On a machine learning setup for discrete determinantal structures

Dissertation submitted for the degree of
Master of Science in Interdisciplinary Mathematics

Johannes Müller

June 15, 2018

University of Warwick
Department of Mathematics

Acknowledgements

I would like to thank the dreamteam for being such an eggcelent friendgroup and for all the pun we had during the last year.

Abstract

Determinantal point processes are random subsets that exhibit a diversifying behaviour in the sense that the randomly selected points tend to be not similar in some way. This repellent structure first arrose in theortical physics and pure mathematics, but they have recently been used to model a variety of many real world scenarios in a machine learning setup. We aim to give an overview over the main ideas of this approach which is easily accessible even without prior knowledge in the area of machine learning and sometimes omit technical calculations in order to keep the focus on the concepts.

Contents

Chapter I

Introduction and motivating examples

I.1 Motivation

I.2 Previous work

I.3 Aim and outline of the dissertation

Chapter II

Determinantal points processes: Basic notions and properties

II.1 Historical remarks

II.2 Definitions and properties

Let in the following $\mathcal{Y}$ be a finite set, which we call the *ground set* and $N := |\mathcal{Y}|$. A *point process* on $\mathcal{Y}$ is a random subset of $\mathcal{Y}$, i.e. a random variable with values in the powerset $2^{\mathcal{Y}}$. Usually we will identify this random variable with its law $\mathbb{P}$ and thus refer to probability measures $\mathbb{P}$ on $2^{\mathcal{Y}}$ as a point processes and not distinguish those objects. Let in the following $\mathbf{Y}$ be a random subset drawn according to $\mathbb{P}$, then we call $\mathbb{P}$ a *determinantal point process*, or in short a DPP, if we have

$$\mathbb{P}(A \subseteq \mathbf{Y}) = \det(K_A) \quad \text{for all } A \subseteq \mathcal{Y} \tag{2.1}$$

where K is a symmetric matrix indexed by the elements in $\mathcal{Y}$ and K_A denotes the restriction of the matrix K to indices in A. We call K the *marginal kernel* of the DPP and note immediately that K is necessarily non negative definite. Further it can be shown (cf. page 3 in [Borodin, 2009]) that also the complement of a DPP is a DPP with marginal kernel $I - K$ where I is the identity matrix, i.e.

$$\mathbb{P}(A \subseteq \mathbf{Y}^c) = \det(I_A - K_A)$$

and thus $I - K \geq 0$ and therefore $0 \leq K \leq I$. This actually turns out to be sufficient for K to define a DPP through (2.1) (cf. [Kulesza et al., 2012]). We call the elements of $\mathcal{Y}$ *items* and by choosing $A = \{i\}$ and $A = \{i, j\}$ for $i, j \in \mathcal{Y}$ and using (2.1) we obtain the probabilities of their occurrence

$$\begin{aligned} \mathbb{P}(i \in \mathbf{Y}) &= K_{ii} \quad \text{and} \\ \mathbb{P}(i, j \in \mathbf{Y}) &= K_{ii}K_{jj} - K_{ij}^2 = \mathbb{P}(i \in \mathbf{Y}) \cdot \mathbb{P}(j \in \mathbf{Y}) - K_{ij}^2, \end{aligned} \tag{2.2}$$

Thus the appearance of the two items i and j are always negatively correlated. This negative correlation is exactly what causes the diversifying behaviour of determinantal point processes. In practice one usually models the negative correlations to be high between items that are similar in some way. For example in a spatial setting being similar could mean being close together and in this case the selected items would tend to be not very close together. This is repulsive behaviour can be seen in Figure In this light the fact that also $\mathbf{Y}^c$ exhibits negative correlations

insert picture!

becomes less surprising, because since the set **Y** tends to spread out due to the repulsion in (2.2), the complement, which is nothing but the gaps that are left after eliminating the elements in **Y**, tend to show a non clustering behaviour.

L-ensembles

Let us now introduce an important subclass of DPPs, namely the ones where not only the marginal probabilities can be expressed through a suitable kernel, but also the elementary probabilities. Because this will be convenient for us later on we will restrict ourselves to this case from now on. If we have even $K < I$, then we define the *elementary kernel*

$$L := K(I - K)^{-1}$$

which specifies the elementary probabilities since one can check

$$\mathbb{P}(A = \mathbf{Y}) = \frac{\det(L_A)}{\det(I + L)} \quad \text{for all } A \subseteq \mathcal{Y}. \tag{2.3}$$

Conversely for any $L \geq 0$ a DPP can be defined via (2.2) and the corresponding marginal kernel is given by

$$K = L(I + L)^{-1}.$$

We call DPPs which arise this way L *ensembles*.

The quality diversity parametrisation

Note that any symmetric, positive semidefinite matrix L can be written as a Gram matrix

$$L = B^T B$$

where $B \in \mathbb{R}^{D \times N}$ whenever D is larger than the rank $\mathrm{rk}(L)$ of L. For example one could take the spectral decomposition $L = U^T C U$ of L and set $B := \sqrt{C} U$ and eventually drop some zero rows from $\sqrt{C}$. Let B_i denote the i-th column of B and write it as the product $q_i \cdot \phi_i$ where $q_i \geq 0$ and $\phi_i \in \mathbb{R}^D$ such that $\|\phi_i\| = 1$. This yields the representation

$$L_{ij} = q_i \phi_i^T \phi_j q_j =: q_i S_{ij} q_j$$

and we call q_i the *quality* of the item $i \in \mathcal{Y}$ and ϕ_i the *diversity feature vector* of i and S the *similarity matrix*. Since we will use this decomposition multiple times, we fix its properties.

Proposition 2.1 (Quality diversity decomposition). *Let $D \in \mathbb{N}$ and let $\mathbb{S}_D$ denote the sqhere in $\mathbb{R}^D$. Further let $\mathbb{R}^{N \times N}_{sym,+}$ be the symmetric positive semidefinite $N \times N$ matrices. The quality diversity parametrisation is a continuous and surjective mapping*

its not a bijection!!!

$$\Psi \colon \mathbb{R}^N_+ \times \mathbb{S}^N_D \to \left\{ L \in \mathbb{R}^{N \times N}_{sym,+} \mid \mathrm{rk}(L) \leq D \right\}, \quad (q, \phi) \mapsto \left(q_i \phi_i^T \phi_j q_j \right)_{1 \leq i,j \leq N}.$$

Remark 2.2. (i) In the case $D = N$ the quality diversity decomposition gives a parametrisation of the whole symmetric positive definite $N \times N$ matrices.

(ii) Note that this parametrisation is not unique, i.e. Ψ is not injective. For example the identity matrix I can be parametrised by any orthonormal system ϕ and $q = (1, \ldots, 1)^T$.

(*iii*) One can without any problems consider diversity features ϕ_i in an abstract Hilbert space $\mathcal{H}$. However we will not need this in the remainder and thus restrict ourselves to the easier Euklidean diversity features.

(*iv*) We call every preimage of L under Ψ the *quality diversity decomposition* of L.

The quality diversity decomposition will provide some useful expressions. For example the elementary probabilities take the form

$$\mathbb{P}(A = \mathbf{Y}) \propto \det\big((B^T B)_A\big) == \left(\prod_{i \in A} q_i^2\right) \cdot \det(S_A) \quad \text{for all } A \subseteq \mathcal{Y}. \tag{2.4}$$

An intuitive understanding of the quality diversity decomposition will play a central role later on if one wants to model real world phenomena as DPPs. To get this we can think of $q_i \geq 0$ as a measure of how important or high in quality the item is and the diversity feature vector $\phi_i \in \mathbb{R}^D$ can be thought of as some kind of state vector that consists of internal quantities that describe the item i in some way. Further we interpret the scalar product $\phi_i^T \phi_j \in [0, 1]$ as a measure of similarity between the items i and j which justifies the name similarity matrix for S. Note that if i and j are perfectly similar or antisimilar, i.e. $\phi_i^T \phi_j = \pm 1$, then they can not occur at the same time, since

$$\mathbb{P}(i, j \in \mathbf{Y}) = \det \begin{pmatrix} 1 & \pm 1 \\ \pm 1 & 1 \end{pmatrix} = 0.$$

If we identify i with the vector $B_i = q_i \phi_i \in \mathbb{R}^D$, we can obtain a geometric interpretation of (2.4) since $\det\big((B^T B)_A\big)$ is the volume that is spanned by the columns $B_i, i \in A$, which is visualised in II.1. This volume increases if the lengths of the edges that correspond to the quality increase and decrease when the similarity feature vectors point into more similar directions.

One last property of DPPs that we shall mention is the fact that the negative correlations of the DPP posses a transient property in the sense, that if i and j and j and k are similar, then i and k are also similar. This is due to the fact

$$\left\|\phi_i - \phi_j\right\|^2 = \|\phi_i\|^2 + \left\|\phi_j\right\|^2 - 2\phi_i^T \phi_j = 2(1 - \phi_i^T \phi_j)$$

and thus

$$\sqrt{1 - \phi_i^T \phi_k} = \frac{1}{2} \|\phi_i - \phi_k\| \leq \frac{1}{2}\big(\left\|\phi_i - \phi_j\right\| + \left\|\phi_j - \phi_k\right\| \big) = \sqrt{1 - \phi_i^T \phi_j} + \sqrt{1 - \phi_j^T \phi_k}.$$

reformulate that part!

II.3 Variations of DPPs

Conditional DPPs

A *conditional DPP* is a collection of DPPs indexed by $X \in \mathcal{X}$, where X is called the *input* of the conditional DPP. Thus for every $X \in \mathcal{X}$ we get a finite set $\mathcal{Y}(X)$ and a determinantal point

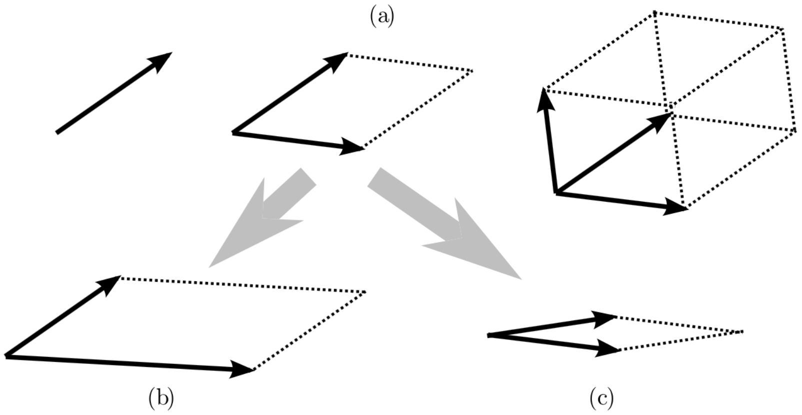

Figure II.1.: Taken from [Kulesza et al., 2012]; the first line (a) illustrates the volumes spanned by vectors, and in the second line it can be seen how this volume increases if the length – associated with the quality – increases (b) and decreases if they become more similar in direction which we interpret as two items becoming more similar (c)

process $\mathbb{P}(\cdot \mid X)$ on $\mathcal{Y}(X)$ which is given by the elementary kernel $L(X)$, i.e.

$$\mathbb{P}(A|X) \propto \det\left(L_A(X)\right) \quad \text{for all } A \subseteq \mathcal{Y}(X).$$

Further we denote the quality and diversity features of the conditional DPP by $q_i(X)$ and $\phi_i(X)$ respectively.

It is not immediately clear why one would want to model a family of DPPs as a conditional DPP rather than as seperate DPPs. The reason for this is that one wants to estimate the kernels $L(X)$ for every $X \in \mathcal{X}$. However if we would do this naively we would need to observe each of the DPPs $\mathbb{P}(\cdot \mid X)$ individually which is often not possible. Thus one hopes to not only memorise the kernels $L(X)$ for every single input $X \in \mathcal{X}$ but rather to learn the mapping L that assigns every input X its elementary kernel $L(X)$. If one achieved this task, one would be able to simulate and predict a DPP that one has not observed so far just by the knowledge about some DPPs that belong to the same conditional DPP. Of course this can only work if we assume some regularity or a certain structure of the function L which we will do in the third chapter where we put those consideration into a precise framework.

Fixed size or k-DPPs

Structured DPPs

We call a DPP *structured DPP* or short sDPP if the ground set is the cartesian product of some other set $\mathcal{M}$, which we will call the *set of parts*, i.e. if we have

Say something about number of parameters

$$\mathcal{Y} = \mathcal{M}^R = \left\{y_i = (y_i^r)_{r=1,\dots,R} \mid i = 1, \dots, N\right\}$$

where R is a natural number, $M = |\mathcal{M}|$ and $N = M^R$. The quality diversity decomposition of L take the form

$$L_{ij} = q(y_i)\phi(y_i)^T\phi(y_j)q(y_j)$$

and since $N = M^R$ is typically very big, it is impractical to define or store the quality and diversity features for every item $y_i \in \mathcal{Y}$. To deal with this problem we will assume that they admit factorisations and are thus a combination of only a few qualities and diversities.

More precisely we call $F \subseteq 2^{\{1,\dots,R\}}$ a *set of factorisations* and for a *factor* $\alpha \in F$, y_α denotes the subtupel of $y \in \mathcal{Y}$ that is indexed by α. Further we will work with the decompositions

$$\begin{aligned} q(y) &= \prod_{\alpha \in F} q_\alpha(y_\alpha) \\ \phi(y) &= \sum_{\alpha \in F} \phi_\alpha(y_\alpha) \end{aligned} \tag{2.5}$$

for a suitable set of factorisations F and qualities and diversities q_α and ϕ_α for $\alpha \in F$. Note that so far this is neither a restriction of generality – we could simply choose $F = \big\{\{1,\dots,R\}\big\}$ – nor a simplification – in that case we have the exact same number of qualities and diversities. However we are interested in the case where F consists only of small subsets of $\{1,\dots,R\}$. For example suppose that F is the set of all subsets with one or two elements, then we only have

$$R \cdot M + \binom{R}{2} \cdot M^2 = O(R^2 M^2)$$

quality and diversity features instead of

$$M^R = O(M^R).$$

This reduction of variables will make modelling, storing and estimating them feasible again in a lot of cases where naive approaches are foredoomed because of their shear size.

II.4 The magic properties of DPPs

II.5 The mode problem

Chapter III

Learning setups

III.1 What does learning mean and why is it interesting?

III.2 Assymptotic reconstruction of the kernel

In this section we want to see how we can estimate the marginal kernel from an increasing number of observations $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ that are distributed according to $\mathbb{P}$. For this we will sketch the procedure in [Urschel et al., 2017]. Let $\hat{\mathbb{P}}_n$ be the empirical distribution

$$\hat{\mathbb{P}}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{\mathbf{Y}_i}.$$

We can identify the probability measures on a finite set, in our case $2^{\mathcal{Y}}$, with the probability simplex

$$\left\{ x \in \mathbb{R}^{2^{\mathcal{Y}}} \;\middle|\; x_A \in [0,1] \text{ for all } A \in 2^{\mathcal{Y}} \text{ and } \sum_{A \in 2^{\mathcal{Y}}} x_A = 1 \right\}.$$

Doing this the strong law of large numbers yields that the empirical distributions converge to $\mathbb{P}$ almost surely if the sequence $(\mathbf{Y}_k)_{k \in \mathbb{N}}$ of observations is independent. Therefore we can consistently estimate all principle minors of K, since

cite

explain consistency

$$\hat{\mathbb{P}}_n(A \subseteq \mathbf{Y}) \xrightarrow{n \to \infty} \mathbb{P}(A \subseteq \mathbf{Y}) = \det(K_A) \quad \text{almost surely.}$$

Thus the question naturally arrises whether we can reconstruct the kernel K from the knowledge of all of its principle minors. This is known as the *principle minor assignment problem* and has been studied extensively (cf. [Griffin and Tsatsomeros, 2006] and [Urschel et al., 2017]) and an computationally efficient algorithm has been proposed for the problem in [Rising et al., 2015]. It is in fact possible to retain the matrix from its principle minors up to an equivalence relation which identifies matrices with each other, that have the same principle minors. Obviously this is sufficent for the task of learning a DPP, because those matrices are exactly those who give rise to the same point process. To see roughly how this reconstruction works we note that the diagonal is given by

$$K_{ii} = \det(K_{\{i\}})$$

and the absolute value of the off diagonal can be obtained through

$$K_{ij}^2 = K_{ii} K_{ii} - \det\left(K_{\{i,j\}}\right).$$

The reconstruction of the signs of the entries K_{ij} turns out to be the main difficulty, but this can be done analysing the cycles of the adjacency graph G_K corresponding to K. The adjacency graph has $\mathcal{Y}$ as its vertex set and the set of edges consists of the pairs $\{i, j\}$ such that $K_{ij} \neq 0$. The reconstruction now relies on the analysis of the cycles of this graph and it has been shown, that one only needs to know all the principle minors up to the order of the cycle sparsity of G_K (cf. [Urschel et al., 2017]). Following this method it is possible to compute estimators $\hat{K}_n$ of K in polynomial time and give a bound on the speed of convergence in some suitable metric.

see whether this proof can be done in a simplified way without considering the sparsity l

state and explain the result

In completely analogue fashion one can learn the elementary kernel L and those estimations can be used to sample from a DPP that was observed. This procedure might be sufficient in some scenarios, but this approach lacks the ability to extrapolate the knowledge one has of specific DPPs onto some new, unobserved DPPs which is exactly the point that would distinguish the procedure from classical statistics and would allow for far more interesting applications. To achieve this, we introduce the notion of conditional DPPs in the following section which are customised to describe families of DPPs with kernels that are in some way similar.

III.3 Maximum likelihood estimation using optimisation techniques

general blabla about MLE

Kernel estimation

rewrite it in a MLE rather than ML fashion

The approach described above is clearly of traditional statistical type and we want to touch on how the kernel estimation can be put into a machine learning task (cf. [Affandi et al., 2014]). For this we assume that we have a training set, i.e. a number of subsets $Y_1, \dots, Y_T \subseteq \mathcal{Y}$ drawn independently and according to the DPP $\mathbb{P}$. We aim to establish a quantity that gives an intuitive measure of how well a different DPP describes the training set and then we want to find the elementary kernel L for which the associated DPP $\mathbb{P}_L$ optimises this quantity. A widely used choice in the machine learning community for this is the so called *log likelihood function*

$$\log \left(\prod_{t=1}^{T} \mathbb{P}_L(Y_t) \right)$$

and we will work with its negative

$$\mathcal{L}(L) := -\log \left(\prod_{t=1}^{T} \mathbb{P}_L(Y_t) \right) = -\sum_{t=1}^{T} \log \left(\det(L_{Y_t}) \right) + T \log \left(\det(L + I) \right)$$

where high values of $\mathcal{L}$ correspond to kernels L where at least one element Y_t of our training set is very unlikely. Thus it is natural to minimise the loss function $\mathcal{L}$ over all positive semidefinit $L \in \mathbb{R}^{N \times N}$. Note that we have $\mathcal{L}(L) = \infty$ if and only if an observation Y_t of our training set is impossible under the DPP $\mathbb{P}_L$, i.e. we do not consider those kernels in our estimation. We note that the loss function is smooth and the gradient of this can be explicitly expressed, at least on the domain $\{\mathcal{L} < \infty\}$. This is due to the fact that the determinants of the submatrices are polynomials in the entries of L and the composition of those with the smooth function $\log\colon (0, \infty) \to \mathbb{R}$ stays smooth. This property allows the use of gradient methods but they face the problem that the loss function is non convex and thus those algorithms will generally not converge to a global minimiser (cf. [Affandi et al., 2014]). Nevertheless it is worth investigating whether those local minima turn out to produce good results in real world scenarios.

Explain why it is non convex

Learning the quality

Let now $\{Y_t\}_{t=1,\dots,T}$ be a training set where $Y_t \subseteq \mathcal{Y}$ for every $t = 1, \dots, T$. Unlike earlier we will not try to estimate the whole kernel L but only the qualities q_i of the items $i \in \mathcal{Y}$. More precisely we can parametrise the positive definite symmetric matrices L using the quality diversity decomposition, i.e. we consider the bijection

whats the domain?

$$(q, S) \mapsto L \quad \text{where } L_{ij} = q_i S_{ij} q_j .$$

Now we fix a similarity kernel S_0, that usually comes from modelling, and only try to estimate the quality $q \in \mathbb{R}_+^N$. This means that we optimise the likelihood function over a smaller set of kernels, namely the ones that arise from (q, S_0) for $q \in \mathbb{R}_+^N$ and thus the maximal likelihood that can be achieved will be lower compared to the general kernel estimation.

$$q_i(X) = g(f_i(X)), \quad \phi_i(X) = G(f_i(X))$$

where $f_i(X) \in \mathcal{Z}$ is being modelled and $g\colon \mathcal{Z} \to [0, \infty)$ and $G\colon \mathcal{Z} \to \mathbb{R}^D$ will be learned based on the observations. We will assume that $\mathcal{Z}$ is a subset of a vector space and therefore we call $f_i(X)$ the *feature vector*. If it is possible to estimate the quality and diversity as above, we would be able to sample from every DPP $\mathbb{P}(\cdot \mid X)$ and even from those that we haven't observed so far – just by the knowledge about DPPs with a similar structure.

Let us again illustrate this procedure in the example of the human point selection and we will restrict ourselves to learn the function g that determines the quality function, we might have a reason to be absolutely sure that we have modelled the diversity features $\phi_i(X)$ perfectly, so there is no need to learn, i.e. optimise them any further. However we are not convinced any more that humans really do not prefer some points over others – maybe we have the feeling that they lean more towards the points located in the center of the square. Therefore it is natural to assume that the quality, which is nothing but the popularity of a point, depends on the distance to the centre point of the square $m = (1/2, 1/2)$, i.e.

$$q_i(n) = g(\|i - m\|) = g(f_i(n))$$

where we want to learn g with respect to some loss function over a given family $\mathcal{F}$ of functions.

To put this back into the general setting we note that $g \in \mathcal{F}$ gives rise to a different conditional DPP which we will denote by $\mathbb{P}_g(\cdot \mid X)$. Just like in the case of simple DPPs we will work with the negative of the log likelihood function

$$\mathcal{L}(g) := -\log \left(\prod_{t=1}^{T} \mathbb{P}_g(Y_t \mid X_t) \right)$$

and seek a minimiser of the loss function $\mathcal{L}$. Thus we obtain an optimisation problem over a family of functions and in practice it is convenient to restrict ourselves to a parametric family

$$\mathcal{F} = \left\{ g_\theta \mid \theta \in U \subseteq \mathbb{R}^M \right\}.$$

In this case we write $\mathbb{P}_\theta(\cdot \mid X)$ for the conditional DPP that is induced by g_θ and the kernels become functions of θ and thus we write $L(\theta; X)$ and $K(\theta; X)$ for the kernel associated with the parameter θ. In analogue fashion we denote the loss function by

$$\mathcal{L}(g_\theta) = \mathcal{L}(\theta) = -\sum_{t=1}^{T} \log \left(\mathbb{P}_\theta(Y_t \mid X_t) \right).$$

Properties of the loss function $\mathcal{L}$

We want to see how the log likelihood approach naturally leads to a log linear model in θ for the quality features if one wants to obtain a convex loss function. Of course the motivation for a convex loss function is given by the nice properties of convex optimisation tasks described earlier. In order to see in which cases the loss function is convex, we use (**??**) to obtain

$$\begin{aligned} -\log\left(\mathbb{P}_\theta(Y_t \mid X_t)\right) &= -\log(\det(L_Y(\theta; X))) + \log\big(\det(L(\theta; X_t) + I)\big) \\ &= -2 \cdot \sum_{i \in Y_t} \log\left(g_\theta(f_i(X_t))\right) - \log\left(\det\left(S_{Y_t}(X_t)\right)\right) \\ &\quad + \log\left(\sum_{A \subseteq \mathcal{Y}(X_t)} \left(\prod_{i \in A} g_\theta(f_i(X_t))^2\right) \det(S_A(X_t))\right). \end{aligned} \tag{3.1}$$

This expression is well defined in $[0, \infty]$ if we adapt the common convention $\det(S_\emptyset(X)) = 1$. In order to give some criteria for the convexity and coercivity of the loss function, we say that a function f *log concave*, *log convex* or *logarithmically (affine) linear* if $\log(f)$ has the respective property.

Proposition 3.1 (Coercivity and convexity of the loss function). (i) *The rate function is coercive for all possible training sets if and only if*

$$\mathbb{P}_\theta(Y \mid X) \xrightarrow{|\theta| \to \infty} 0 \quad \textit{for all } Y \subseteq \mathcal{Y}(X) \textit{ and } X \in \mathcal{X}. \tag{3.2}$$

(ii) *The rate function is convex for all possible training sets if $g_\theta(f_i(X_t))$ is log concave in θ for all $i \in \mathcal{Y}(X)$, $X \in \mathcal{X}$ and if*

$$\prod_{i \in B} g_\theta^2(f_i(X_t))$$

is log convex in θ for all $B \subseteq \mathcal{Y}(X))$ and $X \in \mathcal{X}$.

(iii) *The conditions in (ii) are satisfied if and only if $g_\theta(f_i(X))$ is logarithmically affine linear in θ for every $i \in \mathcal{Y}(X)$ and $X \in \mathcal{X}$.*

Proof. (i) It is clear that under (3.2) we have

$$\exp\left(-\mathcal{L}(\theta)\right) = \prod_{t=1}^{T} \mathbb{P}_\theta(Y_t \mid X_t) \xrightarrow{|\theta| \to \infty} 0$$

for every possible training set and thus $\mathcal{L}$ is coercive. If on the other hand $\mathcal{L}$ is coercive for every training set we could also choose (Y, X) arbitrary as our training set and immediately obtain (3.2).

(ii) This condition for the convexity of the loss function can be directly derived from the fact that linear combination of log convex functions are log convex and formula (3.1).

(iii) If $g_\theta(f_i(X))$ is logarithmically affine linear, then it is also log convex and

$$\log\left(\prod_{i \in B} g_\theta^2\left(f_i(X)\right)\right) = 2 \sum_{i \in B} \log\left(g_\theta(f_i(X))\right)$$

is convex. On the other side if (*ii*) holds, then all functions $\log\big(g_\theta(f_i(X))\big)$ are concave and $\sum\limits_{i\in B} \log\big(g_\theta(f_i(X))\big)$ is convex and thus $\log\big(g_\theta(f_i(X))\big)$ has to be affine linear.

□

The result above shows that logarithmically affine linear models are the natural fit for the parametric family $\mathcal{F}$ that we want to optimise over. However they can be easily transformed into log linear models through a simple parameter shift if we assume $f_i(X) \neq 0$ and thus we can assume without loss of generality that the functions g_θ have the form

$$g_\theta(f_i(X)) = \exp\left(\frac{1}{2}\theta^T f_i(X)\right) \quad \text{for all } i \in \mathcal{Y}(X) \text{ and } X \in \mathcal{X}.$$

This structure can be used to derive some explicit expression for this case. Of course this log linear model is only well defined if the feature space $\mathcal{Z}$ is a subset of $\mathbb{R}^M$ which we will assume from now on. We note that this is no restriction if we assume a log linear model, because otherwise we could just replace the feature functions f_i by the log linearity constants $\hat{f}_i(X) \in \mathbb{R}^M$. First we can apply the explicit structure to the elementary probabilities and get

$$\mathbb{P}_\theta(A \mid X) \propto \exp\left(\theta^T f_A(X)\right) \det(S_A(X))$$

where $f_A(X) := \sum_{i\in A} f_i(X)$. Using this we get that the single summands of the loss function are equal to

$$-\theta^T f_Y(X) - \det(S_Y(X)) + \log\left(\sum_{A\subseteq\mathcal{Y}(X)} \exp\left(\theta^T f_A(X)\right)\det(S_A(X))\right) \tag{3.3}$$

Since a lot of numerical optimisation algorithms depend on the gradient of the function, it is worth noting that an explicit expression for the gradient of the loss function $\mathcal{L}$ can be derived from this formula, since differentiating (3.3) with respect to θ gives

$$\begin{aligned} -f_Y(X) + \frac{\sum_{A\subseteq\mathcal{Y}(X)} f_A(X) L_A(\theta; X)}{\sum_{A\subseteq\mathcal{Y}(X)} L_A(\theta; X)} &= -f_Y(X) + \sum_{A\subseteq\mathcal{Y}(X)} f_A(X)\mathbb{P}_\theta(A \mid X) \\ &= -f_Y(X) + \sum_{i\in\mathcal{Y}(X)} f_i(X) \sum_{i\in A\subseteq\mathcal{Y}(X)} \mathbb{P}_\theta(A \mid X) \\ &= -f_Y(X) + \sum_{i\in\mathcal{Y}(X)} f_i(X)\mathbb{P}_\theta(i \in \mathbf{Y} \mid X) \\ &= -f_Y(X) + \sum_{i\in\mathcal{Y}(X)} f_i(X) K_{ii}(\theta; X). \end{aligned} \tag{3.4}$$

The later expression of this gradient has the advantage that it can be efficiently computed in contrary to the evaluation of the exponentially large sum in the first line.

Obviously the loss function is not coercive in general, since for $f_i(X) = 0$ the probability $\mathbb{P}_\theta(\{i\} \mid X)$ is constant in θ. However it is not straight forward whether it becomes coercive under the assumption $f_i(X) > 0$ entrywise for every $i \in \mathcal{Y}(X)$ and $X \in \mathcal{X}$ and this could be investigated further.

Estimating the mixture coefficients of k-DPPs

Learning kernels of conditional DPPs

III.4 A Bayesian approach to the kernel estimation

Chapter IV

Toy examples and experiments

IV.1 Minimal example?

IV.2 Points on the line

IV.3 Points in the square

IV.4 Toy example for quality learning

Chapter V

Summary and conclusion

Chapter A

Generated code

Bibliography

[Affandi et al., 2014] Affandi, R. H., Fox, E., Adams, R., and Taskar, B. (2014). Learning the parameters of determinantal point process kernels. In *International Conference on Machine Learning*, pages 1224–1232.

[Benard and Macchi, 1973] Benard, C. and Macchi, O. (1973). Detection and“emission”processes of quantum particles in a“chaotic state”. *Journal of Mathematical Physics*, 14(2):155–167.

[Borodin, 2009] Borodin, A. (2009). Determinantal point processes. *arXiv preprint arXiv:0911.1153*.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[Griffin and Tsatsomeros, 2006] Griffin, K. and Tsatsomeros, M. J. (2006). Principal minors, part ii: The principal minor assignment problem. *Linear Algebra and its applications*, 419(1):125–171.

[Higham, 1990] Higham, N. J. (1990). Exploiting fast matrix multiplication within the level 3 blas. *ACM Transactions on Mathematical Software (TOMS)*, 16(4):352–368.

[Kulesza et al., 2012] Kulesza, A., Taskar, B., et al. (2012). Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286.

[Magen and Zouzias, 2008] Magen, A. and Zouzias, A. (2008). Near optimal dimensionality reductions that preserve volumes. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 523–534. Springer.

[Rising et al., 2015] Rising, J., Kulesza, A., and Taskar, B. (2015). An efficient algorithm for the symmetric principal minor assignment problem. *Linear Algebra and its Applications*, 473:126–144.

[Samuel, 1959] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229.

[Urschel et al., 2017] Urschel, J., Brunel, V.-E., Moitra, A., and Rigollet, P. (2017). Learning determinantal point processes with moments and cycles. *arXiv preprint arXiv:1703.00539*.