



Proposition de stage de fin d'étude

Sujet de stage : Etude mathématique de méthodes d'échantillonnage préférentiel en grande dimension

Localisation : ISAE-SUPAERO, Département d'Ingénierie des Systèmes Complexes

Encadrant principal : Florian Simatos (ISAE-SUPAERO), florian.simatos@isae.fr

Co-encadrant : Jérôme Morio (ONERA), jerome.morio@onera.fr

Contexte général

L'**échantillonnage préférentiel** est une méthode statistique classique de simulation et d'estimation. Elle consiste à approcher une densité f sur \mathbb{R}^d à l'aide d'un échantillon (X_1, \dots, X_n) tiré de manière indépendante selon une loi auxiliaire g en associant à chaque tirage X_i un poids d'importance w_i . Dans le cas classique, par exemple en statistique bayésienne, où f n'est connue qu'à une constante près, $f \propto \pi$, alors f est approchée par $\hat{f} = \sum_i w_i \delta_{X_i}$, où les poids auto-normalisés w_i sont donnés par

$$w_i = \frac{\tilde{w}_i}{\sum_j \tilde{w}_j} \quad \text{avec} \quad \tilde{w}_i = \frac{\pi(X_i)}{g(X_i)}.$$

L'**échantillonnage préférentiel adaptatif** (AIS) consiste à choisir la densité auxiliaire dans une famille paramétrique $\{\pi_\theta\}$ en **estimant** un "bon" paramètre, par exemple qui satisfait un critère théorique d'optimalité. Un cas représentatif est celui où π_θ est une loi gaussienne de paramètre $\theta = (\mu, \Sigma)$ avec $\mu \in \mathbb{R}^d$ la moyenne et $\Sigma \in \mathbb{R}^{d \times d}$ la matrice de variance-covariance. Dans ce cas, θ est typiquement mis à jour par estimation empirique de la moyenne et de la variance de l'estimation courante \hat{f} de f .

Cette méthode est très efficace lorsque la dimension d est faible mais lorsque d augmente, la précision se détériore très fortement pour deux raisons. D'une part, on observe un problème de **dégénérescence des poids** qui se traduit par le fait que tous les poids w_i sont presque nuls, sauf un qui porte presque toute la masse. D'autre part, estimer une matrice de covariance nécessite d'estimer un grand nombre de termes (d'ordre d^2).

Récemment, plusieurs méthodes ont été proposées afin d'améliorer l'efficacité de ces méthodes en grande dimension [1, 2, 3]. Elles cherchent à réduire la dimension du problème en projetant les paramètres à estimer dans un sous-espace de faible dimension et apportent une amélioration numérique significative. **Le but de ce stage est d'essayer de prouver des résultats théoriques pour quantifier cette amélioration, et comprendre les gains que l'on peut espérer.**

Une thèse de doctorat dans la continuité de ce stage est proposée par l'ISAE SUPAERO et l'ONERA Toulouse.

La candidate ou le candidat devront avoir une bonne maîtrise de statistique mathématique, notamment concernant les théorèmes limites, les vecteurs gaussiens et les méthodes classiques d'estimation paramétrique.



Références

- [1] Maxime El Masri, Jérôme Morio, and Florian Simatos. Optimal projection to improve parametric importance sampling in high dimension. arXiv 2107.06091.
- [2] Maxime El Masri, Jérôme Morio, and Florian Simatos. Improvement of the cross-entropy method in high dimension for failure probability estimation through a one-dimensional projection without gradient estimation. *Reliability Engineering & System Safety*, 216:107991, 2021.
- [3] Felipe Uribe, Iason Papaioannou, Youssef M. Marzouk, and Daniel Straub. Cross-entropy-based importance sampling with failure-informed dimension reduction for rare event simulation. arXiv 2006.05496, 2020.