ISAE
Institut Supérieur de l'Aéronautique et de l'Espace
SUPAERO

Sujet de stage de fin d'étude

Sujet : Projection optimale pour l'échantillonnage préférentiel en grande dimension

Localisation : ISAE-SUPAERO, Département d'Ingénierie des Systèmes Complexes
Encadrant principal : Florian Simatos (ISAE-SUPAERO), florian.simatos@isae-supaero.fr
Co-encadrant : Jérôme Morio (ONERA), jerome.morio@onera.fr

Contexte général

La méthode d'**échantillonnage préférentiel** consiste à estimer une espérance $m = E(\psi(X))$ avec X de loi f à partir d'un échantillon d'une autre loi, appelée loi auxiliaire, puis à estimer la moyenne recherchée en incorporant des rapports de vraisemblance. Ainsi, si les X_i sont tirés selon une loi auxiliaire g, alors

$(1/N)\ \Sigma_i\ \psi(X_i)\ L(X_i)$ avec $L = f/g$ est un estimateur de m.

L'estimation de ce type d'espérance apparait en fait autour de nombreux systèmes physiques schématiquement décrits par une relation du type $Y = \psi(X)$, où l'entrée multidimensionnelle X est supposée aléatoire et où la sortie Y est déterminée via la fonction déterministe ψ. Un exemple proéminent d'application est l'analyse d'un code de calcul boîte noire : φ représente alors un code de calcul, tel que des calculs de contraintes sur des structures mécaniques complexes et X les conditions extérieures dans lesquelles ce calcul est effectué. On peut notamment penser à un code de type éléments finis, dont la complexité rend impossible toute étude analytique de la fonction φ et donc de la sortie Y.

Problématique de recherche

La méthode d'**entropie croisée** permet ainsi de trouver une bonne densité auxiliaire pour l'estimation de m parmi une famille paramétrique $\{g_\theta\}$ en cherchant le paramètre θ qui minimise la divergence de Kullback-Leibler entre g_θ et la densité optimale g^* [1]. Néanmoins, il est bien connu que cette méthode n'est pas efficace quand la dimension du paramètre θ est grande, par exemple dans le cas classique où g_θ est la loi gaussienne de matrice de variance-covariance θ [2] : dans ce cas, on assiste à un **phénomène d'effondrement de l'estimation de la matrice de variance-covariance** [3]. Pour pallier cette difficulté, de récents développements [4,5] ont été proposés en estimant une matrice de covariance dans un sous-espace de dimension réduite à l'aide d'une projection. La direction de projection est alors un élément important du processus d'estimation de la matrice de covariance. L'approche décrite dans [4] pour déterminer un sous-espace est particulièrement précise sans être optimale mais elle nécessite la connaissance du gradient de ψ, ce qui n'est pas toujours envisageable. En fait, les **directions optimales de projection** sont des vecteurs propres de la matrice de covariance [5]. Néanmoins, l'estimation de ces directions optimales requiert pour l'instant une estimation de la matrice de covariance; or la matrice de covariance empirique considérée dans [5], ne résistant pas à l'effet de la dimension, ne permet donc pas toujours une estimation fine des directions de projection optimales.

Objectifs du stage

L'objectif de ce stage est de proposer une approche robuste pour estimer des directions optimales de projection pour l'échantillonnage préférentiel en grande dimension. Dans ce but, un état de l'art sur les techniques d'estimation de matrice de covariance en grande dimension et de ses vecteurs propres afin d'améliorer l'estimation des directions optimales de projection devra être effectué par l'étudiant.e. Parmi les approches potentiellement pertinentes, on pourra citer notamment [6] et [7].

Dans un second temps, l'étudiant.e devra implémenter l'approche retenue sur des fonctions ψ de

différentes complexités afin d'évaluer leur efficacité pour l'estimation d'une espérance en grande dimension. Enfin, selon les résultats obtenus, l'étudiant.e analysera dans quelle mesure cette solution peut être adaptée à l'échantillonnage préférentiel adaptatif et aux distributions optimales multi-modales en la combinant aux algorithmes récemment proposés [3,4].

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

Connaissances souhaitées : Mathématiques appliquées, probabilité et statistiques. Très bonne maîtrise d'au moins un outil de programmation « numérique »standard (Python, R. . .).

[1] R. Y. Rubinstein and D. P. Kroese, The cross-entropy method, Springer-Verlag, New York, 2004.

[2] T. Bengtsson, P. Bickel and B. Li, Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems, IMS collections, 2008.

[3] Y. El-Laham, V. Elvira and M. F. Bugallo, Robust Covariance Adaptation in Adaptive Importance Sampling, IEEE Signal Processing Letters, 2018.

[4] F. Uribe, I. Papaioannou, Y. Marzouk and D. Straub, Cross-entropy-based importance sampling with failure-informed dimension reduction for rare event simulation, SIAM/ASA Journal on Uncertainty Quantification, 2021.

[5] M. El Masri, J. Morio and F. Simatos, Optimal projection to improve parametric importance sampling in high dimension. arXiv preprint arXiv:2107.06091, 2021.

[6] O. Ledoit and M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices. Journal of multivariate analysis, 2004.

[7] K. Ashurbekova, A. Usseglio-Carleve F. Forbes and S. Achard, Optimal shrinkage for robust covariance matrix estimators in a small sample size setting, hal-02378034, 2020.