

# Les processus ponctuels déterminantaux en apprentissage automatique

Guillaume GAUTIER,<sup>†\*</sup> Rémi BARDENET,<sup>†</sup> Michal VALKO<sup>\*‡</sup>

<sup>†</sup>Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRISAL

<sup>\*</sup>INRIA Lille - Nord Europe, <sup>‡</sup>DeepMind Paris

G.GAUTIER@INRIA.FR, REMI.BARDENET@GMAIL.COM, MICHAL.VALKO@INRIA.FR

**Résumé** – Pour la session «Outils mathématiques pour l’apprentissage automatique», nous proposons une courte revue des processus ponctuels déterminantaux, un modèle et un outil probabiliste populaire en apprentissage et prometteur en traitement du signal.

**Abstract** – For the session “Mathematical tools in machine learning”, we propose a short survey of determinantal point processes, a popular probabilistic model and tool in machine learning, which already has promising applications in signal processing.

Un processus ponctuel définit un ensemble aléatoire de points. Ces points vivent typiquement dans  $\mathbb{R}^d$  ou dans un ensemble fini. Un processus ponctuel *déterminantal* (DPP) est un type de processus ponctuel dont les points se repoussent. La répulsion est encodée par les sous-déterminants d’un *noyau*, au sens des méthodes à noyau en apprentissage artificiel. Cette forme algébrique particulière associée à un choix judicieux du noyau, confère aux DPP de nombreux avantages statistiques et computationnels. En effet, chose rare pour des processus répulsifs, les quantités-clés (e.g., la constante de normalisation) admettent des expressions explicites et les calculs-clés sont d’ordre polynomial.

Les DPP ont été introduits par Odile Macchi [15] comme modèle en optique fermionique. Ils sont ensuite devenus un outil pour l’étude de matrices aléatoires [10], et ont trouvé des applications aussi variées qu’en statistiques spatiales [13], intégration Monte Carlo [1], traitement du signal et apprentissage artificiel, où ils sont utilisés à la fois comme modèles [11] et comme outil d’échantillonnage intelligent [18, 3].

Nous proposons une courte revue des rôles joués par les DPP en apprentissage et des problèmes de recherche associés. Nous introduisons le formalisme des DPP à espace d’états fini en Section 1. En Sections 2 et 3, nous discutons quelques applications des DPP respectivement comme modèles et outils en apprentissage. En Sections 4 et 5, nous présentons les verrous méthodologiques que sont l’échantillonnage et l’inférence.

## 1 Encoder la répulsion avec un noyau

Pour les détails de cette section, nous référons à [11]. En particulier, nous nous limitons par simplicité au cas où les points du processus appartiennent à un ensemble fini  $[n] := \{1, \dots, n\}$ . On peut penser, par exemple, que chaque élément de  $[n]$  est l’indice d’une phrase dans un grand corpus de textes. Pour former un résumé de ce corpus, il s’agit de tirer un ensemble de phrases portant chacune de l’information utile, mais qui sont différentes les unes des autres. Nous allons voir qu’on peut penser un DPP, ou sa variante un  $k$ -DPP, comme une recette qui crée aléatoirement de tels résumés.

### 1.1 Les DPP

Une façon de définir un ensemble aléatoire est de fixer ce qu’on appelle les probabilités d’inclusion.

**Définition 1** Soit une matrice  $\mathbf{K} \in \mathbb{R}^{n \times n}$ . On appelle DPP de noyau  $\mathbf{K}$  l’ensemble aléatoire  $\mathcal{X} \subset [n]$  défini par

$$\forall S \subset [n], \quad \mathbb{P}[S \subset \mathcal{X}] = \det(\mathbf{K}_S), \quad (1)$$

où  $\mathbf{K}_S$  est la sous-matrice  $[\mathbf{K}_{ij}]_{i,j \in S}$ , avec  $\det(\mathbf{K}_\emptyset) := 1$ .

Pour une matrice  $\mathbf{K}$  donnée, il n’est pas évident que (1) définisse bien une distribution de probabilité. Dans la suite, nous supposons que le noyau  $\mathbf{K}$  est symétrique ; un résultat fondateur garantit alors l’existence du DPP lorsque ses valeurs propres sont comprises dans l’intervalle  $[0, 1]$ . L’hypothèse de symétrie est plus qu’une commodité mathématique permettant de diagonaliser  $\mathbf{K}$ , elle rend aussi le modèle facilement interprétable et lui confère son caractère répulsif. En effet, pour deux objets  $i \neq j \in [n]$ , (1) donne alors

$$\begin{aligned} \mathbb{P}[\{i, j\} \subset \mathcal{X}] &= \mathbf{K}_{ii}\mathbf{K}_{jj} - \mathbf{K}_{ij}\mathbf{K}_{ji} \\ &= \mathbb{P}[i \in \mathcal{X}]\mathbb{P}[j \in \mathcal{X}] - \mathbf{K}_{ij}^2 \end{aligned} \quad (2)$$

$$\leq \mathbb{P}[i \in \mathcal{X}]\mathbb{P}[j \in \mathcal{X}]. \quad (3)$$

On interprète (2) de la façon suivante : la probabilité de trouver deux éléments  $i$  et  $j$  dans un même échantillon est d’autant plus faible que leur «similarité»  $\mathbf{K}_{ij}$  est grande. Autrement dit, la présence de coefficients hors diagonaux dans  $\mathbf{K}$  traduit l’écart du DPP aux tirages indépendants ayant les mêmes probabilités marginales, cf. (3).

Pour reprendre l’exemple des résumés de texte mentionné en début de section, un DPP qui génère de bons résumés correspond intuitivement à un noyau avec des coefficients  $\mathbf{K}_{ii}$  d’autant plus grands que la phrase  $i$  est marginalement pertinente (*contient-elle des mots appartenant au champ lexical d’intérêt ?*), et des coefficients  $\mathbf{K}_{ij}$  dont la valeur absolue croît avec la similarité des phrases  $i$  et  $j$  (*utilisent-elles les mêmes mots ?*). Ainsi, les résumés retournés par le DPP comportent avec grande probabilité des phrases à la fois marginalement pertinentes et conjointement diverses.

Lorsque  $\mathbf{K}$  a ses valeurs propres dans  $\{0, 1\}$ , on parle de DPP de projection ou *élémentaire*. Ce sont les briques de base des DPP. En effet, tout DPP peut s'écrire comme un mélange de DPP de projection, une propriété cruciale pour l'échantillonnage, cf. Section 4. En outre, les tirages d'un DPP de projection ont (presque sûrement) toujours le même nombre d'éléments, égal au rang de  $\mathbf{K}$  [11, Lemma 2.7].

En apprentissage, les DPP sont souvent définis à travers leur noyau de vraisemblance  $\mathbf{L}$  plutôt que  $\mathbf{K}$ .

**Proposition 2** Soit une matrice  $\mathbf{L} \in \mathbb{R}^{n \times n}$  semi-définie positive. L'ensemble aléatoire  $\mathcal{X} \subset [n]$  défini par

$$\forall S \subset [n], \quad \mathbb{P}[\mathcal{X} = S] = \det(\mathbf{L}_S) / \det(\mathbf{I} + \mathbf{L}), \quad (4)$$

correspond au DPP de noyau  $\mathbf{K} = \mathbf{L}(\mathbf{I} + \mathbf{L})^{-1}$ .

Définir un DPP via la Proposition 2 laisse plus de liberté sur le choix de  $\mathbf{L}$  que la Définition 1, mais elle ne couvre pas des cas importants. En particulier, les noyaux  $\mathbf{K}$  de projection ne peuvent être représentés. Pour pallier ce manque et retrouver des modèles de cardinal fixé, Kulesza & Taskar [11, Section 5] ont introduit les  $k$ -DPP.

## 1.2 Les $k$ -DPP

Au-delà des DPP de projection dont le cardinal est lié au noyau de façon intrinsèque, il existe une autre façon naturelle de définir un processus répulsif de cardinal  $k$  prescrit.

**Définition 3** Soit une matrice  $\mathbf{L} \in \mathbb{R}^{n \times n}$  semi-définie positive. L'ensemble aléatoire  $\mathcal{X} \subset [n]$  défini par

$$\forall S \subset [n], \quad \mathbb{P}[\mathcal{X} = S] \propto \mathbb{1}_{\{|S|=k\}} \det(\mathbf{L}_S), \quad (5)$$

est appelé  $k$ -DPP de noyau  $\mathbf{L}$ .

Ainsi les  $k$ -DPP peuvent-ils être compris comme des DPP de noyau  $\mathbf{L}$  conditionnés à avoir exactement  $k$  éléments. Et même si un déterminant apparaît toujours dans la définition, ce ne sont pas toujours des DPP au sens de la Définition 1 : seuls les DPP de projection sont aussi des  $k$ -DPP.

Les  $k$ -DPP sont populaires en apprentissage, alors que les DPP caractérisés par (1) ont la faveur des autres domaines, ce qui occasionne des confusions. La raison principale de l'adoption des  $k$ -DPP en apprentissage repose sur la liberté de construction du noyau  $\mathbf{L}$  : toute matrice semi-définie positive qui possède au moins un mineur principal de taille  $k$  non nul définit bien un  $k$ -DPP.

A contrario, définir un DPP par (1) requiert de vérifier que le noyau satisfait certaines propriétés, par exemple, que  $\mathbf{K}$  est symétrique et à valeurs propres dans  $[0, 1]$ . À ce prix, les DPP offrent une plus grande interprétabilité et une boîte à outils mathématique plus importante.

## 2 Les DPP comme modèles

La principale application des DPP et  $k$ -DPP en apprentissage concerne la sélection d'éléments d'un grand corpus sous critère

de diversité. Ceci dans le but de générer des résumés de texte, faire des recommandations (plateformes média, sites marchands, etc.), ou encore diversifier les réponses d'un moteur de recherche d'images; cf. [11] pour des pointeurs.

Dans la plupart de ces applications, le processus de modélisation, i.e., le choix du noyau, est similaire. Considérons l'exemple d'un utilisateur qui effectue la recherche «bolt» dans son moteur de recherche d'images sur la toile. La Figure 1 donne deux exemples de sorties : la première ligne contient des images marginalement pertinentes, mais sans diversité. Pour désambiguïser la requête et afficher des résultats représentatifs de la polysémie du mot «bolt» en anglais, on préfère renvoyer la seconde ligne d'images, qui préserve la pertinence et inclut de la diversité.

En pratique, on affecte à chaque image une pertinence  $q_i \geq 0$  qui croît par exemple avec le nombre d'occurrences du mot «bolt» sur la page dont est extraite l'image, la popularité de cette page, etc. Ensuite, on définit une notion de similarité entre images. On procède souvent en construisant, pour chaque image  $i$ , un vecteur de descripteurs  $\phi_i \in \mathbb{R}^d$  tel que  $|\phi_i^\top \phi_j|$  est d'autant plus grand que les deux images sont similaires. Des descripteurs classiques en traitement d'images sont l'histogramme des couleurs présentes, des descripteurs SIFT, etc. On normalise ensuite chacun de ces  $n$  vecteurs d'attributs  $\phi_i$ . On groupe enfin la pertinence  $q_i$  et les attributs  $\phi_i$  de tous les objets dans une matrice  $\Phi \in \mathbb{R}^{n \times d}$ , dont la ligne  $\Phi_i$  égale  $\sqrt{q_i} \phi_i$ . La normalisation des attributs a pour effet de rendre la norme de chaque ligne de  $\Phi$  égale à sa pertinence  $q_i$ . Enfin, on définit le noyau

$$\mathbf{L} = \Phi \Phi^\top. \quad (6)$$

L'interprétation géométrique est la suivante. Que ce soit pour le DPP ou le  $k$ -DPP de noyau  $\mathbf{L}$ , chaque sous-ensemble  $S$  est tiré proportionnellement au carré du volume du parallétope de  $\mathbb{R}^{|S|}$  dont les arêtes sont les lignes  $\Phi_i$  pour  $i \in S$ . La Figure 1(b) illustre deux couples d'images et les parallétopes associés : si le noyau est bien choisi, le volume augmente avec la diversité souhaitée. Ainsi, le DPP et le  $k$ -DPP de noyau  $\mathbf{L}$  vont bien favoriser des images à grand  $q_i$ , mais avec des attributs  $\phi_i$  les plus orthogonaux possibles. Toutefois, il n'est pas nécessairement vrai que  $\mathbb{P}[i \in \mathcal{X}] = q_i$ . Cette marginale est en fait assez subtile [11, Section 5.2.3] et n'est pas directement reliée avec ce que l'utilisateur voulait modéliser.

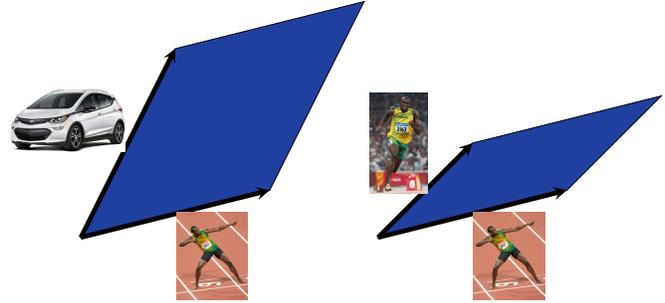
Définir le noyau  $\mathbf{K}$  d'un DPP comme en Définition 1 donne des marginales plus interprétables. En effet, utilisons la même matrice d'attributs dilatés  $\Phi$ . En supposant cette matrice de rang plein, on définit alors la matrice de projection sur l'espace engendré par les colonnes de  $\Phi$ , i.e., les attributs:

$$\mathbf{K} = \Phi [\Phi^\top \Phi]^{-1} \Phi^\top. \quad (7)$$

Pour le DPP de noyau  $\mathbf{K}$ , la probabilité marginale  $\mathbf{K}_{ii}$  de tirer l'objet  $i$  correspond au levier (*leverage score* en anglais) du vecteur  $\Phi_i$ . Le levier est une mesure familière en régression linéaire qui identifie les points influents de l'ensemble d'entraînement : une petite variation de l'observation associée



(a) Deux exemples de réponses du moteur de recherche.



(b) Le volume du parallélotope des attributs définit la répulsion.

Figure 1: Illustration de l'utilisation d'un DPP dans un moteur de recherche d'images, pour la requête «bolt».

à une ligne à fort levier aura une grande répercussion sur la régression. Concrètement, si les colonnes de  $\Phi$  sont orthogonales, le levier de l'élément  $i$  est bien sa pertinence  $q_i$ . Sinon, la marginale  $\mathbb{P}[i \in \mathcal{X}]$  grandit avec la mesure de pertinence  $q_i$ , mais aussi avec le «non-alignement» de  $\phi_i$  avec les autres vecteurs d'attributs. Là encore, on obtiendra des échantillons représentés par des vecteurs qui ont à la fois une grande norme (pertinence) et sont presque orthogonaux (diversité).

En revanche, les DPP sont plus rigides sur le choix du nombre de points : si l'utilisateur veut nécessairement un ensemble à  $k$  éléments à partir de (7), il faut soit choisir le nombre de descripteurs  $d = k$ , soit ne retenir qu'un sous-ensemble des  $d \geq k$  points tirés. Une solution naturelle consiste à retenir chaque point indépendamment avec probabilité  $k/d$ , on obtient alors un DPP de noyau proportionnel au noyau initial, mais qui n'est plus de projection [13].

Malgré les avantages d'interprétabilité des DPP, la simplicité de définition des  $k$ -DPP les rend majoritaires dans les publications en apprentissage. Peut-être est-ce appelé à changer, notamment lorsque les DPP sont utilisés comme outils.

### 3 Les DPP comme outils

La répulsion «algébrique» des DPP en fait un outil naturel lorsqu'il s'agit de choisir aléatoirement des vecteurs qui doivent jouer le rôle de base d'un espace vectoriel, un problème récurrent en apprentissage comme en traitement du signal. Dans ce cas, les DPP servent d'outil d'échantillonnage intelligent plutôt que de modèle statistique. Un choix de noyau judicieux couplé à l'arsenal mathématique des DPP peut alors donner des garanties théoriques fortes. En traitement du signal sur graphes, par exemple, les DPP sont utilisés pour choisir les nœuds auxquels évaluer un signal à bande limitée [18].

Nous illustrons ici les DPP comme outils pour la sélection d'attributs en apprentissage [3], les détails faisant l'objet d'une autre soumission GRETSI. Soit  $\mathbf{X} \in \mathbb{R}^{n \times d}$  une matrice de  $n$  données en dimension  $d$ . Lorsque  $d \gg 1$ , il est parfois souhaitable de réduire la dimension de représentation à  $k \leq d$ . L'analyse en composantes principales (ACP) utilise la décomposition en valeurs singulières  $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$ , et

préconise de projeter les données sur les colonnes de la matrice  $\mathbf{V}_{:[k]}$  associée aux  $k$  plus grandes valeurs singulières de  $\mathbf{X}$ . Cependant les attributs ainsi obtenus n'ont plus d'interprétation claire. Si l'interprétation est clé, on préfère à l'ACP des méthodes de sélection de colonnes, i.e., d'attributs. Belhadji et al. [3] proposent de sélectionner des colonnes par un tirage du DPP de projection de noyau  $\mathbf{K} = \mathbf{V}_{:[k]}\mathbf{V}_{:[k]}^T$ . Ils montrent que l'erreur moyenne d'approximation de la matrice  $\mathbf{X}$  par sa projection sur les colonnes sélectionnées est proche de l'erreur (optimale) de la PCA. Sans hypothèse supplémentaire, leurs bornes atteignent l'état de l'art (établi par un  $k$ -DPP !), et se révèlent meilleures dès que les leviers des attributs sont parcimonieux. Les DPP permettent donc une projection aléatoire contrainte qui s'adapte à la parcimonie des leviers. Des garanties sur l'erreur de prédiction de la régression faite à partir des données projetées sont également données.

Parmi tous les processus ponctuels répulsifs qui peuvent être utilisés comme outils, les DPP de propriétés computationnelles attractives. En particulier, ils peuvent être échantillonnés, i.e., simulés, en temps polynomial.

### 4 Échantillonner un DPP

Il existe des procédures génériques permettant de simuler tout DPP de manière exacte. Sans prétraitement du noyau [17], chaque échantillon coûte  $\mathcal{O}(n^3)$  dans le cas général, à l'exception des DPP de projection qui, rappelons le, constituent les briques élémentaires des DPP. Ces derniers peuvent être simulés en  $\mathcal{O}(nr^2)$ , où  $r$  est le rang du noyau  $\mathbf{K}$ . Sinon, la décomposition spectrale ou de Cholesky [12] du noyau sous-jacent est requise. Une fois ce prétraitement initial d'ordre  $\mathcal{O}(n^3)$  effectué, chaque échantillon coûte  $\mathcal{O}(nm^2)$ , où  $m$  est le nombre moyen d'éléments du DPP considéré.

À l'instar des autres méthodes à noyaux, ces procédures ne passent pas à l'échelle lorsque  $n \gg 1$ . Cet obstacle a orienté la communauté apprentissage vers des procédures exactes moins coûteuses dans certains régimes [11, 5], ainsi que des méthodes approchées, e.g., approximation de faible rang du noyau, projection aléatoire [11] ou des méthodes à base de chaînes de Markov [14, 7].

Afin de remettre l’accent sur l’interprétabilité des DPP, nous décrivons les deux grandes étapes du schéma d’échantillonnage exact d’un DPP de noyau  $\mathbf{K}$  fourni par [9, Theorem 7]. Tout DPP peut s’écrire comme un mélange de DPP de projection. Plus précisément, pour sélectionner une composante du mélange, on commence par tirer des Bernoullis indépendantes de paramètres les valeurs propres de  $\mathbf{K}$ . Soit  $d$  le nombre de 1 ainsi obtenus. On rassemble les vecteurs propres correspondants aux Bernoullis égales à 1 dans une matrice  $\Phi \in \mathbb{R}^{n \times d}$ . Il reste enfin de simuler du DPP de projection de noyau  $\tilde{\mathbf{K}} = \Phi\Phi^\top$ . Ce dernier ayant un cardinal presque sûrement fixé à  $d$ , on sait donc déjà que les réalisations d’un DPP comportent  $m = \text{Tr } \mathbf{K}$  éléments en moyenne.

Pour finir, on peut tirer séquentiellement les  $d$  points du DPP de projection de noyau  $\tilde{\mathbf{K}}$ . À la fin de l’étape  $i < d$ , on a déjà sélectionné un sous-ensemble  $S$  de taille  $i$ . À l’étape suivante, on ajoute à  $S$  un nouvel élément  $t \in [n]$  proportionnellement à  $\det(\tilde{\mathbf{K}}_{S \cup \{t\}}) / \det(\tilde{\mathbf{K}}_S) = \text{distance}^2(\Phi_t, \text{Vect}\{\Phi_s\}_{s \in S})$ . (8) L’intuition géométrique est là : les éléments avec une grande chance d’être ajoutés à l’échantillon dans (8) sont ceux dont le vecteur d’attributs latents  $\Phi_t$  est de grande norme et est le plus orthogonal possible aux attributs déjà sélectionnés. Le tirage d’un DPP s’apparente via (8) à une orthogonalisation de Gram-Schmidt randomisée.

Finalement, mentionnons que l’échantillonnage des  $k$ -DPP est plus subtil car il fait intervenir les polynômes élémentaires symétriques évalués en les valeurs propres du noyau. Pour plus de détails sur l’échantillonnage, nous renvoyons à [19] et à la documentation  du projet DPPy.

## 5 Inférer le noyau d’un DPP

Lorsque les DPP sont utilisés comme modèles, le noyau doit être appris à partir de données observées. Se pose alors la question de la méthode d’inférence. Celles basées sur la vraisemblance (4) sont les plus populaires ; les fréquentistes cherchent à la maximiser [8] tandis que les bayésiens la multiplient par une loi a priori pour ensuite l’intégrer [2, 20].

Une méthode de point fixe a été proposée pour maximiser la vraisemblance sans paramétrer  $\mathbf{L}$ , mais l’interprétation du noyau appris est difficile [16]. Au prix d’une perte de généralité, il est plus commun de paramétrer les entrées de  $\mathbf{K}$  ou  $\mathbf{L}$  par un petit nombre de quantités interprétables [13].

Cependant, si l’on paramètre le noyau  $\mathbf{K}$ , chaque évaluation naïve de la vraisemblance requiert de calculer  $\mathbf{L}$  et la constante de normalisation dans (4), ce qui coûte  $\mathcal{O}(n^3)$ . Si la paramétrisation porte sur  $\mathbf{L}$ , on évite d’inverser  $\mathbf{I} - \mathbf{K}$ , mais on perd en interprétabilité et il reste nécessaire d’évaluer la constante de normalisation en  $\mathcal{O}(n^3)$ . En pratique, pour gagner en efficacité il est courant d’imposer une structure particulière au noyau : rang faible, parcimonie, etc. [6]. Pour plus de références sur l’inférence et sur les propriétés de l’estimateur du maximum de vraisemblance, nous renvoyons à [4].

 <https://dppy.readthedocs.io>

 <https://github.com/guilgautier/DPPy>

## 6 Discussion

Les DPP sont une machine à noyau populaire comme modèle en apprentissage. Des procédures de modélisation, d’échantillonnage et d’inférence sont disponibles et impliquent des manipulations algébriques classiques. Des procédures approchées qui passent à l’échelle voient le jour régulièrement. On peut toutefois objecter qu’avec le temps, d’autres modèles répulsifs – moins élégants mais moins coûteux – donneront des résultats similaires pour les applications comme les systèmes de recommandation. Parallèlement, l’expressivité des DPP pourrait augmenter avec l’étude de noyaux non-symétriques.

Par ailleurs, les liens avec les notions d’orthogonalité et de volume font des DPP un outil naturel pour la sélection d’attributs en régression [3], de nœuds d’un graphe [18], de nœuds d’intégration Monte Carlo [1], etc. Ainsi, nous pensons que les DPP ont un rôle très spécial à jouer en tant qu’outils.

## Références

- [1] Bardenet, R. and Hardy, A. Monte Carlo with determinantal point processes. *arXiv*, 2016.
- [2] Bardenet, R. and Titsias, M. K. Inference for determinantal point processes without spectral knowledge. *NIPS*, 2015.
- [3] Belhadji, A., Bardenet, R., and Chainais, P. A determinantal point process for column subset selection. *arXiv*, 2018.
- [4] Brunel, V.-E., Moitra, A., Rigollet, P., and Urschel, J. Maximum likelihood estimation of DPPs. *arXiv*, 2017.
- [5] Dereziński, M., Calandriello, D., Valko M. Exact sampling of DPPs with sublinear time preprocessing. *arXiv*, 2019.
- [6] Gartrell, M., Paquet, U., and Koenigstein, N. Bayesian Low-Rank Determinantal Point Processes. *RecSys*, 2016.
- [7] Gautier, G., Bardenet, R., and Valko, M. Zonotope Hit-and-run for Efficient Sampling from Projection DPPs. *ICML*, 2017.
- [8] Gillenwater, J. Approximate inference for determinantal point processes. Ph.D. thesis, University of Pennsylvania, 2014.
- [9] Hough, J. B., Krishnapur, M., Peres, Y., and Virág, B. Determinantal Processes and Independence. *Prob. Surv.* 2006.
- [10] Johansson, K. Random matrices and determinantal processes. *Les Houches Summer School Proceedings*, 2006.
- [11] Kulesza, A. and Taskar, B. Determinantal Point Processes for Machine Learning. *FTML*, 2012.
- [12] Launay, C., Galerne, B., and Desolneux, A. Exact Sampling of DPPs without Eigendecomposition. *arXiv*, 2018.
- [13] Lavancier, F., Møller, J., and Rubak, E. Determinantal point process models and statistical inference. *J. R. Stat. Soc.*, 2012.
- [14] Li, C., Jegelka, S., and Sra, S. Fast Mixing Markov Chains for SR Measures, DPPs, and Constrained Sampling. *NIPS*, 2016.
- [15] Macchi, O. The coincidence approach to stochastic point processes. *Adv. Appl. Probab.*, 1975.
- [16] Mariet, Z. and Sra, S. Fixed-point algorithms for learning determinantal point processes. *ICML*, 2015.
- [17] Poulson, J. High-performance sampling of generic Determinantal Point Processes. *arXiv*, 2019.
- [18] Tremblay, N., Amblard, P.-O., and Barthelme, S. Graph sampling with determinantal processes. *EUSIPCO*, 2017.
- [19] Tremblay, N., Barthelme, S., and Amblard, P.-O. Optimized Algorithms to Sample DPPs. *arXiv*, 2018.
- [20] Xu, Y., Müller, P., and Telesca, D. Bayesian inference for latent biologic structure with DPPs. *Biometrics*, 2016.