6.883 Learning with Combinatorial Structure
Note for Lecture 21
Author: Swati Gupta

The last few lectures have been about online combinatorial optimization, learning bandits, etc. Today we will switch gears to talk about probabilistic models of diversity. These models increase the probability of sets that are more spread out and diverse. For example, consider the problem of detecting where the people are in a given picture. Ideally, you would expect people to be spatially far from each other, and not standing on top of each other. We want probabilistic models that capture negative correlations, i.e., including an element a makes it less likely to include another element b.

1 Determinantal Point Processes

Determinantal point processes (DPP) are elegant probabilistic models that capture negative correlation and admit efficient algorithms for sampling, marginalization, conditioning, etc. These processes were first studied by [8], as *fermion processes*, to model the distribution of fermions at thermal equilibrium. The specific term 'determinantal' was coined by [2] and has since then been used widely. A point process $\mathcal{P}$ on the ground set $\mathcal{V}$ is a probability measure on the power set of $\mathcal{V}$, i.e., $2^{\mathcal{V}}$.

Definition 1 A point process $\mathcal{P}$ is called a determinantal point process, if Y is a random subset drawn according to $\mathcal{P}$, then we have for every $S \subseteq Y$,

$$P(S \subseteq Y) = \det(K_S), \tag{1}$$

for some similarity matrix $K \in \mathbb{R}^{n \times n}$ that is symmetric, real and positive semidefinite. Here, K_S denotes the submatrix of K obtained by restricting to rows and columns indexed in S.

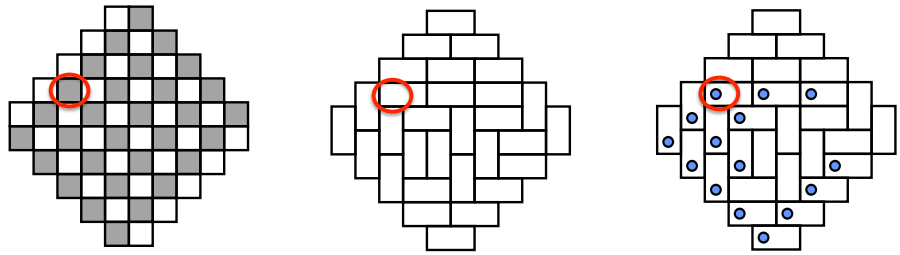
Since the marginal probabilities of any set $S \subseteq \mathcal{V}$ must lie in $[0, 1]$, we require every principal minor of $\det(K_S)$ of K must be nonnegative, hence K must be positive semidefinite. It can be shown that the eigenvalues of K must lie between 0 and 1, i.e., $0 \preceq K \preceq 1$. (In fact, these conditions are sufficient: any K such that $0 \preceq K \preceq 1$ defines a DPP). K is often referred to as the *marginal kernel*.

Note that $P(e_i \in Y) = K_{ii}$, i.e., the marginal probability of including that element. Now, the marginal probability of including two elements, say e_i and e_j, is given by $P(e_i, e_j \in Y) = K_{ii}K_{jj} - K_{ij}^2 = P(e_i \in Y)P(e_j \in Y) - K_{ij}^2$. Since K is positive semidefinite, K_{ij} is non-negative and hence we've modeled repulsion. A larger value of K_{ij}

implies that i and j are more likely to *not* appear together. Further, note that if the kernel matrix K is diagonal, then $P(S \subseteq Y) = \det(\text{diag}(p_1, p_2, \ldots, p_{|S|})) = \Pi_{i \in S} p_i$. Thus, we can model independent point processes using DPPs. However, the correlation between any two elements is always non-negative for DPPs and one cannot model positive correlation.

1.1 Examples

- **Descents in random sequences** [1] Consider a sequence of n random numbers drawn uniformly and independently from a finite set, say $0, \ldots, k$. Location of numbers where the current number is less than the previous number are distributed as a DPP. Note that these locations are a subset of $\{2, \ldots, k\}$.
- **Aztec Diamond Tilings** [6] The Aztec diamond is a diamond-shaped union of lattice squares, as shown in Figure 1, with half the squares colored grey in a checkerboard pattern. Consider a random domino tiling of the Aztec diamond (drawn uniformly from all possible domino tilings). Then the subset of lattice squares that are grey as well as lie in the left half of a horizontal tile or the bottom of a vertical tile are distributed as a DPP!



(a) Aztec diamond (checkered) (b) Domino tiling (c) DPP

Figure 1: Aztec Diamond Tilings with checkerboard coloring and position of lattice squares that give a sample from the corresponding DPP.

- **Random Matrices** [4, 9] Construct a random matrix M by drawing each of its entries independently from a complex normal distribution. The eigenvalues of M (subset of complex plane) are distributed as a DPP.

1.2 L-ensembles

L-ensembles, introduced by Borodin and Rains in 2005 [3], are a slightly restricted but a useful class of DPPs. They are defined using a real, symmetric matrix L indexed by the elements of $\mathcal{V}$. The unnormalized probability of sampling a set of $Y \subseteq \mathcal{V}$ is given by

$$P_L(Y) \propto \det(L_Y). \tag{2}$$

This immediately justifies that $L \preceq 0$. The natural question at this point is, *can we compute the normalization constant for an L-ensemble?* The normalization constant is simply the sum of the unnormalized probabilities over all subsets of the $\mathcal{V}$, i.e. $\sum_{S \subseteq \mathcal{V}} \det(L_S)$, and the following theorem shows that normalization is tractable (as opposed to graphical models, computing normalization reduces to linear algebra for DPPs).

Theorem 1. *For any $A \subseteq \mathcal{V}$:*

$$\sum_{A \subseteq Y \subseteq \mathcal{V}} \det(L_Y) = \det(L + I_{\bar{A}}),$$

where $I_{\bar{A}}$ is a diagonal matrix such that $I_{ii} = 0$ for indices $i \in A$ and $I_{ii} = 1$ for $i \in \bar{A}$.

Proof. By induction on the size of A and expanding the determinant in the induction step. See Theorem 2.1, on Page 8 of Kulesza and Taskar's survey [7] on Determinantal point processes for machine learning for the full proof. □

Setting $A = \phi$, we obtain the following corollary:

Corollary 1. *The normalization constant of an L-ensemble is given by*

$$\sum_{S \subseteq \mathcal{V}} \det(L_S) = \det(L + I_{\mathcal{V}}).$$

Equivalence of the two definitions There is a very close relation between DPPs defined using the marginal kernel K and L-ensembles, and it is precisely quantified by the following result of Macchi [8].

Theorem 2. *An L-ensemble is a DPP, and its marginal kernel is $K = L(L+I)^{-1} = I - (L+I)^{-1}$.*

Proof. Using Theorem 1, the marginal probability of a set $A \subseteq \mathcal{V}$ under the L-ensemble is

$$P_L(A \subseteq Y) = \frac{\sum_{Y:A\subseteq Y\subseteq\mathcal{V}} \det(L_Y)}{\sum_{Y\subseteq\mathcal{V}} \det L_Y} \tag{3}$$

$$= \frac{\det(L + I_{\bar{A}})}{\det(L+I)} \tag{4}$$

$$= \det((L + I_{\bar{A}})(L+I)^{-1}) \tag{5}$$

$$= \det(I_{\bar{A}}(L+I)^{-1} + I - (L+I)^{-1}) \quad \dots \text{ using } L(L+I)^{-1} = I - (L+I)^{-1} \tag{6}$$

$$= \det(I_{\bar{A}}(L+I)^{-1} + (I_A + I_{\bar{A}})(I - (L+I)^{-1}) \tag{7}$$

$$= \det(I_{\bar{A}} + I_A K) \tag{8}$$

$$= \begin{vmatrix} I_{|\bar{A}|\times|\bar{A}|} 0 \\ K_{A,\bar{A}} K_A \end{vmatrix} \tag{9}$$

$$= \det(I_{\bar{A}\times\bar{A}}) \det(K_A) = \det(K_A). \tag{10}$$

This implies that all the marginals agree, and thus an L-ensemble is equivalent to a DPP with a marginal kernel $K = L(L+I)^{-1}$. □

Note that an L-ensemble can be constructed using a marginal kernel K by setting $L = K(I-K)^{-1}$. Hence, a corresponding L-ensemble exists only if the $(I-K)$ is invertible. In this sense, DPPs are slightly more general than L-ensembles. (Existence of a inverse is equivalent to the point process assigning some nonzero probability to the empty set).

Eigenvalue Decomposition Suppose the eigendecomposition of $L = \sum_k \lambda_k v_k v_k^T$, then $K = \sum_k \frac{\lambda_k}{\lambda_k+1} v_k v_k^T$.

Geometric View Suppose we have some data points such that $x_1, \dots, x_n$ are the corresponding feature vectors in $\mathbb{R}^d$. We can construct an L-ensemble such that $L_{ij} = x_i^T x_j$. Then,

$$P_L(S) \propto \det(L_S) = \text{Vol}^2(\{x_i\}_{i\in S}).$$

This implies that if a set contains a more diverse set of feature vectors, the volume spanned by them would be greater resulting in a larger probability of sampling such a set. If the dimension of the feature vectors d is less than the number of points, then the samples from the corresponding DPP will only contain at most d points.

2 Working with DPPs

One of the primary advantages of DPPs is that, although the number of possible realizations of $\mathcal{P}$ is exponential in $\mathcal{V}$, many types of inference can be performed in polynomial time. We discuss some here.

- **Complements** If $Y \sim \text{DPP}(K)$, then $\bar{Y} = \mathcal{V} \setminus Y \sim \text{DPP}(I - K)$. In particular,

$$P(A \cap Y = \phi) = \det(I - K_A).$$

- **Conditioning** The conditional probability that a set B is observed such that it contains elements in A^{in} and does not contain elements in A^{out} conditioned on observing A^{in} and A^{out} is given by

$$P_L(Y = A^{in} \cup B | A^{in} \subseteq Y, A^{out} \cap Y = \phi) = \frac{\det(L_{A^{in} \cup B})}{\det(L_{\bar{A}^{out}} + I_{\bar{A}^{in}})}.$$

- **Marginals** Conditional marginal of B given that A has been observed is given by

$$P(B \subseteq Y | A \subseteq Y) = \det(K_B^A)$$

  for $K^A = I - [(L + I_{\bar{A}})^{-1}]_{\bar{A}}$.

- **Scaling** If $K = \gamma K'$, for some $0 \leq \gamma \leq 1$, then for all $A \subseteq \mathcal{V}$, we have $\det(K_A) = \gamma^{|A|} K'_A$.

Computing the mode of a DPP, i.e., finding a set $Y \subseteq \mathcal{V}$ that maximizes $P_L(Y)$, is in general NP-hard. This problem is also sometimes called the maximum a posteriori (MAP) inference. We will next look at how one can sample efficiently from DPPs.

2.1 Sampling

There are two important questions with respect to sampling: (i) how to draw a sample, and (ii) how many points there are in a sample. The main idea in sampling DPPs is to view every DPP as a mixture of *elementary* DPPs. A DPP is called elementary if every eigenvalue of its marginal kernel is in $\{0, 1\}$. P^T denotes an elementary DPP with marginal kernel $K^T = \sum_{v \in T} vv^T$, where T is a set of orthonormal vectors.

We will look at the sampling procedure developed by Hough et al. [5], formally presented in Algorithm 1. This algorithm has two main steps:

- Sample a mixture component P^T with probability π_T,
- Sample Y from P^T.

In the first phase of the algorithm, a subset of the eigenvectors is selected a random, where the probability of selecting each eigenvector v_k is $\frac{\lambda_k}{\lambda_k+1}$. In the second phase, the corresponding elementary DPP (to the eigenvectors selected in the first phase) is sampled for a subset Y. We first note that a DPP can be expressed as a mixture of simpler DPPs.

Lemma 1 (Sampling lemma)**.** *A DPP with kernel $L = \sum_k \lambda_k v_k v_k^T$ is a mixture of elementary DPPs: P^T such that $P_L = \frac{1}{det(L+I)} \sum_{T \subseteq V} \Pi_{k \in T} \lambda_k P^T$ where V is the set of orthonormal eigenvectors of L.*

Proof. The proof shows that the two distributions agree on all the marginal probabilities. See Lemma 2.6 of [7]. □

After the first phase has selected a set of eigenvectors A, we sample a subset Y from this elementary DPP. It is easier to sample from an elementary DPP, as each sample has cardinality = $|A|$.

Lemma 2. *If Y is drawn according to an elementary DPP P^A, then $|Y| = |A|$ with probability 1.*

Proof. We have $\mathbb{E}[|Y|] = \mathbb{E}[\sum_{i=1}^n 1[i \in Y]] = \sum_{i=1}^n P(i \in Y)$, but this probability is just $\sum_{i=1}^n K_{i,i}$. In general, trace of K gives the expected number of points. In case of elementary DPPs, the trace of K is exactly the cardinality of A, i.e., $tr(K) = \sum_{i=1}^n K_{i,i} = |A|$. In the case of elementary DPPs, we don't even fluctuate away from the expectation since the rank is $|A|$, and thus we cannot sample more than $|A|$. That is, for any $Y, |Y| > |A|$: $P(Y) = \det(L_A) = 0$. Hence, we must always sample sets of cardinality exactly $|A|$. □

Now, for sampling an elementary DPP, we start with an empty set $Y = \phi$. In each iteration i, we pick a point i conditioned on the previous $\{1, \ldots, i-1\}$ selected points. Using the geometric intuition, this conditioning is equivalent to projecting the remaining eigenvectors onto a subspace orthogonal to $\{v_1, \ldots, v_{i-1}\}$.

3 Applications

3.1 Sampling points on a plane

Figures 2 and 3 show the results of sampling from a DPP constructed using a Gaussian kernel with $s_{ij} = \exp(-\frac{1}{2\sigma^2}||x_i - x_j||^2)$. Note that there are fewer clusters of points in the DPP plots, as opposed to uniform iid sampling of the same number of points (as obtained in the DPP sample) in the plots to the right. Also, Figure 3 shows that increasing

Algorithm 1: Sampling from a DPP

Input: eigendecomposition $\{(v_n, \lambda_n)\}_{n=1}^{k}$ of L
$J \leftarrow \phi$;
for $n = 1, 2, \ldots, k$ **do** `/* Sample a mixture component` P^T `*/`
  $A \leftarrow A \cup \{n\}$ with probability $\frac{\lambda_n}{\lambda_n + 1}$;
end
$V \leftarrow \{v_n\}_{n \in A}$;
$Y \leftarrow \phi$;
while $|V| > 0$ **do** `/* Sample` Y `from` P^T `*/`
  Select i from V with $P(i) = \frac{1}{|V|} \sum_{v \in V} (v^T e_i)^2$;
  $Y \leftarrow Y \cup i$;
  $V \leftarrow V_\perp$, an orthonormal basis for the subspace of V orthogonal to e_i;
end
Output Y.

the variance σ^2 decreases the number of points sampled, since any new point is now more similar to already sampled points.

3.2 Pose Estimation

Using a quality model for the likelihood of body parts at different locations and orientations, and a similarity model based on locations, Figure 4 shows how DPPs can be used successfully to detect people in a picture, and their poses.

References

[1] Borodin, A., Diaconis, P., and Fulman, J. (2010). On adding a list of numbers (and other one-dependent determinantal processes). *Bulletin of the American Mathematical Society*, 47(4):639–670.

[2] Borodin, A. and Olshanski, G. (2000). Distributions on partitions, point processes and the hypergeometric kernel. *Communications in Mathematical Physics*, 211(2):335–358.

[3] Borodin, A. and Rains, E. M. (2005). Eynard–mehta theorem, schur process, and their pfaffian analogs. *Journal of statistical physics*, 121(3-4):291–317.

[4] Ginibre, J. (1965). Statistical ensembles of complex, quaternion, and real matrices. *Journal of Mathematical Physics*, 6(3):440–449.

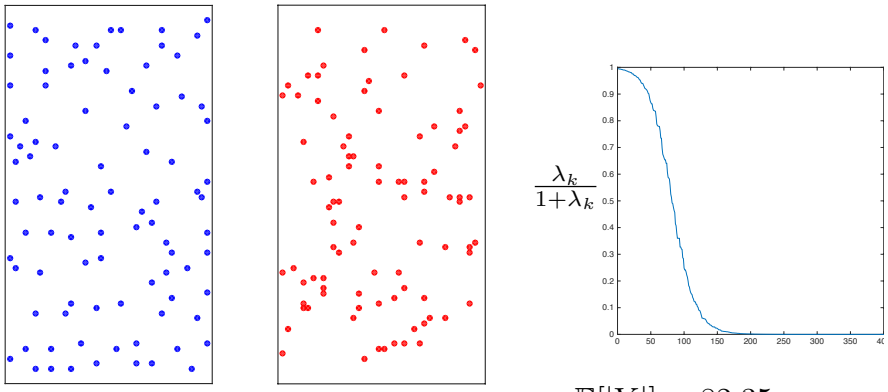


Figure 2: Plot on the left shows points sampled on a plane with a DPP corresponding to a Gaussian kernel matrix L with $\sigma^2 = 35$. The plot on the right shows a uniform iid sample of the same number of points.

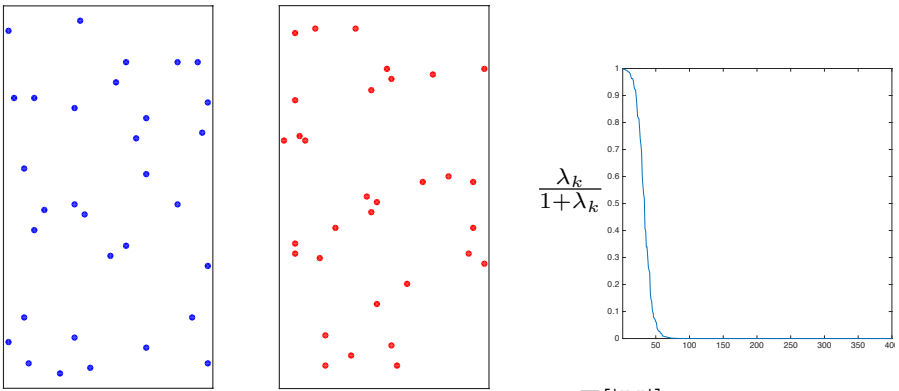


Figure 3: Plot on the left shows points sampled on a plane with a DPP corresponding to a Gaussian kernel matrix L with $\sigma^2 = 135$. The plot on the right shows a uniform iid sample of the same number of points.

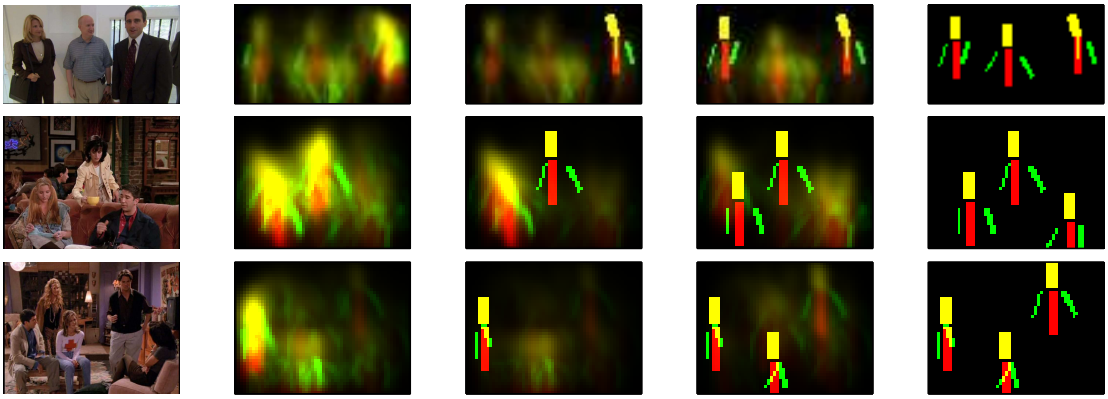


Figure 4: Pose estimation using filtering with a DPP based on the location similarity model.

[5] Hough, J. B., Krishnapur, M., Peres, Y., Virág, B., et al. (2006). Determinantal processes and independence. *Probability Survey*, 3:206–229.

[6] Johansson, K. (2005). The arctic circle boundary and the airy process. *Annals of probability*, pages 1–30.

[7] Kulesza, A. and Taskar, B. (2012). Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*.

[8] Macchi, O. (1975). The coincidence approach to stochastic point processes. *Advances in Applied Probability*, pages 83–122.

[9] Mehta, M. L. and Gaudin, M. (1960). On the density of eigenvalues of a random matrix. *Nuclear Physics*, 18:420–427.