Sujet de stage

Etude de problèmes d'optimisation pour l'échantillonage préférentiel

Localisation : ISAE-SUPAERO, Département d'Ingénierie des Systèmes Complexes
Co-encadrants : Emilien Flayac (ISAE-SUPAERO), emilien.flayac@isae.fr et Florian Simatos (ISAE-SUPAERO), florian.simatos@isae.fr

Contexte général

L'**échantillonnage préférentiel** est une méthode classique de simulation et d'estimation qui consiste à approcher une densité f sur $\mathbb{R}^d$ à l'aide d'un échantillon i.i.d. $(X_1, \dots, X_n)$ tiré selon une loi auxiliaire g en pondérant chaque tirage X_i d'un poids d'importance w_i. Dans le cas de l'estimation d'une intégrale I de la forme

$$I = \mathbb{E}_f(\varphi(X)) = \int_{\mathbb{R}^d} \varphi(x) f(x) \mathrm{d}x$$

pour une certaine fonction $\varphi : \mathbb{R}^d \to \mathbb{R}$, l'estimateur d'échantillonage préférentiel est donné par

$$\hat{I}_g = \frac{1}{n} \sum_{i=1}^{n} w_i \varphi(X_i) \quad \text{avec} \quad w_i = \frac{f(X_i)}{g(X_i)}.$$

Il est bien connu que le choix de g gouverne la qualité de l'estimateur $\hat{I}_g$, qui selon g peut être plus ou moins efficace que l'estimateur Monte-Carlo habituel (correspondant au cas $g = f$). Dans le cas où $\varphi \geq 0$, le choix optimal de densité auxiliaire est donné par $g_{\mathrm{opt}} = \varphi f / I$, qui n'est pas utilisable en pratique car les poids d'importance $w_i = I/\varphi(X_i)$ ne sont alors pas calculables.
Une méthode classique pour choisir g est donnée par la **méthode d'entropie croisée**. Dans ce cas, on se donne une famille $\mathcal{G}$ de densités auxiliaires, et l'on cherche à choisir $g \in \mathcal{G}$ qui soit la plus proche possible de g_{opt} au sens de la divergence de Kullback–Leibler, i.e., on cherche g solution de

$$\arg\min_{g \in \mathcal{G}} D(g_{\mathrm{opt}}, g) \quad \text{avec} \quad D(h, g) = \int_{\mathbb{R}^d} h(x) \ln\left(\frac{h(x)}{g(x)}\right) \mathrm{d}x. \tag{CE}$$

Néanmoins, cette approche mène souvent à des densités auxiliaires g telles que les variables $w_i \varphi(X_i)$ ont un second moment infini, i.e., avec

$$\mathbb{E}_g(w_i^2 \varphi(X_i)^2) = \int \left(\frac{f}{g}\right)^2 \varphi^2 g = \int \frac{(\varphi f)^2}{g} = \infty,$$

ce qui dégrade la vitesse de convergence de l'estimateur $\hat{I}_g$.

Objectifs du stage

L'objectif principal du stage est de formuler et étudier des problèmes d'optimisation inspirés de (CE) qui viseraient à définir des densités auxiliaires proches de g_{opt} au sens de la divergence de Kullback–Leibler, mais sous des contraintes de moments finis. On peut par exemple penser aux deux problèmes suivants :

$$\arg\min \left\{ D(g_{\mathrm{opt}}, g) : g \in \mathcal{G} \quad \text{et} \quad \int \frac{(\varphi f)^2}{g} \leq K \right\} \tag{CE-1}$$

et

$$\arg\max \left\{ \beta : g \in \mathcal{G}, \ \int \frac{(\varphi f)^\beta}{g^{\beta-1}} < \infty \quad \text{et} \quad D(g_{\mathrm{opt}}, g) \leq K \right\}. \tag{CE-2}$$

Le premier problème (CE-1) vise à sélectionner une densité g la plus proche de g_{opt}, mais sous une contrainte sur le second moment de $w_i\varphi(X_i)$. Et le second problème (CE-2) vise à sélectionner une densité g ayant le plus de moments finis possibles pour $w_i\varphi(X_i)$, mais sous une contrainte de rester proche de g_{opt}.
Le stage comportera une dimension d'optimisation, puisqu'il s'agira de formuler des problèmes tels que (CE-1) et (CE-2) et d'essayer de les résoudre pour différentes familles de loi $\mathcal{G}$.
Une première difficulté vient du fait que la contrainte de moments qui apparaît dans les deux problèmes est non-linéaire en la densité g qui est la variable du problème. Une deuxième difficulté provient du fait que, d'un point de vue général, le choix d'une famille paramétrique engendre un problème d'optimisation non-linéaire pour lesquels des méthodes numériques spécifiques sont nécessaires. En particulier, on s'intéressera au cas gaussien multivarié qui mène à des problèmes d'optimisation sur l'espace des matrices de covariance i.e. symétriques définies positives. On essaiera de généraliser au cas des lois exponentielles connues pour être la classe de lois optimales en présence de contraintes de moments linéaires [2].
Dans un second temps, l'étudiant.e devra mobiliser des résultats de **statistique mathématique** pour déterminer les propriétés de convergence des estimateurs d'échantillonage préférentiel obtenus avec les densités auxiliaires solutions de (CE-1) et (CE-2).
Si les résultats en dimension finie s'avèrent concluants, nous essaierons aussi d'étudier ce problème en grande dimension, i.e., dans le régime asymptotique $d \to \infty$. En effet, des résultats récents ont montré que les densités solutions de (CE) étaient optimales en terme de consistance [1], mais en pratique, ces densités n'ont en général pas de second moment fini et convergent donc lentement. A l'inverse, forcer des densités avec un second moment semble une bonne idée du point de vue de la vitesse de convergence, mais peut être sous-optimal en terme de consistance. Ainsi, il s'agira de comprendre les compromis réalisables entre consistance et vitesse de convergence dans le régime de grande dimension.

Référence

[1] Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.
[2] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, Hoboken, NJ, 2001.

ISAE SUPAERO - Institut Supérieur de l'Aéronautique et de l'Espace
10, avenue Edouard-Belin - BP 54032 - 31055 Toulouse CEDEX 4 - FRANCE
Tél : 33 (0)5 61 33 80 80 - Télécopie : 33 (0)5 61 33 83 30 - Site internet : www.isae-supaero.fr