0.1 RESUMÉ EN FRANÇAIS

Le sous-échantillonnage est une tâche récurrente en mathématiques appliquées. Ce paradigme a des applications en traitement du signal, l'analyse des données, l'apprentissage automatique ou bien les statistiques: la discrétisation des signaux analogiques, le calcul approché des intégrales, la réduction de dimension, la réduction du budget d'étiquetage des algorithmes d'apprentissage... Alors qu'ils paraissent différents, ces problèmes peuvent être abordés avec la même stratégie: chercher les éléments les plus représentatifs d'un ensemble. Un bon sous-ensemble de représentants doit éviter de contenir des informations redondantes. Pour certains problèmes à structure linéaire, l'ensemble peut être plongé dans un espace vectoriel et la redondance d'un sous-ensemble peut se mesurer à l'aide du volume du polytope engendré par ce sous-ensemble. Il se trouve qu'il existe une famille de modèles probabilistes qui définissent des sous-ensembles aléatoires avec une propriété de répulsion: d'une façon informelle, la probabilité d'apparition d'un sous-ensemble est proportionnelle au volume qu'il engendre dans cet espace vectoriel. Ces modèles sont connus sous le nom des processus ponctuels déterminantaux et ils ont été étudiés dans plusieurs domaines: les matrices aléatoires, l'optique quantique, les statistiques spatiales, le traitement des images, l'apprentissage automatique et récemment l'intégration numérique.

Cette thèse est consacrée à l'étude de la pertinence des DPPs pour certaines tâches de sous-échantillonnage. Dans un premier temps, nous avons considéré le problème de sélection d'attributs: pour une matrice qui représente des données exprimées sur un système d'attributs, on cherche à sélectionner les attributs les plus représentatifs. En particulier, nous avons étudié l'échantillonnage volumique, un algorithme bien connu dans la littérature, à travers la théorie des DPPs. Nous avons proposé un algorithme impliquant un DPP avec de meilleures garanties théoriques et de meilleures performances empiriques. Le choix de ce DPP était motivé par une nouvelle interprétation géométrique que nous avons mise en évidence: un DPP définit naturellement un sous-espace vectoriel aléatoire qui "flotte" autour d'un sous-espace de référence.

A l'aide de cette nouvelle interprétation, nous avons réussi à étudier un autre problème d'approximation: l'approximation d'intégrales de fonctions qui vivent dans un espace à noyau, aussi appelé le problème de quadrature à noyau.

Pour ce problème, nous avons proposé une nouvelle classe de quadratures: les quadratures à noyau optimisées et basées sur des noeuds qui suivent la distribution d'un DPP. La définition de ce DPP est basée sur les fonctions propres de l'opérateur d'intégration correspondant. Nous avons montré que les taux de convergence de cette classe de quadratures dépendent des valeurs propres de cet opérateur: plus le noyau est régulier, meilleure est la convergence de la quadrature. Néanmoins, les expériences numériques montrent que ces taux de convergence sont pessimistes pour certains espaces fonctionnels.

Cette observation a motivé l'extension de l'échantillonnage volumique au domaine continu. Nous avons étudié le problème de quadrature à noyau ainsi que le problème d'interpolation à noyau pour des noeuds qui suivent cette nouvelle distribution. En particulier, nous avons démontré des formules closes de l'espérance de l'erreur sous cette distribution répulsive. Ces formules ont permis de démontrer l'optimalité de l'échantillonnage volumique pour cette classe de problèmes d'approximation. De plus,

cette nouvelle distribution peut être approchée par un algorithme MCMC qui peut être implémenté sans le recours à la décomposition spectrale de l'opérateur d'intégration.

0.2 RÉSUME EN ANGLAIS

Subsampling is a recurrent task in applied mathematics. This paradigm has many applications in data analysis, signal processing, machine learning and statistics: continuous signal discretization, numerical integration, dimension reduction, learning on a budget, preconditioning... Seemingly unrelated, these problems can be tackled using the same strategy: looking for the most representative elements in a set. A good subset of representatives would capture the essential information avoiding any unnecessary redundancy. For some problems with linear structure, the set can be embedded in a vector space, and the redundancy may be measured by the volume spanned by these elements. Intuitively, a subset would be redundant if it defines a polytope with small volume, and it would be non-redundant if it defines a polytope with large volume. It turns out that there exists a family of probabilistic models that define random subsets with a repulsion property: informally, the probability of appearance of a subset is proportional to the volume it defines in the vector space. These models are called determinantal point processes. They were the topic of intense research in various fields: random matrices, quantum optics, spatial statistics, image processing, machine learning and recently numerical integration. This thesis is devoted to investigating the relevance of determinantal point processes in some popular subsampling tasks.

First, we considered the column subset selection problem. In this problem we look for selecting the most representative columns, or features, of a matrix that represents a dataset. In particular, we investigated volume sampling, a well-known algorithm for this task, through the lens of DPPs. We also proposed an alternative algorithm, based on a DPP, with better theoretical guarantees and empirical performance. The choice of this DPP was motivated by a new geometric interpretation: a DPP naturally defines a random linear subspace that "hovers" around a reference subspace.

Using this new interpretation of DPPs, we provided the theoretical analysis of another approximation problem, namely quadrature. These quadratures are suitable for the approximation of integrals of functions living in a reproducing kernel Hilbert space. We proposed a new class of quadratures: optimal kernel quadrature based on nodes that follow the distribution of a DPP. This DPP is defined through the eigenfunctions of the corresponding integration operator. We showed that the rates of convergence of this class of quadratures depend on the eigenvalues of the integration operator: the smoother is the kernel, the faster is the convergence of the quadrature. However, empirical investigations showed that these theoretical rates were pessimistic with respect to the empirical rates that are observed numerically.

This observation motivated the extension of volume sampling to continuous domains. Indeed, we studied both kernel quadrature and kernel interpolation under this distribution. In particular, we proved tractable formulas of the expected value of the error under this repulsive distribution. These tractable formulas were useful to derive sharp upper bounds for kernel quadrature, which scale as the existing lower bounds. Moreover, the continuous volume sampling distribution has the advantage to be amenable to sampling via a fully kernelized MCMC algorithm. In other words, the implementation of this algorithm relies on evaluating the kernel and does not require the spectral decomposition

of the integration operator. This makes volume sampling a promising approach to kernel quadrature, with both sharp error bounds and an avenue for tractable sampling algorithms.

0.3 RÉSUME VULGARISÉ EN FRANÇAIS

Les processus ponctuels déterminantaux sont des modèles probabilistes de répulsion. Ces modèles ont été étudié dans différents domaines: les matrices aléatoires, l'optique quantique, les statistiques spatiales, le traitement d'images, l'apprentissage automatique et récemment les quadratures. Dans cette thèse, on étudie l'échantillonnage des sous-espaces à l'aide des processus ponctuels déterminantaux. Ce problème se trouve à l'intersection de trois branches de la théorie d'approximation: la sous sélection dans les ensembles discrets, la quadrature à noyau et l'interpolation à noyau. On étudie ces questions classiques à travers une nouvelle interprétation de ces modèles aléatoires: un processus ponctuel déterminantal est une façon naturelle de définir un sous-espace aléatoire. En plus de donner une analyse unifiée de l'intégration et l'interpolation numériques sous les DPPs, cette nouvelle approche permet de développer les garanties théoriques de plusieurs algorithmes à base de DPPs, et même de prouver leur optimalité pour certains problèmes.

0.4 RÉSUME VULGARISÉ EN ANGLAIS

Determinantal point processes are probabilistic models of repulsion. These models were studied in various fields: random matrices, quantum optics, spatial statistics, image processing, machine learning and recently numerical integration. In this thesis, we study subspace sampling using determinantal point processes. This problem takes place within the intersection of three sub-domains of approximation theory: subset selection, kernel quadrature and kernel interpolation. We study these classical topics, through a new interpretation of these probabilistic models: a determinantal point process is a natural way to define a random subspace. Beside giving a unified analysis to numerical integration and interpolation under determinantal point processes, this new perspective allows to work out the theoretical guarantees of several approximation algorithms, and to prove their optimality in some settings.