IMPROVING SAMPLE COMPLEXITY OF CORESETS WITH DETERMINANTAL POINT PROCESSES

By

Hugo Simon-Onfroy

A Master MVA internship report

under the supervision of

Rémi Bardenet — CNRS & Université de Lille
Subhroshekhar Ghosh — National University of Singapore

Abstract

The last several years have seen the rise of data sets of extraordinary scale across a variety of scientific fields. Such massive data sets bring new challenges, because of the high cost of collecting, maintaining, and interpreting them. Existing well-proven methods can become computationally impractical when dealing with millions or even billions of data points, and when they may no longer fit on a single system but instead need to be stored on clusters of servers. New algorithms are therefore required to scale to this massive data setting.

While one could concentrate on specific machine learning issues and develop many new algorithms, we concentrated in this internship on a more comprehensive strategy. We investigate coresets, which are compressed descriptions of potentially enormous data sets, and that still guarantee provable computational properties.

Thanks to the careful selection of sample distribution, i.i.d. random sampling has proven to be one of the most effective techniques for creating coresets. However, independent samples can sometimes be extremely redundant, and one could hope that enforcing diversity would yield better performances.

Because proving coreset property in a general non-i.i.d. framework is difficult, we rely on determinantal point processes (DPPs), a class of point processes that naturally introduce diversity through repulsiveness. DPPs are interesting because they are a rare example of repulsive point processes with tractable theoretical properties, while still being sufficiently expressive to address a variety of problems.

This report is an attempt to provide improved sample complexity results on coresets by the use of well chosen DPP sampling. We start with an inventory on coreset literature, then we bring to light determinantal point processes as an improvement way. Thirdly, we qualitatively justify the use of DPPs over existing sampling methods. We finish by quantitative results, conclusion, and perspectives.

Contents

Chapter 1

Introduction to coresets

1.1 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 $\mathcal{F}$ be a space of functions called queries defined on $\mathcal{X}$, and f an element of $\mathcal{F}$. Classical learning problem aims to find a solution f^* in $\mathcal{F}$ that minimizes a cost function L over the given data $\mathcal{X}$. In this work, we focus on cost functions that are positive and additively decomposable, i.e. we consider cost functions of the form

$$L(f) := \sum_{x \in \mathcal{X}} f(x), \tag{1.1}$$

where the queries $f \in \mathcal{F} \subseteq \mathbb{R}_+^{\mathcal{X}}$ are positive functions defined on $\mathcal{X}$.

A large amount of machine learning problems falls into the framework eq. (1.1), including support vector machines, logistic regression, linear regression and k-means clustering.

Example 1 (k-means)**.** The goal of Euclidean k-means clustering is to find a set of k cluster centers $\mathcal{C} \subseteq \mathbb{R}^d$ minimizing the quantization error

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

In this case, f is the squared distance to the nearest cluster center q in a set of cluster centers $\mathcal{C}$. Formally,

$$\mathcal{F} = \left\{ f : x \mapsto \min_{q \in \mathcal{C}} \|x - q\|_2^2 \mid \mathcal{C} \in \binom{\mathcal{X}}{k} \right\}$$

where $\binom{\mathcal{X}}{k}$ denotes "from $\mathcal{X}$ choose k", the set of all subsets of $\mathcal{X}$ of size k.

Example 2 (linear regression)**.** The goal of ordinary least squares linear regression is to find a vector $a \in \mathbb{R}^d$ and a scalar $b \in \mathbb{R}$ that minimizes the sum of squares

$$L(f) = \sum_{(y,z) \in \mathcal{X}} (a^\top y + b - z)^2$$

where the data is $\mathcal{X} = \{x_i := (y_i, z_i) \mid i \in [\![1, n]\!]\}$ with for all $i \in [\![1, n]\!]$, $y_i \in \mathbb{R}^d$ and $z_i \in \mathbb{R}$. In this case, we have

$$\mathcal{F} = \left\{(y, z) \mapsto (a^\top y + b - z)^2 \mid a \in \mathbb{R}^d, b \in \mathbb{R}\right\}$$

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 many 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 coresets in both the streaming and offline settings. We will focus here on the offline setting.

1.2 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 $\mathcal{X}$, and to any element $x \in \mathcal{S}$, associate a weight $\omega(x) \in \mathbb{R}^+$ that only depends on the value $x \in \mathcal{X}$. Define the estimated cost based on the weighted multiset $\mathcal{S}$ as

$$\hat{L}_{\mathcal{S}}(f) := \sum_{x \in \mathcal{S}} \omega(x) f(x).$$

The aim of this estimator is "to approximate" $L(f)$ defined in eq. (1.1). Depending on the context, there are plenty of meaning "to approximate" can get. We focus especially on the coreset property.

Definition 1 (Coreset)**.** Let $\varepsilon \in]0, 1]$ and $f \in \mathcal{F}$. We say $\mathcal{S}$ is an ε-coreset for f if the estimated cost based on $\mathcal{S}$ is equal to the exact cost up to a relative error ε. Formally

$$\left|\frac{\hat{L}_{\mathcal{S}}(f)}{L(f)} - 1\right| \leq \varepsilon. \tag{1.2}$$

We say $\mathcal{S}$ is an ε-coreset for $\mathcal{F}$ if it is a ε-coreset for any $f \in \mathcal{F}$. Formally

$$\forall f \in \mathcal{F}, \left|\frac{\hat{L}_{\mathcal{S}}(f)}{L(f)} - 1\right| \leq \varepsilon. \tag{1.3}$$

An important consequence of the coreset property is the following

Theorem 1. *Let $\mathcal{S}$ be an ε-coreset for $\mathcal{F}$. Define $f^* := \min_{f\in\mathcal{F}} L(f)$ and $\hat{f}^* := \min_{f\in\mathcal{F}} \hat{L}_{\mathcal{S}}(f)$. Then $L(\hat{f}^*)$ is an $(1+3\varepsilon)$-approximation of $L(f^*)$, i.e.*

$$L(f^*) \leq L(\hat{f}^*) \leq (1+3\varepsilon)L(f^*).$$

Proof. $\mathcal{S}$ being a ε-coreset for $\mathcal{F}$ yields eq. (1.3), which is equivalent to

$$\forall f \in \mathcal{F},\ (1-\varepsilon)L(f) \leq \hat{L}_{\mathcal{S}}(f) \leq (1+\varepsilon)L(f).$$

In particular, this is true for f^* and $\hat{f}^*$ thus

$$(1-\varepsilon)L(f^*) \leq (1-\varepsilon)L(\hat{f}^*) \leq \hat{L}_{\mathcal{S}}(\hat{f}^*) \leq \hat{L}_{\mathcal{S}}(f^*) \leq (1+\varepsilon)L(f^*), \tag{1.4}$$

and moreover

$$L(f^*) \leq L(\hat{f}^*) \leq \frac{(1+\varepsilon)}{(1-\varepsilon)} L(f^*) \leq (1+3\varepsilon)L(f^*).$$

□

A key consequence of that theorem is that one can minimize on the estimated loss $\hat{L}_{\mathcal{S}}$ and still guarantee a low error on the true loss, even when the size of $\mathcal{S}$ is small before the size of $\mathcal{X}$. Therefore, it makes coreset very relevant in a machine learning context, and inscribes them into a more general learning framework that is PAC learning.

1.3 Coresets and 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 Valiant 1984. 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.

1. The richness of the considered class of prediction function can be too small to embrace the complexity of the studied phenomena.
2. The risk optimizer algorithm could struggle finding the minimizing function, for instance only finding local minima, or yielding high computational complexity.
3. 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.

The use of coresets tackles the last two issues. It aims to reduce the number of samples required to compute an optimal prediction function, and still control the error of the op-

timization step. The computational complexity of an optimizer being often linked to the number of samples, it is by the way reduced. In general, if the time complexity for an optimization algorithm to optimize on n data points is $\mathcal{O}(a_n)$, and that it takes $\mathcal{O}(b_m)$ time to sample an ε-coreset which is of size $m \leq n$, then we have interest in building coreset as soon as $\mathcal{O}(a_n) \geq \mathcal{O}(b_m) + \mathcal{O}(a_m)$.

1.4 State-of-the-art results on coresets

The existence of coresets is trivial, the original data set $\mathcal{X}$ itself being a 0-coreset, taking all its elements weighted by 1. The key question is the existence of small coresets where the coreset size is sublinear, if not independent, in the number of data points n, while at the same time having slow rate with respect to other parameters, in particular d the dimension of data, ε the desired error, and in the case where the coreset is obtained probabilistically, δ the probability bound of not being a coreset.

1.4.1 Importance multinomial sampling

In the stochastic case, a well-established approach to coreset construction is importance multinomial sampling. Given any distribution q on $\mathcal{X}$, one can sample a sequence $\mathcal{S} \in \mathcal{X}^m$ from the multinomial distribution of size m based on q, $\mathcal{S} \sim \mathcal{M}(m, q)$, i.e. m i.i.d. sampling of X such that $\forall x \in \mathcal{X},\ \mathbb{P}[X = x] = q(x)$.

By the importance sampling trick, an unbiased estimator of $L(f)$ is then

$$\hat{L}_{\text{iid}}(f) := \sum_{x \in \mathcal{S}} \frac{f(x)}{mq(x)}.$$

And its variance is

$$\mathbb{V}\text{ar}_{\text{iid}}[f] := \frac{1}{m}\,\mathbb{V}\text{ar}\left[\frac{f(x)}{q(x)}\right] = \frac{1}{m}\sum_{x \in \mathcal{X}} \frac{f(x)^2}{q(x)} - \frac{1}{m}L(f)^2$$

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

Now observe that for any query $f \in \mathcal{F}$, the variance is reduced to 0 by taking

$$q_f := \frac{f}{L(f)} = x \mapsto \frac{f(x)}{L(f)}.$$

Of course, attempting to sample from q_f is quiet limited in practice. First, we would prefer not having to make our sampling depend on the query function f. Second and main obstacle is that using q_f implies already knowing $L(f)$, for which we are supposedly looking an approximation for via building coreset. In section 1.4.2 we see one way to bypass these two limitations.

1.4.2 Sensitivity sampling

Intuitively, in order to build a coreset of small size, we want to only select data points that are relevant. This means that for a given $x \in \mathcal{X}$, we want to make its probability to be sampled as small as possible, unless it plays a relevant role in the evaluation of $L(f)$ for some f, which translates to $\frac{f}{L(f)}$ being high.

The idea of Langberg and Schulman 2010 is thus to take for every x, the sampling probability q_f in the worst case f, i.e. for which x is the most relevant in the evaluation of $L(f)$. Formally, they define the following notion of sensitivity.

Definition 2 (Sensitivity)**.** The sensitivity $\sigma(x)$ of a data point $x \in \mathcal{X}$ with respect to $\mathcal{F}$ is defined as

$$\sigma(x) = \sup_{f \in \mathcal{F}} \frac{f(x)}{L(f)} \in [0, 1].$$

and the total sensitivity with respect to $\mathcal{F}$ as $\mathfrak{S} = \sum_{x \in \mathcal{X}} \sigma(x) \in [1, n]$.

If we were now too sample x from a distribution proportional to the sensitivity $\mathbb{P}[X = x] \propto \sigma(x)$, it would not depend on f. However, we would still need to know $L(f)$ for every f, in order to compute it.

But assume that we know an upper bound s on sensitivity σ i.e. $\forall x \in \mathcal{X}, s(x) \geq \sigma(x)$, and define $S := \sum_{x \in \mathcal{X}} s(x) \geq \mathfrak{S}$. We can then sample x from a distribution proportional to s, i.e. $\mathcal{S} \sim \mathcal{M}(m, s/S)$, and we have the following result.

Theorem 2 (Hoeffding bound for fixed query)**.** *Let $m \in \mathbb{N}$ and $\mathcal{S} \sim \mathcal{M}(m, s/S)$. Then for all $\varepsilon > 0$ and all $f \in \mathcal{F}$*

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

Moreover, for all $\delta > 0$

$$m \geq \frac{S^2}{2\varepsilon^2} \log\frac{2}{\delta} \implies \mathcal{S} \text{ is an } \varepsilon\text{-coreset for } f \text{ w.p. } 1 - \delta.$$

Proof. Define $g_s(x) := \frac{f(x)}{s(x)L(f)} \in [0, 1]$. Because of boundedness, we can apply Hoeffding inequality, and have for any $f \in \mathcal{F}$ and $\varepsilon > 0$

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{x \in \mathcal{S}} g_s(x) - \mathbb{E}\left[g_s(x)\right]\right| > \varepsilon/S\right] \leq 2\exp\left(-2m\varepsilon^2/S^2\right).$$

Furthermore, $\mathbb{E}[g_s(x)] = \frac{1}{S}$ and $\frac{1}{m}\sum_{x \in \mathcal{S}} g_s(x) = \frac{\hat{L}_{\text{iid}}(f)}{SL(f)}$, thus

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

Hence, $\mathcal{S}$ satisfies eq. (1.2) i.e. $\mathcal{S}$ is an ε-coreset for f with probability at least $1 - \delta$, if we choose

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

□

The required number of samples depends quadratically on S the upper bound on total sensitivity $\mathfrak{S}$. Hence, the tighter the bound on the sensitivity $s \geq \sigma$ is, the less samples are required. If one take the looser bound $s = 1$, so $S = n$, this implies $m \gtrsim n^2$ which is useless in practice.

Note that total sensitivity $\mathfrak{S} \in [1, n]$ depends on the richness of function space $\mathcal{F}$.

- $\mathfrak{S} = 1$ if and only if for all $x \in \mathcal{X}$, $\sigma(x) = 1/n$. This means $\mathcal{F}$ only contains constant functions.
- $\mathfrak{S} = n$ if and only if for each $x \in \mathcal{X}$, there exists some $f \in \mathcal{F}$ such that for any other element y of $\mathcal{X}$, $f(y) = 0$.

Fortunately, it is possible to compute tight bounds on the sensitivity for many machine learning problems. For instance, Lucic, Bachem, and Krause 2016 show that the sensitivity bound for k-means problem is $\mathfrak{S} = \Theta(k)$. When $k = n - 1$, the query set $\mathcal{F}$ contains all functions that are zero on $n - 1$ points, taking $n - 1$ points to be centers of their own cluster, and we recover the worst case $S = n$.

1.4.3 Extension to all queries

In previous section, we obtained a sample complexity bound on obtaining coreset for a single query f. But in order to obtain an ε-coreset for $\mathcal{F}$, the ε-coreset for f must holds simultaneously for all queries $f \in \mathcal{F}$.

Intuitively, we would like to invoke a union bound argument/Bonferroni correction over all queries $f \in \mathcal{F}$. In the case where $\mathcal{F}$ is finite, applying union bound on the previous bound from theorem 2 yields

$$\mathbb{P}\left[|\hat{L}_{\text{iid}}(f) - L(f)| > \varepsilon L(f)\right] \leq 2|\mathcal{F}| \exp\left(-2m\varepsilon^2/S^2\right).$$

Hence, for all $\delta > 0$

$$m \geq \frac{S^2}{2\varepsilon^2} \log \frac{2|\mathcal{F}|}{\delta} \implies \mathcal{S} \text{ is an } \varepsilon\text{-coreset for } f \text{ w.p. } 1 - \delta$$

However, in the case $\mathcal{F}$ is not finite, this bound diverge. We give here the intuition how a finite bound can still be obtained and we will detail this method in section 4.4.

The key idea is to construct a set $\mathcal{F}'_\varepsilon \subseteq \mathcal{F}$ which still is finite, and that approximate the set $\mathcal{F}$ with ε granularity. It would further imply that if $\mathcal{S}$ is an ε-coreset for $\mathcal{F}$, then it would be an ε'-coreset for $\mathcal{F}'_\varepsilon$, for controlled ε', and then one could apply union bound on $\mathcal{F}'_\varepsilon$ instead.

This would summon the size of $\mathcal{F}'_\varepsilon$ into the obtained bound, which depends on the richness of the set $\mathcal{F}$ it approximate. Richness in that case can be quantified through pseudo-dimension, which generalizes the VC-dimension to real-valued function sets.

Definition 3 (pseudo-dimension)**.** The pseudo-dimension of a set $\mathcal{H}$ of functions defined on $\mathcal{X}$, denoted by $\operatorname{pdim}\mathcal{H}$, is the largest d' such that

- there exists $(x_i)_{i\in[\![1,d']\!]} \subseteq \mathcal{X}^{d'}$, a sequence of d' elements from $\mathcal{X}$,
- there exists $(t_i)_{i\in[\![1,d']\!]} \subseteq \mathbb{R}^{d'}$ a sequence of d' real thresholds,
- such that for each $(b_i)_{i\in[\![1,d']\!]} \subseteq \{0,1\}^{d'}$
- there is an $f \in \mathcal{H}$ such that $\forall i \in [\![1,d']\!]$, we have $f(x_i) \geq r_i \iff b_i = 1$.

Put differently it always exists functions in $\mathcal{H}$ to have values above or below some threshold for every $2^{d'}$ combinations of above/below.

Pseudo-dimension can also be defined through VC-dimension. Indeed, considering the function

$$\begin{aligned}\operatorname{above}_f \colon \mathcal{X}\times\mathbb{R} &\to \{0,1\}\\ (x,r) &\mapsto \mathbb{1}\{f(x)\geq r\}\end{aligned}$$

we have

$$\operatorname{pdim}\mathcal{H} := \operatorname{VCdim}\{\operatorname{above}_f \mid f\in\mathcal{H}\} \tag{1.5}$$

Let now introduce the function space

$$\mathcal{G}_s := \frac{\mathcal{F}}{sL(\mathcal{F})} = \left\{x\mapsto \frac{f(x)}{s(x)L(f)} \mid f\in\mathcal{F}\right\} \subseteq [0,1]^{\mathcal{X}}. \tag{1.6}$$

It can be shown the size of previously mentioned $\mathcal{F}'_\varepsilon$ is $\mathcal{O}(\varepsilon^{-d'})$ where $d' := \operatorname{pdim}\mathcal{G}_s$. Applying union bound then leads to

$$m \gtrsim \frac{S^2}{\varepsilon^2}(d'\log\frac{1}{\varepsilon} + \log\frac{1}{\delta})$$

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.

This result presented here follows from seminal works on theoretic generalizations of the PAC model from Haussler 1992. Improving that scheme, more recent results from Li, Long, and Srinivasan 2001 involving chaining arguments leads to

$$m \gtrsim \frac{S^2}{\varepsilon^2}(d' + \log\frac{1}{\delta}).$$

Moreover, they shown this bound to be tight in the i.i.d. sampling framework, with respect to ε and δ. Finally, Braverman, Feldman, Lang, Statman, and Zhou 2016 improved it with respect to S by showing that under the same framework

$$m \gtrsim \frac{S}{\varepsilon^2}(d'\log S + \log\frac{2}{\delta}). \tag{1.7}$$

As an example, we saw that in the k-means case, $\mathfrak{S} = \Omega(k)$, and it can be shown that for d-dimensional data, $d' = \mathcal{O}(dk\log k)$, which leads to

$$m \gtrsim \frac{k}{\varepsilon^2}(dk\log^2 k + \log\frac{2}{\delta}).$$

We refer to Bachem, Lucic, and Krause 2017 for further insights on sensitivity and pseudo-dimension bounding methods.

Chapter 2

Determinantal Point Processes

We saw in chapter 1 the uniform bound on sample complexity of coresets. These bounds are based on Probabilistically Approximately Correct (PAC) learning theory results, where sample complexity bounds (with respect to δ and ε) are known and tight in the i.i.d. framework. One could then be tempted to extend to the case where samples are drawn dependently.

Whereas the space of distributions for m i.i.d samples has size that does not depend on m, the way m samples can be correlated grows exponentially with m. In order to tackle this space of correlations, more recent results restricted to the cases of martingales or β-mixing processes, e.g. Gao, Niu, and Zhou 2016.

In the current chapter, we introduce Determinantal Point Processes (DPPs), another restriction of correlated sampling, which admits useful tractability properties, while still maintaining expressiveness into the sub-category of negatively correlated sampling. This negative correlation is expected to be a key property in order to perform better sample complexity.

2.1 Some intuition

A Determinantal Point Processes (DPP) is a random sampling over subsets of a given ground set. Noticeably, its distribution is entirely encoded by a given positive kernel, which can be tuned to a range of specific contexts. In a sense, DPPs can be said to be the kernel machine of point processes, as they allow both tractability and flexibility.

An 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.

2.2 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 $\mathbb{P}[\{x_1, x_2, x_3\}]$ 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 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. We refer to Hough, Krishnapur, Peres, and Virág 2006 for a review of DPPs in the continuous case.

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 4 (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 $A \subseteq \mathcal{X}$,

$$\mathbb{P}[A \subseteq \mathcal{S}] = \det K_A \tag{2.1}$$

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

Here, $K_A := [K_{xy}]_{x,y \in A}$ denotes the submatrix of K indexed by elements of A, and we adopt $\det K_\emptyset = 1$. 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 $\det K_A$ of K must be positives, and thus K itself must be positive (for Loewner order $\preceq$). 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 from Macchi 1975,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 4. If $A = \{x\}$ is a singleton, then we have

$$\mathbb{P}[x \in \mathcal{S}] = K_{xx} \tag{2.2}$$

also denoted $K(x, x)$. 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 $A = \{x, y\}$ is a two-element set, then

$$\begin{aligned} \mathbb{P}[\{x, y\} \subseteq \boldsymbol{X}] &= \begin{vmatrix} K_{xx} & K_{xy} \\ K_{yx} & K_{yy} \end{vmatrix} \\ &= K_{xx}K_{yy} - K_{xy}K_{yx} \\ &= \mathbb{P}[x \in \boldsymbol{X}]\,\mathbb{P}[y \in \boldsymbol{X}] - K_{xy}^2 \end{aligned} \tag{2.3}$$

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

eq. (2.3) demonstrates why DPPs 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 $K_{xy} = \sqrt{K_{xx}K_{yy}}$, 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 DPPs cannot represent distributions where elements are more likely to co-occur than if they were independent: correlations are always negative.

A DPP of kernel K can have a random sample size. From eq. (2.2), we know that the average number of samples is equal to the trace of K, Formally

$$\mathbb{E}\left[|\mathcal{S}|\right] = \mathbb{E}\left[\sum_{x\in\mathcal{X}} \mathbb{1}\{x \in \mathcal{S}\}\right] = \sum_{x\in\mathcal{X}} \mathbb{P}\left[x \in \mathcal{S}\right] = \sum_{x\in\mathcal{X}} K_{xx} = \operatorname{Tr} K.$$

But in many cases, one prefers to specify deterministically the number of samples, instead of having a random number of them. This leads to the definition of m-DPP.

Definition 5 (m-DPP)**.** A m-DPP is a DPP conditioned to a fixed sample size m. Therefore, it is a probability distribution supported on $\binom{\mathcal{X}}{m}$ only.

Finally, we introduce a subclass of DPPs called projective DPPs that admits useful properties.

Definition 6 (Projective DPP)**.** A DPP is projective if every eigenvalue of its kernel is in $\{0, 1\}$. This is equivalent to the kernel being a projection matrix.

Projective DPPs are sometimes called elementary DPPs, because it turns out any DPP can be written as a mixture of projective DPP, see Lemma 2.6. from Kulesza and Taskar 2012.

For a DPP of kernel K, being a projective DPP is equivalent to having deterministic sample size. We know this size is equal to $\operatorname{Tr} K$ which in that case is by definition the rank of K. As a corollary, the set of projective DPPs of rank m are precisely the intersection of the set of DPPs and the set of m-DPPs.

2.3 Examples

DPPs 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, DPPs 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.

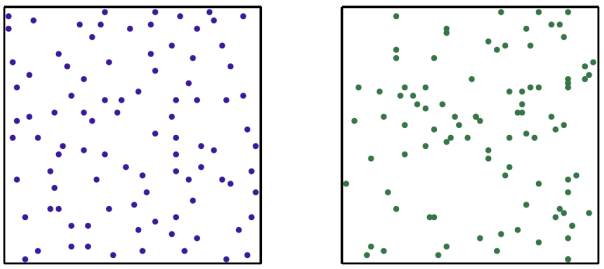

Figure 2.1: (left) A set of points in the plane drawn from a DPP, with K_{xy} inversely related to the distance between points i and j. (right) The same number of points sampled independently using a Poisson point process , which results in random clumping.

2.4 Geometric interpretation

DPPs are defined on determinants, that have an intuitive geometric interpretation. Since a DPP kernel K is symmetric, there exists $r \leq n$ and $V \in \mathbb{R}^{r \times n}$ such that $K = V^\top V$. Denote the columns of V by (V_x) for $x \in \mathcal{X}$. Then $\forall A \subseteq \mathcal{X}$

$$\mathbb{P}\left[A \subseteq \mathcal{S}\right] = \mathrm{Vol}^2(V_A) \tag{2.4}$$

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

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 definition 4 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 fig. 2.2.

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.

2.5 Sampling from a DPP

2.5.1 Exact DPP sampling

Although DPPs 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 Hough et al. 2006.

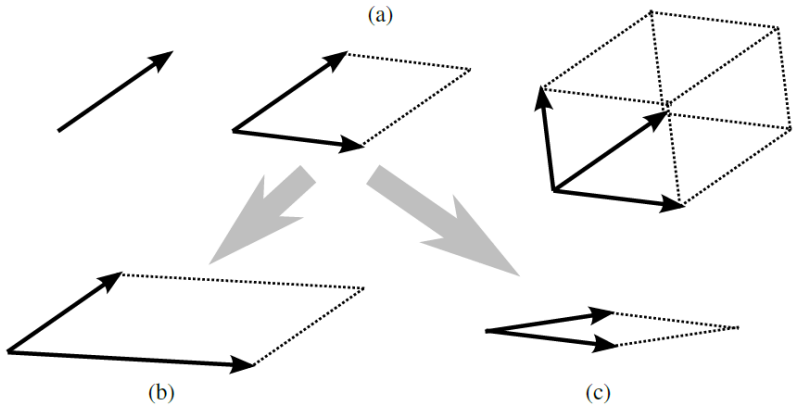

Figure 2.2: from Kulesza and Taskar 2012. A geometric interpretation of a DPP relates each column of V to an element of $\mathcal{X}$. (a) The probability of inclusion of a subset A is the square of the volume spanned by its associated feature vectors. (b) As the magnitude of an item's feature vector 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.

It leverages the fact that DPPs are mixtures of projective DPPs 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 $V \in \mathbb{R}^{r \times n}$. This can often be the computational bottleneck since it generally requires $O(n^3)$ time. Note however that for some DPPs 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 $O(nm^2)$ time, where m is the number of elements sampled. This means $O(n(\operatorname{Tr} K)^2)$ time on average. If K is a projective kernel, $m = \operatorname{Tr} K = r$ 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 Gillenwater, Kulesza, Mariet, and Vassilvtiskii 2019 improved somewhat this complexity. Based on the still needed eigendecomposition, it implements a binary tree structure storing appropriate summary statistics of the eigenvectors, requiring $O(nr^2)$ to build, but can then generate repeated samples in $O(\log(n)m^2r^2 + r^3)$ time, hence $O(\log(n)r^4)$ for a projective kernel. Therefore, this method becomes a viable alternative to the spectral method when the total number of items n is large and when the dimensionality r of the features and the expected sample size $\operatorname{Tr} K$ are small compared to n.

2.5.2 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 a given DPP kernel, using random projections such as in Kulesza et al. 2012, or low-rank factorization techniques.

A second class involves Monte Carlo Markov Chain (MCMC). This is often down in an inexact fashion using target distribution close but different from a DPP one. Noticeably, Gautier, Bardenet, and Valko 2017 proposed an exact MCMC sampler for projective DPPs.

Chapter 3

Correlated importance sampling

We saw in chapter 2 that DPPs are a restriction of correlated sampling that admits useful tractability properties. Moreover, DPPs still maintain expressiveness into the sub-category of negatively correlated sampling, which is the kind of processes we expect to perform better for sample complexity. The intuition is that negatively correlated sampling can eliminate redundancy in sampling sets, an independent sampling can not.

In this chapter, we present current results on coreset sampling with DPPs, and show qualitative results on variance reduction from DPPs.

3.1 A first result with DPPs

Tremblay, Barthelmé, and Amblard 2018 first introduce DPPs into the coreset problems, based on the idea of diversity sampling. Their results holds for both DPPs and m-DPPs. Since projection DPPs are precisely the intersection of both DPPs and m-DPPs, all results apply to them. For the sake of conciseness, we state here their result for m-DPPs and we refer to their article for the DPP case.

Theorem 3 (Tremblay et al. 2018)**.** *Let $m \in \mathbb{N}$, K_m a m-DPP kernel and let sample $\mathcal{S} \sim \mathcal{DPP}(K_m)$. Assume that the query space $\mathcal{F}$ is parametrized by some $\theta \in \Theta$, and that all Lipschitz constant with respect to θ of $f_\theta \in \mathcal{F}$ are bounded by some $\boldsymbol{\ell} := \sup_{x\in\mathcal{X}} \operatorname{Lip}\{\theta \mapsto f_\theta(x)\}$.*

If the minimal sensitivity satisfies $\min_{x\in\mathcal{X}} \sigma(x) \geq 1/n$, then for all $\varepsilon, \delta \in]0,1]$

$$m \geq \frac{32}{\varepsilon^2}\left(\max_{x\in\mathcal{X}} \frac{m\sigma(x)}{K_m(x,x)}\right)^2 \log\frac{4\eta}{\delta} \implies \mathcal{S} \text{ is an } \varepsilon\text{-coreset for } \mathcal{F} \text{ w.p. } 1-\delta$$

where η is the minimal number of balls of radius $\frac{\varepsilon \inf_f L(f)}{6n\boldsymbol{\ell}}$ necessary to cover Θ.

Note first that the fraction $\frac{m}{K_m(x,x)}$ appearing in the right hand side of the bound is due to the correlated importance sampling framework. This fraction does not appear in the i.i.d. framework because the numerator m cancel with the marginal intensity $mq(x)$. In practice, this fraction can be bounded uniformly on m, because $K_m(x,x)$ would typically grow linearly with m.

Also note that typically $\log \eta = \mathcal{O}\left(d' \log \frac{n}{\varepsilon \inf_f L(f)}\right)$ with $d' = \operatorname{pdim} \mathcal{F}$, and therefore depends on n and ε.

Thus, the obtained bound for DPPs of theorem 3 does not improve the sample complexity bound on coreset size in i.i.d. framework, from Braverman et al. 2016. There are two reasons for this.

- First, the result crucially relies on a concentration inequality for strongly Rayleigh measures (especially DPPs) from Pemantle and Peres 2011, which does not improve Hoeffding bound used in chapter 1.

  However, one important fact is that it doesn't rely on more advanced concentration for projective DPPs from Breuer and Duits 2013 that involves the variance of the estimator. Since recent results from Bardenet and Hardy 2020, it is known DPPs can improve variance rate, and we hope this result to be leveraged into an improved bound on coreset for fixed query.
- Second, the argument to generalize to all queries from Tremblay et al. 2018 introduce a $\log \varepsilon^{-1}$ term, and foremost, a dependency in n through η. If not tackled, this could ruin the effort finding improved bound for fixed queries. An improvement way would be to extend classical VC theory arguments in a correlated context.

Despite these mitigated results on concentrations, DPPs has already been shown to perform variance reduction, e.g. Bardenet et al. 2020. In the following section 3.2, we present qualitative variance reductions in favour of DPP and m-DPP sampling, against Bernoulli process sampling and multinomial sampling.

3.2 Variance arguments

We express variance formulas in four sampling cases: multinomial, DPP, Bernoulli process, and m-DPP. Then we compare these variances under a domination criteria.

3.2.1 Four sampling cases

In the multinomial case, we have $\mathcal{S} \sim \mathcal{M}(m, q)$. Then an unbiased estimator of L is

$$\hat{L}_{\text{iid}}(f) := \sum_{x \in \mathcal{S}} \frac{f(x)}{m q(x)}$$

and its variance is

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

where $\boldsymbol{f} := (f(x))_{x \in \mathcal{X}}$, $Q := \operatorname{diag}(q)$ and $\boldsymbol{J} := \boldsymbol{j}\boldsymbol{j}^\top$ the matrix full of ones.

In the DPP case, we have $\mathcal{S} \sim \mathcal{DPP}(K)$, and for all $x \in \mathcal{X}$, we denote its marginals $\pi_x := K_{xx}$. Then an unbiased estimator of L is

$$\hat{L}_{\text{DPP}}(f) := \sum_{x \in \mathcal{S}} \frac{f(x)}{\pi_x}$$

Its variance can be computed using ε_x as the counting variable for x

$$\mathbb{V}\mathrm{ar}_{\mathrm{DPP}}[f] := \sum_{x,y\in\mathcal{X}} \mathbb{E}[\varepsilon_x\varepsilon_y] \frac{f(x)f(y)}{\pi_x\pi_y} - L(f)^2$$

$$\text{with} \quad \mathbb{E}[\varepsilon_x\varepsilon_y] = \begin{cases} \det K_{\{x,y\}} = \pi_x\pi_y - K_{xy}^2, & \text{if } x \neq y \\ \mathbb{E}[\varepsilon_x] = \pi_x, & \text{if } x = y \end{cases}$$

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

$$\mathbb{V}\mathrm{ar}_{\mathrm{DPP}}[f] = \sum_{x\in\mathcal{X}} \left(\frac{1}{\pi_x} - 1\right) f(x)^2 - \sum_{x\neq y} \frac{K_{xy}^2}{\pi_x\pi_y} f(x)f(y) = \boldsymbol{f}^\top(\Pi^{-1} - \tilde{K})\boldsymbol{f} \tag{3.1}$$

In the Bernoulli process case, where for all $x \in \mathcal{X}$, $\mathbb{P}[x \in \mathcal{S}] = \pi_x$ independently, we have a special case of DPP, where the kernel reduces to its diagonal, i.e. $K = \Pi$ and then $\tilde{K} = I$. We denote its variance $\mathbb{V}\mathrm{ar}_{\mathrm{diag}}[f] := \boldsymbol{f}^\top(\Pi^{-1} - I)\boldsymbol{f}$.

In the m-DPP case, we have $\mathcal{S} \sim \mathcal{DPP}(K) \mid |S| = m$, and we denote its marginals $b_x := \mathbb{E}[\varepsilon_i]$, that admit an analytic form one can find in Kulesza et al. 2012. Then an unbiased estimator of L is

$$\hat{L}_{\mathrm{mDPP}}(f) := \sum_{x\in\mathcal{S}} \frac{f(x)}{b_x}$$

and its variance is

$$\mathbb{V}\mathrm{ar}_{\mathrm{mDPP}}[f] := \sum_i \left(\frac{1}{b_x} - 1\right) f(x)^2 + \sum_{x\neq y} C_{xy} f(x)f(y) \tag{3.2}$$

where $C_{xy} := \frac{\mathbb{E}[(\varepsilon_x - b_y)(\varepsilon_y - b_y)]}{\mathbb{E}[\varepsilon_i]\mathbb{E}[\varepsilon_j]} = \frac{\mathbb{E}[\varepsilon_x\varepsilon_y]}{b_x b_y} - 1$

Observe that if the m-DPP kernel is reduced to its diagonal ($C_{xy} = 0$), we recover $\mathbb{V}\mathrm{ar}_{\mathrm{diag}}$, the variance of a Bernoulli process with same marginals ($\pi_x = b_x$), though the former has fixed sample size m, and the latter not.

In order to benefit from some variance reduction, one should find a m-DPP where $\forall x \neq y$, $C_{xy}f(x)f(y) < 0$.

Zhang, Kjellstrom, and Mandt 2017 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 ($C_{xy} < 0$) and their value have positive scalar product ($f(x)f(y) > 0$). This provides variance reduction.

Reversely, they argued that 2 dissimilar points should have positive correlation, and their value show "no tendency to align" hinting $f(x)f(y) < 0$, and again providing variance reduction. However, properties of strong Rayleigh measures implies always $C_{xy} \leq 0$ (see Pemantle et al. 2011). But we could more conservatively consider that, whether DPP or m-DPP, two dissimilar points tend toward independence. Thus the induced variance change, whether positive or negative depending on the sign of $f(x)f(y)$, would in either case be small.

3.2.2 Variance comparison

In the following, we compare processes with the same marginals, and therefore set $\Pi = mQ$. Also, since m-DPP marginals admits analytic but complicated form, we drop the m-DPP case comparison. We show in section 3.2.1 that $\mathbb{V}\mathrm{ar}_{\text{iid}}$, $\mathbb{V}\mathrm{ar}_{\text{diag}}$ and $\mathbb{V}\mathrm{ar}_{\text{DPP}}$ are quadratic forms of $\boldsymbol{f}$ associated with respective matrices

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

This allows to compare samplings through the Loewner ordering ($\preceq$) of the variance associated matrices. For instance, DPP variance strictly dominates Bernoulli process variance i.e. it uniformly yields lower variance, if and only if $\tilde{K}$ is strictly greater than identity. Formally

$$\forall f \in \mathbb{R}^{\mathcal{X}},\ \mathbb{V}\mathrm{ar}_{\text{DPP}}[f] < \mathbb{V}\mathrm{ar}_{\text{diag}}[f] \iff \tilde{K} \succ I.$$

Massaging some linear algebra thus gives

Proposition 1 (Variance comparison)**.**
DPP variance dominates Bernoulli process variance on positive-valued functions

$$\forall f \in \mathbb{R}_+^{\mathcal{X}},\ \mathbb{V}\mathrm{ar}_{DPP}[f] \leq \mathbb{V}\mathrm{ar}_{diag}[f]. \tag{3.3}$$

In the general case of real-valued functions, DPP variance does not dominate Bernoulli process variance but does up to a factor three

$$\exists f \in \mathbb{R}^{\mathcal{X}},\ \mathbb{V}\mathrm{ar}_{DPP}[f] \geq \mathbb{V}\mathrm{ar}_{diag}[f] \tag{3.4}$$

$$\forall f \in \mathbb{R}^{\mathcal{X}},\ \mathbb{V}\mathrm{ar}_{DPP}[f] \leq 3\,\mathbb{V}\mathrm{ar}_{diag}[f]. \tag{3.5}$$

Moreover, if the DPP is projective, then

$$\forall f \in \mathbb{R}^{\mathcal{X}},\ \mathbb{V}\mathrm{ar}_{DPP}[f] \leq \mathbb{V}\mathrm{ar}_{iid}[f]. \tag{3.6}$$

Proof of:

eq. (3.3) Assume $f \in \mathbb{R}_+^{\mathcal{X}}$. Then $\boldsymbol{f}^\top(\tilde{K} - I)\boldsymbol{f} = \sum_{x \neq y} \frac{K_{xy}^2}{\pi_x \pi_y} f(x) f(y) \geq 0$ and therefore $\mathbb{V}\mathrm{ar}_{\text{DPP}}[f] \leq \mathbb{V}\mathrm{ar}_{\text{diag}}[f]$.

eq. (3.4) $\tilde{K} = \Pi^{-1} K^{\odot 2} \Pi^{-1}$ is a symmetric positive matrix and by Hadamard inequality $\det(\tilde{K}) \leq \prod_{x \in \mathcal{X}} \tilde{K}_{xx} = 1$. Therefore at least one of its eigenvalue is lower than 1, hence $\tilde{K} \not\succ I \iff \exists f \in \mathbb{R}^{\mathcal{X}},\ \mathbb{V}\mathrm{ar}_{\text{DPP}}[f] \geq \mathbb{V}\mathrm{ar}_{\text{diag}}[f]$.

eq. (3.5) For all $f \in \mathbb{R}^{\mathcal{X}}$, let denote by $f = f_+ - f_-$ its decomposition into its positive and negative part, which both belong in $\mathbb{R}_+^{\mathcal{X}}$. Then we have

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_{\mathrm{DPP}}[f] &= \mathbb{V}\mathrm{ar}_{\mathrm{DPP}}[f_+] + \mathbb{V}\mathrm{ar}_{\mathrm{DPP}}[f_-] - 2\,\mathbb{C}\mathrm{ov}_{\mathrm{DPP}}[f_+, f_-] && \text{Al-Kashi}\\
&\leq \mathbb{V}\mathrm{ar}_{\mathrm{DPP}}[f_+] + \mathbb{V}\mathrm{ar}_{\mathrm{DPP}}[f_-] + 2\sqrt{\mathbb{V}\mathrm{ar}_{\mathrm{DPP}}[f_+]\,\mathbb{V}\mathrm{ar}_{\mathrm{DPP}}[f_-]} && \text{Cauchy-Schwartz}\\
&\leq \mathbb{V}\mathrm{ar}_{\mathrm{diag}}[f_+] + \mathbb{V}\mathrm{ar}_{\mathrm{diag}}[f_-] + 2\sqrt{\mathbb{V}\mathrm{ar}_{\mathrm{diag}}[f_+]\,\mathbb{V}\mathrm{ar}_{\mathrm{diag}}[f_-]} && \text{eq. (3.3)}\\
&\leq 3\,\mathbb{V}\mathrm{ar}_{\mathrm{diag}}[f]
\end{aligned}$$

where we lastly use that $\mathbb{V}\mathrm{ar}_{\mathrm{diag}}[f] = \mathbb{V}\mathrm{ar}_{\mathrm{diag}}[f_+] + \mathbb{V}\mathrm{ar}_{\mathrm{diag}}[f_-]$ since its associated matrix is diagonal.

eq. (3.6) K being symmetric positive of rank $r \in [\![0, n]\!]$, there 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 $v \in \mathbb{R}^r$, Copenhaver, Kim, Logan, Mayfield, Narayan, Petro, and Sheperd 2013 define its diagram vector

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

concatenating all the $\frac{r(r-1)}{2}$ differences of squares and $\frac{r(r-1)}{2}$ products.

Then introduce $\tilde{V} = (\tilde{V}_i)_{i\in[\![1,n]\!]}$, the matrix whose columns are diagram vectors of matrix V columns. It allows us to rewrite $\tilde{K} = \frac{\boldsymbol{J}}{r} + \frac{r-1}{r}\tilde{V}^\top\tilde{V}$ thus $\tilde{K} - \frac{\boldsymbol{J}}{m} = (\frac{1}{r} - \frac{1}{m})\boldsymbol{J} + \frac{m-1}{m}\tilde{V}^\top\tilde{V}$. Then in order to have

$$\tilde{K} - \frac{\boldsymbol{J}}{m} \succeq 0 \iff \forall f \in \mathbb{R}^{\mathcal{X}},\ \mathbb{V}\mathrm{ar}_{\mathrm{DPP}}[f] \leq \mathbb{V}\mathrm{ar}_{\mathrm{iid}}[f]$$

it is sufficient to have DPP kernel K such that $r \leq m$. On the other hand, we know its average number of samples is $\operatorname{Tr} K = \operatorname{Tr}\Pi^{-1} = m$, because we fixed its marginals. Moreover $\operatorname{Tr} K \leq r$ holds for every DPP, this implies $\operatorname{Tr} K = r$, and therefore it is a projective DPP. Put differently, for any multinomial sampling, we have a projective DPP that beats it uniformly.

□

Note that eq. (3.5) use the general inequality

$$\mathbb{V}\mathrm{ar}[f] \leq \mathbb{V}\mathrm{ar}[f_+] + \mathbb{V}\mathrm{ar}[f_-] + 2\sqrt{\mathbb{V}\mathrm{ar}[f_+]\,\mathbb{V}\mathrm{ar}[f_-]}$$

which justifies that in many cases, we can restrict ourselves to controlling variances of positive-valued functions without loss of generality.

In the case of positive valued functions, proposition 1 shows that for any Bernoulli process or multinomial sampling, taking any projective DPP sampling with same marginals would yield lower variance. This is a strong qualitative argument for the use of projective DPPs for the coreset problem, that we will now try to quantify.

Chapter 4

Improving concentration with DPPs

4.1 Quantitative results on variance

Based on the previous chapter 3, we are interested in quantifying the variance DPPs can yield, in order to leverage it into an improved sample complexity for coresets. We recall classical results on variance of Monte Carlo integration.

4.1.1 Monte Carlo integration

Denote $\mathcal{I} := [-1, 1]$ so that $\mathcal{I}^d$ is an hypercube of dimension d. Given some integrand $h\colon \mathcal{I}^d \to \mathbb{R}$, we are interested in computing $\int_{\mathcal{I}^d} h(x)\mathrm{d}\lambda(x)$, its integral on $\mathcal{I}^d$. The idea of Monte Carlo integration is to sample elements from $\mathcal{I}^d$ and to build an estimator $\hat{L}(h)$ based on these elements.

We know from central limit theorem that as long as variance under sampling distribution exists, i.i.d. sampling Monte Carlo yields

$$\mathbb{V}\mathrm{ar}\left[\hat{L}(h)\right] \lesssim m^{-1} \tag{4.1}$$

Moreover, we know from Bakhvalov 1959 that when the integrand h belongs to the Sobolev space $W^{s,2}(\mathcal{I})$, i.e. when its Fourier coefficients decay sufficiently rapidly,

$$\mathbb{V}\mathrm{ar}\left[\hat{L}(h)\right] \gtrsim m^{-(1+\frac{2s}{d})} \tag{4.2}$$

We refer to Novak 2014 for a more recent inventory on complexity of numerical integration.

Of course, there is plenty of room between this two variance rates. One would be tempted to get out of the i.i.d. framework, without assuming too much regularity of the integrand, and find an estimator that is as close as possible to the lower bound eq. (4.2). We present next a recent result that falls into this description.

4.1.2 Improved variance rate with DPPs

Assume data $\mathcal{X}$ is strictly included in the hypercube $\mathcal{X} \subset \mathcal{I}^d$. For some integrand h, we are interested in constructing an estimator of the mean value of h on $\mathcal{X}$.

Bardenet, Ghosh, and Lin 2021 show the existence of a sequence of DPP kernels $(\tilde{\boldsymbol{K}}_m)_{m\in\mathbb{N}}$, whose induced estimator has asymptotic variance $\mathcal{O}(m^{-(1+\frac{1}{d})})$. We describe here the main key steps of the kernel construction, and introduce needed notations.

Kernel construction. Assume data $\mathcal{X}$ is generated by random samplings from a distribution γ supported in the interior of the hypercube $\mathcal{I}^d$. Then, one can define its kernel density estimation (KDE)

$$\tilde{\gamma}(y) := \frac{1}{n\Delta^d} \sum_{x\in\mathcal{X}} k\left(\frac{x-y}{\Delta}\right)$$

where $\Delta > 0$, and k is a kernel unrelated to any DPP kernel, chosen so that $\int k \mathrm{d}\lambda = 1$.

Let now ω be a strictly positive probability density function on $\mathcal{I}$, and denote by $q = q$ its tensor product density, supported on $\mathcal{I}^d$. Applying Gram-Schmidt algorithm in $L^2(q\mathrm{d}\lambda)$ on multivariate monomials returns a sequence of orthonormal polynomial functions $(\varphi_k)_{k\in\mathbb{N}}$, the multivariate orthonormal polynomials with respect to density q. We refer to Gautschi 2004 for a review of orthogonal polynomials.

From there, define the multivariate Orthogonal Polynomial Ensemble (OPE) kernel associated to q

$$K_q^{(m)}(x,y) := \sum_{k=1}^{m} \varphi_k(x)\varphi_k(y) \colon \mathcal{I}^d \times \mathcal{I}^d \to \mathbb{R} \tag{4.3}$$

which is the outer product of the m first multivariate orthonormal polynomials. Because of orthonormality property, OPE kernel are projective kernels, and so it induces a projective DPP.

Noticeably, we obtained an OPE kernel associated to q, that we can correct to obtain an OPE kernel associated to the KDE distribution $\tilde{\gamma}$, by defining

$$K_{q,\tilde{\gamma}}^{(m)}(x,y) := \sqrt{\frac{q(x)}{\tilde{\gamma}(x)}} K_q^{(m)}(x,y) \sqrt{\frac{q(y)}{\tilde{\gamma}(y)}}. \tag{4.4}$$

However, this projective DPP kernel is still defined on $\mathcal{I}^d \times \mathcal{I}^d$ but we are interested in sampling elements from $\mathcal{X}$, not $\mathcal{I}^d$. A last operation we not detail allows to restrict this kernel into $\tilde{\boldsymbol{K}}_m$, a projective DPP kernel defined on $\mathcal{X} \times \mathcal{X}$, or equivalently, a matrix of size $n \times n$ indexed by the elements of $\mathcal{X}$.

Improved variance rate. It turns out an estimator based on $\tilde{\boldsymbol{K}}_m$ yields improved variance. In particular, Proposition 4. and Equation (S14) from Bardenet et al. 2021 state that

Theorem 4 (Bardenet et al. 2021). *Let $m \in \mathbb{N}$ and $\mathcal{S} \sim \mathcal{DPP}(\tilde{\boldsymbol{K}}_m)$. Define for all $h \in \mathbb{R}^{\mathcal{X}}$ the correlated importance sampling estimator*

$$\hat{L}_{\mathcal{S}}(h) := \sum_{x \in \mathcal{S}} \frac{h(x)}{\tilde{\boldsymbol{K}}_m(x,x)}. \tag{4.5}$$

If $\boldsymbol{\ell}_h := \operatorname{Lip}\left\{\frac{mh}{K^{(m)}_{q,\tilde{\gamma}}}\right\}$ i.e. the Lipschitz constant of $x \mapsto \frac{mh(x)}{K^{(m)}_{q,\tilde{\gamma}}(x,x)}$ is defined, then

$$\mathbb{V}\mathrm{ar}\left[\hat{L}_{\mathcal{S}}(h)\right] = \boldsymbol{\ell}_h^2 \mathcal{O}(m^{-(1+\frac{1}{d})}) + \mathcal{O}(n^{-1/2}). \tag{4.6}$$

It is worth to say this variance generalize to the continuous case. Actually, pre-existing results from Bardenet et al. 2020 treated this case, with same variance improvement. Since the same continuous arguments are invoked in both continuous and discrete case, it explains why the construction of $\tilde{\boldsymbol{K}}_m$ relies first on constructing continuous DPP kernels, though it is not known if that detour can be shortcut in the discrete case.

Also, note that the higher d is, the less the variance gain in eq. (4.6). Intuitively, this is because as the dimension increase, all points become already repelled from each others, so the addition of repulsiveness is less and less effective, or put differently, there is less and less redundancy in an independent sampling that we can get rid of by a negatively correlated sampling.

We now introduce some notations

- We denote by $\mu_{\mathcal{S}} := \frac{1}{|\mathcal{S}|}\sum_{x \in \mathcal{S}} \delta_x$ the empirical measure based on sample $\mathcal{S} \subseteq \mathcal{X}$.
- Hence for all function h, we denote by $\mu_{\mathcal{S}}(h) := \int_{\mathcal{X}} h(x) d\mu_{\mathcal{S}}(x) = \frac{1}{|\mathcal{S}|}\sum_{x \in \mathcal{S}} h(x)$ the expectation of h with respect to $\mu_{\mathcal{S}}$. Furthermore, given a distribution $\mathbb{P}$ on $\mathcal{S}$, we denote by $\mu(h) := \mathbb{E}\left[\mu_{\mathcal{S}}(h)\right]$, its expectation with respect to $\mathbb{P}$.
- Finally, the induced $L^1(\mu_{\mathcal{S}})$ distance between two functions h and h' is denoted by $d_{L^1(\mu_{\mathcal{S}})}(h,h') := \mu_{\mathcal{S}}(|h - h'|)$.

so that we can reformulate the result from theorem 4 into the following corollary.

Corollary 1. *Let $m \in \mathbb{N}$ and $\mathcal{S} \sim \mathcal{DPP}(\tilde{\boldsymbol{K}}_m)$. For all function $g \in \mathbb{R}^{\mathcal{X}}$, we have*

$$\mathbb{V}\mathrm{ar}\left[\mu_{\mathcal{S}}(g)\right] = \boldsymbol{\ell}_g^2 \mathcal{O}(m^{-(1+\frac{1}{d})}) + \mathcal{O}(n^{-1/2})$$

where $\boldsymbol{\ell}_g := \operatorname{Lip}\left\{\frac{\tilde{\boldsymbol{K}}_m}{K^{(m)}_{q,\tilde{\gamma}}} g\right\}$ is the Lipschitz constant of $x \mapsto \frac{\tilde{\boldsymbol{K}}_m(x,x)}{K^{(m)}_{q,\tilde{\gamma}}(x,x)} g(x)$.

Proof. We simply apply theorem 4 with $h = \frac{\tilde{\boldsymbol{K}}_m g}{m}$, such that

$$\mathbb{V}\mathrm{ar}\left[\mu_{\mathcal{S}}(g)\right] = \mathbb{V}\mathrm{ar}\left[\hat{L}_{\mathcal{S}}(h)\right] = \boldsymbol{\ell}_h^2 \mathcal{O}(m^{-(1+\frac{1}{d})}) + \mathcal{O}(n^{-1/2}),$$

where $\boldsymbol{\ell}_h$ is the Lipschitz constant of $x \mapsto \frac{mh(x)}{K^{(m)}_{q,\tilde{\gamma}}(x,x)} = \frac{\tilde{\boldsymbol{K}}_m(x,x)}{K^{(m)}_{q,\tilde{\gamma}}(x,x)} g(x)$.

□

4.2 Regularity assumptions

In order to translate the improved variance rate from corollary 1 into a concentration inequality and then a sample complexity bound for coreset, several assumptions are made and are discussed.

Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be the function space on which we want the coreset property to hold. With the notations of chapter 1, define the following function space

$$\mathcal{G}_m := \frac{m\mathcal{F}}{\tilde{\boldsymbol{K}}_m L(\mathcal{F})} = \left\{ x \mapsto \frac{mf(x)}{\tilde{\boldsymbol{K}}_m(x,x)L(f)} \mid f \in \mathcal{F} \right\} \subseteq \mathbb{R}^{\mathcal{X}}.$$

Note then that for any $f \in \mathcal{F}$, we can define $g := \frac{mf}{\tilde{\boldsymbol{K}}_m L(f)} \in \mathcal{G}_m$ which verifies

$$\frac{\hat{L}_{\mathcal{S}}(f)}{L(f)} = \mu_{\mathcal{S}}(g) \quad \text{and} \quad \mu(g) = \mathbb{E}\left[\mu_{\mathcal{S}}(g)\right] = \frac{\mathbb{E}\left[\hat{L}_{\mathcal{S}}(f)\right]}{L(f)} = 1. \tag{4.7}$$

In the following sections, we crucially assume the functions sets $\mathcal{G}_m$ verify boundedness in Lipschitz constant, infinite norm, and pseudo-dimension. Formally, we assume there exists positive reals $\boldsymbol{\ell}, M \in \mathbb{R}_+$ and $d' \in \mathbb{N}$, such that for all $m \in \mathbb{N}$

1. $\forall g \in \mathcal{G}_m,\ \mathrm{Lip}\left\{ \frac{\tilde{\boldsymbol{K}}_m}{K^{(m)}_{q,\tilde{\gamma}}} g \right\} \leq \boldsymbol{\ell}$
2. $\forall g \in \mathcal{G}_m,\ \|g\|_\infty \leq M$
3. $\mathrm{pdim}\, \mathcal{G}_m \leq d'$

We justify these assumptions by invoking asymptotic behaviour of sets $\mathcal{G}_m$. With notations from previous section 4.1, we first have that for large n, the KDE estimation is close to the true density, namely $\tilde{\gamma} \xrightarrow[n\to+\infty]{} \gamma$. Second, next theorem describes the asymptotic behaviour of any OPE kernel, and in particular confirms that $K^{(m)}_q$ is of order m.

Theorem 5 (Simon 2010)**.** *Assume q is continuous. Then, for every $\varepsilon > 0$, we have uniformly for $x \in [-1+\varepsilon, 1-\varepsilon]^d$*

$$\frac{m}{K^{(m)}_q(x,x)} \xrightarrow[m\to+\infty]{} \frac{q(x)}{q_{\mathrm{eq}}(x)} \tag{4.8}$$

where $q_{\mathrm{eq}} = x \mapsto \prod_{k=1}^d \frac{1}{\pi\sqrt{1-x_k^2}}$.

Finally, the restriction of $K^{(m)}_{q,\tilde{\gamma}}$ into $\tilde{\boldsymbol{K}}_m$ is such that for large n, we have $K^{(m)}_{q,\tilde{\gamma}} \simeq \tilde{\boldsymbol{K}}_m$. Put together, the asymptotic behaviour of sets $\mathcal{G}_m$ is

$$\mathcal{G}_m \to \mathcal{G} := \frac{\mathcal{F}\gamma}{L(\mathcal{F})q_{\mathrm{eq}}} = \left\{ x \mapsto \frac{f(x)\gamma(x)}{L(f)q_{\mathrm{eq}}(x)} \mid f \in \mathcal{F} \right\} \subseteq \mathbb{R}^{\mathcal{X}}. \tag{4.9}$$

Then, assumptions we made on $\mathcal{G}_m$ for every $m \in \mathbb{N}$ translates to assumptions on $\mathcal{G}$. Indeed, if $\mathcal{G}$ verifies boundedness in Lipschitz constant, infinite norm, and pseudo-dimension, we have

1. $\forall g \in \mathcal{G}_m,\ \mathrm{Lip}\left\{\frac{\tilde{\boldsymbol{K}}_m}{K^{(m)}_{q,\tilde{\gamma}}} g\right\} \lesssim \sup_{f\in\mathcal{F}} \mathrm{Lip}\left\{\frac{f\gamma}{L(f)q_{\mathrm{eq}}}\right\} =: \boldsymbol{\ell}$

2. $\forall g \in \mathcal{G}_m,\ \|g\|_\infty \lesssim \|\frac{f\gamma}{L(f)q_{\mathrm{eq}}}\|_\infty \leq \|\frac{\gamma}{q_{\mathrm{eq}}}\|_\infty =: M$

3. $\mathrm{pdim}\,\mathcal{G}_m \lesssim \mathrm{pdim}\,\mathcal{G} =: d'$

Noticeably the bound M does not depend on $\mathcal{F}$, but only on the underlying distribution of data. Besides, $\mathcal{G}$ plays a similar role as $\mathcal{G}_s$ we introduced in eq. (1.6). More than that, when $\mathcal{F}$ is sufficiently rich, the pseudo-dimension of both $\mathcal{G}$ and $\mathcal{G}_s$ is expected to be driven by the one of $\mathcal{F}$ in a similar fashion. Moreover, just like $\mathcal{G}_s$, $\mathcal{G}$ is a function space whose properties influence the sample complexity bound, as we will now show.

4.3 Concentration for fixed query

4.3.1 Chebyshov bound

For any $\varepsilon > 0$ and $m \in \mathbb{N}$, define the Chebyshov concentration bound

$$\mathcal{R}_{\text{Cheb}}(\varepsilon, m) := \frac{1}{\varepsilon^2} \sup_{\mathcal{G}\in\{\mathcal{G}_m | m\in\mathbb{N}\}} \sup_{g\in\mathcal{G}} \mathbb{V}\mathrm{ar}\left[\mu_{\mathcal{S}}(g)\right]. \tag{4.10}$$

Remark that from corollary 1 and the assumption 1 on Lipschitz constant that $\mathcal{R}_{\text{Cheb}}(\varepsilon, m) = \frac{1}{\varepsilon^2}\left(\boldsymbol{\ell}^2\mathcal{O}(m^{-(1+\frac{1}{d})}) + \mathcal{O}(n^{-1/2})\right)$. From that, it follows

Theorem 6 (Chebyshov bound for fixed query)**.** *Let $m \in \mathbb{N}$ and $\mathcal{S} \sim \mathcal{DPP}(\tilde{\boldsymbol{K}}_m)$. Then for all $\varepsilon > 0$ and all $f \in \mathcal{F}$,*

$$\mathbb{P}\left[|\hat{L}_{\mathcal{S}}(f) - L(f)| > \varepsilon L(f)\right] \leq \mathcal{R}_{\mathit{Cheb}}(\varepsilon, m)$$

Moreover, for all $\delta > 0$ and for n sufficiently large,

$$m \gtrsim \left(\frac{\boldsymbol{\ell}^2}{\delta\varepsilon^2}\right)^{\frac{1}{1+\frac{1}{d}}} \implies \mathcal{S} \text{ is an } \varepsilon\text{-coreset for } f \text{ w.p. } 1-\delta.$$

Proof. Let $f \in \mathcal{F}$, so that $g := \frac{mf}{\tilde{\boldsymbol{K}}_m L(f)} \in \mathcal{G}_m$ verifies eq. (4.7).

Applying the Bienaymé-Chebyshov inequality, we obtain

$$\begin{aligned}
\mathbb{P}\left[|\hat{L}_{\mathcal{S}}(f) - L(f)| > \varepsilon L(f)\right] &= \mathbb{P}\left[|\frac{\hat{L}_{\mathcal{S}}(f)}{L(f)} - 1| > \varepsilon\right] \\
&= \mathbb{P}\left[|\mu_{\mathcal{S}}(g) - \mu(g)| > \varepsilon\right] \\
&\leq \frac{1}{\varepsilon^2}\,\mathbb{V}\mathrm{ar}\left[\mu_{\mathcal{S}}(g)\right] \\
&\leq \mathcal{R}_{\text{Cheb}}(\varepsilon, m) = \frac{1}{\varepsilon^2}\left(\boldsymbol{\ell}^2\mathcal{O}(m^{-(1+\frac{1}{d})}) + \mathcal{O}(n^{-1/2})\right)
\end{aligned}$$

Hence, a sufficient condition for $\mathcal{S}$ to be an ε-coreset for f w.p. $1-\delta$ is to have

$$\mathcal{R}_{\text{Cheb}}(\varepsilon, m) \leq \delta \iff m^{1+\frac{1}{d}} \gtrsim \frac{\boldsymbol{\ell}^2}{\delta\varepsilon^2 + \mathcal{O}(n^{-1/2})} = \frac{\boldsymbol{\ell}^2}{\delta\varepsilon^2} \frac{1}{1 + \frac{1}{\delta\varepsilon^2}\mathcal{O}(n^{-1/2})}.$$

For sufficiently large n (potentially $n \gtrsim \delta^{-2}\varepsilon^{-4}$), we can control the second factor and thus obtain the bound

$$m \gtrsim \left(\frac{\boldsymbol{\ell}^2}{\delta\varepsilon^2}\right)^{\frac{1}{1+\frac{1}{d}}}.$$

□

We obtained a sample complexity bound with improved dependency in ε compared to the one from theorem 2 for the i.i.d. sampling framework. Our result is $\mathcal{O}(\varepsilon^{-2/(1+1/d)})$ whereas the latter is $\mathcal{O}(\varepsilon^{-2})$. As for the variance, the higher the dimension d is, the less is the gain in sample complexity, because the less there is redundancy to get rid of by a negatively correlated sampling.

On the other hand, the dependency in δ is worsened compared to i.i.d. sampling. Our result is $\mathcal{O}(\delta^{-1/(1+1/d)})$ whereas the latter is $\log \delta^{-1}$. We propose to tackle this dependency, before generalizing the obtained bound to all queries.

4.3.2 Breuer and Duits bound

There actually exists a concentration bound for projective DPPs from Breuer et al. 2013, that still can leverage the increased variance rate from Bardenet et al. 2021, and that is tighter than Chebyshov bound.

Theorem 7 (Breuer et al. 2013)**.** *Let $\varepsilon > 0$, any bounded function h, any projective DPP kernel K, and let $\mathcal{S} \sim \mathcal{DPP}(K)$.*

Then for any linear statistic $X_h := \sum_{x\in\mathcal{S}} h(x)$,

$$\mathbb{P}\left[|X_h - \mathbb{E}\left[X_h\right]| > \varepsilon\right] \leq \begin{cases} 2\exp\left(-\frac{\varepsilon^2}{4A\,\mathbb{V}\text{ar}[X_h]}\right) & \text{if } \varepsilon < \frac{2A\,\mathbb{V}\text{ar}[X_h]}{3\|h\|_\infty} \\ 2\exp\left(-\frac{\varepsilon}{6\|h\|_\infty}\right) & \text{otherwise} \end{cases}$$

where $A \simeq 7819$ so does not depend on h, K or ε.

Note that the second case in this concentration bound is tighter than the first. Indeed,

$$\varepsilon \geq \frac{2A\,\mathbb{V}\text{ar}\left[X_h\right]}{3\|h\|_\infty} \implies \frac{4A\,\mathbb{V}\text{ar}\left[X_h\right]}{\varepsilon^2} \leq \frac{3\|h\|_\infty}{2A\,\mathbb{V}\text{ar}\left[X_h\right]} \frac{4A\,\mathbb{V}\text{ar}\left[X_h\right]}{\varepsilon} = \frac{6\|h\|_\infty}{\varepsilon}.$$

We then expect to have

$$\mathbb{P}\left[|X_h - \mathbb{E}\left[X_h\right]| > \varepsilon\right] \leq 2\exp\left(-\frac{\varepsilon^2}{4A\,\mathbb{V}\text{ar}\left[X_h\right]}\right). \tag{4.11}$$

Based on that remark, we define for any $\varepsilon > 0$ and $m \in \mathbb{N}$

$$\mathcal{R}_{\text{BD}}(\varepsilon, m) := 2\exp\left(\left(-4A\mathcal{R}_{\text{Cheb}}(\varepsilon, m)\right)^{-1}\right)$$

which we call a Breuer and Duits (BD)-type bound, and that verifies what follows.

Theorem 8 (BD-type bound for fixed query)**.** *Let $m \in \mathbb{N}$ and $\mathcal{S} \sim \mathcal{DPP}(\tilde{\boldsymbol{K}}_m)$. Then for all $\varepsilon > 0$ and all $f \in \mathcal{F}$,*

$$\mathbb{P}\left[|\hat{L}_{\mathcal{S}}(f) - L(f)| > \varepsilon L(f)\right] \leq \mathcal{R}_{BD}(\varepsilon, m).$$

Moreover, for all $\delta > 0$ and for n sufficiently large,

$$m \gtrsim \left(\frac{\boldsymbol{\ell}^2}{\varepsilon^2} \log \frac{2}{\delta}\right)^{\frac{1}{1+\frac{1}{d}}} \implies \mathcal{S} \text{ is an } \varepsilon\text{-coreset for } f \text{ w.p. } 1-\delta.$$

Proof. Let $f \in \mathcal{F}$, so that $g := \frac{mf}{\tilde{\boldsymbol{K}}_m L(f)} \in \mathcal{G}_m$ verifies eq. (4.7).

Then we apply the Breuer and Duits bound from theorem 7 taking $h = \frac{g}{m}$.

$$\begin{aligned}
\mathbb{P}\left[|\hat{L}_{\mathcal{S}}(f) - L(f)| > \varepsilon L(f)\right] &= \mathbb{P}\left[|\mu_{\mathcal{S}}(g) - \mu(g)| > \varepsilon\right] \\
&\leq 2\exp\left(\begin{cases} \frac{-\varepsilon^2}{4A\,\mathbb{V}\mathrm{ar}[\mu_{\mathcal{S}}(g)]} & \text{if } \varepsilon < \frac{2Am\,\mathbb{V}\mathrm{ar}[\mu_{\mathcal{S}}(g)]}{3\|g\|_\infty} \\ \frac{-\varepsilon m}{6\|g\|_\infty} & \text{otherwise} \end{cases}\right) \\
&\leq 2\exp\left(-\frac{\varepsilon^2}{4A\,\mathbb{V}\mathrm{ar}\left[\mu_{\mathcal{S}}(g)\right]}\right) \\
&\leq 2\exp\left(\left(-4A\mathcal{R}_{\text{Cheb}}(\varepsilon, m)\right)^{-1}\right) = \mathcal{R}_{\text{BD}}(\varepsilon, m)
\end{aligned}$$

where we firstly used the remark in eq. (4.11), and secondly the definition eq. (4.10) of $\mathcal{R}_{\text{Cheb}}$.

Hence, a sufficient condition for $\mathcal{S}$ to be an ε-coreset for f w.p. $1-\delta$ is to have

$$\mathcal{R}_{\text{BD}}(\varepsilon, m) \leq \delta \iff \left(\log \frac{2}{\delta}\right)^{-1} \geq 4A\mathcal{R}_{\text{Cheb}}(\varepsilon, m)$$

and we know from theorem 6 this implies that for n sufficiently large

$$m \gtrsim \left(\frac{\boldsymbol{\ell}^2}{\varepsilon^2} \log \frac{2}{\delta}\right)^{\frac{1}{1+\frac{1}{d}}}.$$

□

Discussing the tighter case. Seeing theorem 7, one could wonder why we didn't leverage the second tighter case, noticed in eq. (4.11). Applied to our context, it would lead to the bound

$$m \geq \frac{6M}{\varepsilon} \log \frac{2}{\delta}$$

which is better than the one we shown. But for the second case to apply and obtain this bound, one would require

$$\varepsilon \geq \frac{2Am\,\mathbb{V}\mathrm{ar}\left[\mu_{\mathcal{S}}(g)\right]}{3\|g\|_\infty} = \mathcal{O}\left(\frac{\boldsymbol{\ell}^2}{\|g\|_\infty m^{1/d}}\right)$$
$$\iff m \gtrsim \left(\frac{\boldsymbol{\ell}^2}{\varepsilon\|g\|_\infty}\right)^d \implies m \gtrsim \left(\frac{\boldsymbol{\ell}^2}{\varepsilon M}\right)^d.$$

Thereby, the condition on ε translates to a much worse bound on m, ruining the interest of the second case.

Compared to the i.i.d. framework bound from theorem 2, we obtained a better bound on coreset size for fixed query. This corroborate the qualitative results we obtained on variance comparison in section 3.2, by quantifying the effect of increased variance rate on sample complexity of coreset.

Although the concentration bound from Breuer et al. 2013 pre-dates the works of Tremblay et al. 2018, it is only because of the improved variance rate from Bardenet et al. 2021 that it can yields improved concentration result. If one were to apply classical variance rate in $\mathcal{O}(m^{-1})$ to it, it would find no improvements of the i.i.d. framework.

4.4 Extension to all queries

In order to obtain an ε-coreset for $\mathcal{F}$, property eq. (1.2) must hold simultaneously for all queries $f \in \mathcal{F}$. To do so, recall the idea from chapter 1 of constructing a surrogate finite set that approximate well $\mathcal{F}$, then apply union bound on it. We summarize this process into

Theorem 9 (Infinite union bound)**.** *Let $\mathcal{H} \subseteq [0, M]^{\mathcal{X}}$ be a set of bounded functions defined on a base set $\mathcal{X}$, with $d'_{\mathcal{H}} := \operatorname{pdim} \mathcal{H}$ its pseudo-dimension (see definition 3). Let moreover $m \in \mathbb{N}$ and $\mathbb{P}$ a distribution supported on $\mathcal{X}^m$.*

Assume there exists a bounding function $\mathcal{R}$ such that for all $\varepsilon > 0$, function $h \in \mathcal{H}$, integer $m \in \mathbb{N}$, and let $\mathcal{S} \sim \mathbb{P}$, we have the bound

$$\mathbb{P}\left[|\mu_{\mathcal{S}}(h) - \mu(h)| > \varepsilon\right] \leq \mathcal{R}(\varepsilon, m). \tag{4.12}$$

Then for all $\varepsilon \in]0, M]$ and all $m \in \mathbb{N}$ such that $\mathcal{R}(\varepsilon/2, m) \leq 1/2$, it holds

$$\mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}}(h) - \mu(h)| > \varepsilon\right] \leq 8\left(\frac{8eM}{\varepsilon}\right)^{2d'_{\mathcal{H}}} \mathcal{R}(\varepsilon/8, m). \tag{4.13}$$

Contrary to the classical union bound, this theorem allows to convert any reasonable uniform bound $\mathcal{R}$ into a similar bound over an infinite set. We delay its proof of in section 4.5. Applied to the BD-type bound we obtained for fixed queries, it provides our main result on sample complexity of coreset.

Theorem 10 (BD-type bound for all queries)**.** *Let $m \in \mathbb{N}$ and $\mathcal{S} \sim \mathcal{DPP}(\tilde{\boldsymbol{K}}_m)$. Then for all $\varepsilon \in]0, M]$*

$$\mathbb{P}\left[\exists f \in \mathcal{F},\ |\hat{L}_{\mathcal{S}}(f) - L(f)| > \varepsilon L(f)\right] \leq 8\left(\frac{8eM}{\varepsilon}\right)^{2d'} \mathcal{R}_{BD}(\varepsilon/8, m)$$

Moreover, for all $\delta > 0$ and for n sufficiently large

$$m \gtrsim \left(\frac{\boldsymbol{\ell}^2}{\varepsilon^2}\left(d' \log \frac{M}{\varepsilon} + \log \frac{1}{\delta}\right)\right)^{\frac{1}{1+\frac{1}{d}}} \implies \mathcal{S} \text{ is an } \varepsilon\text{-coreset for } \mathcal{F} \text{ w.p. } 1-\delta.$$

Proof. Let $f \in \mathcal{F}$, so that $g := \frac{mf}{\tilde{\boldsymbol{K}}_m L(f)} \in \mathcal{G}_m$ verifies eq. (4.7).

We know from theorem 8 that for all $\varepsilon \in]0, M]$ and $m \in \mathbb{N}$ we have

$$\mathbb{P}\left[|\mu_{\mathcal{S}}(g) - \mu(g)| > \varepsilon\right] \leq \mathcal{R}_{\text{BD}}(\varepsilon, m).$$

Moreover, the assumption 2 implies $\mathcal{G}_m \in [0, M]^{\mathcal{X}}$. The hypotheses of theorem 9 are thus satisfied, and we can apply it taking $\mathcal{H} = \mathcal{G}_m$, which yields that for all $m \in \mathbb{N}$ such that $\mathcal{R}_{\text{BD}}(\varepsilon/2, m) \leq 1/2$, it holds

$$\begin{aligned}\mathbb{P}\left[\exists f \in \mathcal{F},\ |\hat{L}_{\mathcal{S}}(f) - L(f)| > \varepsilon L(f)\right] &= \mathbb{P}\left[\exists g \in \mathcal{G}_m,\ |\mu_{\mathcal{S}}(g) - \mu(g)| > \varepsilon\right] \\ &\leq 8\left(\frac{8eM}{\varepsilon}\right)^{2d'_{\mathcal{G}_m}} \mathcal{R}_{\text{BD}}(\varepsilon/8, m) \leq 8\left(\frac{8eM}{\varepsilon}\right)^{2d'} \mathcal{R}_{\text{BD}}(\varepsilon/8, m)\end{aligned}$$

where we used the assumption 3 that for all $m \in \mathbb{N}$, $\operatorname{pdim} \mathcal{G}_m \leq d'$.

Hence, a sufficient condition for $\mathcal{S}$ to be an ε-coreset for $\mathcal{F}$ w.p. $1-\delta$ is to have

$$\begin{aligned}\mathcal{R}_{\text{BD}}(\varepsilon/8, m) \leq \frac{\delta}{8}\left(\frac{8eM}{\varepsilon}\right)^{-2d'} &\iff m \gtrsim \left(\frac{64\boldsymbol{\ell}^2}{\varepsilon^2} \log\left(\frac{16}{\delta}\left(\frac{8eM}{\varepsilon}\right)^{2d'}\right)\right)^{\frac{1}{1+\frac{1}{d}}} \\ &\iff m \gtrsim \left(\frac{\boldsymbol{\ell}^2}{\varepsilon^2}\left(\log \frac{1}{\delta} + d' \log \frac{M}{\varepsilon}\right)\right)^{\frac{1}{1+\frac{1}{d}}}.\end{aligned}$$

This rate is conditioned to the fact that m is such that $\mathcal{R}_{\text{BD}}(\varepsilon/2, m) \leq 1/2$. But we know it holds as soon as $m \gtrsim \left(\frac{4\boldsymbol{\ell}^2}{\varepsilon^2} \log 4\right)^{\frac{1}{1+\frac{1}{d}}}$, which is trivially implied by the obtained bound.

□

Our results is to be compared to eq. (1.7)

$$m \gtrsim \frac{S}{\varepsilon^2}(\operatorname{pdim} \mathcal{G}_s \log S + \log \frac{2}{\delta}).$$

Discussing assumptions in section 4.2, we remarked that for sufficiently rich query space $\mathcal{F}$, $\operatorname{pdim} \mathcal{G}_s$ can be quiet comparable to d'. Also was noticed that M does not need to depend on $\mathcal{F}$. With this in mind, our bound is quiet comparable to the state-of-the-art, and improve it by a $(1 + 1/d)$-root. However, our result requires a supplementary condition on Lipschitz constant. Perspectives of improvements are discussed in chapter 5

4.5 Proof of theorem 9

We follow a similar proof scheme as in section 9.4 of Haussler 1992. We specifically revisit Lemmas 12. and 13., getting rid of independency hypothesis, and making intermediary results more flexible to further improvements. Importantly, the proof is still in progress, the remaining work being reduced to conjecture 1.

We start by a symmetrisation argument that relates a concentration statement between a sampling and its expectancy, with a concentration statement between two i.i.d. samplings.

Lemma 1 (Symmetrisation)**.** *Assume the hypothesis of theorem 9 about bounding function $\mathcal{R}$. Let furthermore $\varepsilon > 0$, $m \in \mathbb{N}$, and $\mathcal{S}_1, \mathcal{S}_2 \overset{i.i.d.}{\sim} \mathbb{P}$, two sequences of size m independently sampled from the same distribution supported on $\mathcal{X}^m$.*

Then for all $m \in \mathbb{N}$ such that $\mathcal{R}(\varepsilon/2, m) \leq 1/2$

$$\mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu(h)| > \varepsilon\right] \leq 2\mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/2\right]$$

Proof. Let $\varepsilon > 0$ and take $m \in \mathbb{N}$ such that $\mathcal{R}(\varepsilon/2, m) \leq 1/2$. Then let $\mathcal{S}_1$ be sampled such that $\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu(h)| > \varepsilon$. This obviously happens with probability $\mathbb{P}\left[\exists h \in \mathcal{H}, |\mu_{\mathcal{S}_1}(h) - \mu(h)| > \varepsilon\right]$.

For such an h, we then independently sample $\mathcal{S}_2$ such that $|\mu_{\mathcal{S}_2}(h) - \mu(h)| \leq \varepsilon/2$. Because $\mathcal{R}(\varepsilon, m) \leq 1/2$, we know this happens with probability greater than $1 - 1/2 = 1/2$, and we thus have

$$\begin{aligned}
&\frac{1}{2}\mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu(h)| > \varepsilon\right] \\
&\leq \mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu(h)| > \varepsilon \wedge |\mu_{\mathcal{S}_2}(h) - \mu(h)| \leq \varepsilon/2\right] \\
&\leq \mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/2\right]
\end{aligned}$$

where we used the triangular inequality

$$|\mu_{\mathcal{S}_1}(h) - \mu(h)| - |\mu_{\mathcal{S}_2}(h) - \mu(h)| \leq |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)|.$$

□

We then introduce the notion of ε-packing and ε-covering, on which rely the construction of a surrogate query set that approximate $\mathcal{F}$ to an ε granularity.

Definition 7 (Separated set and packing number)**.** Let $(\mathcal{H}, d)$ be a metric space.

- For any $\varepsilon > 0$, a subset $\mathcal{H}' \subseteq \mathcal{H}$ is said to be ε-separated if for all distinct $h'_1, h'_2 \in \mathcal{H}'$, $d(h'_1, h'_2) > \varepsilon$.
- The ε-packing number on $(\mathcal{H}, d)$, denoted by $\mathbf{P}(\varepsilon, \mathcal{H}, d)$, is the cardinality of the largest ε-separated subset $\mathcal{H}'$ of $\mathcal{H}$.

Intuitively, the ε-packing number is the maximal number of balls of radius $\varepsilon/2$ that can fit into $\mathcal{H}$ without intersecting.

Definition 8 (Covered set and cover number)**.** Let $(\mathcal{H}, d)$ be a metric space.

- For any $\varepsilon > 0$, a subset $\mathcal{H}'$ of $\mathcal{H}$ is said to be an ε-cover of $\mathcal{H}$ if for all $h \in \mathcal{H}$, there exists $h' \in \mathcal{H}'$ with $d(h, h') \leq \varepsilon$.
- The ε-covering number on $(\mathcal{H}, d)$, denoted by $\mathbf{C}(\varepsilon, \mathcal{H}, d)$, is the cardinality of the smallest ε-cover of $\mathcal{H}$.

Intuitively, the ε-covering number is the minimal number of balls of radius ε than can fill $\mathcal{H}$, with possible overlaps.

One can easily check that for all $\varepsilon > 0$

$$\mathbf{P}(2\varepsilon, \mathcal{H}, d) \leq \mathbf{C}(\varepsilon, \mathcal{H}, d) \leq \mathbf{P}(\varepsilon, \mathcal{H}, d) \tag{4.14}$$

Moreover, in the case of metric function spaces, packing numbers can be related to the pseudo-dimension of the packed space through

Theorem 11 (Pollard 1984 and Haussler 1995)**.** *For any set $\mathcal{X}$, any probability distribution μ on $\mathcal{X}$, any set $\mathcal{H} \subseteq [0, M]^{\mathcal{X}}$ of μ-measurable positive functions on $\mathcal{X}$ bounded by some real M, and any $\varepsilon \in]0, M]$, one has*

$$\mathbf{P}(\varepsilon, \mathcal{H}, d_{L^1(\mu)}) \leq \min \left\{ 2 \left(\frac{2eM}{\varepsilon} \log \frac{2eM}{\varepsilon} \right)^{d'_{\mathcal{H}}},\ e(d'_{\mathcal{H}} + 1) \left(\frac{2eM}{\varepsilon} \right)^{d'_{\mathcal{H}}} \right\}$$

where $d'_{\mathcal{H}} = \operatorname{pdim} \mathcal{H}$ is the pseudo-dimension of $\mathcal{H}$.

Note that the second bound is better than the first one when ε is sufficiently small compared to $d'_{\mathcal{H}}$ and vice versa.

We introduce the following conjecture, as an attempt to extend Lemma 13. from Haussler 1992 to a non-i.i.d. framework. We discuss in the draft of proof an inequality that would lead to the result. In addition, we discuss how the method used in i.i.d. case can not be applied easily to ours.

Conjecture 1. *Let $\varepsilon \in]0, 1]$, $m \in \mathbb{N}$, and $\mathcal{S}_1, \mathcal{S}_2 \overset{i.i.d.}{\sim} \mathbb{P}$, two sequences of size m independently sampled from the same distribution supported on $\mathcal{X}^m$. Then*

$$\begin{gathered} \mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon\right] \\ \leq \\ \sup_{\mathcal{S} \in \mathcal{X}^{2m}} \mathbf{C}(\varepsilon/4, \mathcal{H}, d_{L^1(\mu_{\mathcal{S}})}) \sup_{h \in \mathcal{H}} \mathbb{P}\left[|\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/2\right] \end{gathered}$$

Draft of Proof. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ sampled such that $\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon$.

We denote their concatenation by $\mathcal{S} := \mathcal{S}_1 \uplus \mathcal{S}_2 \in \mathcal{X}^{2m}$ which is of size $2m$. Let then be taken $\mathcal{H}'_{\mathcal{S}}$, a minimal $\varepsilon/4$-cover of $\mathcal{H}$ for the $d_{L^1(\mu_{\mathcal{S}})}$ distance, then $|\mathcal{H}'_{\mathcal{S}}| = \mathbf{C}(\varepsilon/4, \mathcal{H}, d_{L^1(\mu_{\mathcal{S}})})$. We thus know there exists $h' \in \mathcal{H}'_{\mathcal{S}}$ such that $d_{L^1(\mu_{\mathcal{S}})}(h, h') \leq \varepsilon/4$.

Then triangular inequalities yields that

$$\begin{aligned}|\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| &\leq |\mu_{\mathcal{S}_1}(h') - \mu_{\mathcal{S}_2}(h')| + |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_1}(h')| + |\mu_{\mathcal{S}_2}(h) - \mu_{\mathcal{S}_2}(h')| \\ &\leq |\mu_{\mathcal{S}_1}(h') - \mu_{\mathcal{S}_2}(h')| + \mu_{\mathcal{S}_1}(|h - h'|) + \mu_{\mathcal{S}_2}(|h - h'|) \\ &\leq |\mu_{\mathcal{S}_1}(h') - \mu_{\mathcal{S}_2}(h')| + 2d_{L^1(\mu_{\mathcal{S}})}(h, h').\end{aligned}$$

Because of the $\varepsilon/4$-cover, it implies

$$|\mu_{\mathcal{S}_1}(h') - \mu_{\mathcal{S}_2}(h')| \geq |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| - 2d_{L^1(\mu_{\mathcal{S}})}(h, h') > \varepsilon/2$$

and therefore

$$\mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon\right] \leq \mathbb{P}\left[\exists h \in \mathcal{H}'_{\mathcal{S}},\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/2\right]$$

By the law of total expectation, we obtain

$$\begin{aligned}\mathbb{P}\left[\exists h \in \mathcal{H}'_{\mathcal{S}},\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/4\right] &= \mathbb{E}\left[\mathbb{1}\{\exists h \in \mathcal{H}'_{\mathcal{S}},\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/2\}\right] \\ &= \mathbb{E}\left[\mathbb{P}\left[\exists h \in \mathcal{H}'_{\mathcal{S}},\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/2 \mid \mathcal{H}'_{\mathcal{S}}\right]\right] \\ &= \mathbb{E}\left[\mathbb{P}\left[\bigcup_{h \in \mathcal{H}'_{\mathcal{S}}} \{|\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/2\} \mid \mathcal{H}'_{\mathcal{S}}\right]\right] \\ &\overset{?}{\leq} \sup_{\mathcal{H}'_{\mathcal{S}}} |\mathcal{H}'_{\mathcal{S}}| \sup_{h \in \mathcal{H}} \mathbb{P}\left[|\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/2\right] \\ &= \sup_{\mathcal{S} \in \mathcal{X}^{2m}} N(\varepsilon/4, \mathcal{H}, d_{L^1(\mu_{\mathcal{S}})}) \sup_{h \in \mathcal{H}} \mathbb{P}\left[|\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/2\right]\end{aligned}$$

$\overset{?}{\leq}$ indicates this inequality is still to be proven. It consists of bounding the probability of a union of event over a random set. Since we can bound uniformly both, event probability, and set cardinality, it could seem intuitive this random union bound hold. However, the conditioning by $\mathcal{H}'_{\mathcal{S}}$ makes this bound non trivial and likely requires further assumptions.

Though we make use of the symmetry between $\mathcal{S}_1$ and $\mathcal{S}_2$, we can't leverage it as much as the i.i.d. framework does. Such as in Haussler 1992, methods from PAC learning theory often leverage the fact that swapping samples from $\mathcal{S}_1$ and $\mathcal{S}_2$ preserves the i.i.d. sampling, which is not the case in general for a DPP sampling. For instance, if $\mathcal{X} = \{0, 1\}$, then a 2-DPP will return almost surely $\mathcal{S}_1 = \mathcal{S}_2 = \{0, 1\}$. But swapping 0 with 1 leads to $\mathcal{S}_1 = \{0, 0\}$ and $\mathcal{S}_2 = \{1, 1\}$. This event has zero probability for a 2-DPP sampling, so the sampling has not been preserved by the swapping.

□

Proof Ideas

One could try making cover independent on $\mathcal{S}$ Take $\mathcal{H}'$, a minimal $\varepsilon/8$-cover of $\mathcal{H}$ for the $d := d_{L^1(\mu)}$ distance instead of the $\hat{d}_{\mathcal{S}} := d_{L^1(\mu_{\mathcal{S}})}$ distance.

$$|\mu_{\mathcal{S}_1}(h') - \mu_{\mathcal{S}_2}(h')| \geq |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| - 2d(h,h') + 2(d(h,h') - \hat{d}_{\mathcal{S}_{1,2}}(h,h'))$$

Then

$$\begin{aligned}\mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon \wedge |d(h,h') - \hat{d}_{\mathcal{S}_{1,2}}(h,h')| \leq \varepsilon/8\right] \\ \leq \mathbb{P}\left[\exists h \in \mathcal{H}',\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/2\right].\end{aligned}$$

The right hand side term is a union bound over a finite deterministic set which can be handled. But there is still to control the other case

$$\begin{aligned}\mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon \wedge |d(h,h') - \hat{d}_{\mathcal{S}_{1,2}}(h,h')| > \varepsilon/8\right] \\ \leq \mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon\right] \mathbb{P}\left[\exists h \in \mathcal{H},\ |d(h,h') - \hat{d}_{\mathcal{S}_{1,2}}(h,h')| > \varepsilon/8\right].\end{aligned}$$

One can rewrite

$$\begin{aligned}\mathbb{P}\left[\exists h \in \mathcal{H},\ |d(h,h') - \hat{d}_{\mathcal{S}_{1,2}}(h,h')| > \varepsilon/8\right] = \\ \mathbb{P}\left[\exists h' \in \mathcal{H}',\ \exists h \in \mathcal{B}_d(h', \varepsilon/8),\ |d(h,h') - \hat{d}_{\mathcal{S}_{1,2}}(h,h')| > \varepsilon/8\right].\end{aligned}$$

This leads to controlling the term

$$\begin{aligned}\mathbb{P}\left[\exists h \in \mathcal{B}_d(h', \varepsilon/8),\ |d(h,h') - \hat{d}_{\mathcal{S}_{1,2}}(h,h')| > \varepsilon/8\right] \\ \leq 2\mathbb{P}\left[\exists h \in \mathcal{B}_d(h', \varepsilon/8),\ |\hat{d}_{\mathcal{S}_{1,2}}(h,h') - \hat{d}_{\mathcal{S}_{3,4}}(h,h')| > \varepsilon/16\right]\end{aligned}$$

by symmetrisation.

$$|\hat{d}_{\mathcal{S}_{1,2}}(h,h') - \hat{d}_{\mathcal{S}_{3,4}}(h,h')| \leq |\hat{d}_{\mathcal{S}_{1,2}}(h'',h') - \hat{d}_{\mathcal{S}_{3,4}}(h'',h')| + 2\hat{d}_{\mathcal{S}_{1,2,3,4}}(h'',h)$$

We are back at the same point except $\hat{d}_{\mathcal{S}_{1,2,3,4}}$ replaced $\hat{d}_{\mathcal{S}_{1,2}}$. This does not seem to lead anywhere.

Admitting conjecture 1, we are able to prove theorem 9.

Proof. Let $\varepsilon > 0$ and take $m \in \mathbb{N}$ such that $\mathcal{R}(\varepsilon/2, m) \leq 1/2$. Combining lemma 1 and conjecture 1 gives

$$\begin{aligned}\mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu(h)| > \varepsilon\right] &\leq 2\mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/2\right] \\ &\leq 2 \sup_{\mathcal{S} \in \mathcal{X}^{2m}} \mathbf{C}(\varepsilon/8, \mathcal{H}, d_{L^1(\mu_{\mathcal{S}})}) \sup_{h \in \mathcal{H}} \mathbb{P}\left[|\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/4\right]\end{aligned}$$

In order to bound the last term, first consider applying the union bound

$$\begin{aligned}\mathbb{P}\left[|\mu_{\mathcal{S}_1}(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/4\right] &\leq \mathbb{P}\left[|\mu_{\mathcal{S}_1}(h) - \mu(h)| + |\mu(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/4\right] \\ &\leq \mathbb{P}\left[|\mu_{\mathcal{S}_1}(h) - \mu(h)| > \varepsilon/8 \vee |\mu(h) - \mu_{\mathcal{S}_2}(h)| > \varepsilon/8\right] \\ &\leq 2\mathbb{P}\left[|\mu_{\mathcal{S}_1}(h) - \mu(h)| > \varepsilon/8\right] \\ &\leq 2\mathcal{R}(\mathcal{H}, \varepsilon/8, m)\end{aligned}$$

Second, we know from eq. (4.14) and theorem 11

$$\mathbf{C}(\varepsilon/8, \mathcal{H}, d_{L^1(\mu_{\mathcal{S}})}) \leq \mathbf{P}(\varepsilon/8, \mathcal{H}, d_{L^1(\mu_{\mathcal{S}})}) \leq 2\left(\frac{16eM}{\varepsilon}\log\frac{16eM}{\varepsilon}\right)^{d'_{\mathcal{H}}} \leq 2\left(\frac{8eM}{\varepsilon}\right)^{2d'_{\mathcal{H}}}$$

where we used the fact that $a\log a < (a/2)^2$ whenever $a \geq 5$, which is the case for $\frac{8eM}{\varepsilon} \geq 5$ since $\varepsilon \in]0, M]$.

The two precedent bounds do neither depend on $\mathcal{S}$ nor f, and therefore

$$\mathbb{P}\left[\exists h \in \mathcal{H},\ |\mu_{\mathcal{S}_1}(h) - \mu(h)| > \varepsilon\right] \leq 8\left(\frac{8eM}{\varepsilon}\right)^{2d'_{\mathcal{H}}}\mathcal{R}(\varepsilon/8, m)$$

which is the desired result. □

Chapter 5

Conclusion and perspectives

In this report, we attempted to provide improved sample complexity results on coresets by the use of a well-chosen DPP. We started with an inventory on coreset literature, then we brought to light determinantal point processes as an improvement way. After qualitatively justifying the use of DPPs over existing sampling methods, we established quantitative results.

We show that under the assumption of conjecture 1, still to be proven, a specific projective DPP based on orthogonal polynomials yields an improved sample complexity. Basically, previous state-of-the-art sample complexity is improved by a $(1 + 1/d)$-root. Also, it does not require to compute bounds on sensitivity, which can be a huge drawback of current methods depending on the problem. However, this result comes with some costs.

First, it requires a supplementary regularity condition, namely, controlled Lipschitz constant. This requirement comes from the nature of the method used in Bardenet et al. 2020 to build the OPE kernel, which relies on a detour through continuous DPPs needed to apply continuous arguments. At the moment this report is redacted, it is not known if that detour can be shortcut in the discrete case. One idea could be to build a DPP kernel directly from a discrete OPE, i.e. to build an orthogonal polynomials family with respect to the empirical measure. When the data size is sufficiently large, one could recover the same behaviour.

Second, we saw in section 2.5 that sampling from DPP comes with additional computational cost. In our case, sampling from the OPE kernel can be achieved in $\mathcal{O}(nm^2)$, compared to $\mathcal{O}(nm)$ in the i.i.d. case. One could rely on approximate DPP sampling, though it is not clear how sensible is our analysis to the exactness of the sampling.

Our analysis should rely on current results on concentration inequalities for DPPs, known achievable variance rates, generalization bound in the non-i.i.d. framework, and known DPP sampling methods. We made it sufficiently flexible to adapt to any potential improvements of these. We hope it provided a good insight on the potential improvements in machine learning applications one should expect from negatively correlated sampling, and especially DPPs. This work was intended to be the most complete on the theoretical level. The next step would now be to confirm it empirically via numerical experiments.

Finally, I can't close this report without thanking the team SigMA from CRIStAL, for its good spirit and understanding that has greatly contributed to dilute the worries of a broken arm while biking. In particular, I acknowledge Rémi Bardenet motivation, openness, relevance and the energy he invests in involving everyone into its scientific adventures.

Bibliography

Bardenet Rémi, Ghosh Subhro, and Lin Meixia (2021). *Determinantal point processes based on orthogonal polynomials for sampling minibatches in SGD*. DOI: 10.48550/ARXIV.2112.06007. URL: https://arxiv.org/abs/2112.06007.

Bardenet Rémi and Hardy Adrien (2020). "Monte Carlo with Determinantal Point Processes". In: *Annals of Applied Probability*. URL: https://hal.archives-ouvertes.fr/hal-01311263.

Gillenwater Jennifer, Kulesza Alex, Mariet Zelda, and Vassilvtiskii Sergei (Sept. 2019). "A Tree-Based Method for Fast Repeated Sampling of Determinantal Point Processes". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2260–2268. URL: https://proceedings.mlr.press/v97/gillenwater19a.html.

Tremblay Nicolas, Barthelmé Simon, and Amblard Pierre-Olivier (2018). "Determinantal Point Processes for Coresets". In: DOI: 10.48550/ARXIV.1803.08700. URL: https://arxiv.org/abs/1803.08700.

Bachem Olivier, Lucic Mario, and Krause Andreas (2017). *Practical Coreset Constructions for Machine Learning*. DOI: 10.48550/ARXIV.1703.06476. URL: https://arxiv.org/abs/1703.06476.

Gautier Guillaume, Bardenet Rémi, and Valko Michal (June 2017). "Zonotope Hit-and-run for Efficient Sampling from Projection DPPs". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1223–1232. URL: https://proceedings.mlr.press/v70/gautier17a.html.

Zhang Cheng, Kjellstrom Hedvig, and Mandt Stephan (2017). *Determinantal Point Processes for Mini-Batch Diversification*. DOI: 10.48550/ARXIV.1705.00607. URL: https://arxiv.org/abs/1705.00607.

Braverman Vladimir, Feldman Dan, Lang Harry, Statman Adiel, and Zhou Samson (2016). *New Frameworks for Offline and Streaming Coreset Constructions*. DOI: 10.48550/ARXIV.1612.00889. URL: https://arxiv.org/abs/1612.00889.

Gao Wei, Niu Xin-Yi, and Zhou Zhi-Hua (2016). "Learnability of Non-I.I.D." In: *Proceedings of The 8th Asian Conference on Machine Learning*. Ed. by Robert J. Durrant and Kee-Eung Kim. Vol. 63. Proceedings of Machine Learning Research. The University of Waikato, Hamilton, New Zealand: PMLR, pp. 158–173. URL: https://proceedings.mlr.press/v63/Gao09.html.

Lucic Mario, Bachem Olivier, and Krause Andreas (2016). *Linear-time Outlier Detection via Sensitivity*. DOI: 10.48550/ARXIV.1605.00519. URL: https://arxiv.org/abs/1605.00519.

Novak Erich (2014). *Some Results on the Complexity of Numerical Integration*. DOI: 10.48550/ARXIV.1409.6714. URL: https://arxiv.org/abs/1409.6714.

Breuer Jonathan and Duits Maurice (2013). *The Nevai condition and a local law of large numbers for orthogonal polynomial ensembles*. DOI: 10.48550/ARXIV.1301.2061. URL: https://arxiv.org/abs/1301.2061.

Copenhaver Martin S., Kim Yeon Hyang, Logan Cortney, Mayfield Kyanne, Narayan Sivaram K., Petro Matthew J., and Sheperd Jonathan (2013). *Diagram vectors and Tight Frame Scaling in Finite Dimensions*. DOI: 10.48550/ARXIV.1303.1159. URL: https://arxiv.org/abs/1303.1159.

Kulesza Alex and Taskar Ben (2012). "Determinantal Point Processes for Machine Learning". In: *Foundations and Trends® in Machine Learning* 5.2–3, pp. 123–286. ISSN: 1935-8237. DOI: 10.1561/2200000044. URL: http://dx.doi.org/10.1561/2200000044.

Pemantle Robin and Peres Yuval (2011). *Concentration of Lipschitz functionals of determinantal and other strong Rayleigh measures*. DOI: 10.48550/ARXIV.1108.0687. URL: https://arxiv.org/abs/1108.0687.

Langberg Michael and Schulman Leonard (Jan. 2010). "Universal epsilon-approximators for integrals". In.

Simon Barry (2010). *Szegő's Theorem and its Descendants: Spectral Theory for L2 Perturbations of Orthogonal Polynomials: Spectral Theory for L2 Perturbations of Orthogonal Polynomials*. Princeton University Press.

Hough J. Ben, Krishnapur Manjunath, Peres Yuval, and Virág Bá lint (Jan. 2006). "Determinantal Processes and Independence". In: *Probability Surveys* 3.none. DOI: 10.1214/154957806000000078. URL: https://doi.org/10.1214%2F154957806000000078.

Gautschi Walter (2004). *Orthogonal Polynomials: Computation and Approximation*. Numerical Mathematics and Scientific Computation. Clarendon Press. ISBN: 0198506724. eprint: https://www.cs.purdue.edu/homes/wxg/OPmatlab.pdf.

Li Yi, Long Philip M., and Srinivasan Aravind (2001). "Improved Bounds on the Sample Complexity of Learning". In: *Journal of Computer and System Sciences* 62.3, pp. 516–527. ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.2000.1741. URL: https://www.sciencedirect.com/science/article/pii/S0022000000917410.

Haussler David (1995). "Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension". In: *Journal of Combinatorial Theory, Series A* 69.2, pp. 217–232. ISSN: 0097-3165. DOI: https://doi.org/10.1016/0097-3165(95)90052-7. URL: https://www.sciencedirect.com/science/article/pii/0097316595900527.

– (1992). "Decision theoretic generalizations of the PAC model for neural net and other learning applications". In: *Information and Computation* 100.1, pp. 78–150. ISSN: 0890-5401. DOI: https://doi.org/10.1016/0890-5401(92)90010-D. URL: https://www.sciencedirect.com/science/article/pii/089054019290010D.

Pollard David (1984). *Convergence of stochastic processes*. Springer Science & Business Media.

Valiant L. G. (Nov. 1984). "A Theory of the Learnable". In: *Commun. ACM* 27.11, pp. 1134–1142. ISSN: 0001-0782. DOI: 10.1145/1968.1972. URL: https://doi.org/10.1145/1968.1972.

Macchi Odile (1975). "The coincidence approach to stochastic point processes". In: *Advances in Applied Probability* 7.1, pp. 83–122. DOI: 10.2307/1425855.

Bakhvalov Nikolai Sergeevich (1959). "On the approximate calculation of multiple integrals". In: *Journal of Complexity* 31.4, pp. 502–516. ISSN: 0885-064X. DOI: https:

//doi.org/10.1016/j.jco.2014.12.003. URL: https://www.sciencedirect.com/science/article/pii/S0885064X14001204.