

# PROCESSUS GAUSSIENS ET PLANIFICATION SÉQUENTIELLE D'EXPÉRIENCES NUMÉRIQUES POUR L'ESTIMATION DE QUANTILES

## En bref

Ce document présente une proposition de stage de M2 recherche au Laboratoire des Signaux et Systèmes (unité mixte de recherche CNRS – CentraleSupélec – Univ. Paris-Saclay), dans un environnement stimulant (projet collaboratif, séminaire <https://www.uqsay.org>, etc.) et offrant une possibilité de poursuite en thèse de doctorat.

Le sujet s'adresse à un(e) étudiant(e) avec une solide formation en mathématiques appliquées — probabilités, statistiques, optimisation, machine learning — couplé à un goût pour la programmation et l'expérimentation numérique.

Lisez la suite pour en savoir plus !

## Contexte

L'utilisation de simulations numériques pour concevoir ou étudier un système, quelle que soit sa nature, est un enjeu très important pour de nombreux acteurs d'industrie et du numérique. Des simulations numériques sont, par exemple, utilisées pour maximiser les performances des systèmes, les rendre plus sûrs et respectueux de l'environnement, dans un contexte économique et environnemental exigeant. On peut penser également au domaine de l'intelligence artificielle, dans lequel des simulations intensives sont utilisées pour optimiser les hyper-paramètres des algorithmes d'apprentissage statistique.

Toutefois, la modélisation numérique réaliste d'un système physique, ou l'évaluation des performances d'un algorithme d'apprentissage sur une large collection d'exemples, requiert des ressources informatiques importantes et, malgré l'essor de moyens de calcul très performants, les temps de calcul élevés qui en résultent nécessitent de conduire les simulations numériques avec parcimonie et de manière efficace.

Les méthodes de planification séquentielle d'expériences, également connue sous le nom d'apprentissage actif (*active learning*) dans le domaine de l'apprentissage statistique, permettent de guider le choix des simulations à réaliser en s'appuyant sur des modèles prédictifs, ou méta-modèles, des fonction d'intérêts (sorties du simulateurs). Ces techniques sont au cœur du projet ANR SAMOURAI [13] dans le cadre duquel est proposé ce stage. Partenaires : IFPEN, EDF R&D, Safran Tech, CEA, CentraleSupélec, EMSE et Polytechnique Montréal.

## Objectif du stage

On s'intéresse dans ce stage à l'estimation d'un ou plusieurs quantile d'un simulateur numérique coûteux, qui peut être déterministe ou stochastique. En représentant le simulateur

d'un point de vue extérieur (« boîte noire ») par une fonction  $f : \mathbb{X} \rightarrow \mathbb{R}$ , définie sur un domaine  $\mathbb{X} \subset \mathbb{R}^d$ , il s'agit étant donnée une mesure de probabilité  $\pi$  sur  $\mathbb{X}$  d'estimer la quantité

$$q_\alpha = \inf \{u \in \mathbb{R} \mid p(u) \geq \alpha\},$$

où  $\alpha \in ]0, 1[$  et  $p(u) = \int_{\mathbb{X}} \mathbb{1}_{f \leq u} d\pi$ . Autrement dit, si  $X$  désigne un vecteur aléatoire à valeurs dans  $\mathbb{X}$ , de densité  $\pi$ , on veut estimer le quantile d'ordre  $\alpha$  de la sortie  $Z = f(X)$ . Cette estimation peut se faire, selon les cas, soit à partir d'évaluations ponctuelles exactes ( $Z_i = F(X_i)$ ), soit à partir d'évaluations bruitées ( $Z_i = f(X_i) + \varepsilon_i$ ), cette deuxième situation pouvant correspondre, par exemple, au cas d'un simulateur dit « stochastique ».

Le stage se concentrera sur les méthodes de planification séquentielle (*active learning*) exploitant une modélisation probabiliste (ou « bayésienne ») de la fonction  $f$ , et plus particulièrement une modélisation par processus gaussien. Le travail proposé durant le stage consiste en :

- ☞ une étude bibliographique de l'état de l'art [1, 4, 5, 7–12];
- ☞ le développement d'un algorithme original s'appuyant sur le paradigme SUR de la réduction séquentielle d'incertitude — voir par exemple [2, 3, 6, 14, 15] — et visant à éviter le coût prohibitif des simulations conditionnelles utilisées dans [1, 8].
- ☞ une étude numérique comparative (benchmark) de la méthode proposée avec des méthodes de l'état de l'art.

## Informations pratiques

**Connaissances souhaitées.** Mathématiques appliquées, optimisation, statistiques, apprentissage. Très bonne maîtrise d'au moins un outil de programmation « numérique » standard (Matlab, R, Python...).

**Aptitudes personnelles souhaitées.** Autonomie, ouverture d'esprit, écoute, synthèse.

**Modalités.** La durée du stage est d'au moins 5 mois. L'étudiant travaillera au Laboratoire des Signaux & Systèmes (L2S), sur le campus de CentraleSupélec Paris-Saclay. Rémunération : environ 600 € par mois.

Possibilité de poursuivre par une thèse de doctorat.

**Contact.** Envoyer curriculum vitae, lettre de motivation et relevés de notes à Julien Bect et Emmanuel Vazquez {julien.bect, emmanuel.vazquez}@supelec.fr.

## 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. In *42èmes Journées de Statistique*, Marseille, France, France, 2010.
- [2] J. Bect, François Bachoc, D. Ginsbourger, et al. A supermartingale approach to gaussian process based sequential design of experiments. *Bernoulli*, 25(4A):2883–2919, 2019.
- [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. *Statistics and Computing*, 22(3):773–793, 2012.

- [4] B. Bichon and J. McFarland. Inverse reliability analysis with EGRA. In *52nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference 19th AIAA/ASME/AHS Adaptive Structures Conference 13t*, page 1908, 2011.
- [5] C. Cannamela, J. Garnier, and B. Iooss. Controlled stratification for quantile estimation. *The Annals of Applied Statistics*, 2(4):1554–1580, 2008.
- [6] C. Chevalier, J. Bect, D. Ginsbourger, E. Vazquez, V. Picheny, and Yann Richet. Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4):455–465, 2014.
- [7] M. Jala, C. Lévy-Leduc, É. Moulines, E. Conil, and J. Wiart. Sequential design of computer experiments for the assessment of fetus exposure to electromagnetic fields. *Technometrics*, 58(1):30–42, 2016.
- [8] M. Jala, C. Lévy-Leduc, Eric Moulines, E. Conil, and J. Wiart. Sequential design of computer experiments for parameter estimation with application to numerical dosimetry. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 909–913. IEEE, 2012.
- [9] T. Labopin-Richard and V. Picheny. Sequential design of experiments for estimating quantiles of black-box functions. *Statistica Sinica*, pages 853–877, 2018.
- [10] J. Oakley. Estimating percentiles of uncertain computer code outputs. *J. Roy. Statist. Soc. Ser. C*, 53(1):83–93, 2004.
- [11] N. Razaaly, D. Crommelin, and P. M. Congedo. Efficient estimation of extreme quantiles using adaptive kriging and importance sampling. *International Journal for Numerical Methods in Engineering*, 121(9):2086–2105, 2020.
- [12] R. Schöbi, B. Sudret, and S. Marelli. Rare Event Estimation Using Polynomial-Chaos Kriging. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 3(2), 2016.
- [13] D. Sinoquet et al. SAMOURAI — Simulation Analytics and Meta-model-based solutions for Optimization, Uncertainty and Reliability Analysis. Projet ANR AAPG 2020, PRCE, accepté.
- [14] E. Vazquez and M. Piera-Martinez. 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.
- [15] J. Villemonteix, E. Vazquez, and É. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534, 2009.