# Journal

**Hugo Simon**
hugo.simon@telecom-paris.fr

## Contents

## 1 Motivations

A common if not the standard approach in machine learning is to formulate learning problems as optimization problems.

Let $\mathcal{X} = \{x_i \mid i \in [\![1, n]\!]\}$ be a multiset (possibly with repetitions) of $n$ data points. Let $\Theta$ be a space of parameters, or queries, and $\theta$ an element of $\Theta$. Given the data $\mathcal{X}$ and a space of possible solutions Q, one aims to find a solution $\theta^{\text{opt}}$ that minimizes a cost function $L$. In this work, we focus on cost functions that are additively decomposable, i.e. we consider cost functions of the form

$$L(\theta) := \sum_{x \in \mathcal{X}} f_\theta(x)$$

for some function $f_\theta \in \{f_\theta \mid \theta \in \Theta\}$.

A large amount of machine learning problems falls into that framework, including support vector machines, logistic regression, linear regression and k-means clustering. For example, the goal of the euclidean k-means clustering is to find a set of k cluster centers in $\mathbb{R}^d$ minimizing the quantization error

$$L(\theta) = \sum_{x \in \mathcal{X}} \min_{q \in \theta} \|x - q\|_2^2$$

In this case, $\theta \in \binom{\mathcal{X}}{k}$, the set of all subsets of $\mathcal{X}$ of size $k$, and $f_\theta = \min_{q \in \theta} \|x - q\|_2^2$

In many machine learning applications, the induced optimization problem can be hard to solve. Given a learning task, if an algorithm is too slow on large datasets, one can either speed up the algorithm or reduce the amount of data. The second alternative is theoretically guaranteed by the "coresets" idea. A coreset is a weighted subset of the original data with the assurance that, up to a controlled relative error, the task's estimated cost function on the coreset will match the cost calculated on the complete dataset for any learning parameter.

An elegant outcome of such property is the ability to execute learning algorithms only on the coreset, assuring nearly-equal performance while significantly reducing the computational cost. There are other algorithms that generate coresets, some of which are more specialized and are designed for a particular purpose (such as k-means, k-medians, logistic regression, etc.). Additionally, keep in mind that there are results for the coreset in both the streaming and offline settings. Nevertheless, we will concentrate here on the offline setting.

## 2 The coreset property

The key idea behind coresets is to approximate the original data set $\mathcal{X}$ by a weighted set $\mathcal{S}$ which satisfies the coreset property. Such property then guarantee $1 + \varepsilon$-approximations.

Let $\mathcal{S} = \{x_i \mid i \in [\![1, m]\!]\}$ be a submultiset of $\mathcal{X}$. To each element $x \in \mathcal{S}$, associate a weight $\omega(x) \in \mathbb{R}^+$. Define the estimated cost associated to the weighted submultiset $\mathcal{S}$ as

$$\hat{L}(\theta) := \sum_{x \in \mathcal{S}} \omega(x)\, f_\theta(x)$$

**Definition 2.1** (Coreset)**.** Let $\varepsilon \in ]0, 1[$. $\mathcal{S}$ is a $\varepsilon$-coreset for $L$ if, for any query $\theta$, the estimated cost is equal to the exact cost up to a relative error, i.e. for all $\theta \in \Theta$

$$\left| \frac{\hat{L}(\theta)}{L(\theta)} - 1 \right| \leq \varepsilon \tag{1}$$

An important consequence of the coreset property is the following

**Theorem 1.** *Let be $\mathcal{S}$, a $\varepsilon$-coreset for $L$. Define $\theta^{opt} := \min_{\theta \in \Theta} L(\theta)$ and $\hat{\theta}^{opt} := \min_{\theta \in \Theta} \hat{L}(\theta)$. Then $L(\hat{\theta}^{opt})$ is an $(1+\varepsilon)$-approximation of $L(\theta^{opt})$, i.e.*

$$L(\theta^{opt}) \leq L(\hat{\theta}^{opt}) \leq (1 + 3\varepsilon) L(\theta^{opt})$$

*Proof.* If $\mathcal{S}$ is a $\varepsilon$-coreset for $L$, we have from **??** that

$$(1 - \varepsilon) L(\theta^{\text{opt}}) \leq (1 - \varepsilon) L(\hat{\theta}^{\text{opt}}) \leq \hat{L}(\hat{\theta}^{\text{opt}}) \leq \hat{L}(\theta^{\text{opt}}) \leq (1 + \varepsilon) L(\theta^{\text{opt}}) \tag{2}$$

and moreover

$$L(\theta^{\text{opt}}) \leq L(\hat{\theta}^{\text{opt}}) \leq \frac{(1 + \varepsilon)}{(1 - \varepsilon)} L(\theta^{\text{opt}}) \leq (1 + 3\varepsilon) L(\theta^{\text{opt}}) \tag{3}$$

□

This key property makes coreset very relevant in a machine learning context, and inscribes them into a more general learning framework that is PAC learning.

# 3 Coresets and PAC learning

## 3.1 PAC learning

In computational learning theory, probably approximately correct (PAC) learning is a framework for mathematical analysis of machine learning. It was proposed in 1984 in **valiant1984learnable**. The first idea is that a learning problem can be formulate into an expected risk minimization. In another words, by learning, one is interested in minimizing errors over a distribution of guesses it would have to make. To do so, the learner will receives samples and must select a prediction function based on them. The PAC framework states that the learner ability can be quantified by how probable (the "probably" part) the learner have a low generalization error (the "approximately correct" part) in some sense.

In that framework, several practical issues can occur.

- The richness of the considered class of prediction function can be too small to embrace the complexity of the studied phenomena.
- The risk optimizer algorithm could struggle finding the minimizing function, for instance only finding local minima, or yielding high computational complexity.
- The sample complexity required for reaching a given level of "probable" in the approximately correctness can vary.

However, a learning problem is generally not separable into these three issues. This means their resolution is not independent and had to be tackled jointly. For instance, making more expressive a class of prediction function can make its optimization more difficult, or make the sample complexity required higher. The latter case is well known as overfitting.

### 3.2 Link with coresets

Let us see how coresets can naturally intervene into the PAC framework. Formally, let be given a probability distribution $\mathbb{P}$ generating the data $\mathcal{S}$, and let be $\mathcal{F}$ a family of loss functions. Minimizing the expected loss is equivalent to finding $f^* := \arg\min_{f \in \mathcal{F}} \mathbb{E}f$. Because we are only given $\mathcal{S}$ and not the full distribution, we have to approximate $\mathbb{E}f$ by some estimate $\mathbb{E}f_{\mathcal{S}}$ based on the data, and then minimizing it with $\hat{f}^* := \arg\min_{f \in \mathcal{F}} \mathbb{E}f_{\mathcal{S}}$.

In order to evaluate this scheme, we fix some $\varepsilon > 0$, and we want with the highest probability $1 - \delta$ as possible

$$|\mathbb{E}\hat{f}^* - \mathbb{E}f^*| \leq \varepsilon$$

Put differently we want

$$\mathbb{P}\left[|\mathbb{E}\hat{f}^* - \mathbb{E}f^*| \geq \varepsilon\right] \leq \delta$$

But we know a sufficient condition to control this error, that's the coreset property! Indeed, if we have sample a data set $\mathcal{S}$ such that

$$\mathbb{P}\left[\forall f \in \mathcal{F},\ |\frac{\mathbb{E}f_{\mathcal{S}}}{\mathbb{E}f} - 1| \leq \varepsilon/3\right] \geq 1 - \delta$$

i.e. $\mathcal{S}$ is an $\varepsilon/3$-coreset for $\mathbb{E}$ with probability $1 - \delta$, then by **??** we know

$$\mathbb{E}f^* \leq \mathbb{E}\hat{f}^* \leq (1 + 3\varepsilon/3)\mathbb{E}f^* \iff |\mathbb{E}\hat{f}^* - \mathbb{E}f^*| \leq \varepsilon$$

We thus see that the PAC framework translates to approximating with high probability the evaluation of a function on a data subset, which is guaranteed by the coreset property.

On another hand, the use of coresets leverage one of the three issues that occur in PAC learning, that is to reduce the number of samples required to compute an optimal prediction function, and still controlling the error. If the time complexity for an optimization algorithm to optimize on $n$ data points is $O(a_n)$, and that it takes $O(b_m)$ time to sample an $\varepsilon$-coreset which is of size $m \leq n$, then we have interest in building coreset as soon as $O(a_n) \geq O(b_m) + O(a_m)$.

## 4 State-of-the-art results on coresets

**Definition 4.1** (Sensitivity)**.** The sensitivity $\sigma_i$ of a data point $x_i$ and the total sensitivity $\mathfrak{S}$ of $\mathcal{X}$ are

$$\begin{cases} \sigma_i = \sup_{\theta \in \Theta} q_\theta(x_i) = \sup_{\theta \in \Theta} \frac{f_\theta(x_i)}{L(\theta)} & \in [0, 1] \\ \mathfrak{S} = \sum_{i=1}^n \sigma_i \end{cases}$$

### 4.1 Main proof

Let be $s$ an upper bound on sensitivity $\sigma$ i.e. $\forall i, s_i \geq \sigma_i$, and $S := \sum_{i=1}^n s_i$. Furthermore, let be sampled $\mathcal{S} \sim \mathcal{M}(m, s/S)$, the multinomial sampling case. Define $g_\theta(x_i) := \frac{q_\theta(x_i)}{s_i} = \frac{f_{\theta(x_i)}}{s_i L(\theta)} \in [0, 1]$

By Hoeffding's inequality, we thus have for any $\theta \in \Theta$ and $\varepsilon' > 0$

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{x \in \mathcal{S}} g_\theta(x) - \mathbb{E}\left[g_\theta(x)\right]\right| > \varepsilon'\right] \leq 2\exp\left(-2m\varepsilon'^2\right) \tag{4}$$

and by definition, $\mathbb{E}\left[g_\theta(x)\right] = \frac{1}{S}$ and $\frac{1}{m}\sum_{x \in \mathcal{S}} g_\theta(x) = \frac{\hat{L}_{\text{iid}}(\theta)}{SL(\theta)}$, thus

$$\mathbb{P}\left[|\hat{L}_{\text{iid}}(\theta) - L(\theta)| > \varepsilon' S L(\theta)\right] \leq 2\exp\left(-2m\varepsilon'^2\right)$$

Hence, $\mathcal{S}$ satisfies the $\varepsilon$-coreset property **??** for any single query $\theta \in \Theta$ with probability at least $1 - \delta$, if we choose

$$m \geq \frac{S^2}{2\varepsilon^2} \log \frac{2}{\delta} \tag{5}$$

### 4.2 Extension to all queries

developp sota

See **Uniform guarantee for all queries** in **bachem2017coresetML**. Introducing the pseudo-dimension $d'$, it gives

$$m \geq \mathcal{O}(\frac{S^2}{2\varepsilon^2}(d' + \log\frac{2}{\delta})) \tag{6}$$

See **Theorem 5.5** of **braverman2016coresetsota** for an improved bound.

$$m \geq \mathcal{O}(\frac{S}{2\varepsilon^2}(d' \log S + \log\frac{2}{\delta})) \tag{7}$$

See **bachem2017coresetML**. Determinantal Point Processes

## 5 Some intuition

A Determinantal Point Processes (DPP) is a random sampling over subsets of a fixed ground set.

The first essential characteristic is that a DPP is entirely encompassed by a given positive kernel, which can be tuned to a range of specific contexts. DPP can thus be said to be the kernel machine of point processes as they allow both tractability and a flexibility.

The second essential characteristic of a DPP is that the occurrences of the element of the ground set are negatively correlated, i.e. the inclusion of one item makes the inclusion of other items less likely. The strengths of these negative correlations are derived from a kernel matrix that defines a global measure of similarity between pairs of items, so that more similar items are less likely to co-occur. As a result, DPPs assign higher probability to sets of items that are diverse.

## 6 Definition

Determinantal Point Processes are before all point processes, which can be described as processes for selecting a collection of mathematical points randomly located on a mathematical space. Formally, a point process $\mathbb{P}$ on a ground set $\mathcal{X}$ is a probability measure over "point patterns" or "point configurations" of $\mathcal{X}$, which are subsets of $\mathcal{X}$. For instance, $\mathcal{X}$ could be a continuous region of the euclidean plane in which a scientist injects some quantum particles trapped into a potential well. Then $\mathbb{P}\left(\{x_1, x_2, x_3\}\right)$ characterizes the likelihood of seeing these particles at places $x_1, x_2$, and $x_3$. Depending on the type of the particles, the measurements might tend to cluster together, or they might occur independently, or they might tend to spread out into space. $\mathbb{P}$ captures these correlations.

In the following, we focus on discrete, finite point processes, where we assume without loss of generality that $\mathcal{X} = \{x_i \mid i \in [\![1, n]\!]\}$, in this setting we sometimes refer to elements of $\mathcal{X}$ as items. The discrete setting is computationally simpler and often more appropriate for real-world data.

In the discrete case, a point process is simply a probability measure on $2^{\mathcal{X}}$ i.e. the power set of $\mathbb{P}$ i.e. the set of all subsets of $\mathcal{X}$. A sample from $\mathbb{P}$ might be the empty set, the entirety of $\mathcal{X}$, or anything in between.

**Definition 6.1** (Determinantal Point Process)**.** $\mathbb{P}$ is called a determinantal point process if, when $\mathcal{S}$ is a random subset drawn according to $\mathbb{P}$, we have, for every $A \subseteq \mathcal{X}$,

$$\mathbb{P}(A \subseteq \mathcal{S}) = \det\left(K_A\right) \tag{8}$$

for some real, symmetric matrix $K \in \mathbb{R}^{n \times n}$ indexed by the elements of $\mathcal{X}$.

Here, $K_A \equiv \left[K_{ij}\right]_{i,j \in A}$ denotes the restriction of $K$ to the entries indexed by elements of $A$, and we adopt $\det\left(K_\emptyset\right) = 1$. Note that normalization is unnecessary here, since we are defining marginal probabilities that need not sum to $1$ .

Since $\mathbb{P}$ is a probability measure, all principal minors $\det\left(K_A\right)$ of $K$ must be positives, and thus $K$ itself must be positive. It is possible to show in the same way that the eigenvalues of $K$ are bounded above by one. These requirements turn out to be sufficient. By the Macchi-Soshnikov theorem (**macchi1975dpp**),any $K$ such that $0 \preceq K \preceq I$, defines a DPP.

We refer to $K$ as the marginal kernel since it contains all the information needed to compute the probability of any subset $A$ being included in $\mathcal{S}$. A few simple observations follow from Definition **??**. If $A = \{i\}$ is a singleton, then we have

$$\mathbb{P}(i \in \mathcal{S}) = K_{ii}$$

That is, the diagonal of $K$ gives the marginal probabilities of inclusion for individual elements of $\mathcal{X}$. Diagonal entries close to 1 correspond to elements of $\mathcal{X}$ that are almost always selected by the DPP. Furthermore, if $A = \{i, j\}$ is a two-element set, then

$$\begin{aligned} \mathbb{P}(i, j \in \boldsymbol{X}) &= \begin{vmatrix} K_{ii} & K_{ij} \\ K_{ji} & K_{jj} \end{vmatrix} \\ &= K_{ii}K_{jj} - K_{ij}K_{ji} \\ &= \mathbb{P}(i \in \boldsymbol{X})\mathbb{P}(j \in \boldsymbol{X}) - K_{ij}^2 \end{aligned} \tag{9}$$

Thus, the off-diagonal elements determine the negative correlations between pairs of elements: large values of $K_{ij}$ imply that $i$ and $j$ tend not to co-occur.

Equation **??** demonstrates why DPP are "diversifying". If we think of the entries of the marginal kernel as measurements of similarity between pairs of elements in $\mathcal{X}$, then highly similar elements are unlikely to appear together. If $K_{ij} = \sqrt{K_{ii}K_{jj}}$, then $i$ and $j$ are "perfectly similar" and do not appear together almost surely. Conversely, when $K$ is diagonal there are no correlations and the elements appear independently. Note that DPP cannot represent distributions where elements are more likely to co-occur than if they were independent: correlations are always negative.

## 7 Examples

DPP occur naturally in some simple random models. Obviously, any independent sampling of elements of a set is trivially a (diagonal) DPP. But maybe the simpler non-trivial instance of a DPP is the descents in random sequences.

Take a sequence of $N$ random numbers drawn uniformly and independently from a finite set e.g. the digits, $[\![0, 9]\!]$. The locations in the sequence where the current number is less than the previous number form a subset of $[\![2, N]\!]$. Noticeably, this subset is distributed as a DPP. Intuitively, if the current number is less than the previous number, it is probably not too large, thus it becomes less likely that the next number will be smaller yet. In this sense, the positions of decreases repel one another.

Edges in uniform spanning trees, eigenvalues of random matrices, as well as some quantum experimental models are also well-known instances of DPP. By the way, and for the history, DPP were first identify as a class by Macchi, who called them "fermion process" because they give the distributions of fermion systems at thermal equilibrium. The Pauli exclusion principle states that no two fermions can occupy the same quantum state; as a consequence fermions exhibit what is known as the "anti-bunching" effect. This repulsion is described precisely by a DPP.

## 8 Geometric interpretation

DPP are defined on determinants, that have an intuitive geometric interpretation. Since the DPP kernel $K$ is symmetric, it exists $V \in \mathbb{R}^{d \times n}$ such that $K = VV^\top$. Denote the columns of $V$ by $(V_i)$ for $i \in [\![1, n]\!]$. Then $\forall A \subset \mathcal{X}$

$$\mathbb{P}(A \subset \mathcal{S}) = \mathrm{Vol}^2(V_A) \tag{10}$$

The right hand side is the squared $|A|$-dimensional volume of the parallelepiped spanned by the columns of $V$ corresponding to elements in $A$.

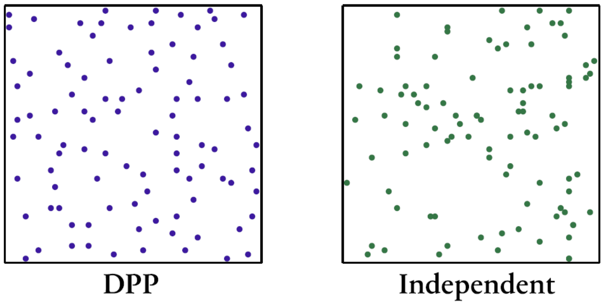

Figure 1: (left) A set of points in the plane drawn from a DPP, with $K_{ij}$ inversely related to the distance between points $i$ and $j$. (right) The same number of points sampled independently using a Poisson point process , which results in random clumping.

Intuitively, we can think of the columns of $V$ as feature vectors describing the elements of $\mathcal{X}$. Then the kernel $K$ measures similarity using dot products between feature vectors, and Equation **??** says that the probability assigned by a DPP to the inclusion of a set $A$ is related to the volume spanned by its associated feature vectors. This is illustrated in Figure **??**.

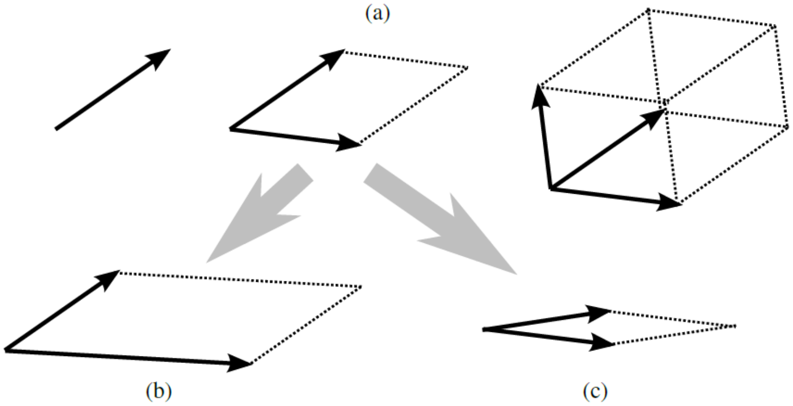

Figure 2: A geometric interpretation: each vector corresponds to an element of $\mathcal{X}$. (a) The probability of inclusion of a subset $A$ is the square of the volume spanned by its associated feature vectors. (b) As the magnitude of an item's feature vector increases, so do the probabilities of sets containing that item. (c) As the similarity between two items increases, the probabilities of sets containing both of them decrease.

This geometric interpretation explain why diverse sets are more probable. It is because their feature vectors are more orthogonal, and hence span larger volumes. Conversely, items with parallel feature vectors are selected together with probability zero, since their feature vectors define a degenerate parallelepiped. Ceteris paribus, items with large magnitude feature vectors are more likely to appear, because the spanned volume for sets containing them evolves linearly with respect to their magnitude, and thus the probability evolves quadratically with respect to it.

## 9 Some useful properties

projection DPP

despite exponential support

# 10 Sampling from a DPP

## 10.1 Exact DPP sampling

Although DPP can be impressively efficient given the exponential number of subsets being sampled from, sampling can be rapidly limited by performance. Except for a few specialised kernels like the edges in uniform spanning trees mentioned previously, the default exact sampler is a spectral algorithm due to **hough2006_hkpv**.

It leverages the fact that DPP are mixtures of projection DPP to generate repeated samples given the spectral content of the kernel. This method is commonly called the spectral method since it requires the spectral/eigendecomposition of the positive kernel.

Formally, if a DPP is defined by a kernel $K$ defined on $n$ data points, one requires the eigendecomposition $K = VV^\top$ where $V \in \mathbb{R}^{d \times n}$. This can often be the computational bottleneck since it generally requires $O(n^3)$ time. Note however that for some DPP based on specific kernels like OPE kernels, $K$ is built via this decomposition and thus it is trivially known.

In any cases when multiple samples are required, this eigendecomposition can be reused. Then each sample from the spectral algorithm requires only $O(nk^2)$ time, where $k$ is the number of elements sampled. This means $O(n(\operatorname{Tr} K)^2)$ time on average. If $K$ is a projection kernel, $k = \operatorname{Tr} K = d$ which is a constant than can be small in many practical applications, e.g. in a recommendation context, $k$ would often be less than 10.

Some recent works from **gillenwater2019_treebased_fast_dpp_sampling** improved somewhat this complexity. Based on the still needed eigendecomposition, it implements a binary tree structure storing appropriate summary statistics of the eigenvectors, requiring $O(nd^2)$ to build, but can then generate repeated samples in $O(\log(n)k^2d^2 + d^3)$ time, hence $O(\log(n)d^4)$ for a projection kernel.

This method becomes a viable alternative to the spectral method when the total number of items $n$ is large and when the dimensionality $d$ of the features and the expected sample size $\operatorname{Tr} K$ are small compared to $n$.

## 10.2 Approximate DPP sampling

Several sampling methods have been developed in the case we only need an approximated DPP sampling.

A first class of methods involves a kernel approximation of an given DPP kernel, using random projections such as in **kulesza2012_dpp_for_ml**, or low-rank factorization techniques.

A second class involves Monte Carlo Markov Chain (MCMC). This is often down in an inexact fashion usigin target distribution close but different from DPP one. Noticeably, **gautier2017_zonotope_for_dpp_sampling** proposed an exact MCMC sampler for projection DPP.

Improving

# 11 Variance arguments

## 11.1 Three sampling cases

### 11.1.1 Multinomial case

In the multinomial case, we have $\mathcal{S} \sim \mathcal{M}(m, q)$ i.e. $m$ i.i.d. categorical sampling with $\mathbb{P}(x_i) = q(x_i)$. Then an unbiased estimator of $L$ is

$$\hat{L}_{\text{iid}}(\theta) = \sum_{x_i \in \mathcal{S}} \frac{f_\theta(x_i)}{mq(x_i)}$$

Its variance is

$$\mathbb{V}\text{ar}_{\text{iid}}(\theta) := \frac{1}{m} \mathbb{V}\text{ar}\left[\frac{f_\theta(x_i)}{q(x_i)}\right] = \frac{1}{m} \sum_{x \in \mathcal{X}} \frac{f_\theta(x)^2}{q(x)} - \frac{1}{m} L(\theta)^2 = f_\theta^\top (\frac{Q^{-1}}{m} - \frac{\boldsymbol{J}}{m}) f_\theta \qquad (11)$$

where $Q = \text{diag}(q)$ and $\boldsymbol{J} = \boldsymbol{j}\boldsymbol{j}^\top$ the matrix full of ones.

For any query $\theta \in \Theta$, the variance is reduced to 0 by

$$q_\theta(x) := \frac{f_\theta(x)}{L(\theta)}$$

### 11.1.2 DPP case

In the DPP case, we have $\mathcal{S} \sim \mathcal{DPP}(K),\ \pi_i := K_{ii}$. Then an unbiased estimator of $L$ is

$$\hat{L}_{\text{DPP}}(\theta) = \sum_{x_i \in \mathcal{S}} \frac{f_\theta(x_i)}{\pi_i}$$

Its variance can be computed using $\varepsilon_i$ as the counting variable for $x_i$:

$$\mathbb{V}\text{ar}_{\text{DPP}}(\theta) = \sum_{i,j} \mathbb{E}\left[\varepsilon_i \varepsilon_j\right] \frac{f_\theta(x_i) f_\theta(x_j)}{\pi_i \pi_j} - L(\theta)^2 \quad \text{with} \quad \mathbb{E}\left[\varepsilon_i \varepsilon_j\right] = \begin{cases} \det(K_{\{i,j\}}) = \pi_i \pi_j - K_{ij}^2, & \text{if } i \neq j \\ \mathbb{E}\left[\varepsilon_i\right] = \pi_i, & \text{if } i = j \end{cases}$$

Introducing $\Pi = \text{diag}(\pi)$ and $\tilde{K} = \Pi^{-1} K^{\odot 2} \Pi^{-1}$, we can rewrite

$$\mathbb{V}\text{ar}_{\text{DPP}}(\theta) = \sum_i \left(\frac{1}{\pi_i} - 1\right) f_\theta(x_i)^2 - \sum_{i \neq j} \frac{K_{ij}^2}{\pi_i \pi_j} f_\theta(x_i) f_\theta(x_j) = f_\theta^\top (\Pi^{-1} - \tilde{K}) f_\theta \qquad (12)$$

For a Bernoulli process where $\mathbb{P}(x_i \in \mathcal{S}) = \pi_i$ independently, the DPP kernel reduces to its diagonal i.e. $K = \Pi$ then $\tilde{K} = I$. We denote its variance $\mathbb{V}\text{ar}_{\text{diag}}$.

### 11.1.3 m-DPP case

In the m-DPP case, we have $\mathcal{S} \sim \mathcal{DPP}(K) \mid |S| = m$, and the marginals $b_i := \mathbb{E}\left[\varepsilon_i\right]$ have an analytic form. Then an unbiased estimator of $L$ is

$$\hat{L}_{\text{mDPP}}(\theta) = \sum_{x_i \in \mathcal{S}} \frac{f_\theta(x_i)}{b_i}$$

Note that we could also be interested in a biased cost function such as the diversified risk introduced by **zhang2017dppminibatch**

$$\tilde{L}(\theta) = \frac{1}{m} \mathbb{E}_{x \sim \text{mDPP}}[f_\theta(x)] = \frac{1}{m} \sum_{x_i \in \mathcal{X}} b_i f_\theta\left(x_i\right)$$

Then an unbiased estimator of $\tilde{L}$ is

$$\hat{\tilde{L}}_{\text{mDPP}}(\theta) = \frac{1}{m} \sum_{x_i \in \mathcal{S}} f_\theta(x_i)$$

We can switch between $L$ and $\tilde{L}$, substituting $f_\theta(x_i)$ by $\frac{b_i f_\theta(x_i)}{m}$.

Returning to the estimation of $L$, we are interested in the variance of $\hat{L}_{\text{mDPP}}$ which is

$$\mathbb{V}\text{ar}_{\text{mDPP}}(\theta) = \sum_i \left(\frac{1}{b_i} - 1\right) f_\theta(x_i)^2 + \sum_{i \neq j} C_{ij} f_\theta(x_i) f_\theta(x_j) \tag{13}$$

where $C_{ij} = \frac{\mathbb{E}[(\varepsilon_i - b_i)(\varepsilon_j - b_j)]}{\mathbb{E}[\varepsilon_i]\mathbb{E}[\varepsilon_j]} = \frac{\mathbb{E}[\varepsilon_i \varepsilon_j]}{b_i b_j} - 1$

Observe that if the m-DPP kernel is reduced to its diagonal ($C_{ij} = 0$), we recover $\mathbb{V}\text{ar}_{\text{diag}}$, the variance of a Bernoulli process with same marginals ($\pi_i = b_i$), though here the number of elements sampled is fixed to $m$.

In order to benefit from some variance reduction, one should want $\forall i \neq j$ , $C_{ij} f_\theta(x_i) f_\theta(x_j) < 0$ for a given m-DPP. **zhang2017dppminibatch** discuss that intuitively, if the m-DPP kernel rely on some similarity measure and that $f$ is smooth for it, then 2 similar points should have both negative correlation ($C_{ij} < 0$) and their value have positive scalar product ($f_\theta(x_i) f_\theta(x_j) > 0$). Reversely, it is argued that 2 dissimilar points should have positive correlation , and their value show "no tendency to align" hinting $f_\theta(x_i) f_\theta(x_j) < 0$. We could more conservatively consider that the induced variance change, whether positive or negative, would in either case be small, as for DPP and m-DPP, 2 dissimilar points tend toward independence.

contradiction with property of strong Rayleigh measures

### 11.2 Variance comparison

In order to compare processes with same marginals, we set $\Pi = mQ$. Then $\mathbb{V}\text{ar}_{\text{iid}}$, $\mathbb{V}\text{ar}_{\text{diag}}$ and $\mathbb{V}\text{ar}_{\text{DPP}}$ are quadratic forms of $f_\theta$ associated with respective matrices

$$\begin{cases} \mathbb{V}\text{ar}_{\text{iid}} \equiv \Pi^{-1} - \frac{\boldsymbol{J}}{m} \\ \mathbb{V}\text{ar}_{\text{diag}} \equiv \Pi^{-1} - I \\ \mathbb{V}\text{ar}_{\text{DPP}} \equiv \Pi^{-1} - \tilde{K} \end{cases}$$

#### 11.2.1 Comparing DPP versus diag

The DPP variance strictly beats uniformly the Bernoulli process variance if $\tilde{K}$ strictly dominates identity i.e.

$$\forall f_\theta, \mathbb{V}\text{ar}_{\text{DPP}} < \mathbb{V}\text{ar}_{\text{diag}} \iff \tilde{K} \succ I \tag{14}$$

But $\tilde{K}$ is a symmetric positive definite matrix and by Hadamard inequality $\det(\tilde{K}) \leq \prod_i \tilde{K}_{ii} = 1$. Therefore at least one of its eigenvalue is lower than 1, hence $\tilde{K} \not\succ I$.

#### 11.2.2 Comparing DPP versus i.i.d.

The DPP variance strictly beats uniformly the multinomial variance if $\tilde{K}$ strictly dominates $\frac{\boldsymbol{J}}{m}$ i.e.

$$\forall f_\theta, \mathbb{V}\text{ar}_{\text{DPP}} < \mathbb{V}\text{ar}_{\text{iid}} \iff \tilde{K} \succ \frac{\boldsymbol{J}}{m} \tag{15}$$

$K$ being symmetric positive of rank $r \in [\![0, n]\!]$, it exists $V \in \mathbb{R}^{r \times n}$ such that $K = V^\top V$, and we denote by $V_i$ its colons, for $i \in [\![1, n]\!]$.

For any vector $v \in \mathbb{R}^r$, **copenhaver2013diagramvectors** define its diagram vector

$$\tilde{v} := \frac{1}{\sqrt{r-1}}((v_k^2 - v_l^2, \sqrt{2r} v_k v_l) \mid k < l)^\top \in \mathbb{R}^{r(r-1)}$$

concatenating all the differences of squares and products.

Then introducing $\tilde{V} = \left(\tilde{V}_i \mid i \in [\![1, n]\!]\right)$ allows us to rewrite $\tilde{K} = \frac{\boldsymbol{J}}{r} + \frac{r-1}{r}\tilde{V}^\top \tilde{V}$ thus $\tilde{K} - \frac{\boldsymbol{J}}{m} = (\frac{1}{r} - \frac{1}{m})\boldsymbol{J} + \frac{m-1}{m}\tilde{V}^\top \tilde{V}$. For having $\tilde{K} - \frac{\boldsymbol{J}}{m} \succeq 0$, it is sufficient to have $m \geq r$. This is exactly the case for a projective DPP with rank $r = m$, because $m \leq r$ holds for every DPP. Therefore, for every multinomial sampling, we have a projective DPP which always beats it uniformly.

## 12 Improving concentration with DPP

**bardenet2021sgddpp** show the existence of a sequence of DPP kernels $(\tilde{\boldsymbol{K}}_m)$, independent of $f$, whose induced estimator has asymptotic variance $\mathcal{O}(m^{-(1+\frac{1}{d})})$. More precisely, equation (S14) yields that

explain how the estimator is built

$$\mathbb{V}\mathrm{ar}[\frac{\hat{L}_{\mathrm{DPP}}(\theta)}{L(\theta)}] = M_\theta \mathcal{O}(m^{-(1+\frac{1}{d})}) + \mathcal{O}(n^{-1/2}) \tag{16}$$

where $\sqrt{M_\theta}$ is the Lipschitz constant of $x \mapsto f_\theta(x)(\frac{1}{m}K^{(m)}_{q,\tilde{\gamma}}(x,x))^{-1}$, supposedly bounded and whose a bound we denote by $M := \sup_{\theta \in \Theta} M_\theta$.

Bienaymé-Tchebychev inequality then gives

$$\mathbb{P}\left[|L(\theta) - \hat{L}_{\mathrm{DPP}}(\theta)| > \varepsilon L(\theta)\right] \leq \frac{\mathbb{V}\mathrm{ar}[\hat{L}_{\mathrm{DPP}}(\theta)]}{L(\theta)^2 \varepsilon^2} = \frac{1}{\varepsilon^2}(M_\theta \mathcal{O}(m^{-(1+\frac{1}{d})}) + \mathcal{O}(n^{-1/2})) \tag{17}$$

Hence, $\mathcal{S} \sim \mathcal{DPP}(\tilde{\boldsymbol{K}}_m)$ satisfies $1-\delta$-surely the $\varepsilon$-coreset property **??** for

$$m^{1+\frac{1}{d}} \gtrsim \frac{M_\theta}{\delta\varepsilon^2 + \mathcal{O}(n^{-1/2})} = \frac{M_\theta}{\delta\varepsilon^2} \frac{1}{1 + \frac{1}{\delta\varepsilon^2}\mathcal{O}(n^{-1/2})} \tag{18}$$

where $y \gtrsim x$ is a transitive notation for $y = \Omega(x)$ i.e. $y$ is lower bounded by $x$ up to a constant factor. Then this means that for sufficiently large $n$ (potentially $n \gtrsim \delta^{-2}\varepsilon^{-4}$), we can control the second factor and thus obtain the bound

$$\boxed{m \gtrsim \left(\frac{M_\theta}{\delta\varepsilon^2}\right)^{\frac{1}{1+\frac{1}{d}}}} \tag{19}$$

make that a Lemma

**Lemma 1.**

$$\sup_{f \in \mathcal{F}} \mathbb{P}\left[|\mathbb{E}f_{\mathcal{S}} - \mathbb{E}f| \geq \varepsilon\right] \leq \delta \tag{20}$$

### 12.1 Extension to all queries

In order to obtain an $\varepsilon$-coreset, the $\varepsilon$-coreset property **??** must holds for all queries, thus the previous result must be generalized to all $\theta \in \Theta$.

For every function $f \in \mathcal{F}$ and multiset $\mathcal{S}$, let the restriction of the function $f$ to the multiset $\mathcal{S}$ be denoted by $f_{\mathcal{S}} := (f(x))_{x \in \mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$, and let its mean be denoted by $\mathbb{E}f_{\mathcal{S}} := \frac{1}{|\mathcal{S}|}\sum_{x \in \mathcal{S}} f(x)$.

**Corollary 1.** *Let be $\mathcal{S} \sim \mathcal{DPP}(K_m)$. Taking $\delta = 1/2$ in precedent result* **??***, we know it exists some constant we denote $c_{1/2}$ such that $\forall \varepsilon > 0$ it holds that*

$$m \geq c_{1/2}\left(\frac{M_{\mathcal{F}}}{\varepsilon^2}\right)^{\frac{1}{1+\frac{1}{d}}} \implies \forall f \in \mathcal{F},\ \mathbb{P}\left[|\mathbb{E}f_{\mathcal{S}} - \mathbb{E}f| \leq \varepsilon\right] \geq \frac{1}{2} \tag{21}$$

We follow a similar proof scheme as in section 9.4 of **haussler1992decisiontheoricgeneralizationofPACmodel**. We specifically revisit Lemma 12. and 13., getting rid of independency hypothesis, and making intermediary results more flexible to further improvements.

state of the art slightly better than haussler, cf. Li, but strongly? require independancy

**Lemma 2.** *Let be $\mathcal{S}_1, \mathcal{S}_2$ two multisets of size $m$ independently sampled from the same distribution supported on $\mathcal{X}^m$. Assume that for a given $\varepsilon > 0$ we can control uniformly on $f$ the concentration $1/2$-surely, formally $\forall f \in \mathcal{F},\ \mathbb{P}\left[|\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f| \leq \varepsilon/2\right] \geq \frac{1}{2}$. Then*

$$\mathbb{P}\left[\exists f \in \mathcal{F},\ |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f| \geq \varepsilon\right] \leq 2\mathbb{P}\left[\exists f \in \mathcal{F},\ |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| \geq \varepsilon/2\right] \tag{22}$$

*Proof.* Let be $\mathcal{S}_1$ sampled such that $\exists f \in \mathcal{F},\ |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f| \geq \varepsilon$. This obviously happens with probability $\mathbb{P}\left[\exists f \in \mathcal{F}, |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f| \geq \varepsilon\right]$.

For such an $f$, we then sample $\mathcal{S}_2$ such that $|\mathbb{E}f_{\mathcal{S}_2} - \mathbb{E}f| \leq \varepsilon/2$. By hypothesis, this happens with probability greater than $1/2$, and we thus have

$$\begin{aligned}
\mathbb{P}\left[\exists f \in \mathcal{F},\ |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f| \geq \varepsilon\right] \frac{1}{2} &\leq \mathbb{P}\left[\exists f \in \mathcal{F},\ |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f| \geq \varepsilon \wedge |\mathbb{E}f_{\mathcal{S}_2} - \mathbb{E}f| \leq \varepsilon/2\right] \\
&\leq \mathbb{P}\left[\exists f \in \mathcal{F},\ |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| \geq \varepsilon/2\right]
\end{aligned}$$

where we lastly used the triangular inequality $|\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f| - |\mathbb{E}f_{\mathcal{S}_2} - \mathbb{E}f| \leq |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}|$ □

**Lemma 3.** *Let be $\mathcal{S}_1, \mathcal{S}_2$ two multisets of size $m$ independently sampled from the same distribution supported on $\mathcal{X}^m$.*

$$\mathbb{P}\left[\exists f \in \mathcal{F},\ |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| \geq \varepsilon\right] \leq \sup_{\mathcal{S} \in \binom{\mathcal{X}}{2m}} N(\varepsilon/8, \mathcal{F}, \mathbb{E}\cdot_{\mathcal{S}}) \sup_{f \in \mathcal{F}} \mathbb{P}\left[|\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| \geq \varepsilon/4\right] \tag{23}$$

recall cover

*Draft of Proof.* For every two multisets $\mathcal{S}_1, \mathcal{S}_2$ of size $m$, we denote their multiset union $\mathcal{S} := \mathcal{S}_1 \uplus \mathcal{S}_2 \subseteq \mathcal{X}^{2m}$.

Let be $\mathcal{S}$ sampled such that $\exists f \in \mathcal{F},\ |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| \geq \varepsilon/2$.

Let then be taken $\mathcal{F}^*_{\mathcal{S}}$, an $\varepsilon/8$-cover of $\mathcal{F}$ for the $\mathbb{E}\cdot_{\mathcal{S}}$ topology, such that $|\mathcal{F}^*_{\mathcal{S}}| = N(\varepsilon/8, \mathcal{F}, \mathbb{E}\cdot_{\mathcal{S}})$. We thus know it exists $f^* \in \mathcal{F}^*_{\mathcal{S}}$ such that $\mathbb{E}|f - f^*|_{\mathcal{S}} \leq \varepsilon/8$.

Applying some triangular inequalities yields that

$$\begin{aligned}
|\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| &\leq |\mathbb{E}f^*_{\mathcal{S}_1} - \mathbb{E}f^*_{\mathcal{S}_2}| + |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f^*_{\mathcal{S}_1}| + |\mathbb{E}f_{\mathcal{S}_2} - \mathbb{E}f^*_{\mathcal{S}_2}| \\
&\leq |\mathbb{E}f^*_{\mathcal{S}_1} - \mathbb{E}f^*_{\mathcal{S}_2}| + \mathbb{E}|f - f^*|_{\mathcal{S}_1} + \mathbb{E}|f - f^*|_{\mathcal{S}_2} \\
&\leq |\mathbb{E}f^*_{\mathcal{S}_1} - \mathbb{E}f^*_{\mathcal{S}_2}| + 2\mathbb{E}|f - f^*|_{\mathcal{S}} \\
\iff \quad |\mathbb{E}f^*_{\mathcal{S}_1} - \mathbb{E}f^*_{\mathcal{S}_2}| &\geq |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| - 2\mathbb{E}|f - f^*|_{\mathcal{S}} \geq \varepsilon/4
\end{aligned}$$

Therefore

$$\mathbb{P}\left[\exists f \in \mathcal{F},\ |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| \geq \varepsilon/2\right] \leq \mathbb{P}\left[\exists f \in \mathcal{F}^*_{\mathcal{S}},\ |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| \geq \varepsilon/4\right] \tag{24}$$

By the law of total expectation, we obtain

$$\begin{aligned}
\mathbb{P}\left[\exists f \in \mathcal{F}^*_{\mathcal{S}},\ |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| \geq \varepsilon/4\right] &= \mathbb{E}\left[\mathbb{1}\{\exists f \in \mathcal{F}^*_{\mathcal{S}},\ |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| \geq \varepsilon/4\}\right] \\
&= \mathbb{E}\left[\mathbb{P}\left[\exists f \in \mathcal{F}^*_{\mathcal{S}},\ |\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| \geq \varepsilon/4 \mid \mathcal{F}^*_{\mathcal{S}}\right]\right] \\
&\leq \sup_{\mathcal{F}^*_{\mathcal{S}}} |\mathcal{F}^*_{\mathcal{S}}| \sup_{f \in \mathcal{F}} \mathbb{P}\left[|\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| \geq \varepsilon/4\right] \\
&= \sup_{\mathcal{S} \in \binom{\mathcal{X}}{2m}} N(\varepsilon/8, \mathcal{F}, \mathbb{E}\cdot_{\mathcal{S}}) \sup_{f \in \mathcal{F}} \mathbb{P}\left[|\mathbb{E}f_{\mathcal{S}_1} - \mathbb{E}f_{\mathcal{S}_2}| \geq \varepsilon/4\right]
\end{aligned}$$

□

Let be $d_{L^1}(x,y) = \frac{1}{m}\sum_{i=1}^{m} |x_i - y_i| = \frac{1}{m}\|x - y\|$ $\mathcal{F}_{\mathcal{S}_1,\mathcal{S}_2} = \{(f_\theta(x))_{x\in\mathcal{S}_1\uplus\mathcal{S}_2} | \theta \in \Theta\} \subseteq \mathbb{R}^{2m}$

For all $x \in \mathbb{R}^{2m}$, let be defined $\mathbb{E}_1[x] = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $\mathbb{E}_2[x] = \frac{1}{m}\sum_{i=m+1}^{2m} x_i$, the empiric mean on the vector first and second half respectively.

$\exists \theta \in \Theta, \frac{1}{m}|L_{\mathcal{S}_1}(\theta) - L_{\mathcal{S}_2}(\theta)| > \varepsilon \iff \exists f \in \mathcal{F}_{\mathcal{S}_1,\mathcal{S}_2}, |\mathbb{E}_1[f] - \mathbb{E}_2[f]| > \varepsilon$

**Lemma 4.**

$$\mathbb{P}_{\mathcal{S}_1}\left[\exists \theta \in \Theta, \frac{1}{m}|\hat{L}_{\mathcal{S}_1}(\theta) - L(\theta)| > \varepsilon\right] \leq 2\mathbb{P}_{\mathcal{S}_1,\mathcal{S}_2}\left[\exists \theta \in \Theta, \frac{1}{m}|L(\theta) - \hat{L}_{\mathcal{S}_1}(\theta)|\right]$$

$$\begin{aligned}
&\mathbb{P}\left[\exists \theta \in \Theta, |L(\theta) - \hat{L}_{\text{DPP}}(\theta)| > \varepsilon L(\theta)\right] \\
\leq &\mathbb{P}\left[\exists \theta \in \Theta^*, |L(\theta) - \hat{L}_{\text{DPP}}(\theta)| > \frac{\varepsilon}{2} L(\theta)\right] \\
\leq &\sum_{\theta\in\Theta^*} \frac{4}{\varepsilon^2}\,\mathbb{V}\text{ar}[\frac{\hat{L}(\theta)}{L(\theta)}] \\
\lesssim &\sum_{\theta\in\Theta^*} \frac{1}{\varepsilon^2}(M_\theta \mathcal{O}(m^{-(1+\frac{1}{d})}) + \mathcal{O}(n^{-1/2})) \\
\leq &\frac{|\Theta^*|}{\varepsilon^2}(M\mathcal{O}(m^{-(1+\frac{1}{d})}) + \mathcal{O}(n^{-1/2}))
\end{aligned}$$

For the $\varepsilon$-coreset property **??** to hold $1-\delta$-surely or equivalently

$$\mathbb{P}\left[\exists \theta \in \Theta, |L(\theta) - \hat{L}_{\text{DPP}}(\theta)| > \varepsilon L(\theta)\right] \leq \delta \tag{25}$$

it suffices to have

$$\delta \geq \frac{|\Theta^*|}{\varepsilon^2}(M\mathcal{O}(m^{-(1+\frac{1}{d})}) + \mathcal{O}(n^{-1/2})) \tag{26}$$

$$\iff m^{1+\frac{1}{d}} \gtrsim \frac{M}{\frac{\delta\varepsilon^2}{|\Theta^*|} + \mathcal{O}(n^{-1/2})} = \frac{M|\Theta^*|}{\delta\varepsilon^2}\frac{1}{1 + \frac{|\Theta^*|}{\delta\varepsilon^2}\mathcal{O}(n^{-1/2})} \tag{27}$$

This means that for sufficiently large $n$ (potentially $n \gtrsim \delta^{-2}\varepsilon^{-4}|\Theta^*|^2$), we can control the second factor and thus obtain the bound

$$\boxed{m \gtrsim \left(\frac{M|\Theta^*|}{\delta\varepsilon^2}\right)^{\frac{1}{1+\frac{1}{d}}}} \tag{28}$$

Assume better variance with DPP, can we improve concentration?

- Can we use the $\sqrt{N^{1+\frac{1}{d}}}$ rate from the SGD paper?
- Concentration inequality for a sum of **dependant** variables?

**Theorem 3.4.** from **pemantle2011rayleighconcentration**: Let $\mathbb{P}$ be a k-homogeneous probability measure on $\mathcal{B}_n$ satisfying the Stochastic Covering Property (SCP). Let $f$ be a 1-Lipschitz function on $\mathcal{B}_n$. Then

$$\mathbb{P}(|f - \mathbb{E}f| \geq a) \leq 2\exp\left(-\frac{a^2}{8k}\right)$$

Bennett inequality: Let be $(X_i)_{i\in[\![1,n]\!]}$ independent and centered real-valued random variables, and $\sigma^2 = \frac{1}{n}\sum_i \mathbb{V}\text{ar}[X_i]$, then for any $t > 0$

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > t\right) \leq \exp\left(-n\sigma^2 h\left(\frac{t}{n\sigma^2}\right)\right)$$

where $h(u) = (1+u)\log(1+u) - u$ for $u \geq 0$.

From **breuer2013nevai**

$$\mathbb{P}\left(|X_f - \mathbb{E}X_f| \geq \varepsilon\right) \leq \begin{cases} 2\exp\left(-\frac{\varepsilon^2}{4A\,\mathbb{V}\mathrm{ar}\,X_f}\right), & \text{if } \varepsilon < \frac{2A\,\mathbb{V}\mathrm{ar}\,X_f}{3\|f\|_\infty} \\ 2\exp\left(-\frac{\varepsilon}{6\|f\|_\infty}\right), & \text{if } \varepsilon \geq \frac{2A\,\mathbb{V}\mathrm{ar}\,X_f}{3\|f\|_\infty} \end{cases}$$

$$P(|L - \hat{L}| \geq L\varepsilon) \leq 2e^{\frac{-L^2\varepsilon^2}{4A\,\mathbb{V}\mathrm{ar}}}$$

$$m \geq \left(\frac{c}{L^2\varepsilon^2}\log\left(\frac{2}{\delta}\right)\left(1 + O\left(n^{-\frac{1}{2}}\right)\right)\right)^{\frac{1}{1+\frac{1}{d}}}$$

## 13 Discrete OPE

> Can we bypass the Kernel Density Estimate (KDE) in SGD paper by using discrete OPE? See **gautschi2004ope**.

## 14 Holydays questions

- Variance for formula for k-DPP, in **zhang2017dppminibatch**.
- How $\tilde{K}$ eigenspaces look like ? When $n \to \infty$ ?
  - How does it compare to **bardenet2020mcdpp** ?
  - If $f$ is given, can I find a $K$ for which $f$ is in "good" eigenspaces (eigenvalue $\geq 1$).
- Defining discrete OPE, because discretized continuous OPE is probably not a DPP. See Gautschi Orthogonal Polynomials, 2004.
  - For making links with SGD paper **bardenet2021sgddpp**
  - Look at the limit e.g. for Jacobi ensembles.

- Take a Bernoulli and beat it with a DPP.
- Focus on metric we could have advantages on, e.g. look how variance decay with coreset size.
- Better with direct applications e.g. on k-means or linear regression.

- Strong raylegh measure positive correlation ?
- $\mathbb{V}\mathrm{ar}\, f = \mathbb{V}\mathrm{ar}\, f_+ + \mathbb{V}\mathrm{ar}\, f_- + 2(\mathbb{E}f_+f_- - \mathbb{E}f_+\mathbb{E}f_-) \leq \mathbb{V}\mathrm{ar}\, f_+ + \mathbb{V}\mathrm{ar}\, f_-$
- Pemantle only for kDPP?
- Hoefding/Benett like proof for $\sum_{x_i \in \mathcal{S}} \varepsilon_i f(x_i)$ using maybe recurisve property of HKPV sampling.