call such a strategy a *stepwise uncertainty reduction* (SUR) strategy. Our numerical experiments show that SUR strategies perform better, on average, than

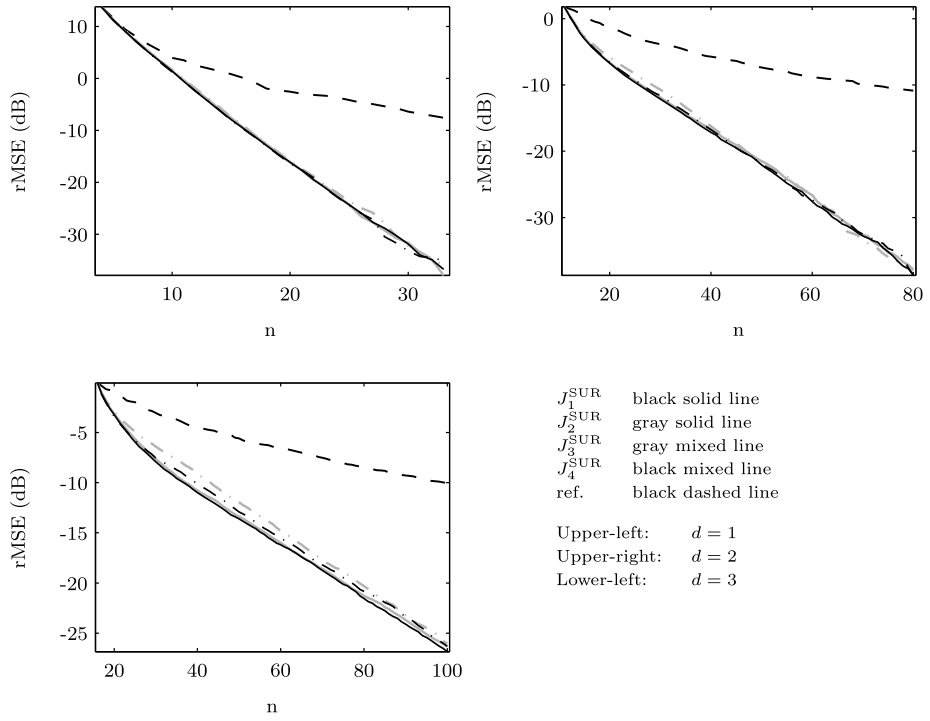


Fig. 7 Relative MSE performance of several SUR strategies

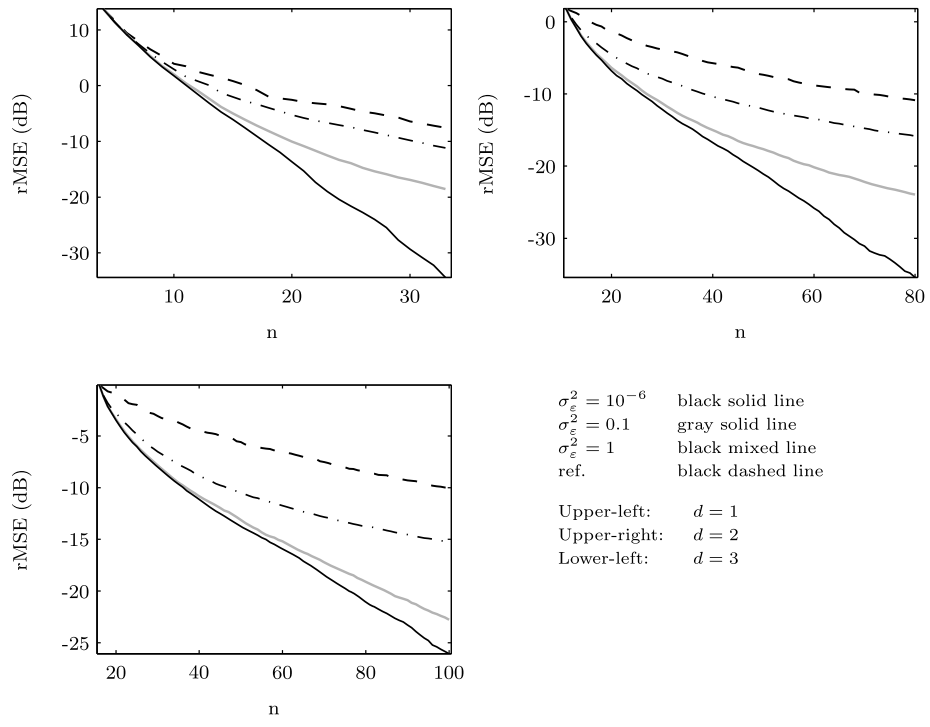


Fig. 8 Relative MSE performance of the tIMSE strategy for several values of its parameter

the other strategies proposed in the literature. However, this comes at a higher computational cost than strategies based only on marginal distributions. The tIMSE sampling criterion, which seems to have a convergence rate comparable to that of the SUR criterions when $\sigma_\varepsilon^2 \approx 0$, also has a high computational complexity.

In which situations can we say that the sequential strategies presented in this paper are interesting alternatives

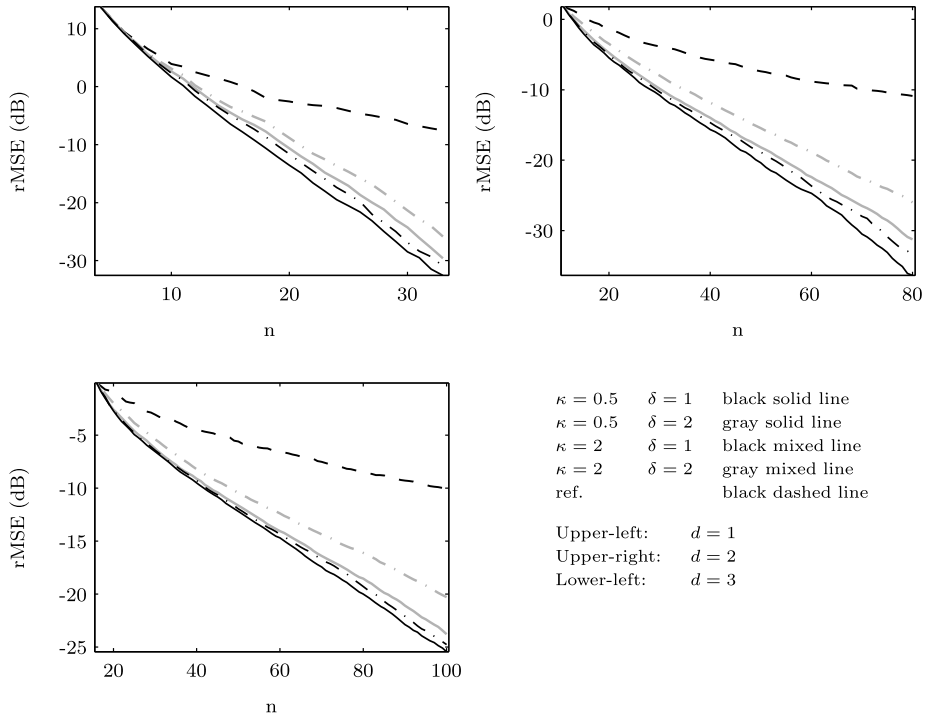


Fig. 9 Relative MSE performance of the J^{RB} criterion, for several values of its parameters

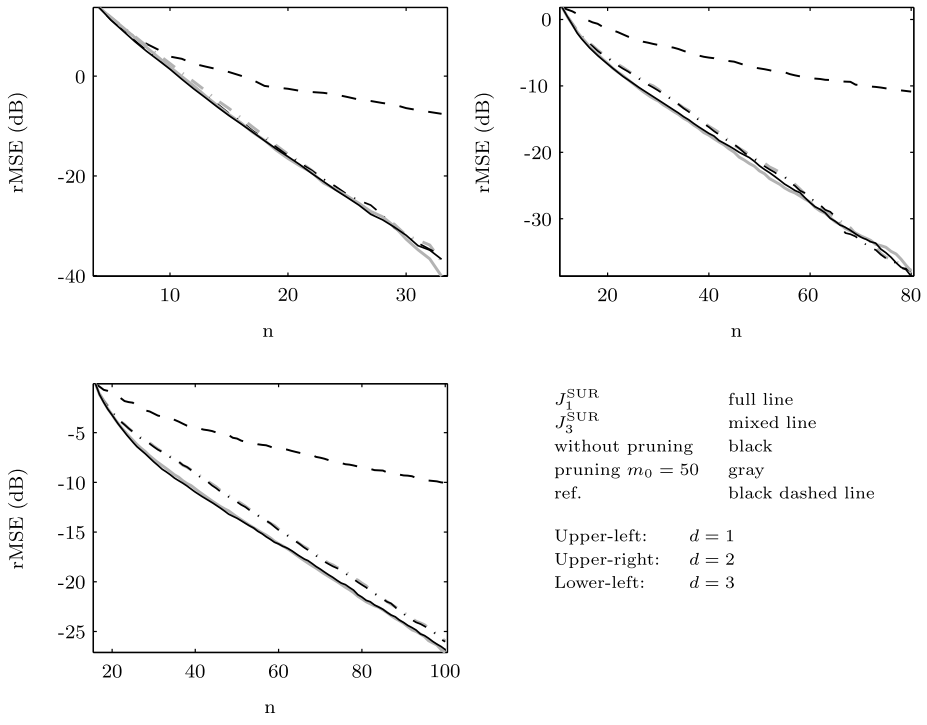


Fig. 10 Relative MSE performance of two SUR criteria, with and without the "pruning trick" described in Sect. 3.4. The *black* and *gray lines* are almost surimposed for each of the criterions J_1^{SUR} and J_3^{SUR}

to classical importance sampling methods for estimating a probability of failure, for instance the subset sampling method of Au and Beck (2001)? In our opinion, beyond the obvious role of the simulation budget N, the answer to this question depends on our capacity to elicit an appropriate prior. In the example of Sect. 5.2, as well as in many other examples using Gaussian processes in the domain of computer experiments, the prior is easy to choose

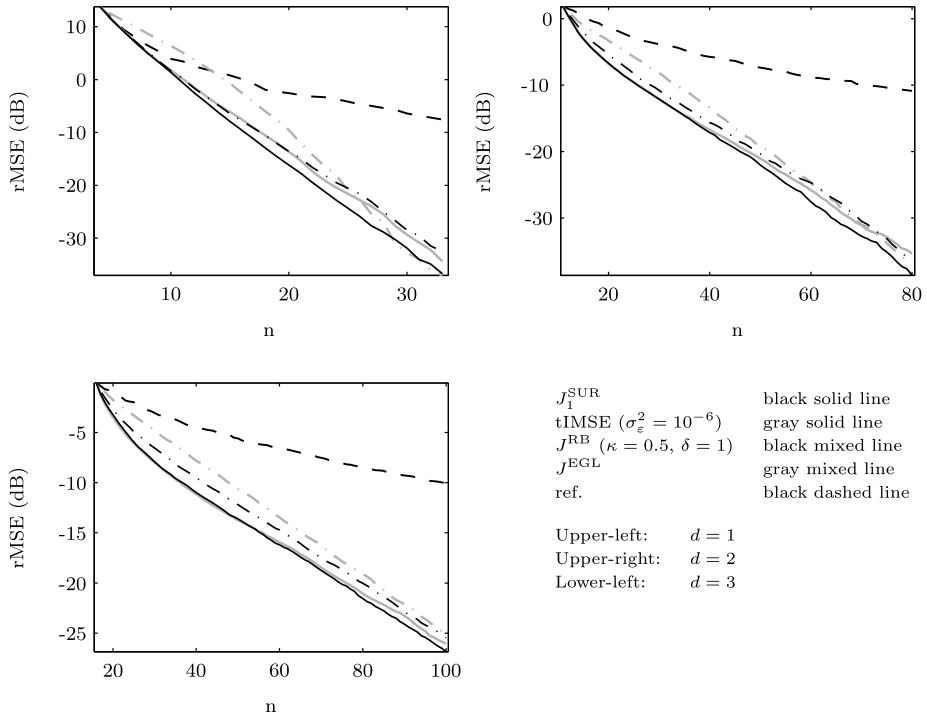


Fig. 11 Relative MSE performance the best strategy in each category

because $\mathbb{X}$ is a low-dimensional space and f tends to be smooth. Then, the plug-in approach which consists of using ML or REML to estimate the parameters of the covariance function of the Gaussian process after each new evaluation is likely to succeed. If $\mathbb{X}$ is high-dimensional and f is expensive to evaluate, difficulties arise. In particular, our sampling strategies do not take into account our uncertain knowledge of the covariance parameters, and there is no guarantee that ML estimation will do well when the points are chosen by a sampling strategy that favors some localized target region (the neighborhood the frontier of the domain of failure in this paper, but the question is equally relevant in the field of optimization, for instance). The difficult problem of deciding the size n_0 of the initial design is crucial in this connection. Fully Bayes procedures constitute a possible direction for future research, as long as they don't introduce an unacceptable computational overhead. Whatever the route, we feel that the robustness of Gaussian process-based sampling strategies with respect to the procedure of estimation of the covariance parameters should be addressed carefully in order to make these methods usable in the industrial world.

Software We would like to draw the reader's attention on the recently published package KrigInv (Picheny and Ginsbourger 2011) for the statistical computing environment R (see Hornik 2010). This package provides an open source (GPLv3) implementation of all the strategies proposed in this paper. Please note that the simulation results presented in this paper were not obtained using this package, that was not available at the time of its writing.

Acknowledgements The research of Julien Bect, Ling Li and Emmanuel Vazquez was partially funded by the French *Agence Nationale de la Recherche* (ANR) in the context of the project OPUS (ref. ANR-07-CIS7-010) and by the French *pôle de compétitivité* SYSTEMATIC in the context of the project CSDL. David Ginsbourger acknowledges support from the French *Institut de Radioprotection et de Sûreté Nucléaire* (IRSN) and warmly thanks Yann Richet.

Appendix A: The Matérn covariance

The exponential covariance and the Matérn covariance are among the most conventionally used stationary covariances of design and analysis of computer experiments. The Matérn covariance class (Yaglom 1986) offers the possibility to adjust the regularity of ξ with a single parameter. Stein (1999) advocates the use of the following parametrization of the Matérn function:

$$\kappa_\nu(h) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(2\nu^{1/2}h\right)^\nu \mathcal{K}_\nu\left(2\nu^{1/2}h\right), \quad h \in \mathbb{R} \quad (35)$$

where Γ is the Gamma function and $\mathcal{K}_\nu$ is the modified Bessel function of the second kind. The parameter $\nu > 0$ controls regularity at the origin of the function. To model

a real-valued function f defined over $\mathbb{X} \subset \mathbb{R}^d$, with $d \geq 1$, we use the following anisotropic form of the Matérn covariance:

$$k_\theta(x, y) = \sigma^2 \kappa_\nu \left(\sqrt{\sum_{i=1}^{d} \frac{(x_{[i]} - y_{[i]})^2}{\rho_i^2}} \right), \quad x, y \in \mathbb{R}^d \quad (36)$$

where $x_{[i]}$, $y_{[i]}$ denote the ith coordinate of x and y, the positive scalar σ^2 is the variance parameter (we have $k_\theta(x, x) = \sigma^2$), and the positive scalars ρ_i are scale or *range* parameters of the covariance, i.e., characteristic correlation lengths. Since $\sigma^2 > 0, \nu > 0, \rho_i > 0$, $i = 1, \ldots, d$, we can take the logarithm of these scalars, and consider the vector of parameters $\theta = \{\log \sigma^2, \log \nu, -\log \rho_1, \ldots, -\log \rho_d\} \in \mathbb{R}^{d+2}$, which is a practical parameterization when σ^2, ν, ρ_i, $i = 1, \ldots, d$, need to be estimated from data.

Appendix B: Proof of Proposition 4

(a) Using the identity $\Phi^{-1}(1 - p) = -\Phi^{-1}(p)$, we get

$$\left|U + \Phi^{-1}(1 - p)\right| = \left|U - \Phi^{-1}(p)\right| \stackrel{\mathrm{d}}{=} \left|U + \Phi^{-1}(p)\right|,$$

where $\stackrel{\mathrm{d}}{=}$ denotes an equality in distribution. Therefore $G_{\kappa,\delta}(1 - p) = G_{\kappa,\delta}(p)$.

(b) Let $S_p = \max(0, \kappa^\delta - |\Phi^{-1}(p) + U|)$. Straightforward computations show that $t \mapsto \mathsf{P}(|t + U| \leq v)$ is strictly decreasing to 0 on $[0, +\infty[$, for all $v > 0$. As a consequence, $p \mapsto \mathsf{P}(S_p < s)$ is strictly increasing to 1 on $[1/2, 1[$, for all $s \in]0, \kappa^\delta[$. Therefore, $G_{\kappa,\delta}$ is strictly decreasing on $[1/2, 1[$ and tends to zeros when $p \to 1$. The other assertions then follow from a).

(c) Recall that $\xi(x) \sim \mathcal{N}(\widehat{\xi}_n(x), \sigma_n^2(x))$ under P_n. Therefore $U := (\xi(x) - \widehat{\xi}_n(x))/\sigma_n(x) \sim \mathcal{N}(0, 1)$ under P_n, and the result follows by substitution in (31).

The closed-form expression of Ranjan et al.'s and Bichon and al.'s criteria (assertions (d) and (e)) are established in the following sections.

B.1 A preliminary decomposition common to both criteria

Recall that $t = \Phi^{-1}(1 - p)$, $t^+ = t + \kappa$ and $t^- = t - \kappa$. Then,

$$\begin{aligned} G_{\kappa,\delta}(p) = G_{\kappa,\delta}(1 - p) &= \mathsf{E}\left(\max\left(0, \kappa^\delta - \left|t - U\right|^\delta\right)\right) \\ &= \int_{\kappa^\delta - |t-u|^\delta \geq 0} \left(\kappa^\delta - |t - u|^\delta\right) \varphi(u)\, \mathrm{d}u \\ &= \int_{t^-}^{t^+} \left(\kappa^\delta - |t - u|^\delta\right) \varphi(u)\, \mathrm{d}u \\ &= \kappa^\delta \left(\Phi(t^+) - \Phi(t^-)\right) - \underbrace{\int_{t^-}^{t^+} |t - u|^\delta \varphi(u)\, \mathrm{d}u}_{\text{Term } A}. \end{aligned} \quad (37)$$

The computation of the integral A will be carried separately in the next two sections for $\delta = 1$ and $\delta = 2$. For this purpose, we shall need the following elementary results:

$$\int_a^b u\varphi(u)\, \mathrm{d}u = \varphi(a) - \varphi(b), \quad (38)$$

$$\int_a^b u^2 \varphi(u)\, \mathrm{d}u = a\varphi(a) - b\varphi(b) + \Phi(b) - \Phi(a). \quad (39)$$

B.2 Case $\delta = 1$

Let us compute the value A_1 of the integral A for $\delta = 1$:

$$\begin{aligned} A_1 &= \int_{t^-}^{t^+} |t - u|\, \varphi(u)\, \mathrm{d}u \\ &= \int_{t^-}^{t} (t - u)\varphi(u)\, \mathrm{d}u + \int_{t}^{t^+} (u - t)\varphi(u)\, \mathrm{d}u \\ &= t\left(\int_{t^-}^{t} \varphi(u)\, \mathrm{d}u - \int_{t}^{t^+} \varphi(u)\, \mathrm{d}u\right) \\ &\quad - \int_{t^-}^{t} u\varphi(u)\, \mathrm{d}u + \int_{t}^{t^+} u\varphi(u)\, \mathrm{d}u \\ &= t\left(2\Phi(t) - \Phi(t^-) - \Phi(t^+)\right) \\ &\quad + 2\varphi(t) - \varphi(t^-) - \varphi(t^+), \end{aligned} \quad (40)$$

where (38) has been used to get the final result. Plugging (40) into (37) yields (33).

B.3 Case $\delta = 2$

Let us compute the value A_2 of the integral A for $\delta = 2$:

$$\begin{aligned} A_2 &= \int_{t^-}^{t^+} (t - u)^2 \varphi(u)\, \mathrm{d}u \\ &= t^2 \int_{t^-}^{t^+} \varphi(u)\, \mathrm{d}u - 2t \int_{t^-}^{t^+} u\varphi(u)\, \mathrm{d}u + \int_{t^-}^{t^+} u^2 \varphi(u)\, \mathrm{d}u \\ &= t^2 \left(\Phi(t^+) - \Phi(t^-)\right) - 2t\left(\varphi(t^-) - \varphi(t^+)\right) \\ &\quad + t^-\varphi(t^-) - t^+\varphi(t^+) + \Phi(t^+) - \Phi(t^-), \end{aligned} \quad (41)$$

where (38) and (39) have been used to get the final result. Plugging (41) into (37) yields (34).

References

Arnaud, A., Bect, J., Couplet, M., Pasanisi, A., Vazquez, E.: Évaluation d'un risque d'inondation fluviale par planification séquentielle d'expériences. In: 42èmes Journées de Statistique (2010)

Au, S.K., Beck, J.: Estimation of small failure probabilities in high dimensions by subset simulation. Probab. Eng. Mech. **16**(4), 263–277 (2001)

Bayarri, M.J., Berger, J.O., Paulo, R., Sacks, J., Cafeo, J.A., Cavendish, J., Lin, C.H., Tu, J.: A framework for validation of computer models. Technometrics **49**(2), 138–154 (2007)

Berry, D.A., Fristedt, B.: Bandit Problems: Sequential Allocation of Experiments. Chapman & Hall, London (1985)

Bertsekas, D.P.: Dynamic Programming and Optimal Control, vol. 1. Athena Scientific, Nashua (1995)

Bichon, B.J., Eldred, M.S., Swiler, L.P., Mahadevan, S., McFarland, J.M.: Efficient global reliability analysis for nonlinear implicit performance functions. AIAA J. **46**(10), 2459–2468 (2008)

Bjerager, P.: On computational methods for structural reliability analysis. Struct. Saf. **9**, 76–96 (1990)

Borri, A., Speranzini, E.: Structural reliability analysis using a standard deterministic finite element code. Struct. Saf. **19**(4), 361–382 (1997)

Box, G.E.P., Draper, N.R.: Empirical Model-Building and Response Surfaces. Wiley, New York (1987)

Bucher, C.G., Bourgund, U.: A fast and efficient response surface approach for structural reliability problems. Struct. Saf. **7**(1), 57–66 (1990)

Chilès, J.P., Delfiner, P.: Geostatistics: Modeling Spatial Uncertainty. Wiley, New York (1999)

Currin, C., Mitchell, T., Morris, M., Ylvisaker, D.: Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. J. Am. Stat. Assoc. **86**(416), 953–963 (1991)

Deheeger, F.: Couplage mécano-fiabiliste : ^{2}SMART—méthodologie d'apprentissage stochastique en fiabilité. Ph.D. thesis, Université Blaise Pascal, Clermont II (2008)

Deheeger, F., Lemaire, M.: Support vector machine for efficient subset simulations: ^{2}SMART method. In: Kanda, J., Takada, T., Furuta, H. (eds.) 10th International Conference on Application of Statistics and Probability in Civil Engineering, Proceedings and Monographs in Engineering, Water and Earth Sciences, pp. 259–260. Taylor & Francis, London (2007)

Echard, B., Gayton, N., Lemaire, M.: Kriging-based Monte Carlo simulation to compute the probability of failure efficiently: AK-MCS method. In: 6èmes Journées Nationales de Fiabilité, Toulouse, France, 24–26 mars 2010a

Echard, B., Gayton, N., Lemaire, M.: Structural reliability assessment using kriging metamodel and active learning. In: IFIP WG 7.5 Working Conference on Reliability and Optimization of Structural Systems (2010b)

Fleuret, F., Geman, D.: Graded learning for object detection. In: Proceedings of the Workshop on Statistical and Computational Theories of Vision of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR/SCTV) (1999)

Frazier, P.I., Powell, W.B., Dayanik, S.: A knowledge-gradient policy for sequential information collection. SIAM J. Control Optim. **47**(5), 2410–2439 (2008)

Ginsbourger, D.: Métamodèles multiples pour l'approximation et l'optimisation de fonctions numériques multivariables. Ph.D. thesis, Ecole nationale supérieure des Mines de Saint-Etienne (2009)

Ginsbourger, D., Le Riche, R., Carraro, L.: Kriging is well-suited to parallelize optimization. In: Hiot, L.M., Ong, Y.S., Tenne, Y., Goh, C.K. (eds.) Computational Intelligence in Expensive Optimization Problems. Adaptation Learning and Optimization, vol. 2, pp. 131–162. Springer, Berlin (2010)

Handcock, M.S., Stein, M.L.: A Bayesian analysis of kriging. Technometrics **35**(4), 403–410 (1993)

Hornik, K.: The R FAQ (2010). http://CRAN.R-project.org/doc/FAQ/R-FAQ.html. ISBN 3-900051-08-9

Hurtado, J.E.: An examination of methods for approximating implicit limit state functions from the viewpoint of statistical learning theory. Struct. Saf. **26**(3), 271–293 (2004)

Hurtado, J.E.: Filtered importance sampling with support vector margin: a powerful method for structural reliability analysis. Struct. Saf. **29**(1), 2–15 (2007)

Jones, D.R., Schonlau, M., William, J.: Efficient global optimization of expensive black-box functions. J. Glob. Optim. **13**(4), 455–492 (1998)

Kennedy, M., O'Hagan, A.: Bayesian calibration of computer models. J. R. Stat. Soc., Ser. B, Stat. Methodol. **63**(3), 425–464 (2001)

Kimeldorf, G.S., Wahba, G.: A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann. Math. Stat. **41**(2), 495–502 (1970)

Kushner, H.J.: 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)

Loeppky, J.L., Sacks, J., Welch, W.J.: Choosing the sample size of a computer experiment: a practical guide. Technometrics **51**(4), 366–376 (2009)

Mockus, J.: Bayesian Approach to Global Optimization. Theory and Applications. Kluwer Academic, Dordrecht (1989)

Mockus, J., Tiesis, V., Zilinskas, A.: The application of Bayesian methods for seeking the extremum. In: Dixon, L., Szego, E.G. (eds.) Towards Global Optimization, vol. 2, pp. 117–129. Elsevier, Amsterdam (1978)

Oakley, J.: Estimating percentiles of uncertain computer code outputs. J. R. Stat. Soc., Ser. C **53**(1), 83–93 (2004)

Oakley, J., O'Hagan, A.: Bayesian inference for the uncertainty distribution of computer model outputs. Biometrika **89**(4) (2002)

Oakley, J., O'Hagan, A.: Probabilistic sensitivity analysis of complex models: a Bayesian approach. J. R. Stat. Soc., Ser. B, Stat. Methodol. **66**(3), 751–769 (2004)

O'Hagan, A.: Curve fitting and optimal design for prediction. J. R. Stat. Soc., Ser. B, Methodol. **40**(1), 1–42 (1978)

Parzen, E.: An approach to time series analysis. Ann. Math. Stat. **32**, 951–989 (1962)

Paulo, R.: Default priors for Gaussian processes. Ann. Stat. **33**(2), 556–582 (2005)

Picheny, V., Ginsbourger, D.: KrigInv: Kriging-based inversion for deterministic and noisy computer experiments, version 1.1 (2011). http://cran.r-project.org/web/packages/KrigInv

Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R.T., Kim, N.H.: Adaptive designs of experiments for accurate approximation of target regions. J. Mech. Des. **132**(7), 1–9 (2010)

Piera-Martinez, M.: Modélisation des comportements extrêmes en ingénierie. Ph.D. thesis, Université Paris Sud, Paris XI (2008)

Pradlwarter, H., Schuëller, G., Koutsourelakis, P., Charmpis, D.: Application of line sampling simulation method to reliability benchmark problems. Struct. Saf. **29**(3), 208–221 (2007). A Benchmark Study on Reliability in High Dimensions

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C. The Art of Scientific Computing, 2nd edn. Cambridge University Press, Cambridge (1992)

Rajashekhar, M.R., Ellingwood, B.R.: A new look at the response surface approach for reliability analysis. Struct. Saf. **12**(3), 205–220 (1993)

Ranjan, P., Bingham, D., Michailidis, G.: Sequential experiment design for contour estimation from complex computer codes. Technometrics **50**(4), 527–541 (2008)

Rubinstein, R., Kroese, D.: The Cross-Entropy Method. Springer, Berlin (2004)

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P.: Design and analysis of computer experiments. Stat. Sci. **4**(4), 409–435 (1989)

Santner, T.J., Williams, B.J., Notz, W.: The Design and Analysis of Computer Experiments. Springer, Berlin (2003)

Schueremans, L.: Probabilistic evaluation of structural unreinforced masonry. Ph.D. thesis, Catholic University of Leuven (2001)

Schueremans, L., Gemert, D.V.: Benefit of splines and neural networks in simulation based structural reliability analysis. Struct. Saf. **27**(3), 246–261 (2005)

Stein, M.L.: Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York (1999)

Vazquez, E., Bect, J.: A sequential Bayesian algorithm to estimate a probability of failure. In: Proceedings of the 15th IFAC Symposium on System Identification, SYSID 2009, Saint-Malo, France (2009)

Vazquez, E., Piera-Martinez, M.: Estimation du volume des ensembles d'excursion d'un processus gaussien par krigeage intrinsèque. In: 39ème Journées de Statistiques Conférence Journée de Statistiques, Angers, France (2007)

Vestrup, E.M.: The Theory of Measures and Integration. Wiley, New York (2003)

Villemonteix, J.: Optimisation de fonctions coûteuses. Ph.D. thesis, Université Paris-Sud XI, Faculté des Sciences d'Orsay (2008)

Villemonteix, J., Vazquez, E., Walter, E.: An informational approach to the global optimization of expensive-to-evaluate functions. J. Glob. Optim. **44**(4), 509–534 (2009)

Waarts, P.H.: Structural reliability using finite element methods: an appraisal of DARS. Ph.D. thesis, Delft University of Technology (2000)

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J., Morris, M.D.: Screening, predicting and computer experiments. Technometrics **34**, 15–25 (1992)

Yaglom, A.M.: Correlation Theory of Stationary and Related Random Functions I: Basic Results. Springer Series in Statistics. Springer, New York (1986)

2.4 Estimation of the volume of an excursion set of a Gaussian process using intrinsic kriging

This article is where the notion of SUR strategy appears first in my work. This a technical report on arXiv.org written in 2006, which makes an analysis of the convergence of a kriging-based estimator of the volume of an excursion set.

Estimation of the volume of an excursion set of a Gaussian process using intrinsic Kriging

Emmanuel Vazquez[a]

E-mail address: emmanuel.vazquez@supelec.fr

Miguel Piera Martínez[a,b]

a: Département Signaux et Systèmes Électroniques, Supélec, 91192 Gif-sur-Yvette, France

b: Laboratoire des Signaux et Systèmes, CNRS, Supélec, Université Paris-Sud, 91192 Gif-sur-Yvette, France

November 9, 2006

Abstract — Assume that a Gaussian process ξ is predicted from n pointwise observations by intrinsic Kriging and that the volume of the excursion set of ξ above a given threshold u is approximated by the volume of the predictor. The first part of this paper gives a bound on the convergence rate of the approximated volume. The second part describes an algorithm that constructs a sequence of points to yield a fast convergence of the approximation. The estimation of the volume of an excursion set is a highly relevant problem for the industrial world since it corresponds to the estimation of the failure probability of a system that is known only through sampled observations.

Keywords – Excursion set; Gaussian process; Intrinsic Kriging; Quantile estimation; Failure probability; Design of experiments

1 Introduction

The problem to be considered in this paper is the estimation of the probability

$$\mathcal{P}_u := \mathsf{P}\{f(X) \geq u\}, \tag{1}$$

where $f(x)$ is a real function defined over an arbitrary set $\mathbb{X}$ ($\mathbb{X} = [0,1]^d$ or $\mathbb{X} = \mathbb{R}^d$, in most situations) endowed with a probability measure μ and $X \in \mathbb{X}$ is a random vector with the distribution μ. In practice, the estimation of (1) is based on a finite sequence of evaluations of f at points $(x_i)_{1 \leq i \leq n}$ in $\mathbb{X}$. Another way of looking at (1) is via the excursion set

$$A_u(f) := \{x \in \mathbb{X} : f(x) \geq u\} \tag{2}$$

of the function f above the level u, since $\mathsf{P}\{f(X) \geq u\}$ is the volume $\mu(A_u(f))$, hereafter denoted by $|A_u(f)|$.

Such a problem is frequently encountered in engineering: the probability that the inputs of the system will generate a level of a function of the outputs that exceeds a specified reference level may be expressed as (1) (where in this case, X is the vector of the inputs of the system and f is a statistic of the outputs). Since to obtain the value of f at a given x may be very expensive in practice, because it may involve heavy computer codes for instance, it is often essential to estimate $\mathcal{P}_u$ using as few evaluations of f as possible.

To overcome the problem of evaluating f many times, one possible approach is to estimate $|A_u(f_n)|$ instead of $|A_u(f)|$, where f_n is an approximation of f constructed from a small set $\{f(x_1), \ldots, f(x_n)\}$ of pointwise evaluations. Such an approximation can be obtained by assuming that f is a sample path of a Gaussian random process ξ and by using a linear predictor ξ_n of ξ constructed from $\xi(x_i)$, $i = 1, \ldots, n$. In this paper, intrinsic Kriging (Matheron, 1973) will be used to obtain ξ_n. We shall show in Section 2 that this method is likely to give faster convergences than the classical Monte Carlo estimators, depending on the regularity of ξ.

A second step is to choose a sequence of evaluation points (x_i) so that $|A_u(\xi_n)| - |A_u(\xi)|$ conditioned on the random variables $\xi(x_i)$, $i \leq n$, converges rapidly to zero. Section 3 presents an acceleration algorithm based on computing an upper bound of the mean square error of volume approximation conditioned on the events $\{\xi(x_i) = f(x_i),\ i = 1 \ldots, n\}$: a point x_{n+1} is selected so that evaluating $f(x_{n+1})$ yields the potential largest decrease of the upper bound. Section 4 provides a numerical example.

2 Excursion set volume estimation by intrinsic Kriging

This section deals with the estimation of the probability $\mathcal{P}_u$ from observations of f at a finite sequence of points $(x_i)_{1\leq i\leq n}$. As mentioned above, $\mathcal{P}_u$ is the volume of $A_u(f)$ under the probability distribution μ. We assume moreover that f is a sample path of a (separable) Gaussian process ξ, with mean $m(x)$, $x \in \mathbb{X}$, and covariance $k(x, y)$, $(x, y) \in \mathbb{X}^2$.

2.1 Monte Carlo estimation

Monte Carlo is a commonly used method to estimate $|A_u(\xi)|$. The volume of excursion of a Gaussian process ξ may be estimated by

$$|A_u(\xi)|_l := \frac{1}{l}\sum_{i=1}^{l} \mathbf{1}_{\{\xi(X_i)\geq u\}} \quad \to_l \quad |A_u(\xi)| \qquad \text{a.s.} \tag{3}$$

where the X_is are independent random variables with distribution μ. The estimator (3) is unbiased, since $E[|A_u(\xi)|_l \mid \xi] = |A_u(\xi)|$, and

$$\mathrm{E}\left[(|A_u(\xi)|_l - |A_u(\xi)|)^2 \mid \xi\right] = \frac{1}{l}|A_u(\xi)|\big(1 - |A_u(\xi)|\big)\,.$$

If evaluating f (a sample path of ξ) at many points of $\mathbb{X}$ is not particularly demanding, then estimating $|A_u(f)|$ is straightforward. However, if $|A_u(f)|$ is small, then the variance of the Monte Carlo estimator is approximately $|A_u(f)|/l$. To achieve a given standard deviation $\kappa|A_u(f)|$, with $\kappa > 0$ small, the required number of evaluations is approximately $1/(\kappa^2|A_u(f)|)$, i.e. it is high. Thus, the convergence of (3) may be too slow in many real applications where doing a lot of evaluations of f may not be affordable (for instance, f may be a complex computer simulation and may take hours or days to run). Of course, many other methods have been proposed to improve the basic Monte Carlo convergence. For instance, methods based on importance sampling, on cross-entropy (Rubinstein, 1999), on the classical extreme value theory (e.g. Embrechts et al., 1997), etc. They are not considered here for the sake of brevity.

2.2 Estimation based on an approximation

An alternative approach is to replace f by an approximation f_n constructed from a set of n point evaluations of f. Provided f_n converges rapidly enough

to f, one expects a good estimation of the excursion sets and their volume using only a few evaluations of f. There are many ways of constructing such an approximation. Let us mention two classical methods: regularized regressions in reproducing kernel Hilbert spaces, e.g. splines or radial basis functions (see for instance Wendland, 2005), and linear prediction of random processes, also known as Kriging (see for instance Chilès and Delfiner, 1999). In this paper, we shall adopt the probabilistic framework[1].

Thus, let us consider that an unbiased linear estimator ξ_n of ξ has been obtained from $\xi(x_1), \ldots, \xi(x_n)$. In particular, we can use ordinary Kriging when the mean of $\xi(x)$ is known and intrinsic Kriging when it is unknown, which is more often the case.

Can we expect a faster convergence when ξ is replaced by ξ_n? Here, we assume the computation time to evaluate $\xi_n(x)$, $x \in \mathbb{X}$, conditioned on $\xi(x_i) = f(x_i)$, $i = 1, \ldots, n$, is small, which means that we can make $|A_u(\xi_n)|_l - |A_u(\xi_n)|$ negligible with respect to $|A_u(\xi_n)| - |A_u(\xi)|$. Thus, we are now interested in the convergence of $|A_u(\xi_n)|$ to $|A_u(\xi)|$. Section 2.2.2 shows how the convergence rate in mean square of $|A_u(\xi_n)|$ to $|A_u(\xi)|$ depends on the fill distance of $\mathbb{X}$ and the regularity of ξ. In Section 3, we shall propose an algorithm to speed up this rate by a sequential choice of the evaluation points.

2.2.1 Intrinsic Kriging basics

In this paper, we use *intrinsic Kriging* (IK) to obtain a linear predictor of ξ based on a finite set of pointwise observations of the process. We recall here the main results (Matheron, 1973). IK extends linear prediction when the mean of $\xi(x)$ is unknown but can be written as a linear parametric function $m(x) = b^{\mathsf{T}} p(x)$. Here, $p(x)$ is a q-dimensional vector of base functions of a vector space $\mathcal{N}$ of translation-stable functions (in practice, all polynomials of degree less or equal to l) and b is a vector of unknown parameters. Intrinsic Kriging assumes that observed values of f are samples from a representation of an intrinsic random function (IRF), a generalized random process defined over a space Λ_l of measures orthogonal to $\mathcal{N}$, and characterized by its stationary generalized covariance $k(h)$ (see the Appendix Section for more details).

Proposition 1 (Intrinsic Kriging, Matheron 1973)**.** *Let ξ_{G} be an IRF(l), with*

[1] In fact, these two classes of methods, which have been studied separately, are equivalent (see for instance Kimeldorf and Wahba (1970)).

generalized covariance $k(h)$. Assume n observations be sample values of the random variables $\xi_{x_i}^{\text{obs}} = \xi(x_i) + N_i$, $i = 1, \ldots, n$, where ξ is an unknown representation of ξ_{G} and the N_is are zero-mean random variables independent of $\xi(x)$, with covariance matrix K_N.

The intrinsic Kriging predictor of $\xi(x)$ based on the observations, is the linear projection $\xi_n(x) = \sum_i \lambda_{i,x} \xi_{x_i}^{\text{obs}}$ of $\xi(x)$ onto $\mathcal{H}_S = \text{span}\{\xi_{x_i}^{\text{obs}}, i = 1, \ldots, n\}$, such that the variance of the prediction error $\xi(x) - \xi_n(x)$ is minimized under the constraint $\delta_x - \sum \lambda_{i,x} \delta_{x_i} \in \Lambda_l$. The coefficients $\lambda_{i,x}$, $i = 1, \ldots, n$, are solutions of a system of linear equations, which can be written in matrix form as

$$\begin{pmatrix} K + K_N & P^{\mathsf{T}} \\ P & 0 \end{pmatrix} \begin{pmatrix} \lambda_x \\ \mu \end{pmatrix} = \begin{pmatrix} k_x \\ p_x \end{pmatrix}, \tag{4}$$

where K is the $n \times n$ matrix of generalized covariances $k(x_i - x_j)$, P is a $q \times n$ matrix with entries ${x_j}^i$ for $j = 1, \ldots, n$ and multi-indexes $i = (i_1, \ldots, i_d)$ such that $|i| := i_1 + \cdots + i_d \leq l$, μ is a vector of Lagrange coefficients, k_x is a vector of size n with entries $k(x - x_i)$ and p_x is a vector of size q with entries x^i, i such that $|i| \leq l$.

The variance of the prediction error is given by $\sigma_n(x)^2 := \text{var}[\xi(x) - \xi_n(x)] = k(0) - \lambda_x^{\mathsf{T}} k_x - \mu^{\mathsf{T}} p_x$.

Proof. See Matheron (1973). □

2.2.2 Asymptotics

In this section, we shall justify that modeling the unknown f by a Gaussian random process ξ and estimating $|A_u(\xi)|$ by $|A_u(\xi_n)|$ is well-founded. Our objective is to establish a mean square convergence when the evaluation points fill $\mathbb{X}$.

Classical results in approximation theory (see for instance Wu and Schaback, 1993 ; Light and Wayne, 1998 ; Narcowich et al., 2003 ; Wendland, 2005) assert that the variance $\sigma_n^2(x)$ of the IK prediction error at x decreases as the sampling density or the regularity of the covariance increases. More precisely, if $\mathbb{X}$ is a bounded domain of $\mathbb{R}^d$, and the Fourier transform of $k(h)$, $h \in \mathbb{R}^d$, satisfies

$$c_1 (1 + \|\omega\|_2^2)^{-\nu} \leq \tilde{k}(\omega) \leq c_2 (1 + \|\omega\|_2^2)^{-\nu}.$$

with $\nu > d/2$, then

$$\|\sigma_n(.)\|_\infty \leq C h_n^{\nu - d/2}, \tag{5}$$

where $h_n = \sup_{y\in\mathbb{X}} \min_i \|y - x_i\|_2$ is a *fill distance* of $(x_1, \ldots, x_n)$ in $\mathbb{X}$.

The following theorem shows that a similar result holds for the process thresholded at a level u.

Theorem 1. *Let ξ be an unknown representation of an IRF(l) ξ_{G}, and $\xi_n(x)$ be the IK predictor of ξ based on observations $\xi(x_i)$, $i = 1, \ldots, n$. Define $\sigma_n(x) := \mathrm{var}\,[\xi(x) - \xi_n(x)]^{1/2}$. Then,*

$$\mathrm{E}\left[(\mathbf{1}_{\xi(x)\geq u} - \mathbf{1}_{\xi_n(x)\geq u})^2\right] = O(\sigma_n(x)|\log(\sigma_n(x))|^{1/2}) \quad \textit{when } \sigma_n(x) \to 0\,.$$

Proof. For all $x \in \mathbb{X}$, $\xi(x) - \xi_n(x)$ is Gaussian with zero-mean and variance $\sigma_n(x)^2$ (but is not orthogonal to $\xi_n(x)$, as would be the case if the mean of ξ were known). Thus, $\forall x \in \mathbb{X}$ and $\forall n \in \mathbb{N}$, we can write $\xi(x)$ as

$$\xi(x) = (1 + a_n(x))\xi_n(x) + b_n(x) + \zeta_n(x)\,, \tag{6}$$

where $a_n(x), b_n(x) \in \mathbb{R}$, $\zeta_n(x)$ is Gaussian and such that $\mathrm{E}[\xi_n(x)\zeta_n(x)] = 0$ and $\mathrm{E}[\zeta_n(x)] = 0$. This decomposition exists and is unique for every n. (To simplify notations, from now on, we shall omit the dependence on x when there is no ambiguity.)

Clearly, $\mathrm{var}[\xi_n]$ is non-decreasing and can be assumed to be strictly positive for n large enough. Since $\mathrm{E}[a_n\xi_n] = -b_n$, we have

$$\sigma_n^2 = \mathrm{var}[a_n\xi_n + b_n + \zeta_n] = \mathrm{E}[(a_n\xi_n + b_n + \zeta_n)^2] = a_n^2\,\mathrm{var}[\xi_n] + E[\zeta_n^2], \tag{7}$$

and thus, the following upper bounds hold for n large enough:

$$\begin{cases} |a_n| \leq K_a\,\sigma_n\,,\ |b_n| \leq K_b\,\sigma_n\,, \\ \tilde{\sigma}_n := \mathrm{E}[\zeta_n^2]^{1/2} \leq \sigma_n\,, \end{cases} \tag{8}$$

for some $K_a, K_b > 0$.

For some threshold $u \in \mathbb{R}$, let α be such that

$$\alpha > |a_n u + b_n| \geq 0\,, \tag{9}$$

and let $N \in \mathbb{N}$ be such that $\forall n > N$, $|a_n| < 1$. For all $n > N$, define

$$\begin{cases} h_n^- &= \dfrac{u - b_n - \alpha}{1 + a_n}\,, \\ h_n^+ &= \dfrac{u - b_n + \alpha}{1 + a_n}\,. \end{cases}$$

Note that $h_n^- < u < h_n^+$ and that

$$h_n^+ - h_n^- = \frac{2\alpha}{1 + a_n}\,.$$

For all $n > N$,

$$\begin{aligned} \mathrm{E}\left[(\mathbf{1}_{\xi(x)\geq u} - \mathbf{1}_{\xi_n(x)\geq u})^2 \mid \xi_n(x)\right] &= \Psi\left(\frac{u-(1+a_n)\xi_n - b_n}{\tilde{\sigma}_n}\right)\mathbf{1}_{\xi_n(x)<u} \\ &+ \Psi\left(-\frac{u-(1+a_n)\xi_n - b_n}{\tilde{\sigma}_n}\right)\mathbf{1}_{\xi_n(x)\geq u}\,, \end{aligned} \tag{10}$$

in which Ψ denotes the tail of the standard Gaussian distribution function. Since

$$\begin{cases} \xi_n < h_n^- & \Rightarrow \quad u-(1+a_n)\xi_n - b_n > \alpha\,, \\ \xi_n > h_n^+ & \Rightarrow \quad -u+(1+a_n)\xi_n + b_n > \alpha\,, \end{cases}$$

and $\tilde{\sigma}_n \leq \sigma_n$, we have

$$\mathrm{E}\left[(\mathbf{1}_{\xi(x)\geq u} - \mathbf{1}_{\xi_n(x)\geq u})^2 \mid \xi_n(x)\right] \leq \Psi\left(\frac{\alpha}{\sigma_n}\right)\mathbf{1}_{\xi_n(x)\in\mathbb{R}\setminus[h_n^-,h_n^+]} + \mathbf{1}_{\xi_n(x)\in[h_n^-,h_n^+]}\,. \tag{11}$$

By integrating with respect to the density of ξ_n, we obtain

$$\begin{aligned} \mathrm{E}\left[(\mathbf{1}_{\xi(x)\geq u} - \mathbf{1}_{\xi_n(x)\geq u})^2\right] &\leq \Psi\left(\frac{\alpha}{\sigma_n}\right) + c_0\frac{2\alpha}{1+a_n} && (12) \\ &\leq \frac{\sigma_n}{\alpha\sqrt{2\pi}}\exp(-\frac{\alpha^2}{2\sigma_n^2}) + c_1\alpha && (13) \end{aligned}$$

where (13) uses a standard Gaussian tail inequality.

The upper bound can be tighten by replacing α with a sequence (α_n) such that

$$\alpha_n := \sqrt{2}\sigma_n|\log(\sigma_n)|^{1/2},$$

which satisfies (9) for n large enough. Therefore,

$$\mathrm{E}\left[(\mathbf{1}_{\xi(x)\geq u} - \mathbf{1}_{\xi_n(x)\geq u})^2\right] \leq O(\sigma_n|\log(\sigma_n)|^{1/2}) \quad \text{when } \sigma_n \to 0. \tag{14}$$

□

Hence, if $\mathbb{X}$ is bounded:

$$\begin{aligned} \mathrm{E}\left[(|A_u(\xi)| - |A_u(\xi_n)|)^2\right] &= \mathrm{E}\left[\left(\int_{\mathbb{X}} \mathbf{1}_{\xi(x)\geq u} - \mathbf{1}_{\xi_n(x)\geq u} d\mu\right)^2\right] \\ &\leq \int_{\mathbb{X}} \mathrm{E}\left[(\mathbf{1}_{\xi(x)\geq u} - \mathbf{1}_{\xi_n(x)\geq u})^2\right] d\mu \\ &\leq C\|\sigma_n(.)\|_\infty|\log\|\sigma_n(.)\|_\infty|^{1/2} \qquad (15) \end{aligned}$$

when $n \to 0$ and $\|\sigma_n(.)\|_\infty \to 0$.

Therefore, this simple result shows that the mean square convergence of $|A_u(\xi_n)|$ to $|A_u(\xi)|$ is related to the mean square convergence of ξ_n to ξ, hence,

due to (5), to the regularity of the covariance and the fill-in distance of $\mathbb{X}$. Informally speaking, we can say that using an approximation will be more efficient than a mere Monte Carlo approach if the regularity of ξ compensates for the slowness of filling $\mathbb{X}$, which of course increases as the dimension d of $\mathbb{X}$ increases. By choosing the x_is on a lattice, the fill distance can be made such that $h_n = O(n^{-1/d})$. Then, the convergence of $|A_u(\xi_n)|$ to $|A_u(\xi)|$ when the x_is fill $\mathbb{X}$ regularly, is faster than Monte Carlo if $\nu > 3d/2$.

3 Convergence acceleration

3.1 Control of convergence

Of course, sampling $\mathbb{X}$ regularly as above may be suboptimal when the evaluations of f are sequential. This section addresses the problem of choosing a sequence $(x_n)_{n\in\mathbb{N}}$ so that the error of volume approximation conditioned on $\xi(x_i) = f(x_i)$, $i = 1, 2, \ldots$ decreases rapidly. More precisely, a desirable strategy would consist in choosing

$$x_n = \underset{x_n\in\mathbb{X}}{\operatorname{argmin}}\ \Upsilon_n(x_n) := \mathrm{E}\left[(|A_u(\xi)| - |A_u(\xi_n)|)^2 \mid Z_{n-1}\right], \tag{16}$$

where for all n, $Z_n = (\xi(x_1), \ldots, \xi(x_n))$. Note that $\Upsilon_n(x_n)$ can also be written as

$$\Upsilon_n(x_n) = \mathrm{E}\left[\,\mathrm{E}\left[(|A_u(\xi)| - |A_u(\xi_n)|)^2 \mid Z_n\right] \mid Z_{n-1}\right]. \tag{17}$$

The distribution of $|A_u(\xi)|$ conditioned on observations is generally unknown (see Adler, 2000, Section 4.4) and therefore, $\mathrm{E}\left[(|A_u(\xi)|-|A_u(\xi_n)|)^2 \mid Z_n\right]$ cannot be easily determined analytically. To overcome this difficulty, we could minimize a Monte Carlo approximation of (16) instead, namely

$$\begin{aligned} x_n = &\underset{x_n\in\mathbb{X}}{\operatorname{argmin}}\ \Upsilon_{n,m}(x_n) := \\ &\mathrm{E}\left[m^{-1}\sum_{i=1}^{m}(|A_u(\xi_n+\zeta_n^i)| - |A_u(\xi_n)|)^2 \;\middle|\; Z_{n-1},\ \{\zeta_n^i, i\le m\}\right], \end{aligned} \tag{18}$$

where the random processes ζ_n^i are m independent copies of ξ conditioned on $Z_n = (0, \ldots, 0)$. The program (18) becomes numerically tractable if we also replace $|A_u(\cdot)|$ by its Monte Carlo estimator $|A_u(\cdot)|_l$. Whereas simulating the conditioned processes ζ_n^i is easy in principle (see Chilès and Delfiner, 1999, chap. 7), it is also computationally intensive since it typically requires $O(l^3)$ operations to simulate ξ at given points $x_1, \ldots, x_l$. Since l has to be high enough

to ensure a degree of accuracy of the estimator $|A_u(\cdot)|_l$, conditional simulations ought to be avoided.

An alternative solution is to approximate $\mathrm{E}\left[(|A_u(\xi)| - |A_u(\xi_n)|)^2 \mid Z_n\right]$ by $\mathrm{E}\left[(|A_u(\xi)|_l - |A_u(\xi_n)|_l)^2 \mid Z_n,\ \{X_i, i \leq l\}\right]$, for l high enough. Then, the Minkowski inequality gives

$$\begin{aligned}\mathrm{E}\left[(|A_u(\xi)|_l - |A_u(\xi_n)|_l)^2 \mid Z_n,\ \{X_i, i \leq l\}\right]^{1/2} \\ \leq \frac{1}{l}\sum_{i=1}^{l} \mathrm{E}\left[(\mathbf{1}_{\xi(X_i)>u} - \mathbf{1}_{\xi_n(X_i)>u})^2 \mid Z_n,\ \{X_i, i \leq l\}\right]^{1/2}. \end{aligned} \tag{19}$$

This makes possible to build a *stepwise uncertainty reduction* algorithm as presented in the next section.

3.2 A stepwise uncertainty reduction algorithm

Denote by $S = \{y_1, \ldots, y_l\}$ a set of l independent sample values of X. Given a finite sequence $(x_i)_{1 \leq i \leq n-1}$ of evaluation points, we wish to obtain a new point x_n that yields the largest decrease of the upper bound of the volume approximation mean-square error obtained in (19), i.e.,

$$x_n = \underset{x_n \in S}{\operatorname{argmin}}\ \Upsilon'_n(x_n) := \frac{1}{l}\sum_{i=1}^{l} \mathrm{E}\left[(\mathbf{1}_{\xi(y_i)>u} - \mathbf{1}_{\xi_n(y_i)>u})^2 \mid B_{n-1}\right]^{1/2}, \tag{20}$$

where B_n denotes the event $\{\xi(x_1) = f(x_1), \ldots, \xi(x_n) = f(x_n)\}$, $n > 0$.

A few steps are needed to transform (20) into a numerically tractable program. First, note that

$$\begin{aligned}\mathrm{E}\left[(\mathbf{1}_{\xi(y_i)>u} - \mathbf{1}_{\xi_n(y_i)>u})^2 \mid B_{n-1}\right] \\ = \int_{z\in\mathbb{R}} \mathrm{E}\left[(\mathbf{1}_{\xi(y_i)>u} - \mathbf{1}_{\xi_n(y_i)>u})^2 \mid \xi(x_n) = z, B_{n-1}\right] \\ \times\ \ p_{\xi(x_n)|B_{n-1}}(z)dz\,, \quad \forall i \in \{1, \ldots, l\}, \end{aligned} \tag{21}$$

where $p_{\xi(x)|B_{n-1}}$ denotes the density of $\xi(x)$ conditionally to B_{n-1}. However, intrinsic Kriging assumes that the mean of ξ is unknown and therefore, for $x \in \mathbb{X}$, $\mathrm{E}\left[(\mathbf{1}_{\xi(x)>u} - \mathbf{1}_{\xi_n(x)>u})^2 \mid \xi(x_n) = z, B_{n-1}\right]$ cannot be determined exactly. Indeed, the values of $a_n(x), b_n(x)$ and $\tilde{\sigma}_n(x)$ in (10) are unknown in practice. Nevertheless, (8) leads to the approximation

$$\mathrm{E}\left[(\mathbf{1}_{\xi(x)>u} - \mathbf{1}_{\xi_n(x)>u})^2 \mid \xi_n(x)\right] \approx \upsilon_n(x) := \Psi\left(\left|\frac{u - \xi_n(x)}{\sigma_n(x)}\right|\right). \tag{22}$$

Finally, define a discretization operator Δ_Q, which can be written for instance as

$$\forall h \in \mathbb{R}\,, \quad \Delta_Q h = z_1 + \sum_{i=2}^{Q} (z_i - z_{i-1}) \mathbf{1}_{]z_i, +\infty[}(h)$$

with $z_1 < z_2 < \cdots < z_Q$. We can now write (20) as a numerically tractable program:

$$x_n = \underset{x_n \in S}{\operatorname{argmin}}\ \Upsilon''_n(x_n) :=$$
$$\frac{1}{l} \sum_{i=1}^{l} \left(\sum_{j=1}^{Q} \mathsf{P}\{\Delta_Q \xi(x_n) = z_j | \, B_{n-1}\} \, \mathrm{E}\left[\upsilon_n(y_i) \mid \xi(x_n) = z_j,\ B_{n-1} \right] \right)^{1/2}. \quad (23)$$

An informal interpretation of (23) is that x_n minimizes the error of prediction of $\mathbf{1}_{\xi(x)>u}$ by $\mathbf{1}_{\xi_n(x)>u}$, which is measured via $\upsilon_n(x)$, averaged on $\mathbb{X}$ under the distribution μ, and conditioned on the observations. When $\Upsilon''_n(x)$ becomes small for all $x \in S$, $|A_u(\xi_n)|_l$ conditioned on observations provides a good approximation of $\mathcal{P}_u$. As will be seen in Section 4, the proposed strategy is likely to achieve very efficient convergences.

4 Example

This section provides a one-dimensional illustration of the proposed algorithm. We wish to estimate (1), where $f(x)$ is a given function defined over $\mathbb{R}$ and $X \sim \mu = \mathcal{N}(0, \sigma^2)$. We assume that f is a sample path of ξ. After a few iterations, the unknown function f (as shown in Figure 1) has been sampled so that the probability of excursion $\mathsf{P}\{\xi(x) > u \mid \xi(x_i) = f(x_i), i = 1 \ldots, n\}$ is determined accurately in the region where the probability density of X is high. This example illustrates the effectiveness of the proposed algorithm. Note that in practice, a parametrized covariance has to be chosen for ξ and its parameters should be estimated from the data, using, for instance, a maximum likelihood approach (e.g. Stein, 1999).

5 Appendix : Intrinsic Random Functions

In this section, we intend to summarize the most important notions about intrinsic random functions (Matheron, 1973). Let $\mathcal{N}$ be a vector space of functions $\{b^\mathsf{T} r(x),\ b \in \mathbb{R}^l\}$ and $\xi(x)$ be a random process with mean $m(x) \in \mathcal{N}$. The

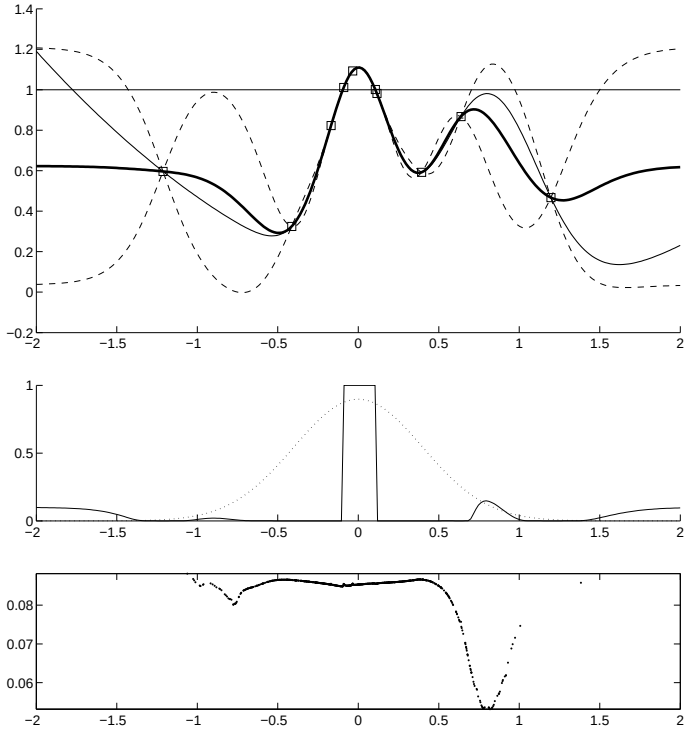


Figure 1. Top: threshold u (horizontal solid line), function f (thin line), n=10 evaluations as obtained by the proposed algorithm using $l = 800$ and $Q = 20$ (squares), IK approximation f_n (thick line), 95% confidence intervals computed from the IK variance (dashed lines). Middle: probability of excursion (solid line), probability density of X (dotted line). Bottom: graph of $\Upsilon_n''(y_i)$, $i = 1, \ldots, l = 800$, the minimum of which indicates where the next evaluation of f should be done (i.e., at approximately 0.75).

main idea of intrinsic random functions is to find some linear transformations of $\xi(x)$ *filtering out* the mean so as to consider a zero-mean process again.

Let $\tilde{\Lambda}$ be the vector space of *finite-support measures*, *i.e.* the space of linear combinations $\sum_{i=1}^{n} \lambda_i \delta_{x_i}$, where δ_x stands for the Dirac measure, such that for any $B \subset \mathbb{X}$, $\delta_x(B)$ equals one if $x \in B$ and zero otherwise. Let $\tilde{\Lambda}_{\mathcal{N}^\perp}$ be the subset of the elements of $\tilde{\Lambda}$ that vanish on $\mathcal{N}$. Thus, $\lambda \in \tilde{\Lambda}_{\mathcal{N}^\perp}$ implies

$$\langle \lambda, f \rangle := \sum_{i=1}^{n} \lambda_i f(x_i) = 0\,, \quad \forall\, f \in \mathcal{N}\,.$$

In the following, we shall restrict ourselves to the case where $\mathcal{N}$ is a vector space of polynomials of degree at most equal to l. Denote by $\mathcal{N}_l$ the linear hull of all multivariate monomials x^i, where $i = (i_1, \ldots, i_d)$ are multi-indexes such that $|i| := i_1 + \cdots + i_d \leq l$, and define $\tilde{\Lambda}_l := \tilde{\Lambda}_{\mathcal{N}_l^\perp}$.

Let $\xi_{\mathrm{G}}(\lambda)$ be a linear map on $\tilde{\Lambda}_l$, with values in $L^2(\Omega, \mathcal{A}, \mathsf{P})$, the space of second-order random variables. Assume that $\mathrm{E}[\xi_{\mathrm{G}}(\lambda)] = 0$ for all λ and that

$$k(\lambda, \mu) := \mathrm{cov}[\xi_{\mathrm{G}}(\lambda), \xi_{\mathrm{G}}(\mu)] = \sum_{i,j} \lambda_i \mu_j k(x_i, y_j)\,,$$

where $k(x, y)$ is a symmetric conditionally positive definite function (i.e. a function such that $k(x, y) = k(y, x)$ and $k(\lambda, \lambda) \geq 0$ for all $\lambda \in \tilde{\Lambda}_l$). Then, $\xi_{\mathrm{G}}(\lambda)$ is a *generalized random process* and $k(x, y)$ is called a *generalized covariance* (note that any covariance is a generalized covariance). Let $\tilde{\mathcal{H}}_l$ be the subspace of $L^2(\Omega, \mathcal{A}, \mathsf{P})$ spanned by $\xi_{\mathrm{G}}(\lambda)$, $\lambda \in \tilde{\Lambda}_l$. Since random variables in $\tilde{\mathcal{H}}_l$ are zero-mean, the inner product of $L^2(\Omega, \mathcal{A}, \mathsf{P})$ can be expressed in $\tilde{\mathcal{H}}_l$ as

$$(\xi_{\mathrm{G}}(\lambda), \xi_{\mathrm{G}}(\mu))_{L^2(\Omega, \mathcal{A}, \mathsf{P})} = k(\lambda, \mu)\,, \quad \lambda, \mu \in \tilde{\Lambda}_l\,.$$

Thus, the bilinear form $k(\lambda, \mu)$ endows $\tilde{\Lambda}_l$ and $\tilde{\mathcal{H}}_{\mathcal{N}^\perp}$ with a structure of pre-Hilbert space. The completions $\mathcal{H}_l$ and Λ_l of $\tilde{\mathcal{H}}_l$ and $\tilde{\Lambda}_l$ under this inner product define isomorphic Hilbert spaces. $\xi_{\mathrm{G}}(\lambda)$ can be extended on Λ_l by continuity. Simplifying hypotheses are introduced in the next paragraph.

Let $\tau_h : \tilde{\Lambda}_l \to \tilde{\Lambda}$ be the translation operator such that for $\lambda = \sum_i \lambda_i \delta_{x_i} \in \tilde{\Lambda}_l$, $\tau_h \lambda = \sum_i \lambda_i \delta_{x_i + h}$. Note that $\tilde{\Lambda}_l$ is stable under translation since $\mathcal{N}_l$ is itself a translation-stable space of functions. Assume further that the generalized covariance $k(x, y)$ is invariant by translation. In the following, we shall write $k(h)$ with $h = x - y$ instead of $k(x, y)$, when the covariance is assumed to be stationary. Then τ_h is continuous and can be uniquely extended on Λ_l.

Definition 1. *Let $\xi_{\mathrm{G}}(\lambda)$ be a zero-mean generalized random process defined on Λ_l, with stationary generalized covariance $k(h)$. The random process $h \mapsto \xi_{\mathrm{G}}(\tau_h\lambda)$, $\lambda \in \Lambda_l$, is therefore weakly stationary. $\xi_{\mathrm{G}}(\lambda)$, $\lambda \in \Lambda_l$, is then an* Intrinsic Random Function *of order l, or IRF(l) in short.*

If $\xi(x)$, $x \in \mathbb{X}$, is a second-order random process, with mean in $\mathcal{N}_l$ and covariance $k(x,y)$, the linear map

$$\begin{array}{rccl} \xi: & \tilde{\Lambda}_l & \to & \mathcal{H} \\ & \lambda = \sum_{i=1}^n \lambda_i \delta_{x_i} & \mapsto & \xi(\lambda) := \sum_{i=1}^n \lambda_i \xi(x_i)\,, \end{array}$$

extends $\xi(x)$ on $\tilde{\Lambda}_l$, where $\mathcal{H}$ stands for the Hilbert space generated by $\xi(x)$, $x \in \mathbb{X}$. Since $k(x,y)$ is positive definite, $(\lambda,\mu)_{\tilde{\Lambda}_l} := (\xi(\lambda),\xi(\mu))_{\mathcal{H}}$ defines an inner product on $\tilde{\Lambda}_l$. Let Λ_l be the completion of $\tilde{\Lambda}_l$ under this inner product and extend $\xi(\lambda)$ on Λ_l by continuity (a generalized random process is thus obtained).

Definition 2. *Let $\xi_{\mathrm{G}}(\lambda)$ be an IRF(l). A second-order random process $\xi(x)$, $x \in \mathbb{X}$, is a* representation *of $\xi_{\mathrm{G}}(\lambda)$ iff*

$$\xi_{\mathrm{G}}(\lambda) = \xi(\lambda), \quad \forall\, \lambda \in \Lambda_l\,.$$

If $\xi_0(x)$ is any representation of $\xi_{\mathrm{G}}(\lambda)$, other representations of $\xi_{\mathrm{G}}(\lambda)$ can be written as

$$\xi(x) = \xi_0(x) + \sum_{i=1}^q B_i p_i(x)\,, \tag{24}$$

where the p_is form a basis of $\mathcal{N}_l$ and the B_is are any second-order random variables. Thus, the representations of an IRF(l) constitute a class of random processes with mean in $\mathcal{N}_l$ (Matheron, 1973).

References

R. J. Adler. On excursion sets, tube formulas and maxima of random fields. *Ann. Appl. Probab.*, 10(1):1–74, 2000.

J.-P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty.* Wiley, New York, 1999.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance.* Applications of Mathematics. Springer-Verlag, Berlin, 1997.

G. S. Kimeldorf and G. Wahba. A correspondance between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41(2): 495–502, 1970.

W. Light and H. Wayne. On power functions and error estimates for radial basis functions interpolation. *J. Approx. Theory*, 92(2):245–266, 1998.

G. Matheron. The intrinsic random functions, and their applications. *Adv. Appl. Prob.*, 5:439–468, 1973.

F. J. Narcowich, J. D. Ward, and H. Wendland. Refined error estimates for radial basis function interpolation. *Constr. Approx.*, 19(4):541–564, 2003.

R. Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodol. Comput.Appl. Probab.*, 2:127–190, 1999.

M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, 1999.

H. Wendland. *Scattered Data Approximation*. Monographs on Applied and Computational Mathematics. Cambridge Univ. Press, Cambridge, 2005.

Z. Wu and R. Schaback. Local error estimates for radial basis function interpolation of scattered data. *IMA J. Numer. Anal.*, 13:13–27, 1993.

2.5 Stepwise Uncertainty Reduction to estimate a quantile

This contribution was presented in the Journées de Statistique in 2010. It follows a collaboration with EDF R&D on assessing the risk of river flooding, which was formalized as a problem of quantile estimation. It details a general framework to carry out sequential Bayesian estimation based on the simulation of sample paths of ξ.

Évaluation d'un risque d'inondation fluviale par planification séquentielle d'expériences

Aurélie Arnaud[1], Julien Bect[2], Mathieu Couplet[1], Alberto Pasanisi[1]
et Emmanuel Vazquez[2,⋆]

1. *EDF R&D, Dépt. Management des Risques Industriels, 78401, Chatou, France*
2. *SUPELEC, Dépt. Signaux & Systèmes Électroniques, 91192 Gif-sur-Yvette, France*

Résumé : Nous nous intéressons au risque d'inondation d'une zone habitable ou industrielle, située à proximité d'un fleuve. Le risque est évalué à partir d'un modèle de la ligne d'eau du fleuve en présence d'incertitudes sur le débit et les caractéristiques du lit fluvial. Comme l'évaluation du modèle de la hauteur d'eau, pour un débit et des caractéristiques du lit fixés, est potentiellement coûteux en temps de calcul, l'estimation d'une probabilité de dépassement de seuil ou d'un quantile de la hauteur d'eau doit en pratique être conduite avec un budget réduit de simulations. Dans cet article, nous nous intéressons spécifiquement à l'estimation d'un quantile et nous proposons une méthode de planification d'expériences séquentielle qui construit une approximation du modèle par krigeage en choisissant les points d'évaluation du modèle de manière à réduire la variance d'estimation du quantile.

Abstract : The risk of river flooding in an inhabitable or industrial area is usually assessed by modeling the water-surface profile of the river, subject to uncertainties on the river discharge and the features of the riverbed. Because a single evaluation of such a model for known discharge and riverbed features is potentially time-consuming, the estimation of a probability of flooding must be achieved with a small budget of simulations. In this paper, we focus on the estimation of a water-level quantile. We propose a sequential Bayesian algorithm that selects relevant simulations to reduce the variance of estimation of the quantile.

1 Introduction

Cette étude concerne l'estimation du risque d'inondation d'une zone habitable ou industrielle, située à proximité d'un fleuve. Soit f la fonction à valeurs réelles, représentant la hauteur de l'eau du fleuve en un point donné, et dont l'argument est un vecteur de facteurs à valeurs dans $\mathbb{X} \subseteq \mathbb{R}^d$. Ces facteurs sont les grandeurs (physiques, morphologiques, etc.) susceptibles d'avoir une influence sur la hauteur d'eau observée. L'ensemble $\mathbb{X}$ est supposé muni d'une mesure de probabilité $\mathsf{P}_{\mathbb{X}}$, qui modélise le fait que les facteurs varient au cours du temps (on peut penser par exemple au débit du fleuve) ou qu'ils sont mal connus (par exemple, les caractéristiques du lit du fleuve). Nous nous intéressons à l'estimation du quantile $q_\alpha(f) = \inf\{u \in \mathbb{R};\, \mathsf{P}_{\mathbb{X}}\{f \leq u\} \geq \alpha\}$, pour une probabilité α donnée et proche de 1. En pratique, la connaissance d'un tel quantile permet de dimensionner la hauteur d'un ouvrage de protection.

La méthode standard pour estimer $q_\alpha(f)$ consiste à simuler un m-échantillon $X_1, \ldots, X_m$ selon la loi $\mathsf{P}_{\mathbb{X}}$, puis à considérer l'estimateur empirique

$$\widehat{q}_{\alpha,m}(f) = \min\left\{y; \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{Y_i \leq y} \geq \alpha\right\} = Y_{\left(\lceil \alpha m \rceil\right)} \tag{1}$$

où $Y_i = f(X_i)$, $i = 1, \ldots, m$, et $Y_{(i)}$ désigne la statistique d'ordre de rang i de l'échantillon $Y_1, \ldots, Y_m$. Il est bien connu que $\sqrt{m}\big(\widehat{q}_{\alpha,m}(f) - q_\alpha(f)\big) \to_m \mathcal{N}\big(0, \sigma^2\big)$, avec $\sigma^2 = \frac{\alpha(1-\alpha)}{p_Y(q_\alpha(f))^2}$,

où p_Y est la densité de $Y = f(X)$, $X \sim \mathsf{P}_{\mathbb{X}}$ (voir par exemple [11]). Il est donc généralement nécessaire de simuler un échantillon de grande taille afin d'obtenir une estimation satisfaisante. Cependant, dans le cas où l'évaluation de la fonction f est coûteuse (par exemple, lorsqu'une évaluation du modèle en un point de l'espace des facteurs nécessite plusieurs heures de calcul), le budget d'évaluations de f sera très limité. La recherche d'estimateurs de quantile avec une faible variance constitue donc un enjeu important pour l'analyse de risque à partir de modèles coûteux. L'échantillonnage d'importance est l'idée la plus naturelle pour atteindre cet objectif [2, 4]. De plus, les techniques d'échantillonnage d'importance peuvent être sensiblement améliorées si l'on est capable de simuler facilement une variable aléatoire auxiliaire, disons Z, fortement corrélée avec Y [2,5]. Une telle variable aléatoire peut être obtenue en construisant une approximation $\widehat{f}$ de f et en posant $Z = \widehat{f}(X)$, $X \sim \mathsf{P}_{\mathbb{X}}$.

Dans cet article, nous proposons une approche fondée sur un algorithme bayésien de planification séquentielle d'expériences, inspiré des algorithmes bayésiens pour l'optimisation globale (voir par exemple [6,7,10]) et d'un algorithme pour estimer des probabilités de défaillance proposé dans [9]. On notera aussi que l'algorithme proposé possède des points communs avec [8]. Supposons que les points $X_i \overset{\text{i.i.d}}{\sim} \mathsf{P}_{\mathbb{X}}$, $i = 1, \ldots, m$ aient été générés mais qu'il n'est pas possible de calculer (1) en raison du coût d'évaluation de f. Notre objectif est de choisir séquentiellement des points d'évaluations $x_1, \ldots, x_n \in \{X_1, \ldots, X_m\}$ de f afin de construire un (méta-)estimateur $\tilde{q}_{\alpha,n}$ de $\widehat{q}_{\alpha,m}(f)$, consistant et rapidement convergent, de telle sorte que l'on puisse avoir $\tilde{q}_{\alpha,n}$ très proche de $\widehat{q}_{\alpha,m}(f)$ avec $n << m$. L'algorithme proposé est exposé dans la section 2. La section 3 fournit une évaluation partielle et empirique des performances de l'algorithme. Enfin la section 4 détaille le contexte applicatif et présente les résultats obtenus.

2 Algorithme séquentiel bayésien pour l'estimation de quantile

Dans [9], l'estimation d'une probabilité de défaillance est formulée comme un problème de planification séquentielle d'expériences dans un cadre bayésien, où l'information provenant des expériences effectuées à un instant est combinée à un a priori sur la fonction f, afin de choisir les expériences futures. Nous adoptons ici le même point de vue pour l'estimation d'un quantile. L'information a priori sur f est spécifiée sous la forme d'un processus aléatoire ξ dont la loi est choisie (ou estimée) par l'utilisateur. En général, on se restreint au cas des processus gaussiens, car il est possible dans ce cas d'écrire la loi a posteriori du processus après n évaluations de f en utilisant le krigeage (voir par exemple [3, 9, 10]).

Dans ce cadre, considérons l'estimateur $\tilde{q}_{\alpha,n} = \mathsf{E}[\,\widehat{q}_{\alpha,m}(\xi) \,|\, \mathcal{F}_n\,]$, où $\mathcal{F}_n$ désigne la σ-algèbre engendrée par les variables aléatoires $\xi(x_1), \ldots, \xi(x_n)$ et les points $X_1, \ldots, X_m$. En pratique, $\tilde{q}_{\alpha,n}$ peut être approché par l'estimateur $\tilde{q}'_{\alpha,n}$ construit de la manière suivante.

A-1 Pour $i = 1, \ldots, N$:

(a) Générer une trajectoire $f^{(n,i)}$ selon la loi de ξ conditionnée par $\xi(x_1), \ldots, \xi(x_n)$, évaluée aux points X_j.

(b) Calculer $q_\alpha^{(n,i)} = \widehat{q}_{\alpha,m}(f^{(n,i)})$, en utilisant l'échantillon $\left\{f^{(n,i)}(X_j)\right\}_{j=1,\ldots,m}$.

A-2 On obtient ainsi un échantillon $q_\alpha^{(n,1)}, \ldots, q_\alpha^{(n,N)}$ distribué selon la loi a posteriori de $\widehat{q}_{\alpha,m}(\xi)$. Définir $\tilde{q}'_{\alpha,n} = \frac{1}{N}\sum_{i=1}^{N} q_\alpha^{(n,i)}$.

Pour l'étape A-1.(a), la technique usuelle est celle du conditionnement par krigeage (pour plus de détails, voir par exemple [3,10]). Notons qu'évaluer $f^{(n,i)}$ en un grand nombre de points est généralement coûteux en temps de calcul. Ceci limite donc la valeur de m que l'on peut considérer en pratique.

Nous cherchons ensuite à réduire l'erreur d'estimation a posteriori de $\widehat{q}_{\alpha,m}(\xi)$ par $\tilde{q}_{\alpha,n}$ en choisissant les points d'évaluation de f (voir [9] pour des explications plus détaillées). Pour ce faire, nous adoptons une stratégie de planification à un pas, consistant à construire la suite $(x_n)_{n\geq 1}$ définie itérativement par

$$x_n = \underset{x\in\{X_1,\ldots,X_m\}}{\operatorname{argmin}} \Upsilon_n(x) := \mathsf{E}\left\{(\widehat{q}_{\alpha,m}(\xi) - \tilde{q}_{\alpha,n})^2 \mid \mathcal{F}_{n-1}\right\}, \tag{2}$$

où $\tilde{q}_{\alpha,n}$ est calculé à partir des observations $\xi(x_i)$, $i = 1, \ldots, n-1$ et de $\xi(x)$ qui n'a pas été observée. Notons que pour tout n, x_n est une fonction de $\xi(x_1), \ldots, \xi(x_{n-1})$. En pratique, le calcul du critère Υ_n en un point x se fait en deux étapes en remarquant que

$$\Upsilon_n(x) = \mathsf{E}\left\{ \mathsf{E}\left\{(\widehat{q}_{\alpha,m}(\xi) - \tilde{q}_{\alpha,n})^2 \mid \mathcal{F}_n\right\} \,\middle|\, \mathcal{F}_{n-1}\right\}.$$

Plus précisément, le calcul numérique de l'espérance conditionnelle intérieure peut se faire d'après la procédure suivante.

B-1 Faire l'étape A-1 ci-dessus en conditionnant les trajectoires par $\xi(x_1), \ldots, \xi(x_{n-1})$, et $\xi(x) = y$

B-2 Définir $\tilde{q}'_{\alpha,n}(x,y) = \frac{1}{N}\sum_{i=1}^{N} q_\alpha^{(n,i)}$ et $\Gamma_n(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}(q_\alpha^{(n,i)} - \tilde{q}'_{\alpha,n}(x,y))^2$.

Le calcul numérique de l'espérance conditionnelle extérieure consiste à approcher l'intégrale $\int_{\mathbb{R}} \Gamma_n(x,y)\, p_{\xi(x)|\mathcal{F}_{n-1}}(y)dy$, où $p_{\xi(x)|\mathcal{F}_{n-1}}$ désigne la densité conditionnelle de $\xi(x)$ par rapport à $\mathcal{F}_{n-1}$. Ceci ne pose pas de problème en pratique.

3 Exemple illustratif

Cette section illustre le comportement de l'algorithme proposé lorsque f est une fonction d'une seule variable scalaire. Nous considérons l'expérience suivante. Nous simulons une trajectoire f d'un processus gaussien sur $\mathbb{X} = \mathbb{R}$, de moyenne nulle et avec une fonction de covariance stationnaire de Matérn (voir par exemple [10]) écrite sous la forme $k(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} s\, \mathcal{K}_\nu(s)$, avec $s = 2\nu^{1/2}h/\rho$, $\sigma^2 = 1$, $\nu = 3$ et $\rho = 1/3$. Nous cherchons à estimer le quantile $q_\alpha(f)$, avec $\alpha = 0.97$, lorsque $\mathbb{X}$ est muni d'une probabilité uniforme sur $[-1, 1]$. Nous choisissons $m = 500$ points et $N = 200$. La figure 1 présente le comportement de l'algorithme de planification séquentielle après $n = 10$ itérations. Nous constatons que les points d'évaluation se concentrent dans les régions où les valeurs de f sont proches de q_α.

Pour compléter cet exemple, nous répétons cette expérience $K = 2000$ fois, et nous calculons les quantiles à 0.005 et 0.995 de l'erreur relative de l'estimateur proposé en fonction du nombre d'itérations. Les résultats sont reportés dans la table 1. On constate empiriquement que l'algorithme a un comportement satisfaisant.

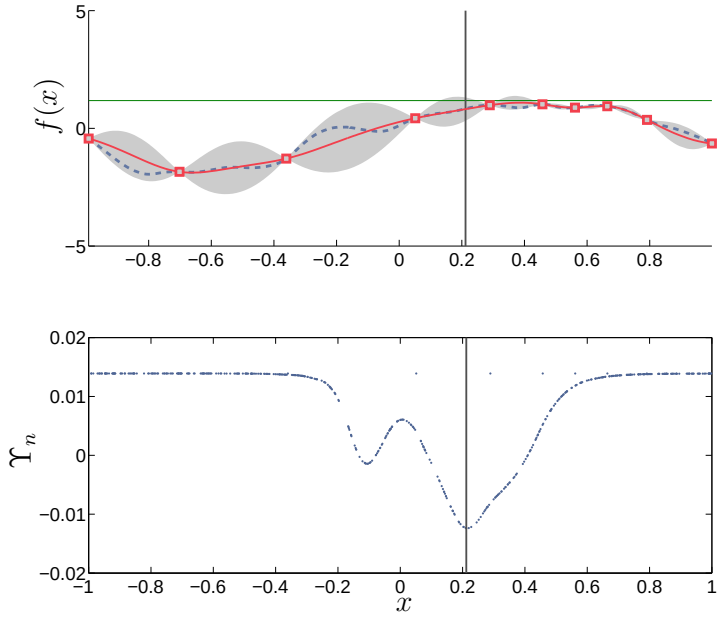


FIGURE 1: Haut : Fonction f (en trait interrompu), n=10 évaluations de f construites d'après l'algorithme proposé (carrés), approximation $\widehat{f}_n$ par krigeage (en trait continu), intervalles de confiance à 95% calculés en utilisant la variance de krigeage (grisé), estimation du quantile à 97% (ligne horizontale). Bas : Graphe de $\Upsilon_n(x_i)$, $i = 1, \ldots, m = 500$. Le minimum de ce graphe indique la position de la prochaine évaluation de f (à environ $x = 0.2$).

n	4	10	15	18	20
$[i^-, i^+]$	$[-12.9, 9.1]$	$[-0.92, 1.31]$	$[-0.12, 0.072]$	$[-2.210^{-2}, 2.3\,10^{-2}]$	$[-4.5\,10^{-3}, 7.0\,10^{-3}]$

TABLE 1: Intervalles empiriques $[i^-, i^+]$ à 99% de l'erreur relative $e = (\tilde{q}_{\alpha,n} - \widehat{q}_{\alpha,m}(\xi))/\widehat{q}_{\alpha,m}(\xi)$ en fonction de n.

4 Application industrielle

On s'intéresse à une portion de la Garonne, d'environ 50 km, comprise entre Tonneins (ville du Lot-et-Garonne, située en aval de la confluence avec le Lot) et La Réole (ville de Gironde, située à la limite de la zone d'influence hydrodynamique de la marée). Bien que cette portion ne présente pas d'installations industrielles importantes, elle se rapproche des configurations fluviales modélisées dans le cadre d'études à plus forts enjeux (notamment de protection d'installations nucléaires), et constitue donc un bon cas-test à la fois pour les études hydrauliques [1] et les analyses de risque.

Nous supposons ici que le problème est unidimensionnel. Plus particulièrement, la grandeur d'intérêt est la ligne d'eau, c'est-à-dire la relation, dépendante du temps, entre la hauteur d'eau et une abscisse curviligne. Le phénomène physique est régi par les équations de Saint-Venant qui lient la hauteur d'eau au débit, à la section mouillée, aux apports de débits latéraux, à la pente du tronçon et aux pertes de charges par frottement entre l'eau

et le lit fluvial. Nous considérons une modélisation dite « en lit composé », avec un débit de crue constant (régime permanent) et des apports latéraux nuls. Pour des raisons de simplicité, nous ne modélisons pas les zones d'expansion de la crue, une fois que la hauteur d'eau a dépassé la côte de la berge. Par conséquent, le modèle aura tendance à surestimer la hauteur d'eau. Les résultats obtenus ont uniquement une valeur d'exemple. Une section fluviale comporte ainsi deux zones : un lit mineur (zone principale d'écoulement) et un lit majeur (zone élargie qui est investie en présence de crues importantes). Ces deux zones sont caractérisées par des sols de nature différente et, par conséquent, par des rugosités différentes, exprimées classiquement par des coefficients de frottement de Strickler. Les coefficients de Strickler permettent d'évaluer, en fonction du débit et de la morphologie de la section fluviale, les pertes de charges dans les équations de Saint Venant. Ils ont la particularité de fournir une mesure décroissante du frottement : plus la valeur du coefficient de Strickler est faible, plus les pertes de charge seront élevées. Le calcul hydraulique est réalisé à l'aide du logiciel Mascaret développé par EDF-R&D et le CETMEF (Centre d'Etudes Techniques Maritimes et Fluviales) et disponible gratuitement.

Dans le cadre de cette étude, les grandeurs physiques supposées incertaines sont le débit, et les coefficients de frottement, affectés par une incertitude de type épistémique, due à un manque de connaissance. La morphologie du cours d'eau et les conditions limites sont considérées connues. La modélisation probabiliste du débit s'avère très facile en pratique car l'historique du débit du fleuve est bien connu. Le maximum annuel du débit peut être modélisé de manière très satisfaisante par une loi de Gumbel. La modélisation de l'incertitude des coefficients de Strickler est en revanche délicate. Ces derniers sont en effet des paramètres du modèle, mais ils ne sont pas directement observables. Il est possible de les estimer indirectement, à partir de couples hauteur-débit, relevés à différents endroits du fleuve. Comme ces estimations sont conduites dans un cadre bayésien, nous avons accès à la loi a posteriori des coefficients de Strickler. Nous choisissons ici de modéliser l'incertitude sur ces facteurs par leur loi a posteriori. Enfin, pour des raisons tenant à l'identifiabilité des coefficients de Strickler à partir des données disponibles, seuls deux coefficients seront considérés (un coefficient global pour le lit mineur du tronçon, et un autre pour le lit majeur). Notons alors que l'espace des facteurs $\mathbb{X}$ est de dimension 3. Nous testons le comportement de l'algorithme proposé pour $\alpha = 0.99$ et $m = 2000$ points. Les résultats sont présentés dans la figure 2 et paraissent satisfaisants.

En conclusion, la technique que nous proposons ici nous semble être intéressante dans le domaine de l'analyse de risque lorsque cette analyse est fondée sur l'évaluation de modèles informatiques coûteux et qui ne permettent pas l'utilisation d'estimateurs empiriques classiques.

Remerciements. Cette étude a été en partie financée par l'Agence Nationale de la Recherche française (ANR) dans le contexte du projet OPUS (réf. ANR-07-CIS7-010 et ANR-07-TLOG-015).

Références

[1] A. Besnard and N. Goutal. Comparison between 1D and 2D models for hydraulic modeling on a flood plain : Case of Garonne river. In *International Conference River Flow*, Cesme-Izmir (Turkey), 2008.

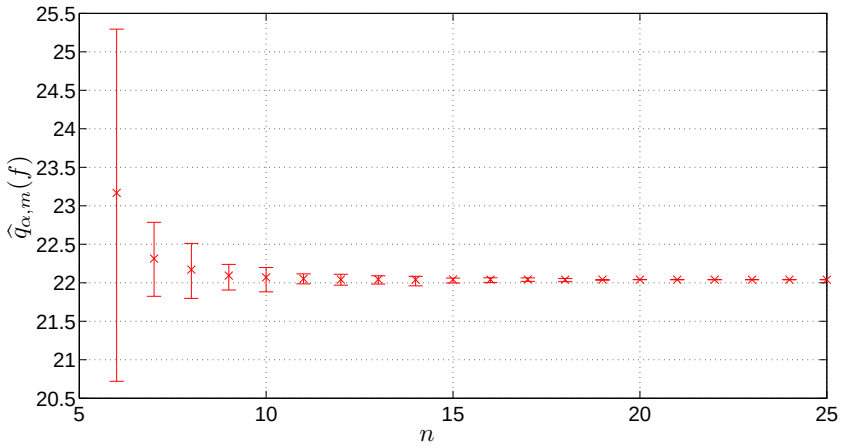


FIGURE 2: Estimation par $\tilde{q}_{\alpha,n}$ de $\widehat{q}_{\alpha,m}(f)$, avec $\alpha = 0.99$ et $m = 2000$, en fonction du nombre n d'évaluations de f. Les barres d'erreurs sont des intervalles a posteriori à 95%. On constate qu'après $n = 25$ évaluations, l'incertitude a posteriori sur $\widehat{q}_{\alpha,m}(f)$ est négligeable. On trouve ainsi $\tilde{q}_{\alpha,n=25} \approx 22.04\,\mathrm{m}$ au dessus du niveau de la mer. À titre indicatif, une estimation par bootstrap d'une région de confiance (fréquentiste) à 95% pour l'estimateur $\widehat{q}_{\alpha,m}(f)$ donne l'intervalle $[21.77, 22.40]$ (calculée à partir de l'approximation par krigeage après $n = 25$ évaluations).

[2] C. Cannamela, J. Garnier, and B. Iooss. Controlled stratification for quantile estimation. *Ann. Appl. Stat.*, 2(4) :1554–1580, 2008.

[3] J.-P. Chilès and P. Delfiner. *Geostatistics : Modeling Spatial Uncertainty*. Wiley, New York, 1999.

[4] P. W. Glynn. Importance sampling for Monte Carlo estimation of quantiles. In *Proceedings of the Second International Workshop on Mathematical Methods in Stochastic Simulation and Experimental Design*, pages 180–185, St. Petersburg, 1996.

[5] T. C. Hesterberg and B. L. Nelson. Control variates for probability and quantile estimation. *Management Science*, 44 :1295–1312, 1998.

[6] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *J. Global Optim.*, 13 :455–492, 1998.

[7] J. Mockus. *Bayesian Approach to Global Optimization : Theory and Applications*. Kluwer Acad. Publ., Dordrecht-Boston-London, 1989.

[8] J. Oakley. Estimating percentiles of uncertain computer code outputs. *J. Roy. Statist. Soc. Ser. C*, 53(1) :83–93, 2004.

[9] E. Vazquez and J. Bect. Sequential bayesian algorithm to estimate a probability of failure. In *15th IFAC Symposium on System Identification, SYSID09*, pages 546–550, Saint-Malo, France, July 6-8 2009.

[10] J. Villemonteix, E. Vazquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *J. Global Optim.*, 44(4) :509–534, 2009.

[11] L. Wasserman. *All of nonparametric statistics*. Springer, 2006.

Chapter 3
Bayesian optimization

3.1 Overview

This chapter consists of a selection of contributions on Bayesian optimization.

Contents

3.2 A new integral loss function for Bayesian optimization

This article is a technical report available on arXiv.org (2014) suggesting the use of a new loss function for Bayesian optimization.

A new integral loss function for Bayesian optimization

Emmanuel Vazquez and Julien Bect

SUPELEC, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France
email: {firstname}.{lastname}@ supelec.fr

Abstract

We consider the problem of maximizing a real-valued continuous function f using a Bayesian approach. Since the early work of Jonas Mockus and Antanas Žilinskas in the 70's, the problem of optimization is usually formulated by considering the loss function $\max f - M_n$ (where M_n denotes the best function value observed after n evaluations of f). This loss function puts emphasis on the value of the maximum, at the expense of the location of the maximizer. In the special case of a one-step Bayes-optimal strategy, it leads to the classical Expected Improvement (EI) sampling criterion. This is a special case of a Stepwise Uncertainty Reduction (SUR) strategy, where the risk associated to a certain uncertainty measure (here, the expected loss) on the quantity of interest is minimized at each step of the algorithm. In this article, assuming that f is defined over a measure space $(\mathbb{X}, \lambda)$, we propose to consider instead the integral loss function $\int_{\mathbb{X}} (f - M_n)_+ \, d\lambda$, and we show that this leads, in the case of a Gaussian process prior, to a new numerically tractable sampling criterion that we call EI^2(for Expected Integrated Expected Improvement). A numerical experiment illustrates that a SUR strategy based on this new sampling criterion reduces the error on both the value and the location of the maximizer faster than the EI-based strategy.

Keywords: Bayesian optimization, computer experiments, Gaussian process, global optimization, sequential design
62L05; 62M20; 62K20; 60G15; 60G25; 90C99

1. Introduction

Let $f : \mathbb{X} \to \mathbb{R}$ be a real-valued continuous function defined on a compact subset $\mathbb{X}$ of $\mathbb{R}^d$, $d \geq 1$. We consider the problem of finding an approximation of the maximum of f,

$$M = \max_{x \in \mathbb{X}} f(x) \,,$$

and of the set of maximizers,

$$x^\star \in \operatorname*{argmax}_{x \in \mathbb{X}} f(x) \,,$$

using a sequence of queries of the value of f at points $X_1, X_2, \ldots \in \mathbb{X}$. At iteration $n+1$, the choice of the evaluation point X_{n+1} is allowed to depend on the results $f(X_1), \ldots, f(X_n)$ of the evaluation of f at $X_1, \ldots, X_n$. Thus, the construction of an optimization strategy $\underline{X} = (X_1, X_2, \ldots)$ can be seen as a sequential decision problem.

We adopt the following Bayesian approach for constructing $\underline{X}$. The unknown function f is considered as a sample path of a random process ξ defined on some probability space $(\Omega, \mathcal{B}, \mathsf{P}_0)$, with parameter $x \in \mathbb{X}$. For a given f, the efficiency of a strategy $\underline{X}$ can be measured in different ways. For instance, a natural loss function for measuring the performance of $\underline{X}$ at iteration n is

$$\varepsilon_n(\underline{X}, f) = M - M_n \,, \tag{1}$$

with $M_n = \max(f(X_1), \ldots, f(X_n))$. The choice of a loss function ε_n, together with a random process model, makes it possible to define the following one-step Bayes-optimal strategy:

$$\left\{ \begin{array}{l} X_1 = x_{\mathrm{init}} \\ X_{n+1} \;=\; \operatorname{argmin}_{x_{n+1} \in \mathbb{X}} \mathsf{E}_n \left(\varepsilon_{n+1}(\underline{X}, \xi) \,\middle|\, X_{n+1} = x_{n+1} \right), \quad \forall n \geq 1, \end{array} \right. \tag{2}$$

Preprint

where E_n denotes the conditional expectation with respect to the σ-algebra $\mathcal{F}_n$ generated by the random variables X_1, $\xi(X_1)$, …, X_n, $\xi(X_n)$. This Bayesian decision-theoretic point of view has been initiated during the 70's by the work of Jonas Mockus and Antanas Žilinskas (see Mockus et al., 1978; Mockus, 1989, and references therein).

For instance, consider the loss defined by (1). Then, at iteration $n+1$, the strategy (2) can be written as

$$\begin{aligned} X_{n+1} &= \operatorname*{argmin}_{x_{n+1}\in\mathbb{X}} \mathsf{E}_n\left(M - M_{n+1} \mid X_{n+1} = x_{n+1}\right) \\ &= \operatorname*{argmax}_{x_{n+1}\in\mathbb{X}} \rho_n\left(x_{n+1}\right), \end{aligned} \tag{3}$$

where $\rho_n(x) := \mathsf{E}_n(\max(\xi(x) - M_n, 0))$ is the *Expected Improvement* (EI) criterion, introduced by Mockus et al. (1978) and later popularized through the EGO algorithm (Jones et al., 1998), both in the case of Gaussian process models (for which $\rho_n(x)$ admits a closed-form expression as a function of the posterior mean and variance of ξ at x).

The contribution of this paper is a new loss function for evaluating the efficiency of an optimization strategy, from which we can derive, in the case of a Gaussian process prior, a numerically tractable sampling criterion for choosing the evaluations points according to a one-step Bayes-optimal strategy. Section 2 explains our motivation for the introduction of a novel loss function, and then proceeds to present the loss function itself and the associated sampling criterion. The numerical implementation of this new sampling criterion is discussed in Section 3. Finally, Section 4 presents a one-dimensional example that illustrates qualitatively the effect of using our new loss function, together with a numerical study that assesses the performance of the criterion from a statistical point of view on a set of sample paths of a Gaussian process.

2. An integral loss function

Observe that (3) can be rewritten as

$$X_{n+1} = \operatorname*{argmin}_{x_{n+1}\in\mathbb{X}} \mathsf{E}_n\left(H_{n+1} \mid X_{n+1} = x_{n+1}\right), \tag{4}$$

with $H_n = \mathsf{E}_n(M - M_n)$. The $\mathcal{F}_{n+1}$-measurable random variable H_{n+1} in the right-hand side of (4) can be seen as a measure of the uncertainty about M at iteration $n+1$: indeed, according to Markov's inequality, $M \in [M_{n+1}; M_{n+1} + H_{n+1}/\delta]$ with probability at least $1-\delta$ under P_{n+1}. Thus, this strategy is actually a special case of *stepwise uncertainty reduction* (Vazquez and Piera-Martinez, 2006; Villemonteix et al., 2009; Bect et al., 2012; Chevalier et al., 2013).

In a global optimization problem, it is generally of interest to obtain a good approximation of *both M and $x^\star$*. The classical loss function $\varepsilon_n = M - M_n$ is not very satisfactory from this respect, since the associated uncertainty measure $H_n = \mathsf{E}_n(M - M_n)$ puts all the emphasis on M, at the expense of $x^\star$. Other uncertainty measures have been proposed recently, which take the opposite approach and focus on $x^\star$ only (Villemonteix et al., 2009; Picheny, 2014a,b).

Assume now that $\mathbb{X}$ is endowed with a finite positive measure λ (e.g., Lebesgue's measure restricted to $\mathbb{X}$), and let us remark that the classical loss function (1) is proportional to $\lambda(\mathbb{X})(M - M_n)$, that is, to the area of the hatched region in Figure 1a. This illustrates that $H_n = \mathsf{E}_n(\varepsilon_n)$ is only a coarse measure of the uncertainty about the pair $(M, x^\star)$. We propose to use instead the integral loss function

$$\varepsilon'_n(\underline{X}, f) = \int_{\mathbb{X}} (f(x) - M_n)_+ \, \lambda(\mathrm{d}x), \tag{5}$$

where $z_+ := \max(z, 0)$. This new loss function is depicted in Figure 1b. The associated uncertainty measure $H'_n = \mathsf{E}_n(\varepsilon'_n)$ should, intuitively, provide a finer measure of the uncertainty about the pair $(M, x^\star)$ and thereby lead to better optimization algorithms. The corresponding stepwise uncertainty reduction strategy can be written as

$$\begin{aligned} X_{n+1} &= \operatorname*{argmin}_{x_{n+1}\in\mathbb{X}} \mathsf{E}_n\left(\int_{\mathbb{X}} (\xi(y) - M_{n+1})_+ \, \lambda(\mathrm{d}y) \;\middle|\; X_{n+1} = x_{n+1}\right) \\ &= \operatorname*{argmin}_{x_{n+1}\in\mathbb{X}} \mathsf{E}_n\left(\int_{\mathbb{X}} \mathsf{E}_{n+1}((\xi(y) - M_{n+1})_+) \, \lambda(\mathrm{d}y) \;\middle|\; X_{n+1} = x_{n+1}\right) \\ &= \operatorname*{argmin}_{x_{n+1}\in\mathbb{X}} \varsigma_n\left(x_{n+1}\right), \end{aligned} \tag{6}$$

where

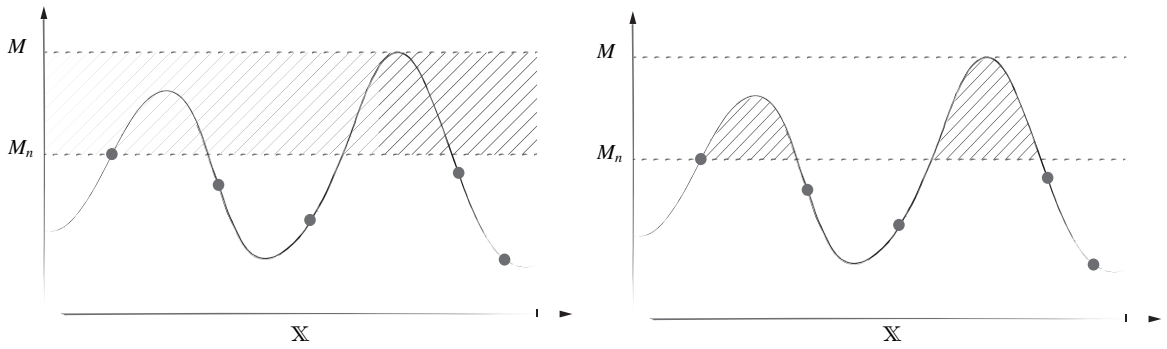
$$\varsigma_n(x_{n+1}) \coloneqq \mathsf{E}_n\left(\int_{\mathbb{X}} \rho_{n+1}(y)\ \lambda(\mathrm{d}y) \,\middle|\, X_{n+1} = x_{n+1}\right) \tag{7}$$

is a new sampling criterion than we call EI^2(for Expected Integrated Expected Improvement). Note that the strategy (6) is very different in spirit from the classical one, associated to the EI criterion. Indeed, while the classical strategy selects a point where the *current* EI is *maximal*, the new strategy selects a point where the integral of the *future* EI is *minimal*, in expectation.

Remark. The sampling criterion defined by (7) is a one-point sampling criterion; that is, a sampling criterion for use in a fully sequential setting. A multi-point sampling criterion can be defined similarly, for use in a batch-sequential setting:

$$\varsigma_{n,r}(x_{n+1},\ldots,x_{n+r}) \coloneqq \mathsf{E}_n\left(\int_{\mathbb{X}} \rho_{n+r}(y)\ \lambda(\mathrm{d}y) \,\middle|\, X_{n+1} = x_{n+1},\ldots,X_{n+r} = x_{n+r}\right) \tag{8}$$

(see Chevalier and Ginsbourger (2013); Chevalier et al. (2013) and references therein for more information on multi-point stepwise uncertainty reduction strategies).



(a) (b)

Figure 1: A diagrammatic interpretation of the loss functions ε_n (left plot) and ε'_n (right plot).

3. Numerical approximation of the sampling criterion

Numerical approximations of the sampling criterion ς_n can be obtained with an acceptable computational complexity when ξ is a Gaussian process. Rewrite (7) as

$$\varsigma_n(x_{n+1}) = \int_{\mathbb{X}} \overline{\rho}_n(y; x_{n+1})\ \lambda(\mathrm{d}y)\,, \tag{9}$$

where $\overline{\rho}_n(y; x_{n+1})$, which we shall call the *Expected Expected Improvement* (EEI) at $y \in \mathbb{X}$ given a new evaluation at $x_{n+1} \in \mathbb{X}$, is defined by

$$\overline{\rho}_n(y; x_{n+1}) \coloneqq \mathsf{E}_n\left(\rho_{n+1}(y) \,\middle|\, X_{n+1} = x_{n+1}\right). \tag{10}$$

(Note that $\overline{\rho}_n(y; x_{n+1}) \neq \rho_n(y)$ because of the implicit dependency of $\rho_{n+1}(y)$ on the future maximum M_{n+1}.)

It turns out that $\overline{\rho}_n(y; x_{n+1})$ can be expressed in closed form, as a function of the posterior mean and covariance of ξ, using the special functions Φ, the cumulative distribution function of the univariate standard normal distribution, and Φ_2, the cumulative distribution function of the bivariate standard normal distribution. To see this, observe that

$$(\xi(y) - M_{n+1})_+ = \widetilde{M}_{n+2} - M_{n+1} = (\widetilde{M}_{n+2} - M_n) - (M_{n+1} - M_n), \tag{11}$$

where $\widetilde{M}_{n+2} = \max(M_{n+1}, \xi(y))$. Therefore, we have

$$\overline{\rho}_n(y; x_{n+1}) = \mathsf{E}_n((\xi(y) - M_{n+1})_+ \mid X_{n+1} = x_{n+1}) = \rho_{n,2}(x_{n+1}, y) - \rho_n(x_{n+1}), \tag{12}$$

where $\rho_{n,r}$ denotes the r-point expected improvement criterion:

$$\rho_{n,r}(x_{n+1}, \ldots, x_{n+r}) \coloneqq \mathsf{E}_n\left(M_{n+r} - M_n \;\middle|\; X_{n+k} = x_{n+k},\ 1 \le k \le r\right). \tag{13}$$

Equation (12) makes it possible to compute $\overline{\rho}_n(y; x_{n+1})$ using the closed-form expression obtained for the multi-point EI by Chevalier and Ginsbourger (2013).

Assuming that $\lambda(\mathbb{X}) < +\infty$, a simple idea for the computation of the integral over $\mathbb{X}$ in (9) is to use a Monte Carlo approximation:

$$\varsigma_n(x_{n+1}) \;\approx\; \frac{\lambda(\mathbb{X})}{m} \sum_{i=1}^{m} \overline{\rho}_n(Y_i; x_{n+1})$$

where $(Y_i)_{1 \le i \le m}$ is a sequence of independent random variables distributed according to $\lambda(\cdot)/\lambda(\mathbb{X})$. Since ς_n has also to be minimized over $\mathbb{X}$, we can also use the sample $(Y_i)_{1 \le i \le m}$ to carry out a simple stochastic optimization. In practice however, we would recommend to use a more advanced sequential Monte Carlo method, in the spirit of that described in Benassi et al. (2012) and Benassi (2013), to carry out both the integration and the optimization steps.

Remark. Equations (9)–(13) are easily generalized to batch sequential optimization. Define a multi-point EEI by

$$\overline{\rho}_{n,r}(y; x_{n+1}, \ldots, x_{n+r}) \coloneqq \mathsf{E}_n\left(\rho_{n+r}(y) \;\middle|\; X_{n+k} = x_{n+k},\ 1 \le k \le r\right).$$

We have

$$\overline{\rho}_{n,r}(y; x_{n+1}, \ldots, x_{n+r}) = \rho_{n,r+1}(x_{n+1}, \ldots, x_{n+r}, y) - \rho_n(x_{n+1}, \ldots, x_{n+r}).$$

Then, we can express a multi-point version of the sampling criterion (7) as

$$\varsigma_{n,r}(x_{n+1}, \ldots, x_{n+r}) = \int_{\mathbb{X}} \overline{\rho}_{n,r}(y; x_{n+1}, \ldots, x_{n+r})\ \lambda(\mathrm{d}y).$$

4. Numerical study

The numerical results presented in this section have been obtained with STK (Bect et al., 2014), a free GPL-licenced Matlab/Octave kriging toolbox.

First, we present a simple one-dimensional illustration, whose aim is to contrast qualitatively the behaviour of a sampling strategy based on the EI^{2}criterion ς_n with that of the classical EI-based strategy. Figure 2 depicts a situation where there is a large expected improvement in a small region of the search domain, and a smaller expected improvement over a large region of the search domain. In such a situation, the new sampling criterion ς_n favors the large region with a smaller expected improvement, thereby inducing a better exploration of the search domain than ρ_n.

Figure 3 represents, for both strategies, the average approximation error obtained on a testbed of 2700 sample paths of a Gaussian process on $\mathbb{R}^d$, $d = 3$, with zero-mean and isotropic Matérn covariance function, simulated on a set of $m = 1000$ points in $[0,1]^d$. The isotropic form of the Matérn covariance on $\mathbb{R}^d$ may be written as $k(x,y) = \sigma^2 r_\nu(\|x-y\|/\beta)$, with $r_\nu : \mathbb{R}^+ \to \mathbb{R}^+$ such that, $\forall h \ge 0$,

$$r_\nu(h) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(2\nu^{1/2}h\right)^\nu \mathcal{K}_\nu\left(2\nu^{1/2}h\right),$$

where Γ is the Gamma function and $\mathcal{K}_\nu$ is the modified Bessel function of the second kind of order ν. Here, $\sigma^2 = 1.0$, $\beta = (4{\cdot}10^{-2}\Gamma(d/2+1)/\pi^{d/2})^{1/d} \approx 0.2$ and $\nu = 6.5$. For each optimization strategy, we use the same covariance function for ξ than that used to generate the sample paths in the testbed. Before running the optimization strategies, an initial evaluation point x_1 is set at the center of $[0,1]^d$. For each sample path f, and each $n \ge 1$, the estimator $x_n^\star$ of $x^\star$ is defined as $x_n^\star = \mathrm{argmax}_{x \in \{x_1, \ldots, x_n\}} f(x)$. Thus, $\|x^\star - x_n^\star\|$ is not a decreasing function of n in general. Figure 3 shows that the approximation errors $M - M_n$ and $\|x^\star - x_n^\star\|$ decrease approximately at the same rate for both strategies; however, the Euclidean distance of $x_n^\star$ to $x^\star$ is significantly smaller in the case of the new strategy.

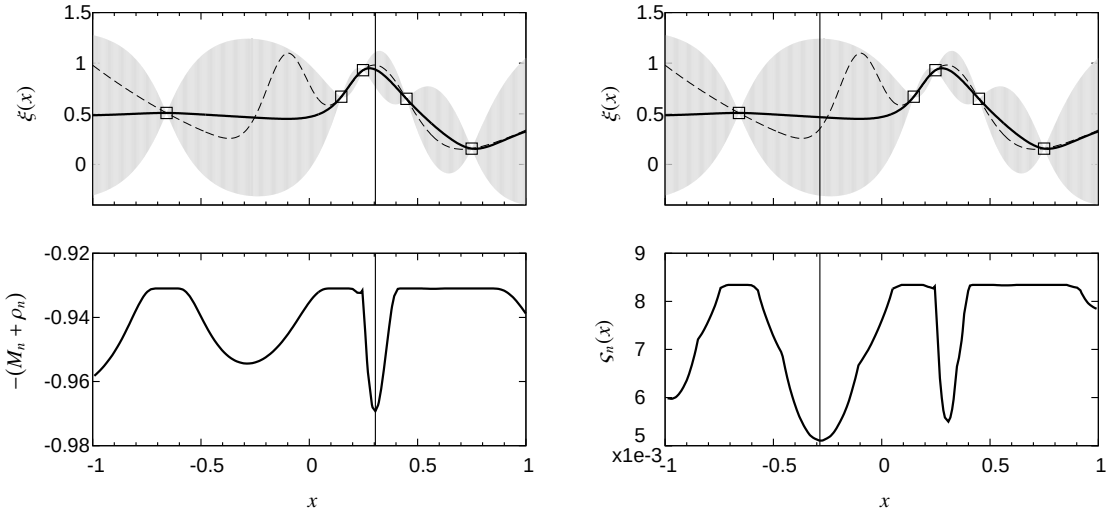


Figure 2: Assessment of the behavior of the sampling criterion ς_n (*bottom, right*) against that of ρ_n (*bottom, left*). The objective is to maximize the function $f : x \in [-1,1] \mapsto (0.8x-0.2)^2 + \exp(-\frac{1}{2}|x+0.1|^{1.95}/0.1^{1.95}) + \exp(-\frac{1}{2}(2x-0.6)^2/0.1) - 0.02$ (*top*, dashed line). Evaluations points are represented by squares; the posterior mean $\widehat{\xi_n}$ is represented by a solid line; 95% credible intervals computed using s_n are represented by gray areas. The next evaluation point will be chosen at the minimum of the sampling criterion (vertical solid line).

References

Bect, J., Ginsbourger, D., Li, L., Picheny, V., Vazquez, E., 2012. Sequential design of computer experiments for the estimation of a probability of failure. Statistics and Computing 22 (3), 773–793.

Bect, J., Vazquez, E., et al., 2014. STK: a Small (Matlab/Octave) Toolbox for Kriging. Release 2.1. URL `http://kriging.sourceforge.net`

Benassi, R., 2013. Nouvel algorithme d'optimisation bayésien utilisant une approche monte-carlo séquentielle. Ph.D. thesis, Supélec.

Benassi, R., Bect, J., Vazquez, E., 2012. Bayesian optimization using sequential Monte Carlo. In: Learning and Intelligent Optimization. 6th International Conference, LION 6, Paris, France, January 16-20, 2012, Revised Selected Papers. Vol. 7219 of Lecture Notes in Computer Science. Springer, pp. 339–342.

Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V., Richet, Y., 2013. Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. Technometrics, 22 pages, accepted for publication, posted online: 21 Nov 2013.

Chevalier, C., Ginsbourger, D., 2013. Fast computation of the multi-points expected improvement with applications in batch selection. In: Nicosia, G., Pardalos, P. (Eds.), Learning and Intelligent Optimization. LNCS. Springer, pp. 59–69.

Jones, D. R., Schonlau, M., Welch, W. J., 1998. Efficient global optimization of expensive black-box functions. J. Global Optim. 13, 455–492.

Mockus, J., 1989. Bayesian approach to Global Optimization: Theory and Applications. Kluwer Acad. Publ., Dordrecht-Boston-London.

Mockus, J., Tiesis, V., Žilinskas, A., 1978. The application of Bayesian methods for seeking the extremum. In: Dixon, L., Szego, G. (Eds.), Towards Global Optimization. Vol. 2. North Holland, New York, pp. 117–129.

Picheny, V., 2014a. Multiobjective optimization using Gaussian process emulators via stepwise uncertainty reduction. arXiv:1310.0732 (to appear in *Statistics and Computing*).

Picheny, V., 2014b. A stepwise uncertainty reduction approach to constrained global optimization. In: Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS), 2014, Reykjavik, Iceland. Vol. 33. JMLR: W&CP, pp. 787–795.

Vazquez, E., Piera-Martinez, M., 2006. Estimation of the volume of an excursion set of a Gaussian process using intrinsic kriging. URL `http://arxiv.org/abs/math.ST/0611273`

Villemonteix, J., Vazquez, E., Walter, E., 2009. An informational approach to the global optimization of expensive-to-evaluate functions. Journal of Global Optimization 44 (4), 509–534.

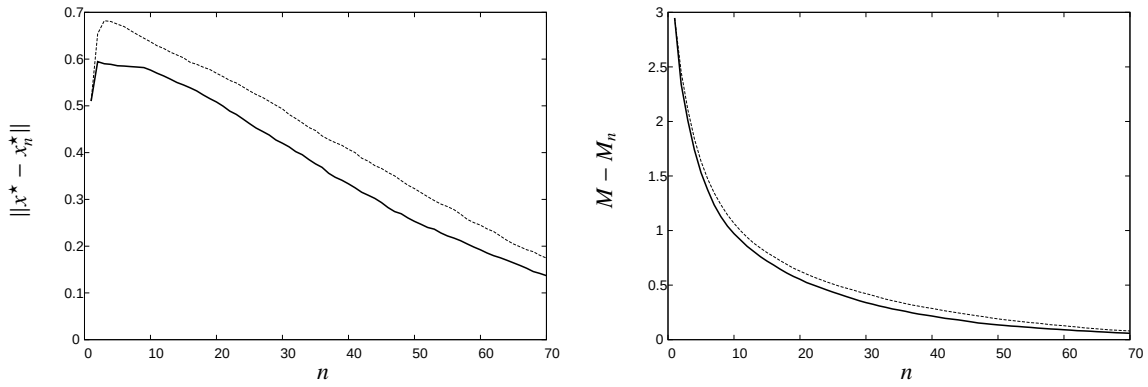


Figure 3: Approximation errors of $x^\star$ (*left*) and M (*right*) using the sampling criteria ς_n (solid line) and ρ_n (dashed line), as a function of the number of evaluations n. More precisely, each plot represents an average approximation error obtained on a testbed of 2700 sample paths of a Gaussian process on $\mathbb{R}^3$, with zero-mean and isotropic Matérn covariance function, simulated on a set of 1000 points in $[0, 1]^3$.

3.3 Convergence properties of the expected improvement algorithm with fixed mean and covariance functions

This article in the Journal of Statistical Planning and Inference (2010) shows that when ξ is a fixed Gaussian process prior satisfying a *no-empty-ball* property, the expected improvement algorithm converges to the global optimum of any f in the RKHS attached to ξ, and almost surely for f drawn from ξ. In 2011, Adam Bull (The Journal of Machine Learning Research, 2011) extends our results with a (seemingly loose) upper-bound of the convergence rate of the expected improvement strategy. In 2012, Dmitry Yarotsky (Journal of Global Optimization, 2012) gives an example of inconsistency of the expected improvement strategy when a Gaussian covariance is used. In 2013, he gives an exponential upper-bound of the convergence rate when f is a univariate analytic function (and a Gaussian covariance is used).

Journal of Statistical Planning and Inference 140 (2010) 3088–3095

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Convergence properties of the expected improvement algorithm with fixed mean and covariance functions

Emmanuel Vazquez *, Julien Bect

SUPELEC, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France

ARTICLE INFO

Article history:
Received 24 July 2009
Received in revised form
13 April 2010
Accepted 15 April 2010
Available online 20 April 2010

Keywords:
Bayesian optimization
Computer experiments
Gaussian process
Global optimization
Sequential design
RKHS

ABSTRACT

This paper deals with the convergence of the expected improvement algorithm, a popular global optimization algorithm based on a Gaussian process model of the function to be optimized. The first result is that under some mild hypotheses on the covariance function k of the Gaussian process, the expected improvement algorithm produces a dense sequence of evaluation points in the search domain, when the function to be optimized is in the reproducing kernel Hilbert space generated by k. The second result states that the density property also holds for P−almost all continuous functions, where P is the (prior) probability distribution induced by the Gaussian process.

1. Introduction

Global optimization is the task of finding the global optima of a real valued function using the results of some pointwise evaluations, which can be chosen sequentially, or in batches, when parallelization is possible. The function to be optimized is generally called *objective function*. In the field of design and analysis of computer experiments, as pioneered by Sacks et al. (1989) and Currin et al. (1991), the objective function—typically an expensive-to-evaluate numerical model of some physical phenomenon—is seen as a sample path of a stochastic process. The stochastic model captures prior knowledge about the objective function and makes it possible to infer the position of the global optima before evaluating the function. This Bayesian decision-theoretic point of view has been largely explored during the 70's and the 80's by the Vilnius school of global optimization led by Mockus (see Mockus et al., 1978; Mockus, 1989; Törn and Zilinskas, 1989; Zilinskas, 1992, and references therein).

In this paper, we consider the *expected improvement* (EI) algorithm, a popular optimization algorithm proposed by Mockus in the 70's and brought to the field of computer experiments by Jones, Schonlau and Welch (Schonlau and Welch, 1996; Schonlau, 1997; Schonlau et al., 1997; Jones et al., 1998). Let $\mathbb{X}$ be a compact subset of $\mathbb{R}^d$, $d \geq 1$, and let ξ be a real valued Gaussian process with parameter $x \in \mathbb{X}$. Our goal is to maximize a given objective function, which is assumed to be a sample path of ξ. The EI algorithm is a sequential planning strategy that constructs a sequence $(x_n)_{n\in\mathbb{N}} \in \mathbb{X}^{\mathbb{N}}$ in such a way that each evaluation point x_n is a function of the previous evaluation points x_i, $i < n$, and the corresponding values of the objective function. Let $M_n = \xi(x_1) \vee \cdots \vee \xi(x_n)$ be the observed maximum at step n; then, a new evaluation point x_{n+1} is chosen in order to maximize the quantity

$$\rho_n(x) := \mathsf{E}[(\xi(x) - M_n)_+ | \xi(x_1), \ldots, \xi(x_n)], \tag{1}$$

* Corresponding author.
E-mail addresses: emmanuel.vazquez@supelec.fr (E. Vazquez), julien.bect@supelec.fr (J. Bect).

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.04.018

where $z_+ = z \vee 0$. Note that this is equivalent to choosing the evaluation point x_{n+1} that maximizes $\mathsf{E}[M_n \vee \xi(x) | \xi(x_1), \ldots, \xi(x_n)]$ with respect to x. The function $\rho_n(x)$, which is called the expected improvement at x, is the conditional mean excess of $\xi(x)$ above the current maximum M_n. It is well known that the expected improvement has a closed-form expression, which can be written using the kriging predictor and its variance (see, e.g., Jones et al., 1998).

This paper addresses the convergence of the EI algorithm, under the assumption that ξ is a Gaussian process with zero mean and known covariance. (Our results still apply if some parameters of the covariance function—for instance, the range and regularity parameters of a Matérn covariance function—are estimated using a first batch of evaluations and held fixed afterward.) It is easily seen that a global optimization algorithm converges for all continuous functions if and only if the sequence of evaluation points produced by the algorithm is dense for all continuous functions (Törn and Zilinskas, 1989, Theorem 1.3). In the case of the EI algorithm, this property was proved by Locatelli (1997), with $d=1$, $\mathbb{X} = [0,1]$ and ξ a Brownian motion. Mockus (1989, Section 4.2) claims a much more general convergence result, but his proof unfortunately contains a severe technical gap.[1]

The main contribution of this paper is a couple of convergence results for the EI algorithm. The first result (Theorem 6) states that the sequence of evaluation points is dense in the search domain provided that the objective function belongs to the reproducing kernel Hilbert space $\mathcal{H}$ attached to ξ, under a non-degeneracy assumption on the covariance function that we call the no-empty-ball (NEB) property. This convergence result is quite natural from the point of view of interpolation theory. The second result (Theorem 7) states that the density property also holds for P-almost all continuous functions, where P is the (prior) probability distribution of the Gaussian process ξ.

The paper is outlined as follows. Section 2 introduces our framework, notations and standing assumptions. Section 3 describes the EI algorithm in greater details and states the main results of the paper. Section 4 provides a sufficient condition for the NEB property, in the case of a stationary covariance function. Section 5 contains the proof of the main theorems. Finally, Section 6 gives our conclusions and discusses future work.

2. Preliminaries

2.1. Framework and standing assumptions

The central mathematical object in global optimization theory is the objective function $\omega : \mathbb{X} \to \mathbb{R}$, defined on some *search space* $\mathbb{X}$. A deterministic search strategy can therefore be seen as a mapping $\underline{X}$ from the set $\Omega = \mathbb{R}^{\mathbb{X}}$ to the set $\mathbb{X}^{\mathbb{N}}$ of all sequences in $\mathbb{X}$,

$$\underline{X}(\omega) \coloneqq (X_1(\omega), X_2(\omega), \ldots), \tag{2}$$

with the property that, for all $n \geq 1$, $X_{n+1}(\omega)$ depends only on the first n evaluations $\omega(X_1(\omega)), \ldots, \omega(X_n(\omega))$. Assuming measurability of the X_ns with respect to the product $\sigma-$algebra $\mathcal{A}$ on Ω (i.e., the $\sigma-$algebra generated by cylinder sets), this can be reformulated in the language of probability theory—although there is no probability measure involved yet. Indeed, let

$$\xi : \mathbb{X} \times \Omega \to \mathbb{R}, \quad (x, \omega) \mapsto \xi(x, \omega) \coloneqq \omega(x), \tag{3}$$

denote the canonical process on the path space $(\Omega, \mathcal{A})$. Then, the above search strategy $\underline{X}$ can be seen as a *random* sequence in $\mathbb{X}$, with the property that X_{n+1} is $\mathcal{F}_n-$measurable, where $\mathcal{F}_n$ is the $\sigma-$algebra generated by $\xi(X_1), \ldots, \xi(X_n)$. It must be stressed that, despite the lexical shift, we are still dealing with *deterministic* algorithms: randomness only comes from the fact that we are now considering the objective function $\xi(\cdot, \omega) = \omega$ as a random element in Ω.

In the Bayesian approach to global optimization, prior information on the objective function is taken into account under the form of a probability measure P on $(\Omega, \mathcal{A})$, which amounts to specifying the probability distribution of the stochastic process ξ. This prior information is then updated at each step of the search, through the computation of the conditional distribution $\mathsf{P}\{\cdot | \mathcal{F}_n\}$. For practical reasons, only Gaussian process priors have been considered in the literature: in this case, the prior is completely specified by the mean $m(x)$ and the covariance function $k(x, x')$, and the process ξ remains Gaussian under the conditional distributions $\mathsf{P}\{\cdot | \mathcal{F}_n\}$, $n \geq 1$. Throughout the paper we shall make the following standing assumptions:

Assumption 1. (i) $\mathbb{X}$ is a compact subset of $\mathbb{R}^d$, for some $d \geq 1$,
(ii) ξ is a centered Gaussian process under P,
(iii) the covariance function k is continuous and positive definite.

Let $\mathcal{H} \subset \Omega$ denote the reproducing kernel Hilbert space (RKHS) that is canonically attached to ξ (also known as the Cameron–Martin space of ξ; see, e.g., Bogachev, 1998). Assumption 1(iii) entails that $\mathcal{H}$ is a space of continuous functions. We shall denote by $(\cdot, \cdot)_{\mathcal{H}}$ the inner product of $\mathcal{H}$ and by $\|\cdot\|_{\mathcal{H}}$ the corresponding norm. It is worth noting that $\mathsf{P}(\mathcal{H}) = 0$

[1] More precisely, the arguments given on page 45 fail to prove the key result claimed in Lemma 4.2.2, i.e., the density of the sequence of evaluation points.

(see, e.g., Lukic and Beder, 2001, Driscoll's theorem). We shall comment on this fact with respect to our convergence result in Section 3.

Remark 2. Unless otherwise specified (see Section 4), it is not assumed that the covariance k is stationary. To the best of our knowledge, however, most practical applications of the EI algorithm have used stationary covariances to model the objective function prior to any evaluation.

2.2. Linear prediction and the no-empty-ball property

For $n \geq 1$, $\underline{x}_n = (x_1, \ldots, x_n) \in \mathbb{X}^n$ and $x \in \mathbb{X}$, we denote by $\hat{\xi}_n(x; \underline{x}_n)$ the conditional expectation of $\xi(x)$ given $\xi(x_1)$, $\xi(x_2), \ldots, \xi(x_n)$. Since ξ is a centered Gaussian process, the conditional expectation is also the best linear predictor in $L^2(\Omega, \mathcal{A}, \mathsf{P})$, and therefore can be written as

$$\hat{\xi}_n(x, \omega; \underline{x}_n) = \sum_{i=1}^{n} \lambda_n^i(x; \underline{x}_n) \xi(x_i, \omega). \tag{4}$$

Let $\sigma_n^2(\underline{x}_n)$ denote the mean-square prediction error, i.e.,

$$\sigma_n^2(x; \underline{x}_n) \coloneqq \mathsf{E}[(\xi(x) - \hat{\xi}_n(x; \underline{x}_n))^2]. \tag{5}$$

(Recall that, since ξ is a Gaussian process, the error of prediction is independent of the $\sigma-$algebra generated by the $\xi(x_i)$s, $1 \leq i \leq n$; see, e.g., Chilès and Delfiner, 1999, Section 3.3.4.)

Definition 3. We shall say that the Gaussian process ξ—or, equivalently, the covariance function k—has the no-empty-ball (NEB) property if, for all sequences $(x_n)_{n \geq 1}$ in $\mathbb{X}$ and all $y \in \mathbb{X}$, the following assertions are equivalent:

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

Since k is assumed continuous, (i) always implies (ii) in Definition 3. The NEB property is therefore equivalent to the assertion that, if the prediction error at y goes to zero, then there can be no "empty ball" centered at y (i.e., for all $\varepsilon > 0$, there exists $n \geq 1$ such that $|y - x_n| < \varepsilon$)—hence its name. A sufficient condition for the NEB property will be given in Section 4. To the best of our knowledge, finding necessary and sufficient condition for the NEB property is an open problem.

2.3. Simplified notations

Since the notations introduced in (4) and (5) would rapidly become cumbersome in the next sections, the following simplified notations will be used

$$\hat{\xi}_n(x, \omega) \coloneqq \hat{\xi}_n(x, \omega; \underline{X}_n(\omega)), \tag{6}$$

$$\sigma_n^2(x, \omega) \coloneqq \mathsf{E}[(\xi(x) - \hat{\xi}_n(x; \underline{X}_n))^2 | \mathcal{F}_n](\omega) = \sigma_n^2(x; \underline{X}_n(\omega)), \tag{7}$$

with $\underline{X}_n = (X_1, \ldots, X_n)$. Remark that $\sigma_n^2(x, \omega)$ is a stochastic process indexed by $\mathbb{X}$. The second equality in (7) follows from the fact that $\hat{\xi}_n(x; \cdot)$ is continuous for all $x \in \mathbb{X}$.

3. Main results

In this paper, we shall consider a generalization of the EI criterion. Define

$$\rho_n(x) = \gamma(\hat{\xi}_n(x) - M_n, \sigma_n^2(x)), \tag{8}$$

where the function $\gamma : \mathbb{R} \times [0; +\infty) \to [0; +\infty)$ satisfies the following requirements:

$$R_1 : \gamma \text{ is continuous,}$$

$$R_2 : \forall z \leq 0,\ \gamma(z, 0) = 0,$$

$$R_3 : \forall z \in \mathbb{R},\ \forall s > 0,\ \gamma(z, s) > 0. \tag{9}$$

The corresponding optimization algorithm can then be written as

$$\begin{cases} X_1 = x_{\text{init}} \in \mathbb{X}, \\ X_{n+1} = \operatorname{argmax}_{x \in \mathbb{X}} \rho_n(x). \end{cases} \tag{10}$$

Remark 4. It is well known (Schonlau and Welch, 1996) that the EI criterion defined by (1) can be rewritten under the form (8). More precisely, let Φ denote the Gaussian cumulative distribution function. Then (8) holds for the EI criterion with

$$\gamma(z,s) = \begin{cases} \sqrt{s}\Phi'\left(\frac{z}{\sqrt{s}}\right) + z\Phi\left(\frac{z}{\sqrt{s}}\right) & \text{if } s > 0, \\ \max(z,0) & \text{if } s = 0. \end{cases} \tag{11}$$

In fact, Eq. (8) with γ thus defined should be taken as the true definition of the EI criterion. Indeed, the exact mathematical meaning of "$\rho_n(x) \coloneqq \mathsf{E}[(\xi(x)-M_n)_+ | \mathcal{F}_n]$" has to be specified, since, for each x, the conditional expectation is only defined up to a $\mathsf{P}-$negligible subset of Ω.

Remark 5. The criterion $x \mapsto \rho_n(x)$ is continuous, but there is no guarantee that the maximizer over $\mathbb{X}$ will be unique. Therefore, a more rigorous statement of the iterative part of (10) would be $X_{n+1} \in \arg\max_{x\in\mathbb{X}} \rho_n(x)$. In this way, instead of a single algorithm, we encapsulate the family of all algorithms that choose (measurably) X_{n+1} among the maximizers of ρ_n. General measurable selection theorems (see, e.g., Molchanov, 2005) ensure that such an algorithm does exist.

The first result of this paper is the following density theorem:

Theorem 6. *Assume that the covariance function k has the NEB property. Then, for all $x_{\text{init}} \in \mathbb{X}$ and all $\omega \in \mathcal{H}$, the sequence $(X_n(\omega))_{n\geq 1}$ generated by (10) is dense in $\mathbb{X}$.*

The fact that Theorem 6 is stated for objective functions in the RKHS $\mathcal{H}$ calls for some comments. From the point of view of interpolation theory, it is indeed quite natural that an algorithm built on the best interpolants $\hat{\xi}_n(\cdot,\omega)$ in a RKHS $\mathcal{H}$ should be provably working, using the tools of RHKS theory, only when ω is in this very space. From the probabilistic point of view, however, the event $\{\xi(\cdot) \in \mathcal{H}\}$ almost never happens according to Driscoll's theorem (Lukic and Beder, 2001). The second result of this paper states that the result of Theorem 6 also holds $\mathsf{P}-$almost surely in Ω.

Theorem 7. *Assume that the covariance function k has the NEB property. Then, for all $x_{\text{init}} \in \mathbb{X}$, the sequence $(X_n)_{n\geq 1}$ generated by (10) is P-almost surely dense in $\mathbb{X}$.*

It is still an important open question to determine whether the algorithm converges for *all* continuous functions, as claimed in Mockus (1989). Another interesting open problem would be to determine whether the NEB assumption can be relaxed.

Remark 8. We have assumed for the sake of simplicity that the optimization algorithm starts after a single evaluation performed at $X_1 = x_{\text{init}}$. In practice, especially when some parameters of the covariance need to be estimated, the algorithm starts with an initial design of several evaluations $x_{\text{init}}^1, \ldots, x_{\text{init}}^{n_0}$. This is equivalent to saying that $\mathcal{F}_1$ is the $\sigma-$algebra generated by $\xi(x_{\text{init}}^1), \ldots, \xi(x_{\text{init}}^{n_0})$. The proofs of Theorems (6) and (7) carry over without modification.

4. A sufficient condition for the NEB property

4.1. Statement of the result

In this section we shall prove that the following assumption is a sufficient condition for the NEB property:

Assumption 9. The process ξ is stationary and has spectral density S, with the property that S^{-1} has at most polynomial growth.

In other words, Assumption 9 means that there exist $C > 0$ and $r \in \mathbb{N}\backslash\{0\}$ such that $S(u)(1+|u|^r) \geq C$, almost everywhere on $\mathbb{R}^d$. This assumption prevents k from being *too regular*. In particular, the so-called *Gaussian* covariance,

$$k(x,y) = \sigma^2 e^{-\alpha\|x-y\|^2}, \quad \sigma > 0,\ \alpha > 0, \tag{12}$$

does not satisfy Assumption 9. However, we are still allowed to consider a large class of covariances. For instance, the exponential covariances

$$k(x,y) = \sigma^2 e^{-\alpha\|x-y\|^s}, \quad \sigma > 0,\ \alpha > 0,\ 0 < s < 2, \tag{13}$$

the class of Matérn covariances (see, e.g., Stein, 1999), and their anisotropic versions, all satisfy Assumption 9. The main result of this section is:

Proposition 10. *Let $(x_n)_{n\geq 1}$ and $(y_n)_{n\geq 1}$ be two sequences in $\mathbb{X}$. Assume that the sequence (y_n) is convergent, and denote by $y^\star$ its limit. Then each of the following conditions implies the next one*:

(i) *$y^\star$ is an adherent point of the set $\{x_n, n \geq 1\}$,*
(ii) *$\sigma_n^2(y_n; \underline{x}_n) \to 0$ when $n \to \infty$,*
(iii) *$\hat{\xi}_n(y_n, \omega; \underline{x}_n) \to \xi(y^\star, \omega)$ when $n \to \infty$, for all $\omega \in \mathcal{H}$.*

Moreover, under Assumption 9, *the three conditions are equivalent and therefore* ξ *has the NEB property*.

Remark 11. As already observed, the Gaussian covariance does not satisfy Assumption 9. In fact, it is known that the Gaussian covariance does not even have the NEB property (Vazquez and Bect, 2010).

4.2. Consequence of Assumption 9 in terms of RKHS

Let $\mathcal{H}'$ denote the RKHS associated to k on $\mathbb{R}^d$. It is well known (Aronszajn, 1950, Section 1.5) that $\mathcal{H}$ embeds isometrically into $\mathcal{H}'$ and that, for all $\omega \in \mathcal{H}'$, the orthogonal projection of ω onto $\mathcal{H}$ is simply its restriction to $\mathbb{X}$.

Under Assumption 9, $\mathcal{H}'$ contains the Sobolev space $H^{r/2}(\mathbb{R}^d)$, and the injection is continuous. Indeed, denoting by $\hat{\omega}$ the Fourier transform of $\omega \in \mathcal{H}'$, we have

$$\int (1+|u|^r)|\hat{\omega}(u)|^2 \,\mathrm{d}u \geq C \int S(u)^{-1}|\hat{\omega}(u)|^2 \,\mathrm{d}u = \|\omega\|^2_{\mathcal{H}'}.$$

A useful consequence is that $\mathcal{H}'$ contains the space $C^\infty_c(\mathbb{R}^d)$ of all compactly supported infinitely differentiable functions on $\mathbb{R}^d$, for any r. In particular, k is a *universal kernel* on $\mathbb{X}$ in the sense of Steinwart (2001), which means that $\mathcal{H}$ is dense in the Banach space $C(\mathbb{X})$ of all continuous functions on $\mathbb{X}$.

4.3. Proof of Proposition 10

(i) $\Rightarrow$ (ii). Assume that $y^\star \notin \{x_n, n \geq 1\}$ (otherwise the result holds trivially). Let (x_{ϕ_k}) be a subsequence of (x_n) converging to $y^\star$ and let $\psi_n = \max\{\phi_k; \phi_k \leq n\}$. Then,

$$\sigma_n^2(y_n;\underline{x}_n) = \mathrm{var}[\xi(y_n) - \hat{\xi}_n(y_n;\underline{x}_n)] \leq \mathrm{var}[\xi(y_n) - \xi(x_{\psi_n})].$$

Since $\psi_n \to \infty$, it follows from the continuity of k that

$$\mathrm{var}[\xi(y_n) - \xi(x_{\psi_n})] = k(y_n,y_n) + k(x_{\psi_n},x_{\psi_n}) - 2k(x_{\psi_n},y_n) \to 0.$$

(ii) $\Rightarrow$ (iii). Using the Cauchy–Schwarz inequality in $\mathcal{H}$, we have

$$|\xi(y_n,\omega) - \hat{\xi}_n(y_n,\omega;\underline{x}_n)| \leq \sigma_n(y_n;\underline{x}_n)\|\omega\|_{\mathcal{H}}.$$

Therefore

$$|\xi(y^\star,\omega) - \hat{\xi}_n(y_n,\omega;\underline{x}_n)| \leq |\xi(y^\star,\omega) - \xi(y_n,\omega)| + |\xi(y_n,\omega) - \hat{\xi}_n(y_n;\underline{x}_n)| \leq |\omega(y^\star) - \omega(y_n)| + \sigma_n(y_n;\underline{x}_n)\|\omega\|_{\mathcal{H}} \to 0,$$

since ω is continuous.

Under Assumption 9, (iii) $\Rightarrow$ (i). Suppose (i) is false. Then, there exists a neighborhood U of $y^\star$ in $\mathbb{R}^d$ that does not intersect $\{x_n, n \geq 1\}$. Besides, it follows from Assumption 9 that there exists $\omega \in \mathcal{H}$ such that supp $\omega \subset U$ and $\omega(y^\star) > 0$ (where supp ω denotes the support of ω). Then, $\hat{\xi}_n(y^\star,\omega;\underline{x}_n) = 0$ for all n, whereas $\xi(y^\star,\omega) = \omega(y^\star) \neq 0$. Therefore (iii) does not hold. □

5. Proofs of the main theorems

5.1. Proof of Theorem 6

Let $v_n = \sup_{x \in \mathbb{X}} \rho_n(x)$, where ρ_n is the criterion defined by Eq. (8). Note that, for all $n \geq 1$,

$$v_n = \rho_n(X_{n+1}) = \gamma(\hat{\xi}_n(X_{n+1}) - M_n, \sigma_n^2(X_{n+1})).$$

Our proof of Theorem 6 will be based on the following result (which does not require the NEB property):

Lemma 12. *For all* $\omega \in \mathcal{H}$, $\liminf_{n\to\infty} v_n(\omega) = 0$.

Proof. Fix $\omega \in \mathcal{H}$. For all $n \geq 1$, set $x_n = X_n(\omega)$, $s_n = \sigma_n^2(x_{n+1},\omega)$ and $z_n = \hat{\xi}_n(x_{n+1},\omega) - M_n(\omega)$, so that $v_n(\omega) = \gamma(z_n,s_n)$. Let $y^\star$ be a cluster point of the sequence (x_n) and let (x_{ϕ_n}) be any subsequence converging to $y^\star$: we are going to prove that $v_{\phi_n - 1}(\omega) \to 0$. It follows from Proposition 10, (i) $\Rightarrow$ (iii), that $\hat{\xi}_{\phi_n - 1}(x_{\phi_n},\omega) \to \omega(y^\star)$. Moreover, $(M_{\phi_n - 1}(\omega))$ is a bounded increasing sequence, with the property that $M_{\phi_n - 1}(\omega) \geq M_{\phi_{n-1}}(\omega) \geq \omega(x_{\phi_{n-1}}) \to \omega(y^\star)$. Therefore $(z_{\phi_n - 1})$ has a finite limit, such that

$$\lim_{n\to\infty} z_{\phi_n - 1} = \lim_{n\to\infty} \hat{\xi}_{\phi_n - 1}(x_{\phi_n},\omega) - \lim_{n\to\infty} M_{\phi_n - 1}(\omega) \leq 0.$$

By Proposition 10, (i) $\Rightarrow$ (ii), we also know that $s_{\phi_n-1} = \sigma^2_{\phi_n-1}(x_{\phi_n},\omega) \to 0$. Therefore, using (R_1) and (R_2),

$$v_{\phi_n-1}(\omega) = \gamma(z_{\phi_n-1}, s_{\phi_n-1}) \to \gamma\left(\lim_{n\to\infty} z_{\phi_n-1}, 0\right) = 0.$$

This completes the Proof of Lemma 12. □

Proof of Theorem 6. Now fix $\omega \in \mathcal{H}$, and suppose that $\{X_n(\omega), n \geq 1\}$ is not dense in $\mathbb{X}$. Then there exist a point $y^\star \in \mathbb{X}$ that is not adherent to $\{X_n(\omega), n \geq 1\}$. This implies, by the NEB property, that

$$\inf_{n\geq 1} \sigma_n^2(y^\star,\omega) > 0.$$

Besides, using the Cauchy–Schwartz inequality in $\mathcal{H}$, we observe that the sequence $(\hat{\xi}_n(y^\star,\omega))$ is bounded. Indeed, we have

$$|\hat{\xi}_n(y^\star,\omega) - \omega(y^\star)|^2 \leq \sigma_n^2(y^\star,\omega)\|\omega\|^2_{\mathcal{H}} \leq k(y^\star,y^\star)\|\omega\|^2_{\mathcal{H}}.$$

The sequence $(M_n(\omega))$ is also obviously bounded by $\|\omega\|_\infty$. Therefore, we obtain as a consequence of (R_1) and (R_3) that

$$\rho_n(y^\star,\omega) \geq \inf_{k\geq 1} \gamma(\hat{\xi}_k(y^\star,\omega) - M_k(\omega), \sigma_k^2(y^\star,\omega)) > 0.$$

This is a contradiction with Lemma 12, since $v_n(\omega) = \max_{x\in\mathbb{X}} \rho_n(x,\omega)$. The proof is thus complete. □

5.2. Proof of Theorem 7

In essence, the structure of the proof of Theorem 7 is the same as that of Theorem 6. The first step is to obtain an almost sure version of Lemma 12.

Lemma 13. $\liminf_{n\to\infty} v_n = 0$ *almost surely*.

Proof. Let $D_n = \min_{1\leq i\leq n} |X_{n+1} - X_i|$ be the distance of X_{n+1} to the set of all previous evaluation points. Define $T_k = \min\{n \geq 1; D_n \leq r_k\}$, with (r_k) a sequence of positive numbers such that $\lim r_k = 0$. Note that each T_k is finite, since the set $\mathbb{X}$ is compact, and is an $(\mathcal{F}_n)$−stopping time since the sequence (D_n) is $(\mathcal{F}_n)$−adapted.

The first step is to see that, as in the proof of Proposition 10,

$$\sigma^2_{T_k}(X_{T_k+1}) \leq \eta_k := \sup_{|x-y|\leq r_k} k(x,x) + k(y,y) - 2k(x,y) \underset{k\to\infty}{\longrightarrow} 0. \tag{14}$$

Note that $(X_{T_k+1})_k$ does not necessarily converge.

The next step is to prove that $\hat{\xi}_{T_k}(X_{T_k+1}) - \xi(X_{T_k+1})$ converges to zero almost surely, for a suitable choice of the sequence (r_k). First, using that T_k is a stopping time, we have

$$\begin{aligned}
\mathsf{E}[(\hat{\xi}_{T_k}(X_{T_k+1}) - \xi(X_{T_k+1}))^2] &= \mathsf{E}\left[\sum_{n\geq 1} \mathbf{1}_{T_k=n} (\hat{\xi}_n(X_{n+1}) - \xi(X_{n+1}))^2\right] \\
&= \sum_{n\geq 1} \mathsf{E}[\mathbf{1}_{T_k=n} \mathsf{E}[(\hat{\xi}_n(X_{n+1}) - \xi(X_{n+1}))^2 | \mathcal{F}_n]] \\
&= \mathsf{E}\left[\sum_{n\geq 1} \mathbf{1}_{T_k=n} \sigma_n^2(X_{n+1})\right] \\
&= \mathsf{E}[\sigma^2_{T_k}(X_{T_k+1})] \leq \eta_k.
\end{aligned}$$

Then, for each $\varepsilon > 0$, it follows from Markov's inequality that

$$\mathsf{P}\{(\hat{\xi}_{T_k}(X_{T_k+1}) - \xi(X_{T_k+1}))^2 > \varepsilon\} \leq \eta_k/\varepsilon.$$

Choosing r_k such that, for instance, $\eta_k = 1/k^2$, ensures that $\hat{\xi}_{T_k}(X_{T_k+1}) - \xi(X_{T_k+1})$ converges to zero almost surely. Therefore, the sequence $(\hat{\xi}_{T_k}(X_{T_k+1}))$ is almost surely bounded. Moreover,

$$\limsup_{k\to\infty} \hat{\xi}_{T_k}(X_{T_k+1}) - M_{T_k} = \limsup_{k\to\infty} \hat{\xi}_{T_k}(X_{T_k+1}) - M_{T_k+1} \leq \lim_{k\to\infty} \hat{\xi}_{T_k}(X_{T_k+1}) - \xi(X_{T_k+1}) = 0 \quad \text{a.s.}, \tag{15}$$

where we have used the fact that (M_n) is convergent.

Finally, using (R_1) and (R_2), the fact that $(\hat{\xi}_{T_k}(X_{T_k+1}) - M_{T_k})$ is almost surely bounded, (14) and (15), we conclude that

$$v_{T_k} = \gamma(\hat{\xi}_{T_k}(X_{T_k+1}) - M_{T_k}, \sigma^2_{T_k}(X_{T_k+1})) \underset{k\to\infty}{\to} 0 \quad \text{a.s.} \quad \square$$

Proof of Theorem 7. Fix $x \in \mathbb{X}$ and define the event $A_x \in \mathcal{A}$ by

$$A_x = \{x \text{ is not an adherent point of the set } \{X_n, n \geq 1\}\}.$$

Then $\inf_{n \geq 1} \sigma_n^2(x) > 0$ on A_x by the NEB property. Moreover, the martingale $\hat{\xi}_n(x) = \mathsf{E}[\xi(x)|\mathcal{F}_n]$ is bounded in L^2 since $\mathsf{E}\hat{\xi}_n(x)^2 \leq k(x,x) < +\infty$, and thus converges almost surely and in L^2 to a random variable $\hat{\xi}_\infty(x)$ (see, e.g., Williams, 1991). As a consequence, the event

$$B_x := \{(\hat{\xi}_n(x) - M_n) \text{ is bounded}\}$$

has probability one, since (M_n) is also convergent. Therefore, we obtain by (R_1) and (R_3) that, on $A_x \cap B_x$,

$$v_n \geq \rho_n(x) \geq \inf_{k \geq 1} \gamma(\hat{\xi}_k(x) - M_k, \sigma_k^2(x)) > 0.$$

Since $\mathsf{P}(B_x) = 1$, it follows from Lemma 13 that $\mathsf{P}(A_x) = 0$.

Finally, let $\tilde{\mathbb{X}}$ be a countable dense subset of $\mathbb{X}$ and let $\Omega_0 = \Omega \setminus \bigcup_{x \in \tilde{\mathbb{X}}} A_x$. Then $\mathsf{P}(\Omega_0) = 1$ and it is straightforward to see that for each $\omega \in \Omega_0$, the set $\{X_n(\omega), n \geq 1\}$ is dense in $\mathbb{X}$. □

6. Discussion

Since Jones et al. (1998), the expected improvement (EI) algorithm has become a very popular algorithm to optimize an expensive-to-evaluate function. Such functions are often encountered in industrial problems, where the function value may be the output of a complex computer simulation, or the result of costly measurements on prototypes. A body of empirical studies, based on optimization test-beds and real applications, have shown that the EI algorithm can lead to significant evaluation savings over traditional optimization methods (see, e.g., Jones, 2001; Huang et al., 2006; Forrester et al., 2008). Yet, making use of an optimization algorithm without knowing its convergence properties is not satisfying, not only theoretically, but also from a practical viewpoint. Indeed, if it turned out that the EI algorithm could not get arbitrarily close to a global optimizer when the number of function evaluations increases, using this algorithm on a restricted budget of function evaluations would hardly be justified.

In this paper, we have provided two important results. The first one is that the EI improvement algorithm behaves consistently provided that the objective function belongs to the reproducing kernel Hilbert space (RKHS) attached to ξ, under a non-degeneracy assumption on the covariance function that we have called the no-empty-ball (NEB) property. This result is obviously interesting from a theoretical viewpoint; it is less so in practice because one seldom knows in advance whether the objective function belongs to a given RKHS. The second main result of this paper, which states that convergence also takes place for P−almost all continuous functions, where P is the (prior) probability distribution of the Gaussian process ξ, is what really matters from a practical point of view.

These results constitute a first step toward a deeper understanding of global optimization algorithms based on the EI criterion, or more generally on criterions satisfying (9). Possible directions for future research include the derivation of pathwise or average convergence rates, the convergence of the algorithm when some parameters of the covariance are re-estimated after each new evaluation, and the extension—possibly under more restrictive assumptions—of our convergence results to all continuous functions.

References

Aronszajn, N., 1950. Theory of reproducing kernels. Trans. Amer. Math. Soc. 68, 337–404.

Bogachev, V.I., 1998. Gaussian Measures. Mathematical Surveys and Monographs, Vol. 62. American Mathematical Society, Providence, RI.

Chilès, J.-P., Delfiner, P., 1999. Geostatistics: Modeling Spatial Uncertainty. Wiley, New York.

Currin, C., Mitchell, T., Morris, M., Ylvisaker, D., 1991. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. J. Amer. Statist. Assoc. 953–963.

Forrester, A., Sóbester, A., Keane, A., 2008. Engineering Design via Surrogate Modelling. Wiley, Chichester.

Huang, D., Allen, T., Notz, W., Zeng, N., 2006. Global optimization of stochastic black-box systems via sequential kriging meta-models. J. Global Optim. 34, 441–466.

Jones, D., 2001. A taxonomy of global optimization methods based on response surfaces. J. Global Optim. 21, 345–383.

Jones, D.R., Schonlau, M., Welch, W.J., 1998. Efficient global optimization of expensive black-box functions. J. Global Optim. 13, 455–492.

Locatelli, M., 1997. Bayesian algorithms for one-dimensional global optimization. J. Global Optim. 10, 57–76.

Lukic, M.N., Beder, J.H., 2001. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. Trans. Amer. Math. Soc. 353 (10), 3945–3969.

Mockus, J., 1989. Bayesian Approach to Global Optimization: Theory and Applications. Kluwer Academic Publishers, Dordrecht, Boston, London.

Mockus, J., Tiesis, V., Zilinskas, A., 1978. The application of Bayesian methods for seeking the extremum. In: Dixon, L., Szego, G. (Eds.), Towards Global Optimization, vol. 2. North Holland, New York, pp. 117–129.

Molchanov, I.S., 2005. Theory of Random Sets. Springer, London.

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P., 1989. Design and analysis of computer experiments. Statist. Sci. 4 (4), 409–435.

Schonlau, M., 1997. Computer experiments and global optimization. Ph.D. Thesis, University of Waterloo, Waterloo, Ontario, Canada.

Schonlau, M., Welch, W.J., 1996. Global optimization with nonparametric function fitting. In: Proceedings of the ASA, Section on Physical and Engineering Sciences. Amer. Statist. Assoc., pp. 183–186.

Schonlau, M., Welch, W.J., Jones, D.R., 1997. A data analytic approach to Bayesian global optimization. In: Proceedings of the ASA, Section on Physical and Engineering Sciences, Amer. Statist. Assoc., pp. 186–191.

Stein, M.L., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Steinwart, I., 2001. On the influence of the kernel on the consistency of support vector machines. J. Mach. Learn. Res. 2, 67–93.

Törn, A., Zilinskas, A., 1989. Global Optimization. Springer, Berlin.

Vazquez, E., Bect, J., 2010. Pointwise consistency of the kriging predictor with known mean and covariance functions. In: Giovagnoli, A., Atkinson, A.C., Torsney, B. (Eds.), mODa 9—Advances in Model-Oriented Design and Analysis, Proceedings of Ninth International Workshop in Model-Oriented Design and Analysis, Springer, Bertinoro, Italy.
Williams, D., 1991. Probability with Martingales. Cambridge University Press, Cambridge.
Zilinskas, A., 1992. A review of statistical models for global optimization. J. Global Optim. 2, 145–153.

3.4 Informational approach to global optimization

This article in the Journal of Global Optimization (2009) presents a Bayesian approach to global optimization where the experiments are chosen in order to minimize the amount of uncertainty in the posterior distribution of the location of the optimum. To quantity this uncertainty, we use Shannon's entropy. This article also discusses the problem of noisy evaluation results.

J Glob Optim (2009) 44:509–534
DOI 10.1007/s10898-008-9354-2

An informational approach to the global optimization of expensive-to-evaluate functions

Julien Villemonteix · Emmanuel Vazquez · Eric Walter

Received: 3 July 2006 / Accepted: 12 August 2008 / Published online: 26 September 2008

Abstract In many global optimization problems motivated by engineering applications, the number of function evaluations is severely limited by time or cost. To ensure that each evaluation contributes to the localization of good candidates for the role of global minimizer, a sequential choice of evaluation points is usually carried out. In particular, when Kriging is used to interpolate past evaluations, the uncertainty associated with the lack of information on the function can be expressed and used to compute a number of criteria accounting for the interest of an additional evaluation at any given point. This paper introduces minimizers entropy as a new Kriging-based criterion for the sequential choice of points at which the function should be evaluated. Based on *stepwise uncertainty reduction*, it accounts for the informational gain on the minimizer expected from a new evaluation. The criterion is approximated using conditional simulations of the Gaussian process model behind Kriging, and then inserted into an algorithm similar in spirit to the *Efficient Global Optimization* (EGO) algorithm. An empirical comparison is carried out between our criterion and *expected improvement*, one of the reference criteria in the literature. Experimental results indicate major evaluation savings over EGO. Finally, the method, which we call IAGO (for Informational Approach to Global Optimization), is extended to robust optimization problems, where both the factors to be tuned and the function evaluations are corrupted by noise.

Keywords Gaussian process · Global optimization · Kriging · Robust optimization · Stepwise uncertainty reduction

J. Villemonteix
Energy Systems Department, Renault S.A., 78298 Guyancourt, France

E. Vazquez ( )
SUPELEC, 91192 Gif-sur-Yvette, France
e-mail: emmanuel.vazquez@supelec.fr

E. Walter
Laboratoire des Signaux et Systèmes, CNRS-SUPELEC-Univ Paris-Sud, 91192 Gif-sur-Yvette, France

1 Introduction

This paper is devoted to global optimization in a context of expensive function evaluation. The objective is to find global minimizers in $\mathbb{X}$ (the factor space, a bounded subset of $\mathbb{R}^d$) of an unknown function $f : \mathbb{X} \to \mathbb{R}$, using a very limited number of function evaluations. Note that the global minimizer may not be unique (any global minimizer will be denoted as $\boldsymbol{x}^*$). Such a problem is frequently encountered in the industrial world. For instance, in the automotive industry, optimal crash-related parameters are obtained using costly real tests and time-consuming computer simulations (a single simulation of crash-related deformations may take up to 24 h on dedicated servers). It then becomes essential to favor optimization methods that use the dramatically scarce information as efficiently as possible.

To make up for the lack of knowledge on the function, surrogate (also called meta or approximate) models are used to obtain cheap approximations [13]. They turn out to be convenient tools for visualizing the function behavior or suggesting the location of an additional point at which f should be evaluated in the search for $\boldsymbol{x}^*$. Surrogate models based on Gaussian processes have received particular attention. Known in geostatistics under the name of *Kriging* since the early 1960s [15], Gaussian process models provide a probabilistic framework to account for the uncertainty stemming from the lack of information on the system. When dealing with an optimization problem, this framework allows the set of function evaluations to be chosen efficiently [12–14].

In this context, several strategies have been proposed, with significant advantages over traditional optimization methods when confronted to expensive-to-evaluate functions. Most of them *implicitly* seek a likely value for $\boldsymbol{x}^*$, and then assume it to be a suitable location for a new evaluation of f. Yet, given existing evaluation results, the most likely location of a global minimizer is not necessarily a good evaluation point to improve our knowledge on $\boldsymbol{x}^*$. As we shall show, by making full use of Kriging, it is instead possible to *explicitly* estimate the probability distribution of the optimum location, which allows an information-based search strategy.

Based on these observations, the present paper introduces minimizers entropy as a criterion for the choice of new evaluation points. This criterion, directly inspired from *stepwise uncertainty reduction* [9], is then inserted in an algorithm similar to the *Efficient Global Optimization* (EGO) algorithm [14]. We call the resulting algorithm IAGO, for *Informational Approach to Global Optimization*.

Sect. 2 recalls the principle of Kriging-based optimization, along with some general ideas on Gaussian process modeling that are used in Sect. 3 to build an estimate of the distribution of the global minimizers. Sect. 4 details the stepwise uncertainty reduction approach applied to global optimization, while Sect. 5 describes the corresponding algorithm and its extensions to noisy problems. Sect. 6 illustrates the behavior of the new algorithm on some simple benchmark problems, along with its performances compared with those of the classical EGO algorithm, chosen for its good compromise between local and global search [17]. Finally, after a conclusion section and to make this paper self-contained, Sect. 8 recalls, as an appendix, some more results on Gaussian process modeling and Kriging.

2 Kriging-based global optimization

When dealing with expensive-to-evaluate functions, optimization methods based on probabilistic surrogate models (and Kriging in particular) have significant advantages over traditional optimization techniques, as they require fewer function evaluations to provide an acceptable

solution. Kriging provides not only a cheap approximation of the function but also an estimate of the potential error in this approximation. Numerous illustrations of this superiority can be found in the literature (see, for instance, [6]) and many variations have been explored (for extensive surveys, see [13] and [17]). As explained in this section, these methods deal with the cost of evaluation using an adaptive sampling strategy, replacing the optimization of the expensive-to-evaluate function f by a series of optimizations of a cheap criterion.

2.1 Gaussian process modeling and Kriging

This section briefly recalls the principle of Gaussian process (GP) modeling, and lays down the necessary notation. A more detailed presentation is available in the appendix (Sect. 8).

When modeling with Gaussian processes, the function f is assumed to be a sample path of a Gaussian random process F, with mean function $m(\boldsymbol{x})$ and covariance function $k(\cdot,\cdot)$ defined over $\mathbb{X}^2$. If we denote $(\Omega, \mathcal{A}, \mathcal{P})$ the underlying probability space, this amounts to assuming that $\exists\, \omega \in \Omega$, such that $F(\omega,\cdot) = f(\cdot)$. Whenever possible, we shall omit the dependence of F in ω to simplify notation.

In particular, given a set of n evaluation points $\mathbb{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ (the *design*), $\forall \boldsymbol{x}_i \in \mathbb{S}$ the evaluation result $f(\boldsymbol{x}_i)$ is viewed as a sample value of the random variable $F(\boldsymbol{x}_i)$. Kriging computes an unbiased linear predictor of $F(\boldsymbol{x})$ in the vector space $\mathbb{H}_{\mathbb{S}} = \text{span}\{F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)\}$, which can be written as

$$\hat{F}(\boldsymbol{x}) = \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{F}_{\mathbb{S}}, \tag{1}$$

with $\boldsymbol{F}_{\mathbb{S}} = [F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)]^{\mathsf{T}}$, and $\boldsymbol{\lambda}(\boldsymbol{x})$ the vector of Kriging coefficients for the prediction at $\boldsymbol{x}$.

Given the covariance function of F, the Kriging coefficients can be computed along with the variance of the prediction error

$$\hat{\sigma}^2(\boldsymbol{x}) = \text{var}(\hat{F}(\boldsymbol{x}) - F(\boldsymbol{x})). \tag{2}$$

The covariance function of F is chosen within a parametrized class (for instance, the Matérn class), and its parameters are either estimated from the data or chosen a priori (see Sect. 8.3.2 for details on the choice of a covariance function).

Once f has been evaluated at all evaluation points in $\mathbb{S}$, the predicted value of f at $\boldsymbol{x}$ is given by

$$\hat{f}(\boldsymbol{x}) = \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{f}_{\mathbb{S}}, \tag{3}$$

with $\boldsymbol{f}_{\mathbb{S}} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]^{\mathsf{T}}$ ($\boldsymbol{f}_{\mathbb{S}}$ is viewed as a sample value of $\boldsymbol{F}_{\mathbb{S}}$). The same results could be derived in a Bayesian framework, where $F(\boldsymbol{x})$ is Gaussian conditionally to the evaluations carried out ($\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}$), with mean $\hat{f}(\boldsymbol{x})$ and variance $\hat{\sigma}^2(\boldsymbol{x})$.

Note that the random processes $F(\boldsymbol{x})$ and $\hat{F}(\boldsymbol{x})$ satisfy

$$\forall\, \boldsymbol{x}_i \in \mathbb{S},\quad \hat{F}(\boldsymbol{x}_i) = F(\boldsymbol{x}_i), \tag{4}$$

and that the prediction at $\boldsymbol{x}_i \in \mathbb{S}$ is $f(\boldsymbol{x}_i)$. When f is assumed to be evaluated exactly, Kriging is thus an interpolation, with the considerable advantage over other interpolation methods that it also provides an explicit characterization of the prediction error (zero-mean Gaussian with variance $\hat{\sigma}^2(\boldsymbol{x})$).

2.2 Adaptive sampling strategies

The general principle of optimization using Kriging is iteratively to evaluate f at a point that optimizes a criterion based on the model obtained using previous evaluation results. The simplest approach would be to choose a minimizer of the prediction $\hat{f}$ as a new evaluation point. However, by doing so, too much confidence would be put in the current prediction and search is likely to stall on a local optimum (as illustrated by Fig. 1). To compromise between local and global search, more emphasis has to be put on the prediction error, which can indicate locations where additional evaluations are needed to improve confidence in the model. This approach has led to a number of criteria to select additional evaluation points based on both prediction and prediction error.

A standard example of such a criterion is *expected improvement* (EI) [18]. As the name suggests, it involves computing how much improvement in the optimum is expected, if f is evaluated at a given additional point. Let $f_{\min}$ be the best function value obtained so far. The improvement expected from an additional evaluation of f at $\boldsymbol{x}$ given $\boldsymbol{f}_{\mathbb{S}}$, the results of past evaluations, can then be expressed as

$$\mathrm{EI}(\boldsymbol{x}) = \mathbb{E}\left[\max\left(f_{\min} - F\left(\boldsymbol{x}\right), 0\right) | \boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}\right].$$

Since $F(\boldsymbol{x})$ is conditionally Gaussian with mean $\hat{f}(\boldsymbol{x})$ and variance $\hat{\sigma}^2(\boldsymbol{x})$,

$$\mathrm{EI}(\boldsymbol{x}) = \hat{\sigma}(\boldsymbol{x})\left[u\Phi(u) + \frac{d\Phi}{du}(u)\right], \tag{5}$$

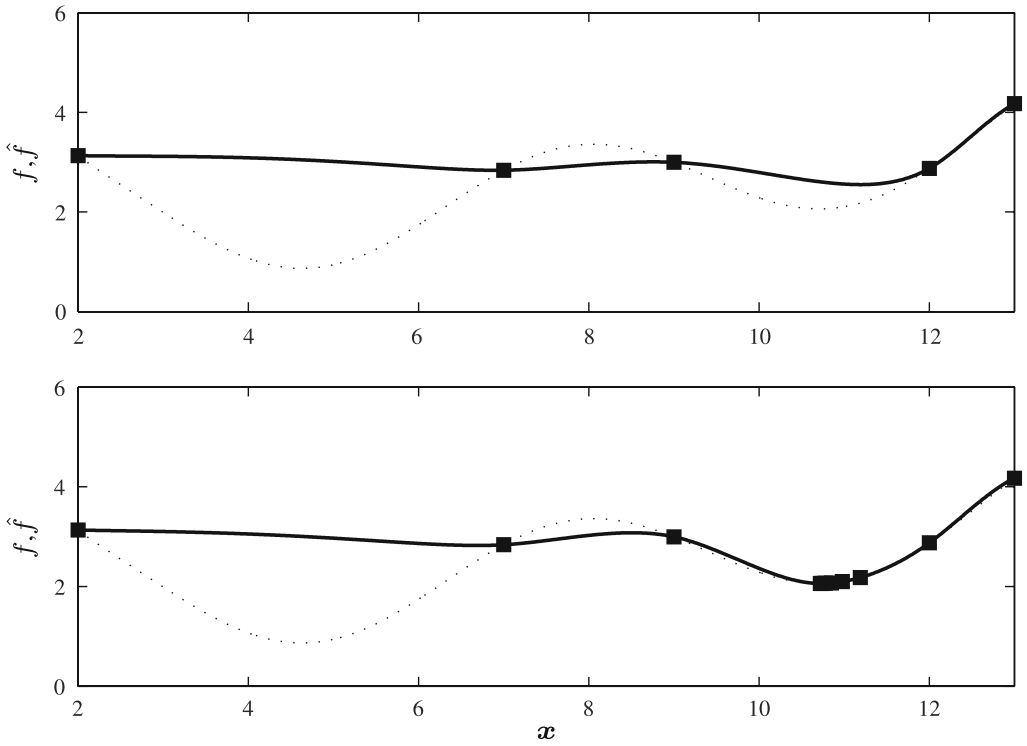


Fig. 1 Naive approach to optimization using Kriging: (*top*) prediction $\hat{f}$ (*bold line*) of the true function f (*dotted line*, supposedly unknown) obtained from an initial design materialized by squares; (*bottom*) prediction after seven iterations minimizing $\hat{f}$

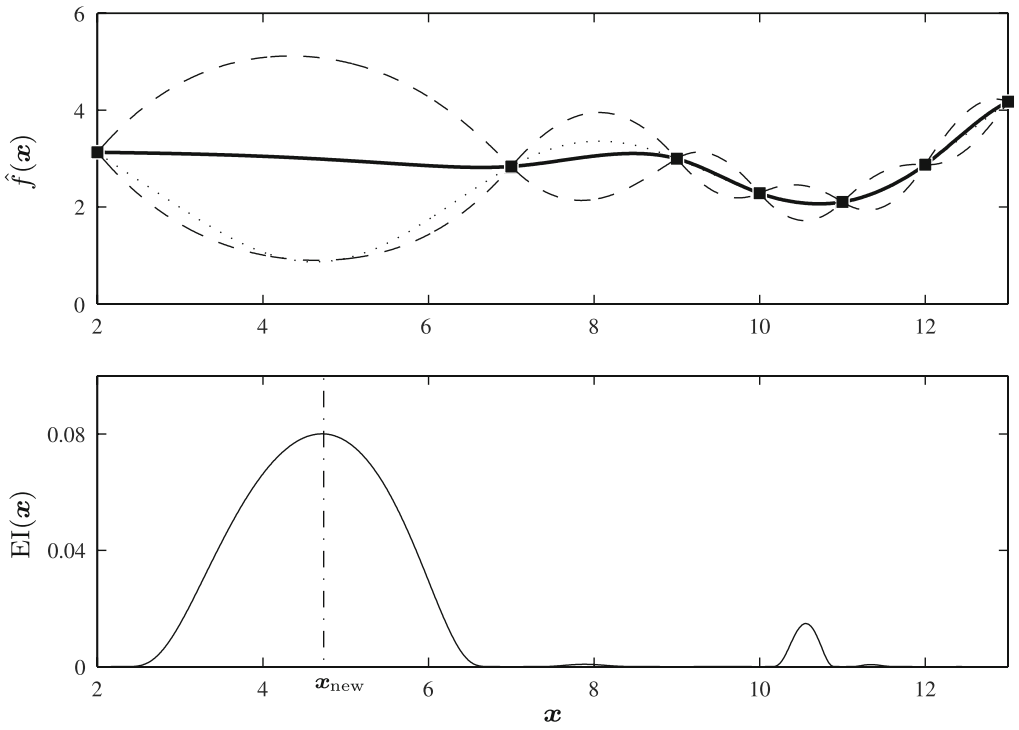


Fig. 2 EI approach to optimization using Kriging: (*top*) prediction $\hat{f}$ (*bold line*), 95% confidence intervals computed using $\hat{\sigma}$ (*dashed line*) and true function f (*dotted line*); (*bottom*) expected improvement

with

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

and Φ the normal cumulative distribution function. The new evaluation point is then chosen as a global maximizer of EI($\boldsymbol{x}$). An example is given on Fig. 2, where the problem that deceived the naive method of Fig. 1 is directly solved with the EI criterion. This method has been used for computer experiments in [17], while modified criteria have been used in [11] and [26] to deal with noisy functions.

In [13] and [24], a fair number of alternative criteria are presented and compared. Although quite different in their formulation, they generally aim to answer the same question: What is the most likely position of $\boldsymbol{x}^*$? Another, and probably more relevant, question is: Where should the evaluation be carried out optimally to improve knowledge on the global minimizers?

In what follows, a criterion that addresses this question will be presented, along with its performances. The reference for comparison will be EI, which is a reasonable compromise between local and global search [17], and has been successfully used in many applications.

3 Estimating the distribution of x^*

Once a Kriging surrogate model $\hat{f}$ has been obtained, any global minimizer of $\hat{f}$ is a natural approximation of $\boldsymbol{x}^*$. However, it might be excessively daring to trust this approximation as it does not take in account the uncertainty of the prediction. A more cautious approach to

estimating $\boldsymbol{x}^*$ is to use the probabilistic framework associated with F. Of course, $\boldsymbol{x}^*$ is not necessarily unique, and we shall focus on describing the set of all global minimizers of f as efficiently as possible.

3.1 Probabilistic modeling of the global minimizers of f

According to the GP model, a global minimizer $\boldsymbol{x}^*$ of f corresponds to a global minimizer of this particular sample path of F. It seems therefore natural to use the GP model of f to obtain a probabilistic model for $\boldsymbol{x}^*$. Consider the *random* set $\mathcal{M}^*_{\mathbb{X}}$ of the global minimizers of F over $\mathbb{X}$, i.e., the set of all global minimizers for each sample path, which for any $\omega \in \Omega$ can be written as

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

To ensure that $\mathcal{M}^*_{\mathbb{X}}(\omega)$ is not empty for all ω, we assume that F has continuous sample paths with probability one. This continuity can be ensured through a proper choice of covariance function (see, e.g., [1]).

Let $\boldsymbol{X}^*$ be a random vector uniformly distributed on $\mathcal{M}^*_{\mathbb{X}}$ (from now on, we omit the dependency of $\mathcal{M}^*_{\mathbb{X}}$ in ω). The probability density function of this random vector conditional to past evaluation results, that we shall thereafter call conditional density of the global minimizers and denote $p_{\boldsymbol{X}^*|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x})$, is of great interest, as it allows one not only to estimate the global minimizers of f (for example, through the maximization of their conditional density), but also to characterize the uncertainty associated with this estimation. In fact, $p_{\boldsymbol{X}^*|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x})$ contains all of what has been assumed and learned about the system. However, no tractable analytical expression for $p_{\boldsymbol{X}^*|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x})$ is available [2,19]. To overcome this difficulty, the approach taken here is to consider a discrete version of the conditional distribution, and to approximate it using Monte Carlo simulations.

Let $\mathbb{G} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be a finite subset of $\mathbb{X}$, $\mathcal{M}^*_{\mathbb{G}}$ be the random set of global minimizers of F over $\mathbb{G}$, and $\boldsymbol{X}^*_{\mathbb{G}}$ be a random vector uniformly distributed on $\mathcal{M}^*_{\mathbb{G}}$. The conditional probability mass function of $\boldsymbol{X}^*_{\mathbb{G}}$ given $\boldsymbol{f}_{\mathbb{S}}$ (or simply minimizers distribution) is then $\forall \boldsymbol{x} \in \mathbb{G}$

$$P_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x}) = \mathrm{P}(\boldsymbol{X}^*_{\mathbb{G}} = \boldsymbol{x} \mid \boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}).$$

It can be approximated using conditional simulations, i.e., simulations of F that satisfy $\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}$. Assuming that non-conditional simulations are available, several methods exist to make them conditional [4]. Conditioning by Kriging seems the most promising of them in the present context and will be presented in the next section.

To keep the presentation simple, we assume in what follows that $\mathbb{S} \subset \mathbb{G}$.

3.2 Conditioning by Kriging

This method, due to G. Matheron, uses the unbiasedness of the Kriging prediction to transform non-conditional simulations into simulations interpolating the results $\boldsymbol{f}_{\mathbb{S}}$ of the evaluations. The idea is to sample from the conditional distribution of the prediction error $F - \hat{F}$ rather than from the conditional distribution of F, which is made easier by the fact that the statistical properties of the prediction error do not depend on the result of the evaluations, nor on the mean $m(\boldsymbol{x})$ of $F(\boldsymbol{x})$.

To present this more formally, let Z be a zero-mean Gaussian process with covariance function k (the same as that of F) and $\hat{Z}$ be its Kriging predictor based on the random variables

$Z(\boldsymbol{x}_i),\ \boldsymbol{x}_i \in \mathbb{S}$, and consider the random process

$$T(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}) + \left[Z(\boldsymbol{x}) - \hat{Z}(\boldsymbol{x}) \right], \tag{6}$$

where $\hat{f}$ is the mean of the Kriging predictor based on the design points in $\mathbb{S}$. Since this Kriging predictor is an interpolator, at evaluation points in $\mathbb{S}$, we have $\hat{f}(\boldsymbol{x}_i) = f(\boldsymbol{x}_i)$. Equation 4 implies that $Z(\boldsymbol{x}_i) = \hat{Z}(\boldsymbol{x}_i)$, which leads to $T(\boldsymbol{x}_i) = f(\boldsymbol{x}_i),\ \forall \boldsymbol{x}_i \in \mathbb{S}$. In other words, T is such that all its sample paths interpolate the known values of f. It is then easy to check that T has the same finite-dimension distributions as F conditionally to past evaluation results [7], simply because the prediction error $Z - \hat{Z}$, for Z, has the same distribution as the prediction error for F, $F - \hat{F}$. Note that the same vector $\boldsymbol{\lambda}(\boldsymbol{x})$ of Kriging coefficients is used to interpolate the data and the simulations at design points. Using (3), one can rewrite (6) as

$$T(\boldsymbol{x}) = Z(\boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \left[\boldsymbol{f}_{\mathbb{S}} - \boldsymbol{Z}_{\mathbb{S}} \right], \tag{7}$$

with $\boldsymbol{Z}_{\mathbb{S}} = [Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)]^{\mathsf{T}}$.

In summary, to simulate F over $\mathbb{G}$ conditionally to past evaluation results $\boldsymbol{f}_{\mathbb{S}}$, we can simulate a zero-mean Gaussian process Z over $\mathbb{G}$, compute the prediction error for each simulation and shift the prediction error around the desired mean $\hat{f}$. This is achieved by the following procedure (illustrated on Fig. 3):

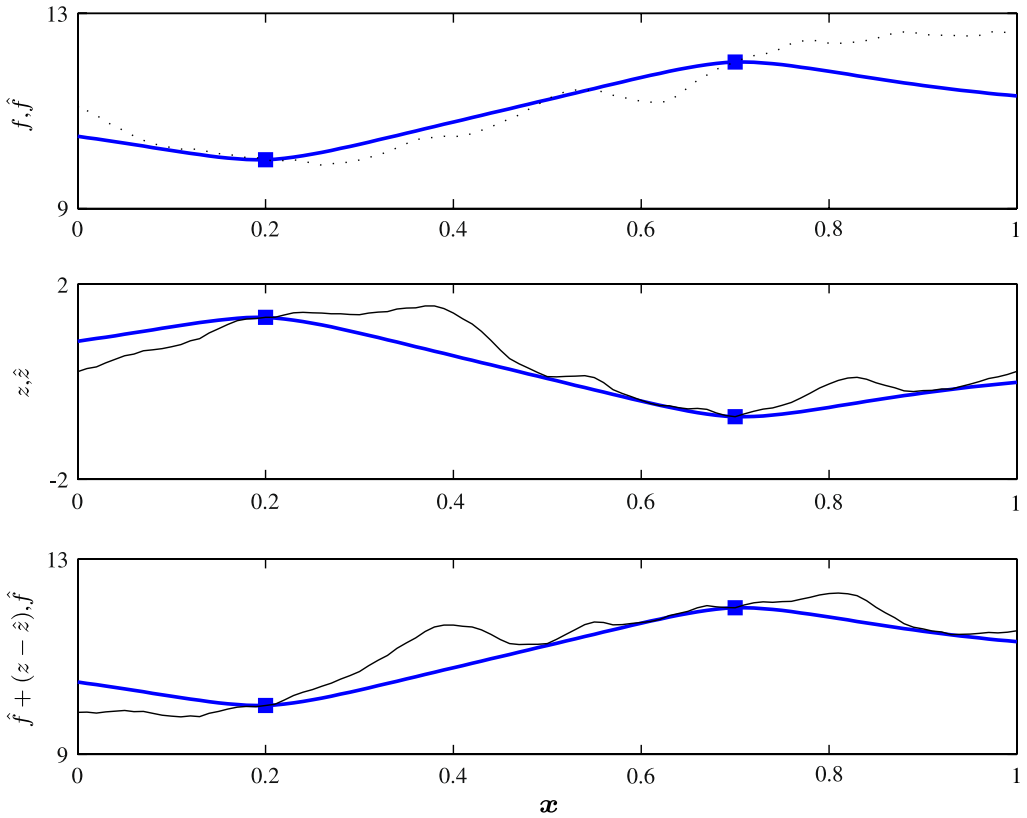


Fig. 3 Conditioning a simulation: (*top*) unknown real curve f (*doted line*), sample points (*squares*) and associated Kriging prediction $\hat{f}$ (*bold line*); (*middle*) non-conditional simulation z, sample points and associated Kriging prediction $\hat{z}$ (*bold line*); (*bottom*) the simulation of the Kriging error $z - \hat{z}$ is picked up from the non-conditional simulation and added to the Kriging prediction to get the conditional simulation (*thin line*)

- compute, for every point in $\mathbb{G}$, the vector of Kriging coefficients based on the design points in $\mathbb{S}$,
- compute the Kriging prediction $\hat{f}(\boldsymbol{x})$ based on past evaluation results $\boldsymbol{f}_{\mathbb{S}}$ for every $\boldsymbol{x}$ in $\mathbb{G}$,
- collect non-conditional sample paths of Z over $\mathbb{G}$ (provided that a Gaussian sampler is available, setting the proper covariance for the simulated vector can be achieved using, for example, the Cholesky decomposition),
- apply (7) for each non conditional simulation and at every point in $\mathbb{G}$. That is, to generate $t(\boldsymbol{x})$, a conditional simulation of $T(\boldsymbol{x})$ from a non-conditional simulation $z(\boldsymbol{x})$ of $Z(\boldsymbol{x})$, apply

$$t(\boldsymbol{x}) = z(\boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}}[\boldsymbol{f}_{\mathbb{S}} - \boldsymbol{z}_{\mathbb{S}}], \tag{8}$$

  where $\boldsymbol{z}_{\mathbb{S}}$ is the sampled valued of Z over $\mathbb{S}$, which is available since $\mathbb{S} \subset \mathbb{G}$.

With this sampling method, it becomes straightforward to estimate $P_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}$. Let $\boldsymbol{x}_i^*$ be a global minimizer of the i-th conditional simulation ($i = 1, \ldots, r$) over $\mathbb{G}$ (if it is not unique, choose one randomly). Then, for any $\boldsymbol{x}$ in $\mathbb{G}$, a classical estimator is

$$\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x}) = \frac{1}{r}\sum_{i=1}^{r} \delta_{\boldsymbol{x}_i^*}(\boldsymbol{x}), \tag{9}$$

with δ the Kronecker symbol. Figure 4 presents the approximation $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}$ for an example where locating a global minimizer is not easy. Knowing the conditional distribution of $\boldsymbol{X}^*_{\mathbb{G}}$ gives valuable information on the areas of $\mathbb{X}$ where a global minimizer might be located, and that ought to be investigated. This idea will be detailed in the next section.

4 The stepwise uncertainty reduction strategy

The knowledge about the global minimizers of f is summarized by $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}$. In order to evaluate the interest of a new evaluation of f at a given point, a measure of the expected information gain is required. An efficient measure is *conditional entropy*, as used in sequential testing [9] in the *Stepwise Uncertainty Reduction* (SUR) strategy. This section extends the SUR strategy to global optimization.

4.1 Conditional entropy

The entropy of a discrete random variable U (expressed in bits) is defined as:

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

$H(U)$ measures the spread of the distribution of U. It decreases as this distribution gets more peaked. In particular:

- $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x}) = 1/N \;\; \forall \boldsymbol{x} \in \mathbb{G} \;\Rightarrow\; H(\boldsymbol{X}^*_{\mathbb{G}}) = \log_2(N)$,
- $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}}}(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \boldsymbol{x} \neq \boldsymbol{x}_0 \\ 1 & \text{if } \boldsymbol{x} = \boldsymbol{x}_0 \end{cases} \;\Rightarrow\; H(\boldsymbol{X}^*_{\mathbb{G}}) = 0$

Similarly, for any event $\mathcal{B}$, the entropy of U relative to the probability measure $\mathrm{P}(.|\mathcal{B})$ is

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

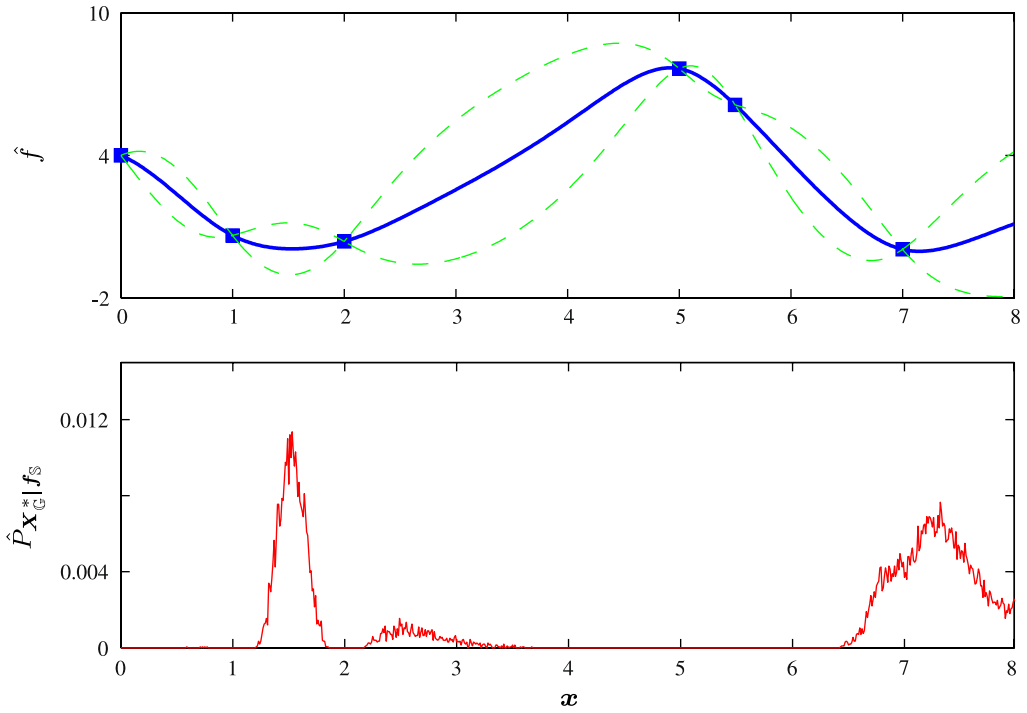


Fig. 4 Estimation of the distribution of $\boldsymbol{X}^*_{\mathbb{G}}$: (*top*) Kriging interpolation, 95% confidence intervals and sample points; (*bottom*) estimated distribution of $\boldsymbol{X}^*_{\mathbb{G}}$ using 10000 conditional simulations of F and a regular grid for $\mathbb{G}$

The conditional entropy of U given another discrete random variable V is

$$H(U|V) = \sum_v \mathrm{P}(V = v) H(U|V = v),$$

and the conditional entropy of U given $\mathcal{B}$ and V is

$$H(U|\mathcal{B}, V) = \sum_v \mathrm{P}(V = v|\mathcal{B}) H(U|\mathcal{B}, V = v). \tag{10}$$

Note that $H(U|V)$ and $H(U|\mathcal{B}, V)$ are, despite the similarity of notation with conditional expectation, deterministic quantities. More details on conditional entropy can be found in [5].

4.2 Conditional minimizers entropy

Let $F_Q(\boldsymbol{x})$ be a discrete version of $F(\boldsymbol{x})$, defined as $F_Q(\boldsymbol{x}) = Q(F(\boldsymbol{x}))$ with Q a quantization operator. Q is characterized by a finite set of M real numbers $\{y_1, \ldots, y_M\}$, and defined $\forall u \in \mathbb{R}$ as

$$Q(u) = y_k \text{ with } k = \arg\min_i |y_i - u|. \tag{11}$$

For optimization problems, the SUR strategy for the selection of the next value of $\boldsymbol{x} \in \mathbb{X}$ at which f will be evaluated will be based on $H(\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}, F_Q(\boldsymbol{x}))$, the conditional entropy of $\boldsymbol{X}^*_{\mathbb{G}}$ given the evaluation results $\{\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}\}$ and $F_Q(\boldsymbol{x})$ (we shall refer to it later on as conditional entropy of the minimizers, or simply minimizers entropy).

Using (10) we can write

$$H(\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}, F_Q(\boldsymbol{x})) \;=\; \sum_{i=1}^{M} \mathrm{P}(F_Q(\boldsymbol{x}) = y_i|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}) H(\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}, F_Q(\boldsymbol{x}) = y_i) \tag{12}$$

with

$$H(\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}, F_Q(\boldsymbol{x}) = y_i) \;=\; -\sum_{\mathbf{u}\in\mathbb{G}} P_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}},y_i}(\mathbf{u}) \log_2 \; P_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}},y_i}(\mathbf{u}),$$

and

$$P_{\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{f}_{\mathbb{S}},y_i}(\boldsymbol{u}) = \mathrm{P}(\boldsymbol{X}^* = \boldsymbol{u}|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}, F_Q(\boldsymbol{x}) = y_i).$$

$H(\boldsymbol{X}^*_{\mathbb{G}}|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}, F_Q(\boldsymbol{x}))$ is a measure of the anticipated uncertainty remaining in $\boldsymbol{X}^*_{\mathbb{G}}$ given the candidate evaluation point $\boldsymbol{x}$ and the result $\boldsymbol{f}_{\mathbb{S}}$ of the previous evaluations. Anticipation is introduced in (12) by considering the entropy of $\boldsymbol{X}^*_{\mathbb{G}}$ resulting from every possible sample value of $F_Q(\boldsymbol{x})$. At each stage of the iterative optimization, the SUR strategy retains for the next evaluation a point that minimizes the expected entropy of the minimizers distribution after the evaluation, i.e., a point that maximizes the expected gain in information about $\boldsymbol{X}^*_{\mathbb{G}}$.

The conditional entropy of the minimizers thus takes in account the conditional statistical properties of F and particularly the covariance function of the model. There lies the interest of the SUR strategy applied to global optimization. It makes use of what has been previously assumed and learned about f to pick up the most informative evaluation point. By contrast, the EI criterion (as most standard criteria) depends only on the conditional mean and variance of F at the design point being considered.

5 Implementing the SUR strategy

5.1 IAGO algorithm

Our algorithm is similar in spirit to the strategy for Kriging-based optimization known as *Efficient Global Optimization* (EGO) [14]. EGO starts with a small initial design, estimates the parameters of the covariance function of F and computes the Kriging model. Based on this model, an additional point is selected in the design space to be the location of the next evaluation of f using the EI criterion. The parameters of the covariance function are then re-estimated, the model re-computed, and the process of choosing new points continues until the improvement expected from sampling additional points has become sufficiently small. The IAGO algorithm uses the same idea of iterative incorporation of the obtained information to the prior on the function, but with a different criterion.

To compute the minimizers entropy using (12), a new quantization operator Q_x is used for each value of $\boldsymbol{x}$ to improve the precision with which the empirical mean of entropy reduction over possible evaluation results is computed. We use the fact that $F(\boldsymbol{x})$ is conditionally Gaussian with mean $\hat{f}(\boldsymbol{x})$ and variance $\hat{\sigma}^2(\boldsymbol{x})$ obtained by Kriging, to select a set of values $\{y_1(\boldsymbol{x}), \ldots, y_M(\boldsymbol{x})\}$, such that

$$P(F_{Q_x}(\boldsymbol{x}) = y_i|\boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}) = \frac{1}{M} \;\forall\, i \in [\![1 : M]\!]\,. \tag{13}$$

Here we used a set of ten possible values ($M = 10$).

Table 1 Selection of a new evaluation point for f

Algorithm
Input: Set $\mathbb{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of evaluation points and corresponding values $\boldsymbol{f}_\mathbb{S}$ of the function f
Output: Additional evaluation point $\boldsymbol{x}_{\text{new}}$
1. Choose $\mathbb{G}$, a discrete representation of $\mathbb{X}$
2. Set covariance parameters either a priori or by maximum-likelihood estimation based on $\boldsymbol{f}_\mathbb{S}$
3. Compute r non-conditional simulations over $\mathbb{G}$
4. Compute $\hat{f}(\boldsymbol{x})$ and $\hat{\sigma}(\boldsymbol{x})$ over $\mathbb{G}$ by Kriging from $\boldsymbol{f}_\mathbb{S}$
5. **while** the set of candidate points has not been entirely explored
6. **do** Take an untried point $\boldsymbol{x}_\text{c}$ in the set of candidate points
7. Compute the parameters $\{y_1, \ldots, y_M\}$ of the quantization operator Q
8. Compute the Kriging coefficients at every point in $\mathbb{G}$ based on evaluation points in $\mathbb{S}$ and $\boldsymbol{x}_\text{c}$
9. **for** $i \leftarrow 1$ **to** M
10. **do** Construct conditional simulations using (7) and assuming that $f(\boldsymbol{x}_\text{c}) = y_i$
11. Find a global minimizer $\boldsymbol{x}_k^*$ of the k-th conditional simulation over $\mathbb{G}$ $(k = 1, \ldots, r)$
12. Estimate $P_{\boldsymbol{X}^*_\mathbb{G}|\boldsymbol{f}_\mathbb{S}, y_i}$ over $\mathbb{G}$ using (9)
13. Compute $H(\boldsymbol{X}^*_\mathbb{G}|\boldsymbol{F}_\mathbb{S} = \boldsymbol{f}_\mathbb{S}, F_Q(\boldsymbol{x}_\text{c}) = y_i)$
14. Compute the minimizers entropy given an evaluation at $\boldsymbol{x}_\text{c}$ using (12)
15. Output $\boldsymbol{x}_{\text{new}}$ that minimizes the conditional entropy over the set of candidate points

For each of these possible values (or hypotheses $F(\boldsymbol{x}) = y_i$), $\hat{P}_{\boldsymbol{X}^*_\mathbb{G}|\boldsymbol{f}_\mathbb{S}, y_i}$ is computed using conditional simulations. The minimizers entropy is then obtained using (12). These operations are carried out on a discrete set of candidate evaluation points (see Sect. 5.2 for some details on the choice of this set), and a new evaluation of f is finally performed at a point that minimizes minimizers entropy. Next, as in the EGO algorithm, the covariance parameters are re-estimated and the model re-computed. The procedure for the choice of an additional evaluation point is described in Table 1.

When the number of additional function evaluations is not specified beforehand, we propose to use as a stopping criterion the conditional probability that the global minimum of the GP model be no further apart of $f_{\min} = \min_{\boldsymbol{x}_i \in \mathbb{S}} f(\boldsymbol{x}_i)$ (the best function value yet obtained) than a given tolerance threshold δ. The algorithm then stops when

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

with $F^* = \min_{\boldsymbol{x} \in \mathbb{G}} F(\boldsymbol{x})$, and $P_{\text{Stop}} \in [0, 1]$ a critical value to be chosen by the user. Proposed in [18], this stopping criterion is well suited here, since evaluating the repartition function of $f(\boldsymbol{x}^*)$ does not require any additional computation. We can indeed use the conditional simulations that have been performed to approximate the conditional distribution of $\boldsymbol{X}^*_\mathbb{G}$ for this purpose, provided that we keep track, for each of them, not only of a global minimizer, but also of the minimum. The histogram thus obtained can then easily be transformed into a simple approximation of the conditional repartition function of the minimum.

5.2 Computational complexity

With the previous notation, n the number of evaluation points, r the number of conditional simulations, N the number of points in $\mathbb{G}$ and M the number of discretized potential evaluation results for an evaluation, the computational complexity for the approximation of the minimizers entropy (Steps 7–14 in Table 1) is as follows:

– computing Kriging coefficients at every point in $\mathbb{G}$ (Step 8): $\mathrm{O}(n^2 N)$, as (20) (to be found in appendix) has to be solved N times. The covariance matrix can be factorized, and Kriging at an untried point is then simply in $\mathrm{O}(n^2)$,

- constructing conditional simulations (Step 10): $O(nrN)$ (M is not involved since the main part of the conditioning procedure described by (8) can be carried out outside the loop on the discretized potential evaluation results),
- locating the global minimizers for each simulation by exhaustive search (Step 11): $O(rNM)$.

Since all other operations are in $O(N)$ at most, evaluating minimizers entropy at any given point requires $O(N)$ operations.

To complete the description of an implementable algorithm, we must specify a choice for $\mathbb{G}$ and a policy for the minimization of minimizers entropy. What follows is just an example of a possible strategy, and many variants could be considered.

The simplest choice for $\mathbb{G}$ is a uniform grid on $\mathbb{X}$. However, as the number of evaluations of f increases, the spread of $P_{\boldsymbol{X}^*_{\mathbb{G}}|f_{\mathbb{S}}}$ diminishes along with the precision for the computation of the entropy. To keep a satisfactory precision over time, $\mathbb{G}$ can be a random sample of points in $\mathbb{X}$, re-sampled after every evaluation of f with the distribution $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|f_{\mathbb{S}}}$. Re-sampling makes it possible to use a set $\mathbb{G}$ with a smaller cardinal and to escape, at least partly, the curse of dimensionality (to resample using $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|f_{\mathbb{S}}}$, any non-parametric density estimator could be used along with a sampling method such as Metropolis-Hastings, see, e.g., [3]).

Ideally, to choose an additional evaluation point for f using IAGO, minimizers entropy should be minimized over $\mathbb{X}$. However, this of course is in itself a global optimization problem, with many local optima. It would be possible to design an ad-hoc optimization method (as in [13]), but this perspective is not explored here. Instead, we evaluate the criterion extensively over a chosen set of candidate points. Note that only the surrogate model is involved at this stage, which makes the approach practical. The idea is, exactly as for the choice of $\mathbb{G}$, to use a space-filling sample covering $\mathbb{X}$ and resampled after each new evaluation. The current implementation of IAGO simply uses a Latin Hyper Cube (LHC) sample, however, it would be easy to adapt this sample iteratively using the conditional distribution of the minimizers $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|f_{\mathbb{S}}}$ as a prior. For instance, areas of the design space where the distribution is sufficiently small could be ignored. After a few evaluations, a large portion of the design space usually satisfies this property, and the computations saved could be used to improve knowledge on the criterion by sampling where $\hat{P}_{\boldsymbol{X}^*_{\mathbb{G}}|f_{\mathbb{S}}}$ is high (using the same approach as for the choice of $\mathbb{G}$).

As dimension increases, trying to cover the factor space while keeping the same accuracy leads to an exponential increase in complexity. However, in a context of expensive function evaluation, the objective is less to specify exactly all global minimizers (which would be too demanding in function evaluations anyway), than to use available information efficiently to reduce the likely areas for the location of these minimizers. This is exactly the driving concept behind IAGO. In practice, within a set of one thousand candidate points, picking an additional evaluation point requires about 3 min with a standard personal computer (and this figure is relatively independent of the dimension of factor space). Moreover, the result obtained can be trusted to be a consistent choice within this set of candidate points, in regard of what has been assumed and learned about f.

5.3 Taking noise in account

Practical optimization problems often involve noise. This section discusses possible adaptations of the optimization algorithm that make it possible to deal with noisy situations, namely noise on the evaluation of f and noise on the factors.

5.3.1 Noise on the evaluation of f

When the results of the evaluations of f are corrupted by noise, the algorithm must take this fact into account. A useful tool to deal with such situations is *non-interpolative Kriging* (see Sect. 8.2).

If the evaluation at $\boldsymbol{x}_i \in \mathbb{S}$ is assumed to be corrupted by an additive Gaussian noise ε_i with known mean and variance, the Kriging prediction should no longer be interpolative. The optimization algorithm remains nearly unchanged, except for the conditional simulations. Sample paths of F, should be built conditionally to evaluation results, i.e., realizations of the random variables $f(\boldsymbol{x}_i) + \varepsilon_i$ for $\boldsymbol{x}_i \in \mathbb{S}$. Since the variance of the prediction error is no longer zero at evaluation points (in other words, there is some uncertainty left on the values of f at evaluation points), we first have to sample, at each evaluation point, from the distribution of F conditionally to noisy evaluation results. An interpolative simulation, based on these samples, is then built using conditioning by Kriging. An example of such a simulation is presented on Fig. 5 for a noise variance of 0.01.

5.3.2 Noise on the factors

In many industrial design problems, the variability of the values of the factors in mass production has a significant impact on performance. One might then want to design a system that optimizes some performance measure while ensuring that performance uncertainty (stemming from noise on the factors) remains under control. These so-called *robust optimization*

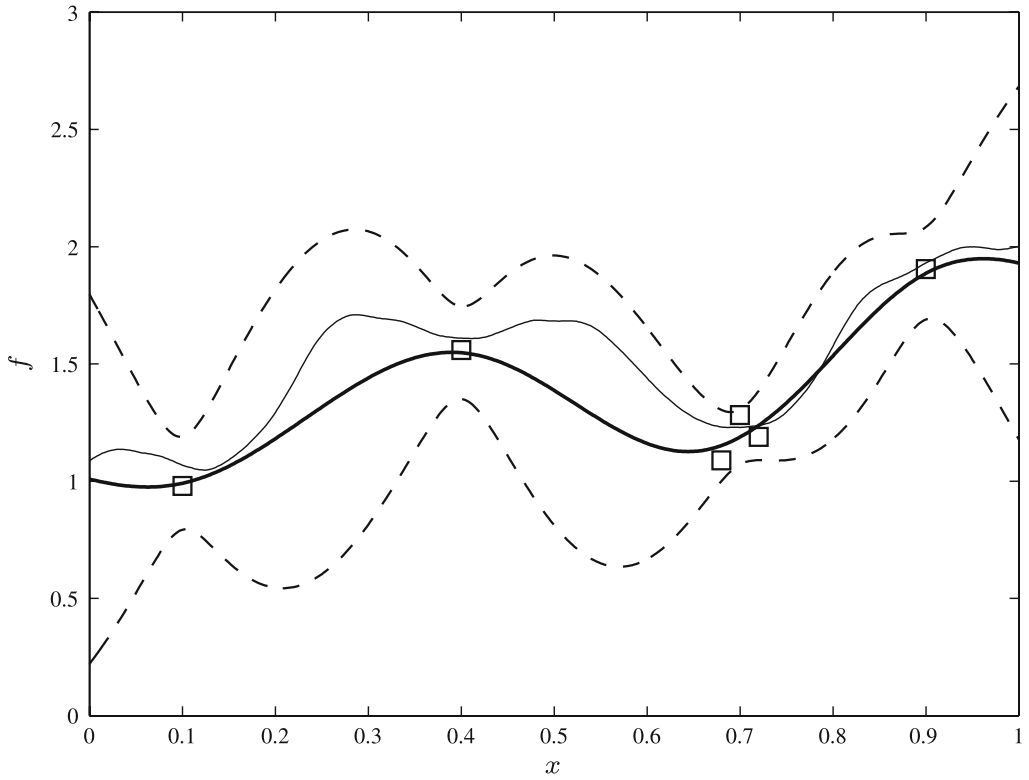


Fig. 5 Example of prediction by Kriging (*bold line*) of noisy measurements represented by squares. Dashed lines represent 95% confidence regions for the prediction and the thin solid line is an example of conditional simulation obtained using the method presented in Sect. 5.3.1

problems can generally be written as

$$\arg\min_{\boldsymbol{x} \in \mathbb{D}} J(\boldsymbol{x}), \tag{14}$$

with $J(\boldsymbol{x})$ a cost function reflecting some statistical property of the corrupted performance measure $f(\boldsymbol{x}+\boldsymbol{\varepsilon})$, where $\boldsymbol{\varepsilon}$ is a random vector accounting for noise on the factors. Classical cost functions are:

- mean: $J(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\varepsilon}}[f(\boldsymbol{x}+\boldsymbol{\varepsilon})]$,
- standard deviation: $J(\boldsymbol{x}) = \sqrt{\mathrm{var}_{\boldsymbol{\varepsilon}}(f(\boldsymbol{x}+\boldsymbol{\varepsilon}))}$,
- linear combination of mean and standard deviation: $J(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\varepsilon}}[f(\boldsymbol{x}+\boldsymbol{\varepsilon})] + \sqrt{\mathrm{var}_{\boldsymbol{\varepsilon}}(f(\boldsymbol{x}+\boldsymbol{\varepsilon}))}$,
- α-quantile: $J(\boldsymbol{x}) = Q^{\alpha}(\boldsymbol{x})$ with $Q^{\alpha}(\boldsymbol{x})$ such that $\mathrm{P}(f(\boldsymbol{x}+\boldsymbol{\varepsilon}) < Q^{\alpha}(\boldsymbol{x})) = \alpha$.

Using, for example, the α-quantile as a cost function, it is possible to adapt our optimization algorithm to solve (14). Given a set of evaluation results $\boldsymbol{f}_{\mathbb{S}}$ at noise-free evaluation points, and if it is possible to sample from the distribution $p_{\boldsymbol{\varepsilon}}$ of $\boldsymbol{\varepsilon}$, a Monte Carlo approximation $\hat{Q}^{\alpha}(\boldsymbol{x})$ of $Q^{\alpha}(\boldsymbol{x})$ is easily obtained by computing $\hat{f}(\boldsymbol{x}+\boldsymbol{\varepsilon})$ over a set sampled from $p_{\boldsymbol{\varepsilon}}$. The global optimization algorithm can then be applied to $Q^{\alpha}(\boldsymbol{x})$ instead of f, using pseudo-evaluations $\hat{\boldsymbol{Q}}^{\alpha}_{\mathbb{S}} = [\hat{Q}^{\alpha}(\boldsymbol{x}_1), \ldots, \hat{Q}^{\alpha}(\boldsymbol{x}_n)]$ (recomputed after each evaluation of f) instead of $\boldsymbol{f}_{\mathbb{S}}$. This naive approach can certainly be improved, but is sufficient to show the feasibility of a robust approach and to illustrate on a simple example (to be presented in the next section) the impact of $\boldsymbol{\varepsilon}$ on the evaluation points to be chosen by IAGO.

It is of course possible to combine these ideas and to deal simultaneously with noise both on the factors and the function evaluations.

6 Illustrations

This section presents some simple examples of global optimization using IAGO, with a regular grid as a set of candidate evaluation points. An empirical comparison with global optimization using expected improvement is also presented. The Matérn covariance class will be used for Kriging prediction, as it facilitates the tuning of the variance, regularity and range of correlation of the underlying random process, but note that any kind of admissible covariance function could have been used. The parameters of the covariance may be estimated from the data using a maximum-likelihood approach (see Sect. 8.3).

6.1 A one-dimensional example

Consider the function with two global minimizers illustrated by Fig. 6 and defined by $f : x \longmapsto 4[1 - \sin(x + 8\exp(x - 7))]$. Given an initial design consisting of three points, the IAGO algorithm is used to compute six additional points iteratively. The final Kriging model is depicted in the left part of Fig. 6, along with the resulting conditional distribution for the minimizers on the right part. After adding some noise on the function evaluations, the variant of IAGO presented in Sect. 5.3.1 is also applied to the function with the same initial design. In both cases, six additional evaluations have significantly reduced the uncertainty associated with the position of the global minimizers. The remaining likely locations reduce to small areas centered on the two actual global minimizers. In the noisy case, larger zones are identified, a direct consequence of the uncertainty associated with the evaluations.

Figure 7 illustrates robust optimization using the same function and initial design, but considering an additive zero-mean Gaussian noise on the factors with a standard deviation of

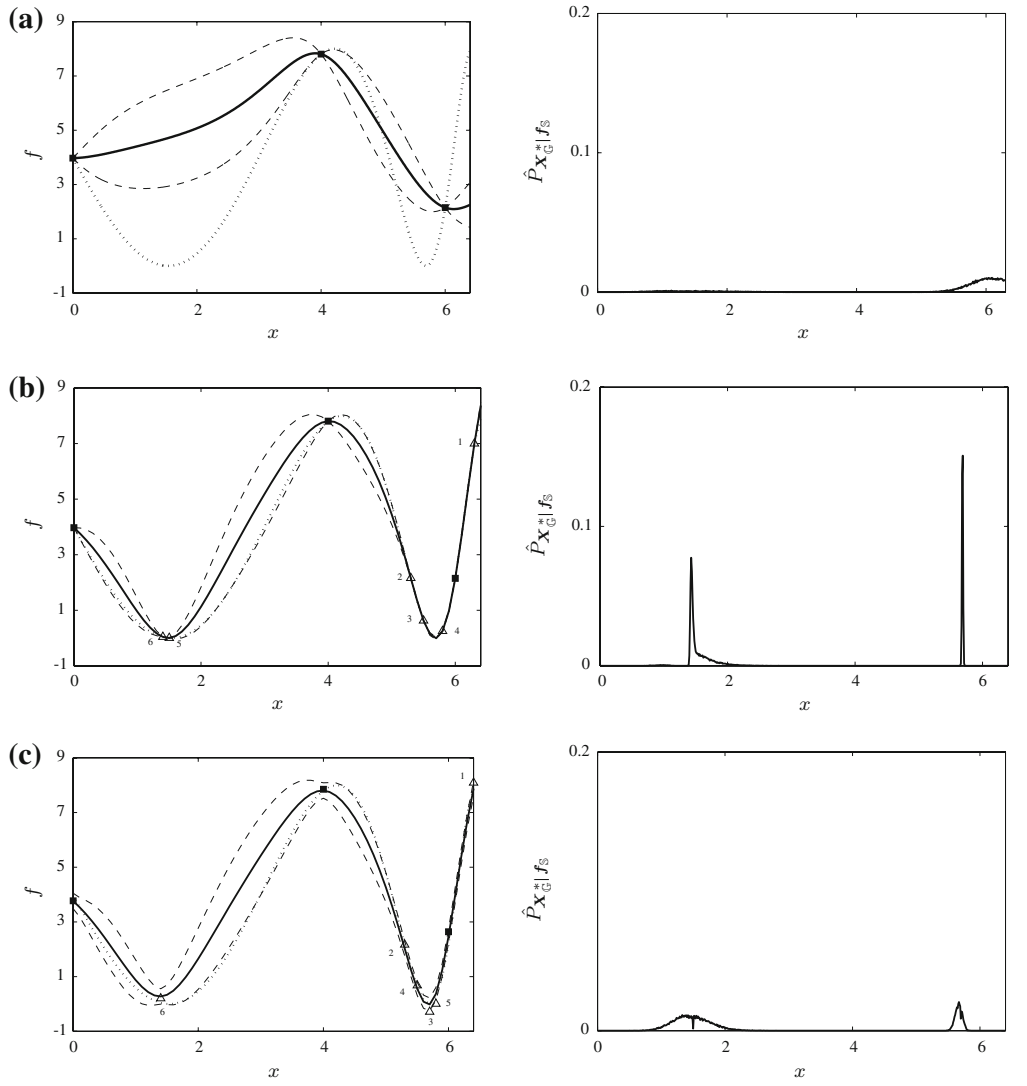


Fig. 6 Example of global optimization using IAGO on a function of one variable (*dotted line*), with an initial design consisting of three points (represented by *squares*). Six additional evaluations are carried out (triangles) using two versions of the IAGO algorithm. The graphs on the *left part* of the figure account for the predictions, while the *right part* presents the corresponding conditional distributions of the global minimizers. **a** Kriging prediction and conditional distribution of the global minimizers based on the initial design. **b** Standard IAGO algorithm (noise free case). **c** IAGO algorithm for noisy evaluations (the additive noise is zero-mean Gaussian with standard deviation 0.2)

0.2. The cost function used is the 90%-quantile $Q^{90\%}$, which is computed on the surrogate model but also, and only for the sake of comparison, on the true function using Monte Carlo uncertainty propagation (the quantile is approximated using 5000 simulations). After six iterations of the robust optimization algorithm, the distribution of the robust minimizers is sufficiently peaked to give a good approximation of the true global robust minimizer.

These result are encouraging as they show that the requirement of fast uncertainty reduction is met. The next section provides some more examples, along with a comparison with EGO, the EI-based global optimization algorithm.

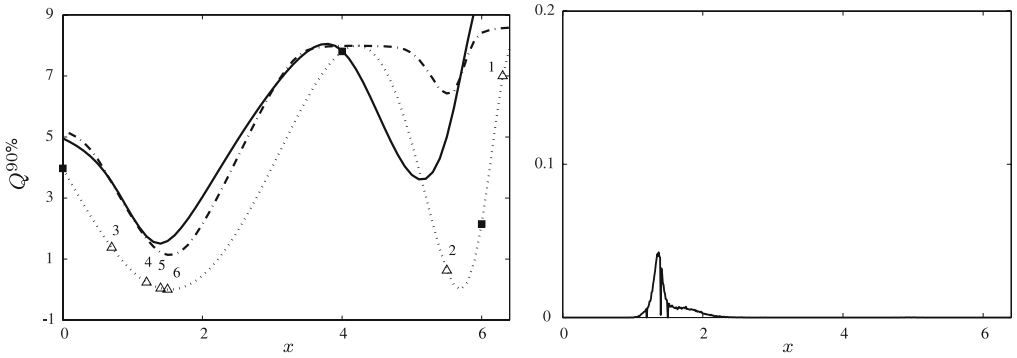


Fig. 7 Example of robust optimization using IAGO and the cost function $Q^{90\%}$. The function f (*dotted line*), corrupted by an additive Gaussian noise on the factor (zero mean with a standard deviation of 0.2), is studied starting from the initial design of three points already used in Fig. 6. Six additional evaluations are carried out (*triangles*), which are used to estimate the cost function based on the Kriging model (*bold line*), along with the conditional distribution of the robust minimizers (*right*). The cost function $Q^{90\%}$ estimated, only for the sake of comparison, from the true function using Monte Carlo uncertainty propagation is also provided (*mixed line*)

6.2 Empirical comparison with expected improvement

Consider first the function described by Fig. 8. Given an initial design of three points, both EI and minimizers entropy are computed. Their optimization provides two candidate evaluation points for f, which are also presented on Fig. 8, along with the post-evaluation prediction and conditional distribution for $\boldsymbol{X}^*_{\mathbb{G}}$. For this example, the regularity parameter of the Matérn covariance is set a priori to a high value (2.5). By taking in account the covariance function of F through conditional simulations, the minimizers entropy uses regularity to conclude faster. The resulting conditional distribution of the minimizers is then generally more peaked using the IAGO algorithm than using the EGO algorithm (as illustrated by Fig. 8b,c).

Consider now the Branin function (see, for instance, [8]), defined as

$$\begin{aligned} f : [-5, 10] \times [0, 15] &\longrightarrow \mathbb{R} \\ (x_1, x_2) &\longmapsto \left(x_2 - \tfrac{5.1}{4\pi^2}x_1^2 + \tfrac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \tfrac{1}{8\pi}\right)\cos(x_1) + 10. \end{aligned}$$

It has three global minimizers $\boldsymbol{x}_1^* \approx (-3.14, 12.27)^\mathsf{T}$, $\boldsymbol{x}_2^* \approx (3.14, 2.27)^\mathsf{T}$ and $\boldsymbol{x}_3^* \approx (9.42, 2.47)^\mathsf{T}$, and the global minimum is approximately equal to 0.4. Given an initial uniform design of sixteen points, fifteen additional points are iteratively selected and evaluated using the IAGO and EGO algorithms. The parameters of the Matérn covariance are estimated on the initial design, and kept unchanged during both procedures. The positions of the evaluation points are presented on Fig. 9 (left), along with the three global minimizers. Table 2 summarizes the results obtained with EGO and IAGO, based on the final Kriging models obtained with both approaches. Note that the EI criterion in EGO is maximized with a high precision, while minimizers entropy in IAGO is computed over a thousand candidate evaluation points located on a regular grid. It appears nevertheless that the algorithm using EI stalls on a single global minimizer, while the minimizers entropy allows a relatively fast estimation of all three of them. Besides IAGO yields a better global approximation of the supposedly unknown function. If twenty additional evaluations are carried out (as presented in the right part of Fig. 9), the final Kriging prediction using minimizers entropy estimates the minimum

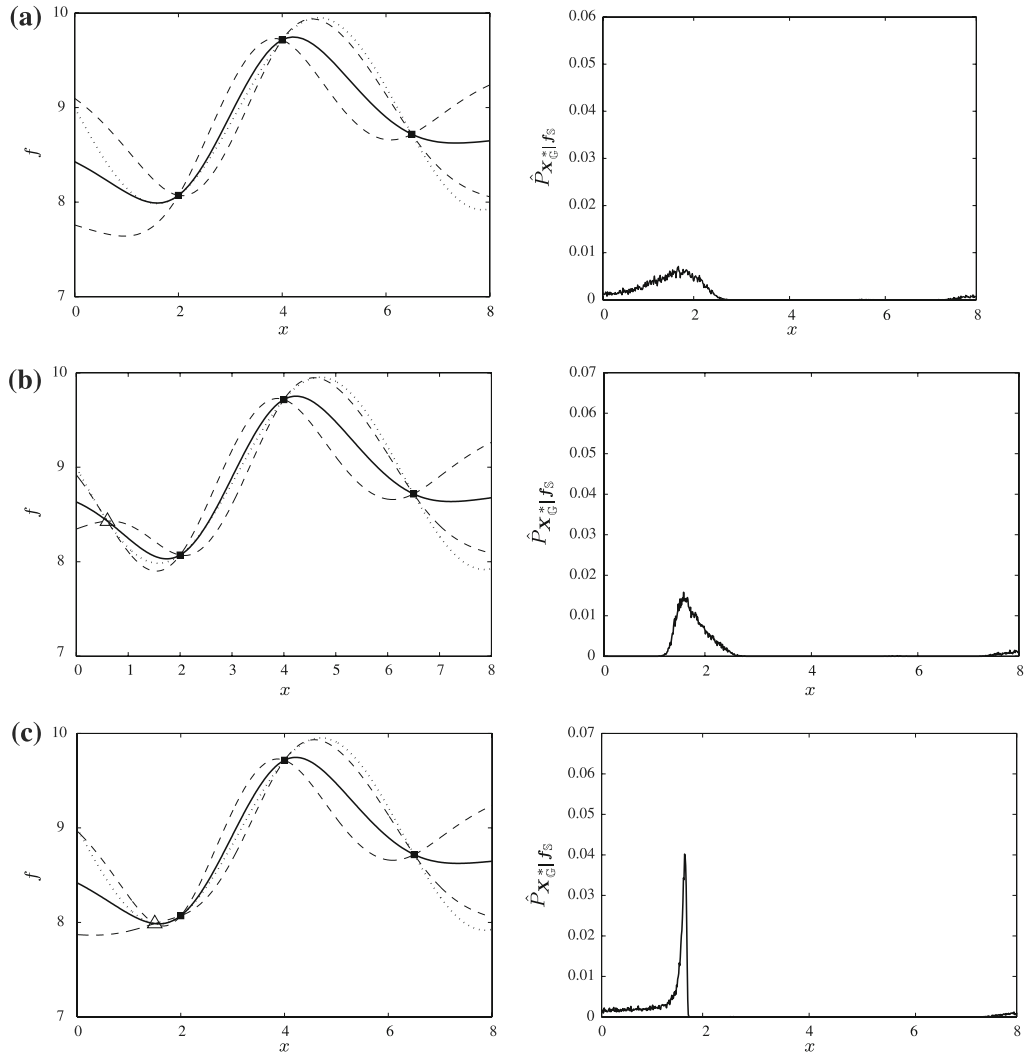


Fig. 8 Comparison between minimizers entropy and EI: the *left* side contains the Kriging predictions before and after an additional evaluation chosen with either EI or minimizers entropy, while the *right* side presents the corresponding conditional distribution of the global minimizers. **a** Initial prediction and minimizers distribution. **b** Prediction and minimizers distribution after an additional evaluation of f chosen with EI. **c** Prediction and minimizers distribution after an additional evaluation of f chosen with minimizers entropy

with an error of less than 0.05 for all three minimizers (cf. Table 2), while the use of EI does not improve the information on any minimizer any further. The difference between the two strategies is clearly evidenced. The EI criterion, overestimating the confidence in the initial prediction, has led to performing evaluations extremely close to one another, for a very small information gain. In a context of expensive function evaluation, this is highly detrimental. The entropy criterion, using the same covariance parameters, does not stack points almost at the same location before having identified the most likely zones for the minimizers. The use of what has been assumed and learned about the function is clearly more efficient in this case, and this property should be highly attractive when dealing with problems of higher dimension.

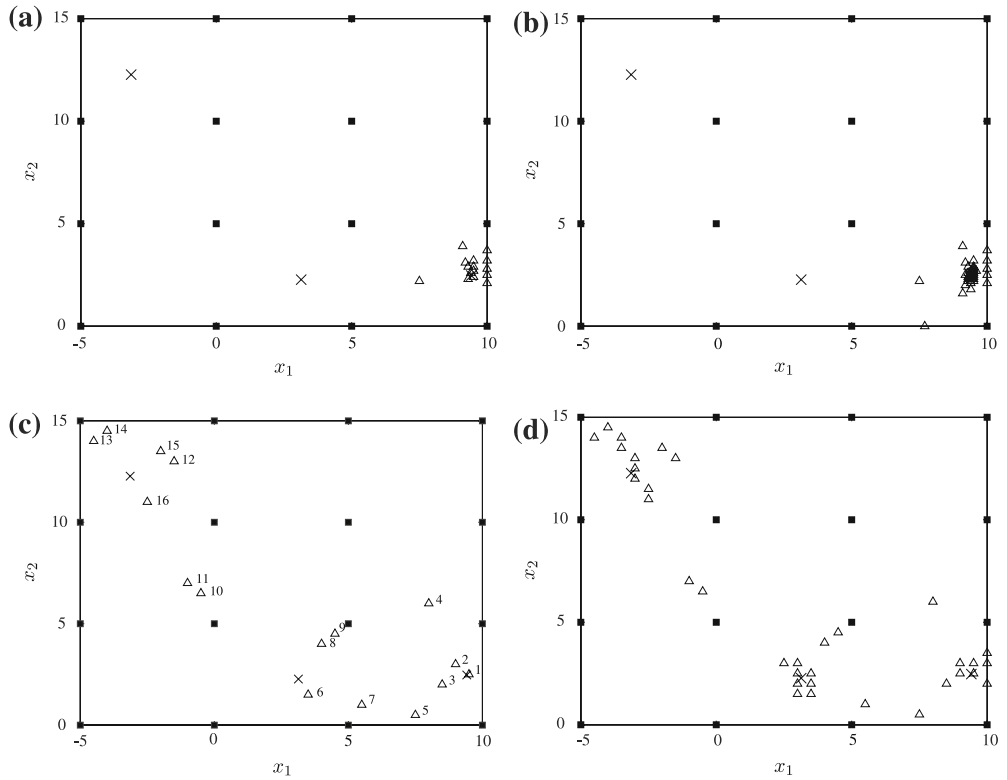


Fig. 9 Fifteen iterations of two optimization algorithms, that differ by their criteria for selecting evaluation points for f, on the Branin function: (*top*) the EI criterion is used, (*bottom*) the minimizers entropy criterion is used with a thousand candidate evaluation points for f set on a regular grid (*squares* account for initial data, *triangles* for new evaluations, and *crosses* give the actual locations of the three global minimizers). **a** 15 iterations using EGO. **b** 35 iterations using EGO. **c** 15 iterations using IAGO. **d** 35 iterations using IAGO

Table 2 Estimation results for the Branin function using the evaluations of Fig. 9

	EGO		IAGO	
	15 Iterations	35 Iterations	15 Iterations	35 Iterations
Euclidean distance between $\boldsymbol{x}_1^*$ and its final estimate	3.22	3.22	**2.18**	**0.23**
Value of the true function at estimated minimizer	17.95	17.95	**2.59**	**0.40**
Euclidean distance between $\boldsymbol{x}_2^*$ and its final estimate	2.40	2.40	**0.44**	**0.18**
Value of the true function at estimated minimizer	13.00	13.00	**0.85**	**0.42**
Euclidean distance between $\boldsymbol{x}_3^*$ and its final estimate	**0.04**	**0.04**	0.82	0.23
Value of the true function at estimated minimizer	**0.40**	**0.40**	1.94	0.44

7 Discussion

7.1 Robustness to uncertainty on the covariance parameters

Jones studied in [13] the potential of Kriging-based global optimization methods such as EGO. One of his most important conclusion, is that these methods "*can perform poorly if the initial sample is highly deceptive*". An eloquent example is provided on page 373 [13],

where a sine function is sampled using its own period, leading to a flat prediction over the domain, associated with a small prediction error.

This potential for deception is present throughout the IAGO procedure, and should not be ignored. To overcome this difficulty, several methods have been proposed (see, e.g., Enhanced Method 4 in [13] or [10]), which achieve some sort of robustness to an underestimation of the prediction error and more generally to a bad choice of covariance function. They seem to perform better than classical algorithms, including EGO.

Comparing the IAGO approach to such methods is an interesting topic for future research. The issue considered here was to demonstrate the interest of the minimizers entropy criterion, and we felt that this had to be done independently from the rest of the procedure.

It is of course essential to make IAGO robust to errors in the estimation of the covariance parameters. In many industrial problems, this can be easily done by using prior knowledge on the unknown function to restrict the possible values for these parameters. For example, experts of the field often have information regarding the range of values attainable by the unknown function. This information can be directly used to restrict the search space for the variance of the modeling process F, or even to choose it beforehand.

More generally, given the probabilistic framework used here, it should be relatively easy to develop a Bayesian or minimax extension of IAGO to guide the estimation of the parameters of the covariance function. A comparison with robust methods such as those detailed in [13] will then be essential.

7.2 Conclusions and perspectives

In this paper, a stepwise uncertainty reduction strategy has been used for the sequential global optimization of expensive-to-evaluate functions. This strategy iteratively selects a minimizer of the conditional minimizers entropy as the new evaluation point. To compute this entropy, a Gaussian random model of the function evaluations is used and the minimizers entropy is estimated through Kriging and conditional simulations. At each iteration, the result of the new evaluation is incorporated in the data base used to re-build the Kriging model (with a possible re-estimation of the parameters of its covariance function).

We have shown on some simple examples that, compared to the classical EI-based algorithm EGO, the method proposed significantly reduces the evaluation effort in the search for global optimizers. The stepwise uncertainty reduction strategy allows the optimization method to adapt the type of search to the information available on the function. In particular, the minimizers entropy criterion makes full use of the assumed regularity of the unknown function to balance global and local searches.

Choosing an adequate set of candidate points is crucial, as it must allow a good estimation of a global minimizer of the criterion, while keeping computation feasible. Promising results have already been obtained with space-filling designs, and adaptive sampling based on the conditional density of the global minimizers should be useful as dimension increases.

Extension to constrained optimization is an obviously important topic for future investigations. When it is easy to discard the candidate points in $\mathbb{X}$ that do not satisfy the constraints, the extension is trivial. For expensive-to-evaluate constraints, the extension is a major challenge.

Finally, the stepwise uncertainty reduction strategy associated with conditioning by Kriging is a promising solution for the robust optimization of expensive-to-evaluate functions, a problem that is central to many industrial situations, for which an efficient product design must be found in the presence of significant uncertainty on the values actually taken by some

factors in mass production. In addition, robustness to the uncertainty associated with the estimation of the parameters of the covariance function should also be sought.

8 Appendix: modeling with Gaussian processes

This section recalls the main concepts used in this paper, namely Gaussian process modeling and Kriging. The major results will be presented along with the general framework for the estimation of the model parameters.

8.1 Kriging when f is evaluated exactly

Kriging [4,15] is a prediction method based on random processes that can be used to approximate or interpolate data. It can also be understood as a kernel regression method, such as *splines* [23] or *Support Vector Regression* [20]. It originates from geostatistics and is widely used in this domain since the 60s. Kriging is also known as the *Best Linear Unbiased Prediction* (BLUP) in statistics, and has been more recently designated as Gaussian Processes (GP) in the 90s in the machine learning community.

As mentioned in Sect. 2.1, it is assumed that the function f is a sample path of a Gaussian random process F. Denote by $m(\boldsymbol{x}) = E[F(\boldsymbol{x})]$ the mean function of $F(\boldsymbol{x})$ and by $k(\boldsymbol{x}, \boldsymbol{y})$ its covariance function, written as

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

Kriging then computes the BLUP of $F(\boldsymbol{x})$, denoted by $\hat{F}(\boldsymbol{x})$, in the vector space generated by the evaluations $\mathbb{H}_{\mathbb{S}} = \mathrm{span}\{F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)\}$. As an element of $\mathbb{H}_{\mathbb{S}}$, $\hat{F}(\boldsymbol{x})$ can be written as

$$\hat{F}(\boldsymbol{x}) = \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{F}_{\mathbb{S}}. \tag{15}$$

As the BLUP, $\hat{F}(\boldsymbol{x})$ must have the smallest variance for the prediction error

$$\hat{\sigma}^2(\boldsymbol{x}) = \mathbb{E}[(\hat{F}(\boldsymbol{x}) - F(\boldsymbol{x}))^2], \tag{16}$$

among all unbiased predictors. The variance of the prediction error satisfies

$$\hat{\sigma}^2(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{K} \boldsymbol{\lambda}(\boldsymbol{x}) - 2\boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{k}(\boldsymbol{x}), \tag{17}$$

with

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

the $n \times n$ covariance matrix of F at evaluation points in $\mathbb{S}$, and

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

the vector of covariances between $F(\boldsymbol{x})$ and $\boldsymbol{F}_{\mathbb{S}}$

The prediction method [16] assumes that the mean of $F(\boldsymbol{x})$ can be written as a finite linear combination

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

where $\boldsymbol{\beta}$ is a vector of fixed but unknown coefficients, and

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

is a vector of known functions of the factor vector $\boldsymbol{x}$. Usually these functions are monomials of low degree in the components of $\boldsymbol{x}$ (in practice, their degree does not exceed two). These functions may be used to reflect some prior knowledge on the unknown function. As we have none for the examples considered here, we simply use an unknown constant.

The Kriging predictor at $\boldsymbol{x}$ is then the best linear predictor subject to the unbiasedness constraint $\mathbb{E}(\hat{F}(\boldsymbol{x})) = m(\boldsymbol{x})$, whatever the unknown $\boldsymbol{\beta}$. The unbiasedness constraint translates into

$$\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{P}^{\mathsf{T}} \boldsymbol{\lambda}(\boldsymbol{x}) = \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{p}(\boldsymbol{x}), \tag{18}$$

with

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

For (18) to be satisfied for all $\boldsymbol{\beta}$, the Kriging coefficients must satisfy the linear constraints

$$\boldsymbol{P}^{\mathsf{T}} \boldsymbol{\lambda}(\boldsymbol{x}) = \boldsymbol{p}(\boldsymbol{x}), \tag{19}$$

called *universality constraints* by Matheron. At this point, Kriging can be reformulated as follows: find the vector of Kriging coefficients that minimizes the variance of the prediction error (17) subject to the constraints (19). This problem can be solved via a Lagrangian formulation, with $\boldsymbol{\mu}(\boldsymbol{x})$ a vector of l Lagrange multipliers for the constraints in (19). The coefficients $\boldsymbol{\lambda}(\boldsymbol{x})$ are then solutions of the linear system of equations

$$\begin{pmatrix} \boldsymbol{K} & \boldsymbol{P} \\ \boldsymbol{P}^{\mathsf{T}} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}(\boldsymbol{x}) \\ \boldsymbol{\mu}(\boldsymbol{x}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{k}(\boldsymbol{x}) \\ \boldsymbol{p}(\boldsymbol{x}) \end{pmatrix}, \tag{20}$$

with $\boldsymbol{0}$ a matrix of zeros. A convenient expression for the variance of the prediction error is obtained by substituting $\boldsymbol{k}(\boldsymbol{x}) - \boldsymbol{P}\boldsymbol{\mu}(\boldsymbol{x})$ for $\boldsymbol{K}\boldsymbol{\lambda}(\boldsymbol{x})$ in (17) as justified by (20), to get

$$\hat{\sigma}^2(\boldsymbol{x}) = \mathbb{E}\left[F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x})\right]^2 = k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{k}(\boldsymbol{x}) - \boldsymbol{p}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{\mu}(\boldsymbol{x}). \tag{21}$$

The variance of the prediction error at $\boldsymbol{x}$ can thus be computed without any evaluation of f, using (20) and (21). It provides a measure of the quality associated with the Kriging prediction. Evaluations of f remain needed to estimate the parameters of the covariance function of F (if any), as will be seen in Sect. 8.3.2.

Once f has been evaluated at all evaluation points, the prediction of the value taken by f at $\boldsymbol{x}$ becomes

$$\hat{f}(\boldsymbol{x}) = \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{f}_{\mathbb{S}}, \tag{22}$$

with $\boldsymbol{f}_{\mathbb{S}} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]^{\mathsf{T}}$ ($\boldsymbol{f}_{\mathbb{S}}$ is viewed as a sample value of $\boldsymbol{F}_{\mathbb{S}}$).

It is easy to check that (20) implies that

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

The prediction of f at $\boldsymbol{x}_i \in \mathbb{S}$ is then $f(\boldsymbol{x}_i)$, so Kriging is an interpolation with the considerable advantage that it also accounts for model uncertainty through an explicit characterization of the prediction error.

Remark The Bayesian framework (see, for instance, [25]) is an alternative approach to derive the BLUP, in which F is viewed as a Bayesian prior on the output. In the case of a zero-mean model, the conditional distribution of the function is then Gaussian with mean

$$\mathbb{E}\left[F(\boldsymbol{x})|\, \boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}\right] = \boldsymbol{k}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{K}^{-1} \boldsymbol{f}_{\mathbb{S}}, \tag{23}$$

and variance

$$\mathrm{Var}\left[F(\boldsymbol{x})|\, \boldsymbol{F}_{\mathbb{S}} = \boldsymbol{f}_{\mathbb{S}}\right] = k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{K}^{-1} \boldsymbol{k}(\boldsymbol{x}),$$

which are exactly the mean (22) and variance (21) of the Kriging predictor for a model F with zero mean. The Kriging predictor can also be viewed as the conditional mean of $F(\boldsymbol{x})$ in the case of an unknown mean, if the universality constraints are viewed as a non-informative prior on $\boldsymbol{\beta}$.

8.2 Kriging when f is evaluated approximately

The Kriging predictor was previously defined as the element of the space $\mathbb{H}_{\mathbb{S}}$ generated by the random variables $F(\boldsymbol{x}_i)$ that minimizes the prediction error. A natural step is to extend this formulation to the case of a function whose evaluations are corrupted by additive independent and identically distributed Gaussian noise variables ε_i with zero mean and variance σ_ε^2. The model of the observations then becomes $F^{\mathrm{obs}}_{\boldsymbol{x}_i} = F(\boldsymbol{x}_i) + \varepsilon_i,\ i = 1, \ldots, n$, and the Kriging predictor for $F(\boldsymbol{x})$ takes the form $\hat{F}(\boldsymbol{x}) = \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{F}^{\mathrm{obs}}_{\mathbb{S}}$ with $\boldsymbol{F}^{\mathrm{obs}}_{\mathbb{S}} = \left[F^{\mathrm{obs}}_{\boldsymbol{x}1}, \ldots, F^{\mathrm{obs}}_{\boldsymbol{x}_n}\right]^{\mathsf{T}}$. The unbiasedness constraint (19) remain unchanged, while the mean-square error (2) becomes

$$\mathbb{E}[\hat{F}(\boldsymbol{x}) - F(\boldsymbol{x})]^2 = k(\boldsymbol{x}, \boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} (\boldsymbol{K} + \sigma_\varepsilon^2 \boldsymbol{I}_n) \boldsymbol{\lambda}(\boldsymbol{x}) - 2\boldsymbol{\lambda}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{k}(\boldsymbol{x}),$$

with $\boldsymbol{I}_n$ the identity matrix. Finally, using Lagrange multipliers as before, it is easy to show that the coefficients $\boldsymbol{\lambda}(\boldsymbol{x})$ of the prediction must satisfy

$$\begin{pmatrix} \boldsymbol{K} + \sigma_\varepsilon^2 \boldsymbol{I}_n & \boldsymbol{P} \\ \boldsymbol{P}^{\mathsf{T}} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}(\boldsymbol{x}) \\ \boldsymbol{\mu}(\boldsymbol{x}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{k}(\boldsymbol{x}) \\ \boldsymbol{p}(\boldsymbol{x}) \end{pmatrix}. \tag{24}$$

The resulting prediction is no longer interpolative, but can still be viewed as the mean of the conditional distribution of F. The variance of the prediction error is again obtained using (21).

8.3 Covariance choice

Choosing a suitable covariance function $k(\cdot, \cdot)$ for a given f is a recurrent and fundamental question. It involves the choice of a parametrized class (or model) of covariance, and the estimation of its parameters.

8.3.1 Covariance classes

The asymptotic theory of Kriging [21] stresses the importance of the behaviour of the covariance near the origin. This behaviour is indeed linked with the quadratic-mean regularity of the random process. For instance, if the covariance function is continuous at the origin, then the process will be continuous in quadratic mean. In practice, one often uses covariances that are *invariant by translation* (or equivalently *stationary*), *isotropic*, and such that regularity can

be adjusted. Non-stationary covariances are seldom used in practice, as they make parameter estimation particularly difficult [4]. Isotropy, however, is not required and can even be inappropriate when the factors are of different natures. An example of an anisotropic, stationary covariance class is $k(\boldsymbol{x}, \boldsymbol{y}) = k(h)$, with $h = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^{\mathsf{T}} \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{y})}$ where $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{X}^2$ and $\boldsymbol{A}$ is a symmetric positive definite matrix.

A number of covariance classes are classically used (for instance, exponential $h \mapsto \sigma^2 \exp(-\theta |h|^{\alpha})$, product of exponentials, or polynomial). The *Matérn covariance* class offers the possibility to adjust regularity with a single parameter [21]. Stein (1999) advocates the use of the following parametrization of the Matérn class:

$$k(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2} h}{\rho} \right)^{\nu} \mathcal{K}_{\nu} \left(\frac{2\nu^{1/2} h}{\rho} \right), \tag{25}$$

where $\mathcal{K}_{\nu}$ is the modified Bessel function of the second kind [27]. This parameterization is easy to interpret, as ν controls regularity, σ^2 is the variance ($k(0) = \sigma^2$), and ρ represents the *range* of the covariance, i.e., the characteristic correlation distance. To stress the significance and relevance of the regularity parameter, Fig. 10 shows the influence of ν on the covariance function, and Fig. 11 demonstrates its impact on the sample paths. Since Kriging assumes that f is a sample path of F, a careful choice of the parameters of the covariance is essential.

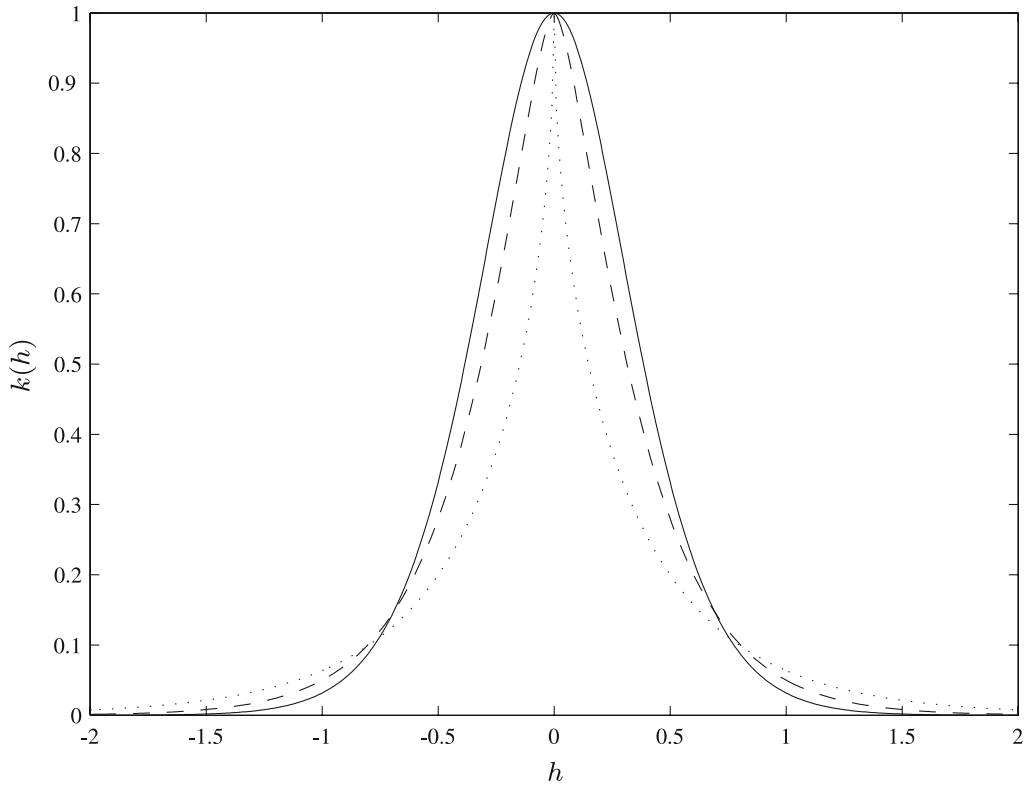


Fig. 10 Matérn covariances with $\rho = 0.5$, $\sigma^2 = 1$. Solid line corresponds to $\nu = 4$, dashed line to $\nu = 1$ and dotted line to $\nu = 0.25$

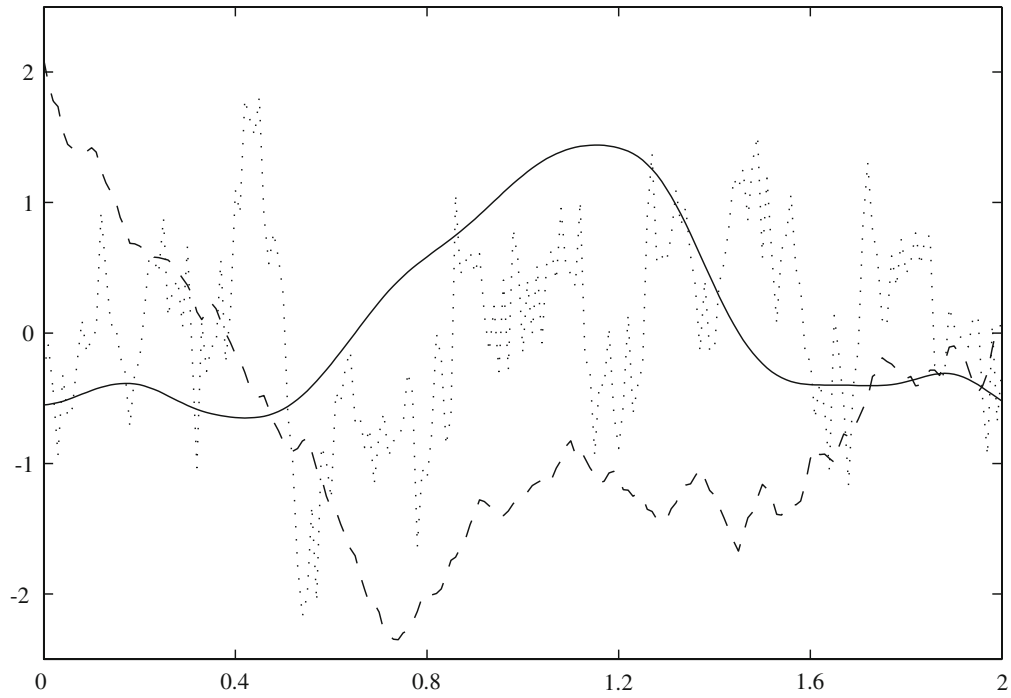


Fig. 11 Three sample paths of a zero-mean Gaussian process with a Matérn covariance. Conventions are as in Fig. 10: $\nu = 4$ for the *solid line*, $\nu = 1$ for the *dashed line* and $\nu = 0.25$ for the *dotted line*

8.3.2 Covariance parameters

The parameters for a given covariance class can either be fixed using prior knowledge on the system, or be estimated from experimental data. In geostatistics, estimation is carried out using the adequacy between the empirical and model covariances [4]. In other areas, cross validation [23] and maximum likelihood [21] are mostly employed. For simplicity and generality reasons [21], the maximum-likelihood method is preferred here. Using the joint probability density of the observed Gaussian vector, and assuming that the mean of $F(\boldsymbol{x})$ is zero for the sake of simplicity, one obtains the maximum-likelihood estimate of the vector $\boldsymbol{\theta}$ of the covariance parameters (see, for instance, [22]) by minimizing the negative log-likelihood

$$l(\boldsymbol{\theta}) = \frac{n}{2}\log 2\pi + \frac{1}{2}\log\det \boldsymbol{K}(\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{f}_{\mathbb{S}}^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{\theta})^{-1}\boldsymbol{f}_{\mathbb{S}}. \tag{26}$$

When the mean for $F(\boldsymbol{x})$ is unknown, the parameters can be estimated, using for example the *REstricted Maximum Likelihood* (REML, see [21]). This is the approach used for the examples in this paper.

Figure 12 illustrates prediction by Kriging with a Matérn covariance, the parameters of which have been estimated by REML. The prediction interpolates the data, and confidence intervals are deduced from the square root of the variance of the prediction error to assess the quality of the prediction between data. Figure 12 also contains a series of conditional simulations (obtained with the method explained in Sect. 3.2), namely sample paths of F that interpolate the data. As implied by (23), the Kriging prediction is the mean of these conditional simulations.

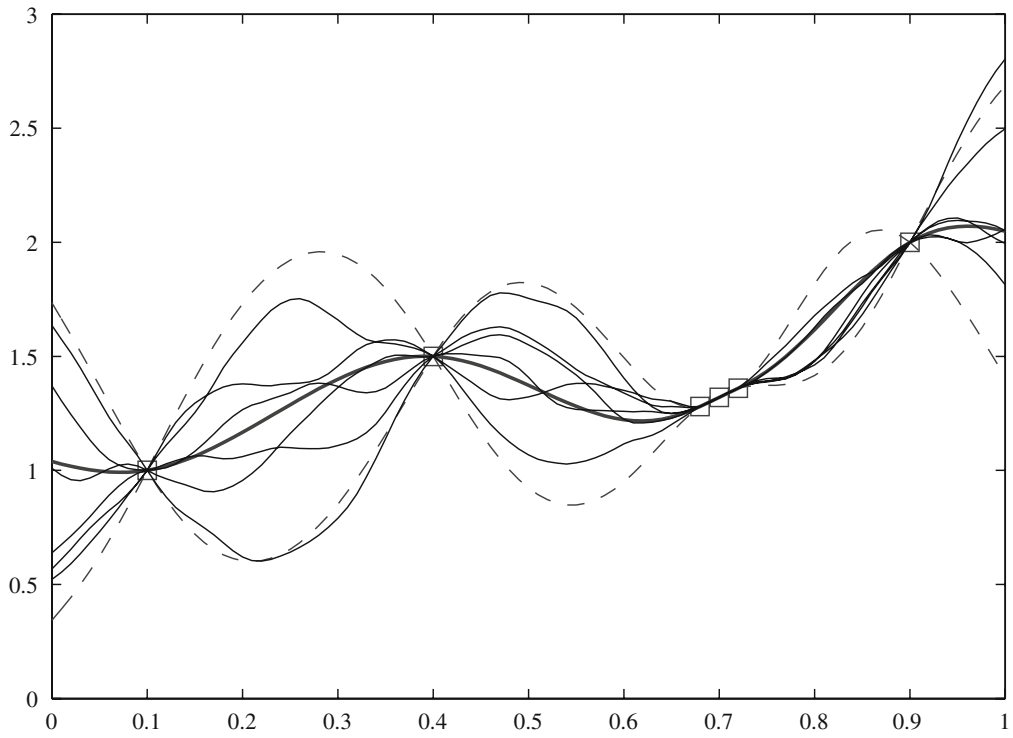


Fig. 12 Example of Kriging interpolation (*bold line*) for a function of one variable. The data are represented by *squares*, and the covariance parameters were estimated by REML. *Dashed lines* delimit 95% confidence region for the prediction. The *thin solid lines* are examples of conditional simulations

Acknowledgements The authors wish to thank Donald R. Jones for his comments that greatly contributed to improving the accuracy and clarity of this paper.

References

1. Abrahamsen, P.: A review of Gaussian random fields and correlation functions. Tech. Rep., Norwegian Computing Center (1997). www.math.ntnu.no/~omre/TMA4250/V2007/abrahamsen2.ps
2. Adler, R.: On excursion sets, tubes formulas and maxima of random fields. Ann. Appl. Prob. **10**(1), 1–74 (2000)
3. Chib, S., Greenberg, E.: Understanding the metropolis-hastings algorithm. Am. Stat. **49**(4), 327–335 (1995)
4. Chilès, J., Delfiner, P.: Geostatistics, Modeling Spatial Uncertainty. Wiley, New York (1999)
5. Cover, T.M., Thomas, A.J.: Elements of Information Theory. Wiley, New York (1991)
6. Cox, D., John, S.: Sdo: a statistical method for global optimization. In: Alexandrov, N., Hussaini, M.Y. (eds.) Multidisciplinary Design Optimization: State of the Art, pp. 315–329. SIAM, Philadelphia (1997). citeseer.ifi.unizh.ch/cox97sdo.html
7. Delfiner, P.: Shift invariance under linear models. Ph.D. thesis, Princetown University, New Jersey (1977)
8. Dixon, L., Szegö, G.: The global optimisation problem: an introduction. In: Dixon, L., Szegö, G. (eds.) Towards Global Optimization 2. North-Holland Publishing Company (1978)
9. Geman, D., Jedynak, B.: An active testing model for tracking roads in satellite images. Tech. Rep. 2757, Institut National de Recherche en Informatique et en Automatique (INRIA) (1995)
10. Gutmann, H.: A radial basis function method for global optimization. J. Glob. Optim. **19**(3), 201–227 (2001)
11. Huang, D.: Experimental planning and sequential Kriging optimization using variable fidelity data. Ph.D. thesis, Ohio State University (2005)

12. Huang, D., Allen, T., Notz, W., Zeng, N.: Global optimization of stochastic black-box systems via sequential Kriging meta-models. J. Glob. Optim. **34**, 441–466 (2006)
13. Jones, D.: A taxonomy of global optimization methods based on response surfaces. J.Glob. Optim. **21**, 345–383 (2001)
14. Jones, D., Schonlau, M., William, J.: Efficient global optimization of expensive black-box functions. J. Glob. Optim. **13**, 455–492 (1998)
15. Matheron, G.: Principles of geostatistics. Econ. Geol. **58**, 1246–1266 (1963)
16. Matheron, G.: Le krigeage universel. In: Cahiers du Centre de Morphologie Mathématique de Fontainebleau. Ecole des Mines de Paris (1969) Fasc.1
17. Sasena, M., Papalambros, P., Goovaerts, P.: Exploration of metamodeling sampling criteria for constrained global optimization. Eng. Opt. **34**, 263–278 (2002)
18. Schonlau, M.: Computer experiments and global optimization. Ph.D. thesis, University of Waterloo (1997)
19. Sjö, E.: Crossings and maxima in Gaussian fields and seas. Ph.D. thesis, Lund Institute of Technology (2000)
20. Smola, A.: Learning with kernels. Ph.D. thesis, Technische Universität Berlin (1998)
21. Stein, M.: Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York (1999)
22. Vechia, A.: Estimation and model identification for continuous spatial processes. J. R. Stat. Soc. **B**(50), 297–312 (1998)
23. Wahba, G.: Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods—Support Vector Learning, pp. 69–87. MIT Press, Boston (1998)
24. Watson, A., Barnes, R.: Infill sampling criteria to locate extremes. Math. Geol. **27**(5), 589–698 (1995)
25. Williams, C., Rasmussen, C.: Gaussian processes for regression. In: Touretzky, D., Mayer, M., Hasselmo, M. (eds.) Advances in Neural Information Processing Systems, vol. 8. MIT Press (1996)
26. Williams, B., Santner, T., Notz, W.: Sequential design of computer experiments to minimize integrated response functions. Stat. Sinica **10**, 1133–1152 (2000)
27. Yaglom, A.: Correlation Theory of Stationary and Related Random Functions I: Basic Results. Vol. 6, Springer Series in Statistics. Springer-Verlag, New-York (1986)

3.5 Constrained multi-objective Bayesian optimization

This article was presented in the Learning and Intelligent OptimizatioN Conference (LION9, 2015). It is a preliminary presentation of a recent contribution on constrained multi-objective optimization. A detailed version is in preparation.

A Bayesian approach to constrained multi-objective optimization

Paul FELIOT, Julien BECT, Emmanuel VAZQUEZ

IRT SystemX, Palaiseau, France & SUPELEC, Gif-sur-Yvette, France

firstname.lastname@irt-systemx.fr or firstname.lastname@supelec.fr

Abstract. This paper addresses the problem of derivative-free multi-objective optimization of real-valued functions under multiple inequality constraints. Both the objective and constraint functions are assumed to be smooth, nonlinear, expensive-to-evaluate functions. As a consequence, the number of evaluations that can be used to carry out the optimization is very limited. The method we propose to overcome this difficulty has its roots in the Bayesian and multi-objective optimization literatures. More specifically, we make use of an extended domination rule taking both constraints and objectives into account under a unified multi-objective framework and propose a generalization of the expected improvement sampling criterion adapted to the problem. A proof of concept on a constrained multi-objective optimization test problem is given as an illustration of the effectiveness of the method.

1 Introduction

This paper addresses the problem of derivative-free multi-objective optimization of real-valued functions under multiple inequality constraints:

$$\begin{cases} \text{Minimize} \quad f(x) \\ \text{Subject to } x \in \mathbb{X} \text{ and } c(x) \leq 0 \end{cases}$$

where $f = (f_j)_{1 \leq j \leq p}$ is a vector of objective functions to be minimized, $\mathbb{X} \subset \mathbb{R}^d$ is the search domain and $c = (c_i)_{1 \leq i \leq q}$ is a vector of constraint functions. Both the objective functions f_j and the constraint functions c_i are assumed to be smooth, nonlinear functions that are expensive to evaluate. As a consequence, the number of evaluations that can be used to carry out the optimization is very limited. This setup typically arises when the values $f(x)$ and $c(x)$ for a given $x \in \mathbb{X}$ correspond to the outputs of a computationally expensive computer program.

In this work, we consider a Bayesian approach to this optimization problem. The objective and constraint functions are modelled using a vector-valued Gaussian process and $\mathbb{X}$ is explored using a sequential Bayesian design of experiments approach [14]. More specifically, we focus on the Expected Improvement (EI) infill sampling criterion. This criterion was originally introduced in the context of single-objective, unconstrained optimization [10,13]. It was later extended to

handle constraints [7,17,19,21,22] and to address unconstrained multi-objective problems [4,18,24,9]. However, to the best of our knowledge, the general case of a constrained multi-objective problem has only been addressed very recently by [23]. In their paper, Shimoyama et al. consider three different Bayesian criteria for unconstrained multi-objective optimization and study the effect of multiplying the criteria by a probability of feasibility in order to handle the constraints.

The approach we propose to handle the constraints is based on an extended domination rule, in the spirit of [6,16,20], which takes both objectives and constraints into account under a unified framework. The extended domination rule makes it possible to derive a new expected improvement criterion to deal with constrained multi-objective optimization problems. Section 2 introduces the proposed method, while Section 3 presents a proof of concept on a classical test case from the literature. Results and future works are briefly discussed at the end of Section 3.

2 An expected improvement criterion for constrained multi-objective optimization

In this section, we present our extended domination rule and introduce a new expected improvement criterion suitable for constrained and unconstrained multi-objective problems. The new criterion is equivalent to the original EI on unconstrained single-objective problems and to Schonlau's extension to the constrained case [22] once a feasible point has been found. It is also similar to the formulation of [24] for unconstrained multi-objective problems and to that of [23] in the constrained case once a feasible point has been found. As such, it can be seen as a generalization of the above-mentioned criteria.

Denote by $\mathbb{F} \subset \mathbb{R}^p$ and $\mathbb{C} \subset \mathbb{R}^q$ the objective and constraint spaces respectively, and let $\mathbb{Y} = \mathbb{F} \times \mathbb{C}$. We shall say that $y_1 \in \mathbb{Y}$ dominates $y_2 \in \mathbb{Y}$, which will be denoted by $y_1 \lhd y_2$, if $\psi(y_1)$ dominates $\psi(y_2)$ in the usual Pareto sense, where

$$\begin{aligned} \psi : \quad \mathbb{F} \times \mathbb{C} \quad &\to \overline{\mathbb{R}}^p \times \mathbb{R}^q \\ (y_f, y_c) \quad &\mapsto \begin{cases} (y_f, 0) & \text{if } y_c \leq 0, \\ (+\infty, \max(y_c, 0)) & \text{otherwise,} \end{cases} \end{aligned}$$

In the above system of equations, $\overline{\mathbb{R}}$ denotes the extended real line. For unconstrained problems, we simply take the usual domination rule on $\mathbb{F}$. Figure 1 illustrates this extended domination rule in different cases.

Assume now that $\mathbb{Y}$ is bounded. Much like [4,24,18], we define the improvement yielded by a new observation as the increase of the dominated hypervolume:

$$I_N\left(x_{N+1}\right) = \left|H_{N+1}\right| - \left|H_N\right|,$$

where H_N is the subset of $\mathbb{Y}$ dominated by the solutions observed so far $\left(f(x_1), c(x_1)\right), \ldots, \left(f(x_N), c(x_N)\right)$ and $|\cdot|$ denotes the usual (Lebesgue) volume measure in $\mathbb{R}^{p+q}$. The corresponding expected improvement criterion can be written as

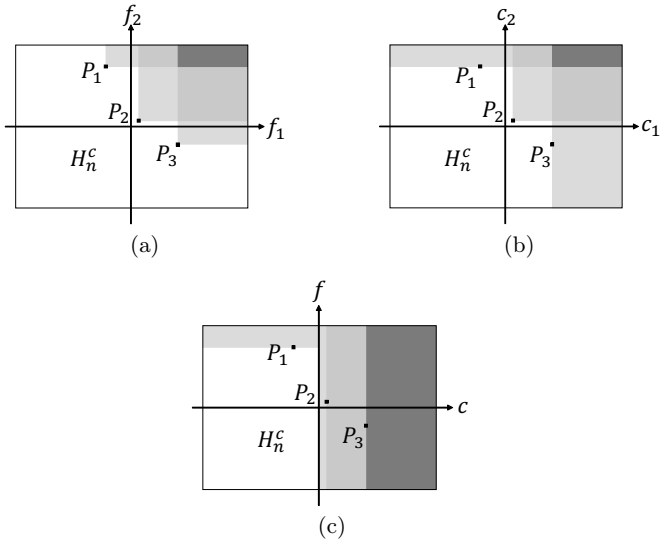


Fig. 1. Illustration of the extended domination rule in different situations. The region dominated by each point is represented by a shaded area. Darker shades of gray indicate overlapping regions. (a) Feasible solutions are compared with respect to their objective values using the usual domination rule in the objective space. (b) Non-feasible solutions are compared component-wise with respect to their constraint violations using the usual domination rule applied in the constraint space. (c) Feasible solutions always dominate non-feasible solutions; other cases are handled as in the first two figures.

$$\begin{aligned}\mathrm{EI}_N(x_{N+1}) &= \mathbb{E}_N\left((I_N(x_{N+1})\right) \\ &= \mathbb{E}_N\left(\int_{\mathbb{Y}\setminus H_N} \mathbb{1}_{\xi(x_{N+1})\lhd y}\,\mathrm{d}y\right) \\ &= \int_{\mathbb{Y}\setminus H_N} \mathbb{P}_N(\xi(x_{N+1})\lhd y)\,\mathrm{d}y\end{aligned}$$

where $\mathbb{P}_N$ denotes the probability conditional to the observations and ξ is a vector-valued Gaussian model for (f, c).

Even though the integrand of the EI formula can be readily computed analytically, its integration is not trivial due to the combinatorial nature of the problem [8,2,5]. To overcome this difficulty, we propose to use a Sequential Monte Carlo (SMC) approximation [3,11,12,1]:

$$\mathrm{EI}_N(x_{N+1}) \approx \sum_{i=1}^{n} w_i\,\mathbb{P}_N(\xi(x_{N+1})\lhd y_i),$$

where $\mathcal{Y}_N = (w_i, y_i)_{1 \leq i \leq n}$ is a weighted sample that targets the uniform density on $\mathbb{Y} \setminus H_N$.

3 Proof of concept

In this paper, we illustrate the behavior of our new optimization strategy using the Osyczka and Kundu test problem [15] for constrained multi-objective optimization ($d = 6$, $p = 2$, $q = 6$). The algorithm is initialized using a Latin Hypercube sample of 18 samples and proceeds using the above mentionned criterion. Figure 2 shows the convergence of the algorithm at different steps of the optimization.

We are also able to report good results on other challenging test cases from the literature and future communications will include a comparison of our method to reference optimization methods. More details about the SMC procedure will also be proposed.

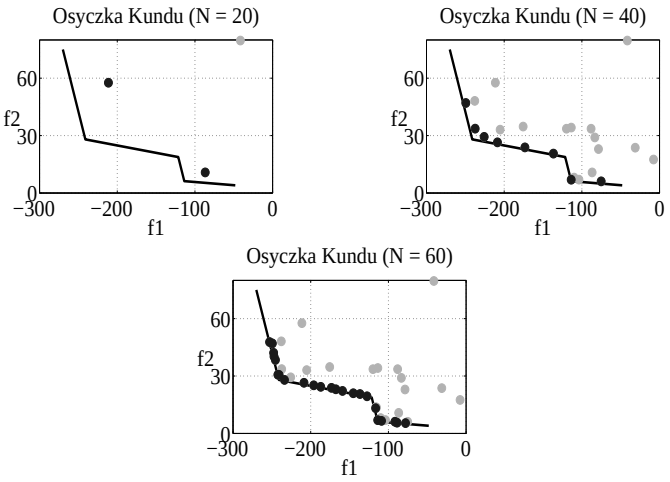


Fig. 2. Test results on Osyczka and Kundu test problem with, from left to right, $N = 20$, 40 and 60 evaluations. Only feasible points are shown on the figures. The dark dots represent non-dominated observations while the light gray dots represent dominated ones. The dark curve represents the target Pareto front.

Acknowledgements. This research work has been carried out in the frame of the Technological Research Institute SystemX, and therefore granted with public funds within the scope of the French Program *Investissements d'Avenir*.

References

1. Au, S.K., Beck, J.L.: Estimation of small failure probabilities in high dimensions by subset simulation. Probabilistic Engineering Mechanics 16(4), 263–277 (2001)
2. Bader, J., Zitzler, E.: Hype: An algorithm for fast hypervolume-based many-objective optimization. Evolutionary Computation 19(1), 45–76 (2011)
3. Benassi, R., Bect, J., Vazquez, E.: Bayesian optimization using sequential Monte Carlo. In: Learning and Intelligent Optimization. 6th International Conference, LION 6, Paris, France, January 16-20, 2012, Revised Selected Papers, Lecture Notes in Computer Science, vol. 7219, pp. 339–342. Springer (2012)
4. Emmerich, M.T.M., Giannakoglou, K.C., Naujoks, B.: Single- and multi-objective evolutionary optimization assisted by Gaussian random field metamodels. IEEE Transactions on Evolutionary Computation 10(4), 421–439 (2006)
5. Emmerich, M., Klinkenberg, J.W.: The computation of the expected improvement in dominated hypervolume of Pareto front approximations. Rapport technique, Leiden University (2008)
6. Fonseca, C.M., Fleming, P.J.: Multiobjective optimization and multiple constraint handling with evolutionary algorithms. I. A unified formulation. IEEE Transactions on Systems, Man and Cybernetics. Part A: Systems and Humans 28(1), 26–37 (1998)
7. Gramacy, R.L., Lee, H.: Optimization under unknown constraints. In: Bayesian Statistics 9. Proceedings of the Ninth Valencia International Meeting. pp. 229–256. Oxford University Press (2011)
8. Hupkens, I., Emmerich, M., Deutz, A.: Faster computation of expected hypervolume improvement. arXiv preprint arXiv:1408.7114 (2014)
9. Jeong, S., Minemura, Y., Obayashi, S.: Optimization of combustion chamber for diesel engine using kriging model. Journal of Fluid Science and Technology 1(2), 138–146 (2006)
10. Jones, D.R., Schonlau, M., Welch, W.J.: Efficient global optimization of expensive black-box functions. Journal of Global Optimization 13(4), 455–492 (1998)
11. Li, L., Bect, J., Vazquez, E.: Bayesian Subset Simulation: a kriging-based subset simulation algorithm for the estimation of small probabilities of failure. In: Proceedings of PSAM 11 & ESREL 2012, 25-29 June 2012, Helsinki, Finland. IAPSAM (2012)
12. Liu, J.S.: Monte Carlo strategies in scientific computing. Springer (2008)
13. Mockus, J.: Application of bayesian approach to numerical methods of global and stochastic optimization. Journal of Global Optimization 4(4), 347–365 (1994)
14. O'Hagan, A., Kingman, J.: Curve fitting and optimal design for prediction. Journal of the Royal Statistical Society. Series B (Methodological) pp. 1–42 (1978)
15. Osyczka, A., Kundu, S.: A new method to solve generalized multicriteria optimization problems using the simple genetic algorithm. Structural Optimization 10(2), 94–99 (1995)
16. Oyama, A., Shimoyama, K., Fujii, K.: New constraint-handling method for multi-objective and multi-constraint evolutionary optimization. Transactions of the Japan Society for Aeronautical and Space Sciences 50(167), 56–62 (2007)
17. Parr, J.M., Keane, A.J., Forrester, A.I.J., Holden, C.M.E.: Infill sampling criteria for surrogate-based optimization with constraint handling. Engineering Optimization 44(10), 1147–1166 (2012)
18. Picheny, V.: Multiobjective optimization using Gaussian process emulators via stepwise uncertainty reduction. Statistics and Computing DOI:10.1007/s11222-014-9477-x, 1–16 (2014)

19. Picheny, V.: A stepwise uncertainty reduction approach to constrained global optimization. In: Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS), 2014, Reykjavik, Iceland. vol. 33, pp. 787–795. JMLR: W&CP (2014)
20. Ray, T., Tai, K., Seow, K.C.: Multiobjective design optimization by an evolutionary algorithm. Engineering Optimization 33(4), 399–424 (2001)
21. Sasena, M.J., Papalambros, P., Goovaerts, P.: Exploration of metamodeling sampling criteria for constrained global optimization. Engineering Optimization 34(3), 263–278 (2002)
22. Schonlau, M., Welch, W.J., Jones, D.R.: Global versus local search in constrained optimization of computer models. In: New Developments and Applications in Experimental Design: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference. IMS Lecture Notes-Monographs Series, vol. 34, pp. 11–25. Institute of Mathematical Statistics (1998)
23. Shimoyama, K., Sato, K., Jeong, S., Obayashi, S.: Updating kriging surrogate models based on the hypervolume indicator in multi-objective optimization. Journal of Mechanical Design 135(9), 094503 (2013)
24. Wagner, T., Emmerich, M., Deutz, A., Ponweiser, W.: On expected-improvement criteria for model-based multi-objective optimization. In: Parallel Problem Solving from Nature, PPSN XI. 11th International Conference, Krakov, Poland, September 11-15, 2010, Proceedings, Part I. Lecture Notes in Computer Science, vol. 6238, pp. 718–727. Springer (2010)

Part III
Conclusions & Future work

Conclusions & Future work

1 Summing up

This manuscript presents a summary of my teaching and research work. The latter represents approximately 20% of my activity during the period 2005–2014.

I would say that my research domain is that of the Bayesian approach to design and analysis of computer experiments. This domain has its roots in contributions such as those of G. S. Kimeldorf and G. Wahba (1970), J. Mockus, V. Tiesis and A. Žilinskas (1978), C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker (1991), M. Locatelli and F. Schoen (1993)... Here, the central object is a computer model of a system, which is regarded as a function $f : \mathbb{X} \to \mathbb{R}$ that maps a vector of input variables to a quantity of interest (a performance, a cost...). The act of running the model is viewed as an experiment, made for the purpose of collecting information about the properties of the system under study. Research in the domain of computer experiments aims at designing algorithms for getting information from computer models as efficiently as possible.

There are several general classes of problems that can be considered: prediction, optimization, level set estimation, sensitivity analysis... In a given real industrial problem several objectives are frequently sought at the same time. For instance, when doing simulation-based car-crash testing, one not only seeks to minimize the probability of injury of the passengers, but also, to minimize the mass of the vehicle, the cost of fabrication, etc. The most fundamental class of problems is probably that of prediction, since it relates to the others classes of problems in addition to being of interest in its own right. In our work, we follow a path traced for over more than two decades and focus on the kriging approach to prediction. The main reason for the central role of kriging in the domain of design and analysis of computer experiments is that this conceptually simple approach is well suited to a Bayesian decision-theoretic framework.

The Bayesian approach to design and analysis of computer experiments has shown to be very useful and effective in practical problems, in particular in the context of expensive computer simulations where Monte Carlo methods are not affordable. A large number of applications can be found in the literature. Yet, there are numerous methodological and theoretical questions still open at present time.

This manuscript does not provide new academic results. It consists of a general presentation of sequential search strategies based on kriging and a selection of articles organized in two parts: a range of articles about the notion of sequential uncertainty reduction, and a specific part about Bayesian optimization. The selec-

tion aims at showing that we try to study both practical and theoretical questions regarding Bayesian strategies for design and analysis of computer experiments.

From a practical point of view, we strive to address effectively industrial needs by taking into account the constraints that appear in real problems. For instance, the idea of considering the problem of the estimation of a probability of failure in 2006 was initiated and nurtured by a series of discussions with people from EADS—now AIRBUS GROUP. Another aspect of our work, which has not been emphasized in this manuscript, is the willingness to provide efficient implementations of Bayesian strategies to solve actual problems. Our motivation is twofold.

The primary motivation is a concern for *reproducible research*. Indeed, the complexity of working implementations of our algorithms is often quite high—for instance, coding a IAGO algorithm from scratch can take several weeks. Besides, the performance of an algorithm is often dependent on its implementation. Today, we try to publish our implementations inside the *Small (Matlab/GNU Octave) Toolbox for Kriging* (STK), which I am contributing to write with my colleague Julien Bect, who is the main developer today. The STK is published under a GPLv3 license and is the tool that we use daily to write and test our algorithms. As such, the STK has become central to our research activities.

The secondary motivation is that developing new algorithms and providing efficient implementations give rise to intrinsically interesting research questions. Since 2009, Julien Bect and I have been supervising PhD students on the problem of the *implementation* of kriging-based sequential strategies, and in particular, on the development of implementations based on sequential Monte Carlo techniques.

2 Ongoing work

At present, our main focus is on the problem of multi-objective optimization of real-valued functions subject to multiple inequality constraints. The problem consists in finding an approximation of the set

$$\Gamma = \{x \in \mathbb{X} : c(x) \leq 0 \text{ and } \nexists x' \in \mathbb{X} \text{ such that } f(x') \prec f(x)\},$$

where $\mathbb{X} \subset \mathbb{R}^d$ is the search domain, $c = (c_i)_{1 \leq i \leq q}$ is a vector of constraint functions ($c_i : \mathbb{X} \to \mathbb{R}$), $c(x) \leq 0$ means that $c_i(x) \leq 0$ for all $1 \leq i \leq q$, $f = (f_j)_{1 \leq j \leq p}$ is a vector of objective functions to be minimized ($f_j : \mathbb{X} \to \mathbb{R}$), and $\prec$ denotes the Pareto domination rule. Our approach so far consists in defining an extended domination rule to handle the constraints and the objectives simultaneously. Then, we define a Bayesian sampling criterion that extends the *expected improvement* sampling criterion in the single-objective unconstrained setting. The calculation and optimization of the criterion are performed using sequential Monte Carlo techniques. In particular, an algorithm similar to our *Bayesian subset simulation* method (L. Li, J. Bect and E. Vazquez; 2012) is used to estimate the expected improvement criterion.

This work is being carried out since 2014 in collaboration with industrial partners and is being supported by the SystemX Institute of Technology.

More recently, we also have started a collaboration with LNE (laboratoire national de métrologie et d'essais) on the design and analysis of multi-fidelity computer experiments applied to fire safety.

3 Perspectives

As probably expected, past studies have stirred at least as many questions they have solved. One topic that has been left out for some time, yet still in my very research interests, is the problem of the convergence of kriging-based sequential strategies. In particular, a first question concerns the possibility to improve on Adam Bull's convergence rates of the expected improvement algorithm, which are seemingly not as tight as expected. A second question concerns the convergence of our new integral loss criterion for optimization (E. Vazquez and J. Bect; 2014). A third question is about the convergence of kriging-based optimization strategies in presence of noisy evaluation results...

New directions related to the choice of priors in sequential search strategies are also of interest. In fact, we feel that the simple framework of stationary Gaussian random process with a Matérn correlation structure has never been entirely satisfying. In particular, the problem becomes particularly difficult when f has local irregularities. Second, this type of model is difficult to use when the dimension of $\mathbb{X}$ becomes high. In fact, in high dimensional problems, it is generally expected that the output of the computer model depends only on a few factors, or maybe a few combinations of them. Studying this type of setting is interesting from both theoretical and practical viewpoints.