Determinantal Point Processes for Coresets

Nicolas Tremblay, Simon Barthelme, Pierre-Olivier Amblard

Determinantal Point Processes for Coresets

Nicolas Tremblay, Simon Barthelmé, and Pierre-Olivier Amblard

Abstract. When one is faced with a dataset too large to be used all at once, an obvious solution is to retain only part of it. In practice this takes a wide variety of different forms, and among them "coresets" are especially appealing. A coreset is a (small) weighted sample of the original data that comes with a guarantee: that a cost function can be evaluated on the smaller set instead of the larger one, with low relative error. For some classes of problems, and via a careful choice of sampling distribution, iid random sampling has turned to be one of the most successful methods for building coresets efficiently. However, independent samples are sometimes overly redundant, and one could hope that enforcing diversity would lead to better performance. The difficulty lies in proving coreset properties in non-iid samples. We show that the coreset property holds for samples formed with determinantal point processes (DPP). DPPs are interesting because they are a rare example of repulsive point processes with tractable theoretical properties, enabling us to prove general coreset theorems. We apply our results to both the k-means and the linear regression problem, and give extensive empirical evidence that the small additional computational cost of DPP sampling comes with superior performance over its iid counterpart.

1. Introduction

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 theory of "coresets" gives theoretical guarantees on the latter option. A coreset is a weighted sub-sample of the original data, with the guarantee that for any learning parameter, the task's cost function estimated on the coreset is equal to the cost computed on the entire dataset up to a controlled relative error.

An elegant consequence of such a property is that one may run learning algorithms solely on the coreset, allowing for a significant decrease in the computational cost while guaranteeing almost-equal performance. There are many algorithms that produce coresets (see for instance [1–4]), with some tailored for a specific task (such as k-means, k-medians, logistic regression, etc.), and others more generic [5]. Also, note that there are coreset results both in the streaming setting and the offline setting: we choose here to focus on the offline setting. The state of the art for many problems consists in tailoring a sampling distribution for the dataset at hand, and then sampling iid from that distribution [4, 6, 7]. However, iid samples are generally inefficient, as an iid process is ignorant of the past, and thus liable to sample similar points repeatedly. An avenue for improvement is to produce samples that are less redundant than what iid sampling produces.

To this end, we focus on Determinantal Point Processes (DPPs), processes known to produce diverse samples, and whose theoretical properties are well-characterized [8]. We show here that DPPs can be used to produce samples that are diverse, and possess the coreset property. Our theorems are quite generic, and assume mostly that the cost functions under study are Lipschitz. We have three main lines of argument: the first is that DPPs do indeed produce coresets, the second is that DPPs should produce *better* coresets (than iid methods) if one uses the right marginal kernel to define the DPP, and the third is an asymptotic rebalancing property of DPPs. We apply our results to both the k-means and the linear regression problems, for which optimal marginal kernels are unfortunately out of reach. We nevertheless propose two tractable approximations: the first one based on a Gaussian kernel and random Fourier features, and a second one based on polynomial features. We show that they both work well in practice, and improve over the iid paradigm on various artificial and real-world datasets.

1.1. Related work

Various coreset construction techniques have been proposed in the past. We follow the recent review [9] to classify them in four categories:

All three authors are with CNRS, Univ Grenoble-Alpes, Gipsa-lab, France.

1) Geometric decompositions [1, 2, 10, 11]. These methods propose to first discretize the ambient space of the data into a set of cells, snap each data point to its nearest cell in the discretization, and then use these weighted cells to approximate the target tasks. In all these constructions, the minimum number of samples required to guarantee the coreset property depends exponentially in the dimensionality of the ambient space, making them less useful in high-dimensional problems.
2) Gradient descent [3, 12–14]. These methods have been originally designed for the smallest enclosing ball problem (*i.e.*, finding the ball of minimum radius enclosing all datapoints), and have been later generalized to other problems. One of the main drawback of these algorithms in the k-means setting for instance is that their running time grows exponentially in the number of classes k [14]. Also, these algorithms provide only so-called *weak* coresets.
3) Random sampling [4,6,7,15,16]. The state of the art for many different tasks such as k-means or k-median is currently via iid random non-uniform sampling. The optimal probability distribution from which to sample datapoints should be set proportional to a quantity known as sensitivity and introduced by Langberg et al. [4]. In practice, it is unpractical to compute sensitivities: state of the art algorithms rely on bi-criteria approximations to find upper bounds, and set the probability distribution proportional to this upper bound. More details on these results are provided in Section 2.4.
4) Sketching and projections [17–24]. Another direction of research regarding data reduction that provably keeps the relevant information for a given learning task is via sketches [18]: compressed mappings (obtained via projections) of the original dataset that are in general easy to update with new or modified data. Sketches are not strictly speaking coresets, and the difference resides in the fact that coresets are subsets of the data, whereas sketches are projections of the data. Note finally that the frontier between the two is permeable and some data summaries may combine both.

1.2. Contributions

Our main contribution falls into the random sampling category, within which we propose to improve over iid sampling by considering negatively correlated point processes, *i.e.*, point processes for which sampling jointly two similar datapoints is less probable than sampling two very different datapoints. We decide to concentrate on a specific type of negatively-correlated sampling: determinantal point processes, known to provide samples that are representative of the "diversity" in the dataset [8]. To the best of our knowledge, we provide the first coreset guarantee using non-iid random sampling.

DPPs are parametrized by a positive semi-definite matrix called L-ensemble and denoted by L. This matrix encodes for the inclusion probabilities of each sample as well as higher order inclusion probabilities defining the correlation between samples. Note that DPPs have in general a random number of samples which in many practical situations is not adapted. This lead authors of [8] to define m-DPPs: DPPs conditioned to output m samples (for precise definitions related to DPPs and m-DPPs, see Section 2.6). It so happens that DPPs are more tractable than m-DPPs and some proofs are easier to show in the DPP context; however, m-DPPs are more useful in practice, especially when one needs to compare with fixed-size sampling methods as we do in this work. The reader should thus be mentally prepared to juggle from one concept to the other throughout the remainder of this paper.

On a theoretical side, our contributions are the following:

- Theorem 3.1 and B.1. Whatever the higher-order inclusion probabilities, if the inclusion probability of each sample of a DPP (or m-DPP) is set proportional to the sensitivity, then the results are at least as good as in the iid case. Technical limitations in controlling the concentration properties of correlated samples currently keep us from deriving exactly the minimum coreset size one may hope for when using DPPs.
- Theorem 3.4. A DPP sample necessarily has a lower variance than its (independent) Poisson counterpart with same inclusion probabilities.

- Theorem 3.5 and its Corollary 3.7. In the fixed-size context: a sample from an m-DPP with a rank m projective L-ensemble (also called projective DPP) necessarily leads to a lower variance than its iid counterpart with same inclusion probabilities.
- Theorem 3.9. Samples from a particular polynomial L-ensemble based on the Vandermonde matrix of the data asymptotically have a rebalancing property, made precise in Section 3.3. For instance in the k-means setting, this rebalancing property means that, asymptotically, such a DPP produces samples in each cluster, even if some are much smaller than others (see Fig.1 for an illustration).
- Lemmas D.1 and D.3. Of independent interest, we provide analytical formulas for the sensitivity in the 1-means and the linear regression cases.

On the applied side, we apply our theorems to both the k-means and the linear regression problems where the initial data consists in n points in $\mathbb{R}^d$. We discuss the ideal choice of L-ensemble L for DPP sampling in these cases. This ideal kernel being untractable in practice, we provide two heuristics: one based on random Fourier features of the Gaussian kernel, the other on polynomial features. We pay particular attention to the computation cost of these two heuristics, and provide implementation details. These heuristics output a coreset sample in respectively $\mathcal{O}(nm^2 + nmd)$ and $\mathcal{O}(nm^2)$ time where m is the number of samples of the coreset. In the k-means context, this is to compare to $\mathcal{O}(nkd)$ the cost of the current state of the art iid sampling algorithm via bi-criteria approximation. m being necessarily larger than d and k to obtain the coreset guarantee in this context, our proposition is computationally heavier, especially as m increases. We provide nonetheless extensive empirical evidence showing that this additional cost stays reasonable given the enhanced performances it provides.

Finally, a Julia toolbox called DPP4Coresets is available on the authors' website[1].

1.3. Organization of the paper

The paper is organized as follows. Section 2 recalls the background: the types of learning problems under consideration, the formal definition of coresets, sensitivities and DPPs. The theoretical Section 3 presents our main theorems on the performance of DPPs for coreset sampling: while Section 3.1 details coreset performance in the usual formulation of coreset theorems, Section 3.2 shows general variance arguments in favor of DPPs, and finally Section 3.3 provides an original asymptotic rebalancing property of DPPs. Section 4.1 shows how these theorems are applicable to both the k-means and the linear regression problems. We provide in Section 5 a discussion on the choice of marginal kernel adapted to these problems, and detail our sampling algorithms. Finally, the empirical Section 6 presents experiments on artificial as well as real-world datasets comparing the performance of DPP sampling to iid sampling. Section 7 concludes the paper. Note that for the sake of readability, many proofs and some implementation details are pushed to the Appendix.

2. Background

Let $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ be a set of n datapoints. Let (Θ, d_Θ) be a metric space of parameters, and θ an element of Θ. We consider cost functions of the form:

$$L(\mathcal{X}, \theta) = \sum_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}, \theta), \tag{1}$$

where f is a non-negative γ-Lipschitz function ($\gamma > 0$) with respect to θ, *i.e.*, $\forall \boldsymbol{x} \in \mathcal{X}$:

$$\forall \theta \in \Theta \qquad f(\boldsymbol{x}, \theta) \geqslant 0, \tag{2}$$

$$\forall (\theta, \theta') \in \Theta^2 \qquad |f(\boldsymbol{x}, \theta) - f(\boldsymbol{x}, \theta')| \leqslant \gamma \; d_\Theta(\theta, \theta'). \tag{3}$$

Many classical machine learning cost functions fall under this model: k-means, k-median, logistic or linear regression, support-vector machines, etc.

[1] or at `https://gricad-gitlab.univ-grenoble-alpes.fr/tremblan/dpp4coresets.jl`

2.1. Problem considered

A standard learning task is to minimize the cost L over all $\theta \in \Theta$. We write:

$$\theta^{\text{opt}} = \operatorname*{argmin}_{\theta\in\Theta} L(\mathcal{X},\theta), \qquad L^{\text{opt}} = L(\mathcal{X},\theta^{\text{opt}}) \qquad \text{and} \qquad \langle f\rangle_{\text{opt}} = \frac{L^{\text{opt}}}{n}. \tag{4}$$

In some instances of this problem, e.g., if n is very large and/or if f is expensive to evaluate and should be computed as rarely as possible, one may rely on sampling strategies to efficiently perform this optimization task.

2.2. Coresets

Let $\mathcal{S} = \{\boldsymbol{x}_{s_1},\dots,\boldsymbol{x}_{s_m}\}$ be a subset of $\mathcal{X}$ (possibly with repetitions). To each element $\boldsymbol{x}_s \in \mathcal{S}$, associate a weight $\omega(\boldsymbol{x}_s) \in \mathbb{R}^+$. Define the estimated cost associated to the weighted subset $\mathcal{S}$ as:

$$\hat{L}(\mathcal{S},\theta) = \sum_{\boldsymbol{x}_s\in\mathcal{S}} \omega(\boldsymbol{x}_s) f(\boldsymbol{x}_s,\theta). \tag{5}$$

Definition 2.1 (Coreset)**.** *Let $\epsilon \in (0,1)$. The weighted subset $\mathcal{S}$ is a ϵ-coreset for L if,* for any parameter θ*, the estimated cost is equal to the exact cost up to a relative error:*

$$\forall\theta\in\Theta \qquad \left|\frac{\hat{L}}{L} - 1\right| \leqslant \epsilon. \tag{6}$$

This is the so-called "strong" coreset definition, as the ϵ-approximation is required for all $\theta \in \Theta$. A weaker version of this definition exists in the literature where the ϵ-approximation is only required for θ^{opt}. In the following, we focus on theorems guaranteeing the strong coreset property.

Let us write $\hat{\theta}^{\text{opt}}$ the optimal solution computed on the weighted subset $\mathcal{S}$: $\hat{\theta}^{\text{opt}} = \operatorname{argmin}_{\theta\in\Theta} \hat{L}(\mathcal{S},\theta)$. An important consequence of the coreset property is the following:

$$(1-\epsilon)L(\mathcal{X},\theta^{\text{opt}}) \leqslant (1-\epsilon)L(\mathcal{X},\hat{\theta}^{\text{opt}}) \leqslant \hat{L}(\mathcal{S},\hat{\theta}^{\text{opt}}) \leqslant \hat{L}(\mathcal{S},\theta^{\text{opt}}) \leqslant (1+\epsilon)L(\mathcal{X},\theta^{\text{opt}}),$$

i.e., running an optimization algorithm on the weighted sample $\mathcal{S}$ will result in a minimal learning cost that is a controlled ϵ-approximation of the learning cost one would have obtained by running the same algorithm on the entire dataset $\mathcal{X}$. Note that the guarantee is over costs only: the estimated optimal parameters $\hat{\theta}^{\text{opt}}$ and θ^{opt} may be different. Nevertheless, if the cost function is well suited to the problem: either there is one clear global minimum and the estimated parameters will almost coincide; or there are multiple solutions for which the learning cost is similar and selecting one over the other is not an issue.

In terms of computation cost, if the sampling scheme is efficient, n is very large and/or f is expensive to compute for each datapoint, coresets thus enable a significant gain in computing time.

2.3. Sensitivity

To define appropriate sampling schemes for coresets, Langberg and Schulman [4] introduce the notion of sensitivity:

Definition 2.2 (Sensitivity)**.** *The sensitivity of a datapoint $\boldsymbol{x}_i \in \mathcal{X}$ with respect to $f : \mathcal{X},\Theta \to \mathbb{R}^+$ is:*

$$\sigma_i = \max_{\theta\in\Theta} \frac{f(\boldsymbol{x}_i,\theta)}{L(\mathcal{X},\theta)} \qquad \in [0,1]. \tag{7}$$

Also, the total sensitivity is defined as :

$$\mathfrak{S} = \sum_{i=1}^{n} \sigma_i. \tag{8}$$

Note that the fraction defining the sensitivity is not defined for $L(\mathcal{X},\theta) = 0$ (that may happen for instance in the 1-means problem, if all $\boldsymbol{x}_i$ and θ are all equal). For simplicity, we suppose that $\forall\theta\in\Theta,\ L(\mathcal{X},\theta) > 0$.

The sensitivity is related to the concept of statistical leverage score [25], which plays a crucial role in iid random sampling theorems in the randomized numerical linear algebra literature [19]. Both notions are similar, but not equivalent. For instance, we show in Lemma D.3 that sensitivities in the linear regression problem are different from the usual definition of leverage score in this context.

In words, the sensitivity σ_i is the worse case contribution of datapoint x_i in the total cost. Intuitively, the larger it is, the larger its "outlierness" [26].

2.4. iid importance sampling and state of the art results

In the iid sampling paradigm, the importance sampling estimator of L is the following. Say the sample set $\mathcal{S}$ consists in m samples drawn iid with replacement from a (discrete) probability distribution $\boldsymbol{p} \in \mathbb{R}^n$ (with p_i the probability of sampling $\boldsymbol{x}_i$ at each draw, and $\sum_i p_i = 1$). Denote by ϵ_i the random variable counting the number of occurences of $\boldsymbol{x}_i$ in $\mathcal{S}$. One may define $\hat{L}_{\text{iid}}$, the so-called importance sampling estimator of L, as :

$$\hat{L}_{\text{iid}}(\mathcal{S}, \theta) = \sum_i \frac{f(\boldsymbol{x}_i, \theta)\epsilon_i}{mp_i}. \tag{9}$$

One can show that $\mathbb{E}(\epsilon_i) = mp_i$, such that $\hat{L}_{\text{iid}}$ is an unbiased estimator of L:

$$\mathbb{E}(\hat{L}_{\text{iid}}(\mathcal{S}, \theta)) = L(\mathcal{X}, \theta). \tag{10}$$

The concentration of $\hat{L}_{\text{iid}}$ around its expected value is controlled by the following state of the art theorem:

Theorem 2.3 (Coresets with iid random sampling)**.** *Let $\boldsymbol{p} \in [0,1]^n$ be a probability distribution over all datapoints $\mathcal{X}$ with p_i the probability of sampling $\boldsymbol{x}_i$ and $\sum_i p_i = 1$. Draw m iid samples with replacement according to $\boldsymbol{p}$. Associate to each sample $\boldsymbol{x}_s$ a weight $\omega(\boldsymbol{x}_s) = 1/mp_s$. The weighted subset obtained is a ϵ-coreset with probability $1-\delta$ provided that:*

$$m \geqslant m^* \tag{11}$$

with

$$m^* = \mathcal{O}\left(\frac{1}{\epsilon^2}\left(\max_i \frac{\sigma_i}{p_i}\right)^2 (d' + \log(1/\delta))\right), \tag{12}$$

where d' is the pseudo-dimension of Θ (a generalization of the Vapnik-Chervonenkis dimension). The optimal probability distribution minimizing m^* is $p_i = \sigma_i/\mathfrak{S}$. In this case, the weighted subset is a ϵ-coreset with probability $1-\delta$ provided that:

$$m \geqslant \mathcal{O}\left(\frac{\mathfrak{S}^2}{\epsilon^2}(d' + \log(1/\delta))\right). \tag{13}$$

For instance, in the k-means setting[2], $d' = dk\log k$ and $\mathfrak{S} = \mathcal{O}(k)$ such that the coreset property is guaranteed with probability $1-\delta$ provided that:

$$m \geqslant \mathcal{O}\left(\frac{k^2}{\epsilon^2}(dk\log k + \log(1/\delta))\right). \tag{14}$$

This theorem is taken from [16] but its original form goes back to [4]. Note that sensitivities cannot be computed rapidly, such that, as it is, this theorem is unpractical. Thankfully, bi-criteria approximation schemes (such as Alg. 2 of [16], or other propositions such as in [15, 28]) may be used to efficiently find an upper bound of the sensitivity for all i: $b_i \geqslant \sigma_i$. Noting $B = \sum b_i$, and setting $p_i = b_i/B$, one shows that the coreset property may be guaranteed in the iid framework provided that $m \geqslant \mathcal{O}\left(\frac{B^2}{\epsilon^2}(d' + \log(1/\delta))\right)$. This idea of using bi-criteria approximations to upper bound the sensitivity also goes back to [4] and has been used in many works on coresets [5, 7, 15, 28].

[2] In the literature [15,27], d' is often taken to be equal to dk. We nevertheless agree with [16] and their discussion in Section 2.6 regarding k-means' pseudo-dimension and thus write $d' = dk\log k$

Note that if one authorizes coresets with negative weights (that is, authorizes negative weights in the estimated cost equation 5), then the above theorem may be further improved [15]. Nevertheless, we prefer to restrict ourselves to positive weights as optimization algorithms such as Lloyd's k-means heuristics [29] are in practice more straightforward to implement on positively weighted sets rather than on sets with possibly negative weights.

Finally, Braverman et al. (Theorem 5.5 of [7]) improve the previous theorem by showing that under the same non-uniform iid framework, the coreset property is guaranteed provided that $m \geqslant \mathcal{O}\left(\frac{\mathfrak{S}}{\epsilon^2}(d' \log \mathfrak{S} + \log(1/\delta))\right)$, thus reducing the term in $\mathfrak{S}^2$ to $\mathfrak{S} \log \mathfrak{S}$.

2.5. Correlated importance sampling

Eq. (9) is not suited to correlated sampling and, in the following, we will use a slightly different importance sampling estimator, more adapted to this case. Consider a point process defined on $\mathcal{X}$ that outputs a random sample $\mathcal{S} \subset \mathcal{X}$. For each data point $\boldsymbol{x}_i$, denote by π_i its inclusion (or marginal) probability:

$$\pi_i = \mathbb{P}\left(\boldsymbol{x}_i \in \mathcal{S}\right). \tag{15}$$

Moreover, denote by ϵ_i the random Boolean variable such that $\epsilon_i = 1$ if $\boldsymbol{x}_i \in \mathcal{S}$, and 0 otherwise. In this paper, we focus on the following definition of the importance sampling cost estimator $\hat{L}$[3]:

$$\hat{L}(\mathcal{S}, \theta) = \sum_i \frac{f(\boldsymbol{x}_i, \theta)\epsilon_i}{\pi_i}. \tag{16}$$

By construction, $\mathbb{E}(\epsilon_i) = \pi_i$, such that $\hat{L}$ is an unbiased estimator of L:

$$\mathbb{E}(\hat{L}(\mathcal{S}, \theta)) = L(\mathcal{X}, \theta). \tag{17}$$

Studying the coreset property in this setting boils down to studying the concentration properties of $\hat{L}$ around its expected value.

2.6. Determinantal Point Processes

In order to induce negative correlations within the samples, we choose to focus on Determinantal Point Processes (DPP), point processes that have recently gained attention due to their ability to output "diverse" subsets within a tractable probabilistic framework (for instance with explicit formulas for marginal probabilities). In the following, $2^{[n]}$ denotes the set of all possible subsets of the n first integers.

The central object is called the L-ensemble, and is nothing else than a positive semi-definite matrix $\mathsf{L} \in \mathbb{R}^{n \times n}$. We will write its eigenvalues $0 \leqslant \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n$.

Definition 2.4 (DPP [8])**.** *Consider a point process, i.e., a process that randomly draws an element $\mathcal{S} \in 2^{[n]}$. It is determinantal with L-ensemble L if*

$$\mathbb{P}(\mathcal{S}) = \frac{\det(\mathsf{L}_\mathcal{S})}{\det(\mathsf{I} + \mathsf{L})},$$

where $\mathsf{L}_\mathcal{S}$ is the restriction of L to the rows and columns indexed by the elements of $\mathcal{S}$.

The following well-known properties are verified (see [8] for details):

- one can indeed show that the normalization is proper: $\sum_\mathcal{S} \det(\mathsf{L}_\mathcal{S}) = \det(\mathsf{I} + \mathsf{L})$.
- all inclusion probabilities, at any order, are explicit:

$$\forall \mathcal{A} \in 2^{[n]} \qquad \mathbb{P}(\mathcal{A} \subseteq \mathcal{S}) = \det(\mathsf{K}_\mathcal{A})$$

  where $\mathsf{K} = \mathsf{L}(\mathsf{I} + \mathsf{L})^{-1} \in \mathbb{R}^{n \times n}$ is called the marginal kernel. In particular, the probability of inclusion of i, π_i, is equal to K_{ii}. Also, to gain insight in the repulsive nature of DPPs, one

[3]Note that in fact $\hat{L}_{\text{iid}}$ and $\hat{L}$ are the same objects if one defines ϵ_i to be the number of times i is sampled (which will be in practice Boolean in the DPP case as the same sample can never be sampled twice in this context) and write $\hat{L}(\mathcal{S}, \theta) = \sum_i \frac{f(\boldsymbol{x}_i, \theta)\epsilon_i}{\mathbb{E}(\epsilon_i)}$. We prefer to introduce both notations to avoid confusions.

may readily see that the joint marginal probability of sampling i and j reads: $\det(\mathsf{K}_{\{i,j\}}) = \pi_i\pi_j - \mathsf{K}_{ij}^2$ and is necessarily smaller than $\pi_i\pi_j$, the joint probability in the case of Poisson uncorrelated sampling. The stronger the "interaction" between i and j (encoded by the absolute value of element K_{ij}), the smaller the probability of sampling both jointly: this determinantal nature thus favors diverse sets of samples.
- K is also positive semi-definite. The eigenvalues of K are $\{\frac{\lambda_i}{1+\lambda_i}\}_i$ and are necessarily between 0 and 1.
- it can be shown that the number of samples of a DPP is itself random and distributed as a sum of Bernoulli parametrized by the eigenvalues of K. In particular, the expected number of samples is $\mu = \mathrm{Tr}(\mathsf{K}) = \sum_i \frac{\lambda_i}{1+\lambda_i}$.

In many cases, one prefers to specify deterministically the number of samples, instead of having a random number of them. This leads to m-DPPs: DPPs conditioned to output m samples.

Definition 2.5 (m-DPP [8])**.** *Consider a point process that randomly draws an element $\mathcal{S} \in 2^{[n]}$. This process is an m-DPP with L-ensemble L if:*

i) $\forall \mathcal{S}$ *s.t.* $|\mathcal{S}| \neq m,\ \mathbb{P}(\mathcal{S}) = 0$
ii) $\forall \mathcal{S}$ *s.t.* $|\mathcal{S}| = m,\ \mathbb{P}(\mathcal{S}) = \frac{1}{Z}\det(\mathsf{L}_{\mathcal{S}})$ *with Z the normalization constant.*

The following properties hold:

- the normalization constant Z is in fact the m-th order elementary symmetric polynomial of the eigenvalues of L:
$$Z = \sum_{\mathcal{S}' \text{ s.t. } |\mathcal{S}'|=m} \det(\mathsf{L}_{\mathcal{S}'}) \ = \ e_m(\lambda_1, \ldots, \lambda_n) = \sum_{1 \leqslant j_1 < j_2 < \cdots < j_m \leqslant n} \lambda_{j_1} \cdots \lambda_{j_m}.$$
- in general, m-DPPs are not DPPs: for instance the probability of including element i, π_i, is no longer K_{ii} in general. In fact, one has $\pi_i = \frac{1}{Z}\sum_{\mathcal{S}' \text{ s.t } |\mathcal{S}'|=m \text{ and } i \in \mathcal{S}'} \det(\mathsf{L}_{\mathcal{S}'})$.
- by construction, $\sum_i \pi_i = m$.

Let us define the specific but important case of projective DPPs.

Definition 2.6 (projective-DPP)**.** *A projective DPP is a m-DPP whose L-ensemble is a projection of rank m:*

$$\mathsf{L} = \mathsf{U}\mathsf{U}^\top \tag{18}$$

where $\mathsf{U} \in \mathbb{R}^{n \times m}$ has orthonormal columns (i.e., $\mathsf{U}^\top\mathsf{U} = \mathsf{I}_m$).

Lemma 2.7 (Lemma 1.3 of [30])**.** *A projective DPP with L-ensemble L is also a DPP, with* marginal kernel L.

In fact, the set of projective DPPs is precisely the intersection between the set of DPPs and the set of m-DPPs. Projective DPPs are very practical objects: they have both the practical convenience of m-DPPs (a fixed number of samples) and the theoretical convenience of DPPs (for instance, π_i is simply L_{ii}, *i.e.*, the sum of squares of the i-th line of U).

3. Coreset theorems

We now detail our main theoretical contributions. In Section 3.1, we present a coreset theorem for m-DPPs providing sufficient conditions on the marginal probabilities $\{\pi_i\}_i$ to guarantee the coreset property. We will see that, similar to the iid case (Theorem 2.3), the optimal marginal probability should be set proportional to the sensitivity. A similar result is derived for DPPs in Appendix B. These theorems are valid for any choice of higher order inclusion probabilities. We further discuss in Section 3.2 how one may take advantage of these additional degrees of freedom encoding the negative correlations of DPPs to improve the coreset performance over iid sampling. Finally, in Section 3.3, we show that a particular polynomial projective DPP asymptotically verifes a rebalancing property, thus making them natural candidates for the coreset problem.

3.1. m-Determinantal Point Processes for coresets

Theorem 3.1 (m-DPP for coresets)**.** *Let $\mathcal{S}$ be a sample from an m-DPP with L-ensemble L, $\epsilon \in (0,1)$, and η the minimal number of balls of radius $\frac{\epsilon \langle f \rangle_{opt}}{6\gamma}$ necessary to cover Θ. $\mathcal{S}$ is a ϵ-coreset with probability larger than $1-\delta$ provided that:*

$$m \geqslant m^* = \max(m_1^*, m_2^*) \tag{19}$$

with:

$$m_1^* = \frac{32}{\epsilon^2} \left(\max_i \frac{\sigma_i}{\bar{\pi}_i} \right)^2 \log \frac{4\eta}{\delta}, \tag{20}$$

$$m_2^* = \frac{32}{\epsilon^2} \left(\frac{1}{n\bar{\pi}_{min}} \right)^2 \log \frac{4\eta}{\delta}, \tag{21}$$

and $\forall i, \bar{\pi}_i = \pi_i / m$.

The proof is provided in Appendix A. Note that m_1^* and m_2^* are not independent of m: they are in fact dependent via $\bar{\pi}_i = \pi_i/m$. While this formulation may be surprising at first, this is due to the fact that in non-iid settings, separating m from π_i is not as straightforward as in the iid case (in Theorem 2.3, m and p_i are independent) . Also, we give this particular formulation of the theorem to mimic classical concentration results obtained with iid sampling.

In order to simplify further analysis, we suppose from now on that $n\sigma_{\min} \geqslant 1$. As shown in the second lemma of Appendix D, this is in fact verified in the k-means case for instance. Nevertheless, the following results may be generalized to cases with unconstrained $\sigma_{\min}$ if needed, with little effects on the main results.

Corollary 3.2. *If $n\sigma_{min} \geqslant 1$, then $m_1^* \geqslant m_2^*$ and the coreset property of Theorem 3.1 is verified if:*

$$m \geqslant m^* = \frac{32}{\epsilon^2} \left(\max_i \frac{\sigma_i}{\bar{\pi}_i} \right)^2 \log \frac{4\eta}{\delta} \tag{22}$$

with $\forall i,\ \bar{\pi}_i = \pi_i/m$.

Proof. Denote by j the index for which $\bar{\pi}_i$ is minimal and, provided that $n\sigma_{\min} \geqslant 1$, one has:

$$\max_i \frac{\sigma_i}{\bar{\pi}_i} n\bar{\pi}_{\min} \geqslant n\sigma_j \geqslant n\sigma_{\min} \geqslant 1,$$

which implies $m^* = \max(m_1^*, m_2^*) = m_1^*$. □

One would like to have the coreset guarantee for a minimal number of samples, that is: to find the marginal probabilities π_i minimizing m^*. A quick glance at Eq. (22) tells us to set $\pi_i = m\sigma_i/\mathfrak{S}$ in order to minimize the bound m^* while satisfying the constraint $\sum_i \pi_i = m$. In practice, however, computing the sensitivities is often untractable. We thus propose to set the marginal probabilities according to the following looser condition.

Corollary 3.3. *If one sets the π_i's such that there exists $\alpha > 0$ and $\beta \geqslant 1$ verifying:*

$$\forall i \qquad \alpha\sigma_i \leqslant \pi_i \leqslant \alpha\beta\sigma_i, \tag{23}$$

$$\text{and} \qquad \frac{\alpha}{\beta} \geqslant \frac{32}{\epsilon^2} \mathfrak{S} \log \frac{4\eta}{\delta}, \tag{24}$$

then $\mathcal{S}$ is a ϵ-coreset with probability at least $1-\delta$. In this case, the number of samples verifies:

$$m \geqslant \frac{32}{\epsilon^2} \beta \mathfrak{S}^2 \log \frac{4\eta}{\delta}.$$

Proof. Let us suppose that the marginal probabilities π_i are set such that there exists $\alpha > 0$ and $\beta \geqslant 1$ verifying:

$$\forall i, \qquad \alpha\sigma_i \leqslant \pi_i \leqslant \alpha\beta\sigma_i.$$

Note that:

$$\left(\max_i \frac{\sigma_i}{\pi_i}\right)^2 m \leqslant \frac{m}{\alpha^2} = \frac{1}{\alpha^2}\sum_i \pi_i \leqslant \frac{\beta}{\alpha}\sum_i \sigma_i = \frac{\beta}{\alpha}\mathfrak{S}.$$

Thus, the inequality

$$\frac{\alpha}{\beta} \geqslant \frac{32}{\epsilon^2}\mathfrak{S}\log\frac{4\eta}{\delta}$$

implies:

$$1 \geqslant \frac{32}{\epsilon^2}\left(\max_i \frac{\sigma_i}{\pi_i}\right)^2 m \log\frac{4\eta}{\delta},$$

that we recognize as the coreset condition (22) by multiplying on both sides by m: $\mathcal{S}$ is indeed a ϵ-coreset with probability larger than $1-\delta$. Moreover, in this case:

$$m = \sum_i \pi_i \geqslant \alpha \sum_i \sigma_i = \alpha\mathfrak{S} \geqslant \frac{32}{\epsilon^2}\beta\mathfrak{S}^2 \log\frac{4\eta}{\delta}.$$

□

Corollary 3.3 is applicable to cases where $\sigma_{\max}$ is not too large. In fact, in order for $\alpha\sigma_i$ to be smaller than π_i, and thus smaller than 1 as π_i is a probability, α should always be set inferior to $\frac{1}{\sigma_{\max}}$. Now, if $\sigma_{\max}$ is so large that $\frac{1}{\sigma_{\max}} \leqslant \frac{32}{\epsilon^2}\mathfrak{S}\log\frac{4\eta}{\delta}$, then, even by setting β to its minimum value 1, there is no admissible α verifying both conditions (23) and (24). We refer to Appendix E for a simple workaround if this issue arises. We will further see in the experimental section (Section 6) that outliers are not an issue in practice.

Similar results are obtained for DPPs (instead of m-DPPs) in Appendix B.

3.2. Links with the iid case and variance arguments

Let us first compare our results with Theorem 2.3 obtained in the iid setting. A few remarks are in order:

1) setting β and α to 1 in Corollary 3.3, that is, setting each π_i exactly to σ_i, the minimum number of required samples is $\frac{32\mathfrak{S}^2}{\epsilon^2}(\log\eta + \log\frac{4}{\delta})$, to compare to $\mathcal{O}(\frac{\mathfrak{S}^2}{\epsilon^2}(d' + \log(1/\delta)))$ of Theorem 2.3, where d' is the pseudo-dimension of Θ. η being the number of balls of radius $\frac{\epsilon\langle f\rangle_{\text{opt}}}{6\gamma}$ necessary to cover Θ, it will typically be $\frac{\epsilon\langle f\rangle_{\text{opt}}}{6\gamma}$ to the power of the ambient dimension of Θ (similar to d'). Both forms are very similar, up to the dependency in ϵ and in $\langle f\rangle_{\text{opt}}$ of the log term. This difference is due to the fact that coreset theorems in the iid case (see for instance [16]) take advantage of powerful results from the Vapnik-Chervonenkis (VC) theory, as detailed in [31]. Unfortunately, these fundamental results are valid in the iid case only, and are not easily generalized to the correlated case. Further work should enable to reduce this small gap, by taking advantage of chaining arguments in correlated contexts such as in [32].
2) Outliers are not naturally dealt with using our proof techniques, mainly due to our multiple use of the union bound that necessarily englobes the worse-case scenario. In fact, in the importance sampling estimator used in the iid case (Eq. 9), outliers are not problematic as they can be sampled several times. In our setting, outliers are constrained to be sampled only once, which in itself makes sense, but complicates the analysis. Empirically, we will see in Section 6 that outliers are not an issue.
3) The DPP coreset theorems obtained are in a sense disappointing: they do not show that the concentration is tighter in the DPP case than in the iid case. They are in fact limited by the current state-of-the-art in concentration of strongly Rayleigh measures [33]. On the bright side, our results take *only* into account singleton inclusion probabilities: the $\{\pi_i\}$'s; meaning that these DPP sampling theorems are valid for any choice of higher-order inclusion probabilities (encoding the correlation between samples). We will now see how these extra

degrees of freedom enable to provably decrease the variance of the cost estimator, compared to the iid case.

3.2.1. A first variance argument: improvement over the Poisson point process

Consider a DPP with marginal kernel K. Build the diagonal kernel K_d with $\mathsf{K}_d(i,i) = \mathsf{K}(i,i)$. Note that a DPP from K_d reduces to a Poisson point process. Note also that marginal probabilities π_i of both processes (and consequently their expected number of samples) are the same. We compare the variance of the estimator $\hat{L}$ obtained with a DPP with marginal kernel K versus its variance obtained with its Poisson uncorrelated counterpart: a DPP with marginal kernel K_d.

Theorem 3.4. *For any admissible marginal kernel* K *(i.e., positive semi-definite with eigenvalues between* 0 *and* 1*), we have:*

$$\forall \theta \in \Theta \qquad \mathrm{Var}(\hat{L}) = \mathrm{Var}_d - \sum_{i \neq j} \frac{\mathsf{K}_{ij}^2}{\pi_i \pi_j} f(\boldsymbol{x}_i, \theta) f(\boldsymbol{x}_j, \theta) \tag{25}$$

where Var_d *is the variance of the estimator based on the diagonal DPP. As the function* f *is positive, the variance of* $\hat{L}$ *via DPP sampling with kernel* K *is thus necessarily smaller than its Poisson counterpart with same inclusion probabilities.*

Proof. We have:

$$\mathrm{Var}(\hat{L}) = \mathbb{E}(\hat{L}^2) - \mathbb{E}(\hat{L})^2 \tag{26}$$

$$= \sum_{i,j} \frac{\mathbb{E}(\epsilon_i \epsilon_j)}{\pi_i \pi_j} f(\boldsymbol{x}_i, \theta) f(\boldsymbol{x}_j, \theta) - L^2. \tag{27}$$

As $\mathcal{S}$ is sampled from a DPP, the following is verified. If $i \neq j$, $\mathbb{E}(\epsilon_i \epsilon_j) = \det(\mathsf{K}_{\{i,j\}}) = \pi_i \pi_j - \mathsf{K}_{ij}^2$. If $i = j$, $\mathbb{E}(\epsilon_i \epsilon_j) = \mathbb{E}(\epsilon_i) = \pi_i$. One obtains:

$$\mathrm{Var}(\hat{L}) = \sum_i \left(\frac{1}{\pi_i} - 1 \right) f(x_i, \theta)^2 - \sum_{i \neq j} \frac{\mathsf{K}_{ij}^2}{\pi_i \pi_j} f(\boldsymbol{x}_i, \theta) f(\boldsymbol{x}_j, \theta). \tag{28}$$

The first term of the rhs is in fact the variance in the case of a diagonal kernel: $\sum_i \left(\frac{1}{\pi_i} - 1 \right) f(\boldsymbol{x}_i, \theta)^2 = \mathrm{Var}_d$, finishing the proof. □

The important message here is that this variance reduction occurs *regardless* of the choice of K's off-diagonal elements: any choice –provided that $0 \preceq \mathsf{K} \preceq 1$ stays true– will reduce the variance.

Proving such a variance reduction when comparing a m-DPP with L-ensemble L versus its conditional Poisson equivalent (a Poisson point process conditioned to m samples, with same $\{\pi_i\}$) is much more involved, and remains open.

3.2.2. A second variance argument: improvement over the iid estimator with replacement

We now compare the variance of the iid estimator with replacement $\hat{L}_{\mathrm{iid}}$ of Eq. (9) and the variance of the DPP estimator $\hat{L}$ of Eq. (16). Consider a DPP with marginal kernel K, with $\forall i \quad \pi_i = \mathsf{K}_{ii}$ the marginal probability of sampling element i such that the expected number of samples $\mu = \sum_i \pi_i$ is an integer. We compare the variance of $\hat{L}$ with such a DPP and the variance of $\hat{L}_{\mathrm{iid}}$ with μ independent draws with replacement with $p_i = \pi_i / \mu$ (in order to have a fair comparison).

Before we state the result, suppose that K is of rank r (with, necessarily, $\mu \leqslant r \leqslant n$). K being positive-semi definite and of rank r, there exists $\mathsf{V} = (\boldsymbol{v}_1 | \boldsymbol{v}_2 | \dots | \boldsymbol{v}_n) \in \mathbb{R}^{r \times n}$ a set of n vectors in dimension r such that $\mathsf{K} = \mathsf{V}^\top \mathsf{V}$. By construction, $\forall i \quad \|\boldsymbol{v}_i\|^2 = \mathsf{K}_{ii} = \pi_i$. For each vector $\boldsymbol{v}$, consider

its diagram vector (see Definition 2.3 of [34]), denoted $\tilde{\boldsymbol{v}}$, defined as:

$$\tilde{\boldsymbol{v}} = \frac{1}{\sqrt{r-1}} \begin{bmatrix} v(1)^2 - v(2)^2 \\ \vdots \\ v(r-1)^2 - v(r)^2 \\ \sqrt{2r}\, v(1)v(2) \\ \vdots \\ \sqrt{2r}\, v(r-1)v(r) \end{bmatrix} \in \mathbb{R}^{r(r-1)}, \tag{29}$$

where the difference of squares $v(i)^2 - v(j)^2$ and the product $v(i)v(j)$ occur exactly once for $i < j, i = 1, 2, \cdots, r-1$.

Theorem 3.5. *One has:*

$$\mathrm{Var}(\hat{L}) = \mathrm{Var}(\hat{L}_{\mathrm{iid}}) + \left(\frac{1}{\mu} - \frac{1}{r}\right) L^2 - \frac{r-1}{r} \left\| \sum_i \frac{f(\boldsymbol{x}_i, \theta)}{\pi_i} \tilde{\boldsymbol{v}}_i \right\|^2. \tag{30}$$

Proof. In the iid case,

$$\mathbb{E}(\hat{L}_{\mathrm{iid}}^2) = \sum_{i=1}^n \sum_{j=1}^n \frac{f(\boldsymbol{x}_i, \theta) f(\boldsymbol{x}_j, \theta) \mathbb{E}(\epsilon_i \epsilon_j)}{\mu^2 p_i p_j} \tag{31}$$

where ϵ_i is not Boolean but counts the number of times i is sampled. One can show that if $i \neq j$, $\mathbb{E}(\epsilon_i \epsilon_j) = p_i p_j (\mu^2 - \mu)$, and if $i = j$, $\mathbb{E}(\epsilon_i \epsilon_j) = p_i \mu + p_i^2 \mu^2 - \mu p_i^2$. Thus:

$$\mathrm{Var}(\hat{L}_{\mathrm{iid}}) = \mathbb{E}(\hat{L}_{\mathrm{iid}}^2) - L^2 = \sum_{i=1}^n \sum_{j \neq i} f(x_i, \theta) f(x_j, \theta)(1 - 1/\mu) + \sum_{i=1}^n f(x_i, \theta)^2 \frac{1 + p_i \mu - p_i}{\mu p_i} - L^2 \tag{32}$$

$$= \frac{1}{\mu} \sum_{i=1}^n \frac{f(\boldsymbol{x}_i, \theta)^2}{p_i} - \frac{1}{\mu} L^2 \tag{33}$$

Moreover:

$$\mathrm{Var}(\hat{L}) = \sum_i \frac{f(x_i, \theta)^2}{\pi_i} - \sum_i \sum_j \frac{f(\boldsymbol{x}_i, \theta) f(\boldsymbol{x}_j, \theta)}{\pi_i \pi_j} \mathsf{K}_{ij}^2. \tag{34}$$

Thus:

$$\mathrm{Var}(\hat{L}) = \mathrm{Var}(\hat{L}_{\mathrm{iid}}) + \frac{1}{\mu} L^2 - \sum_i \sum_j \frac{f(\boldsymbol{x}_i, \theta) f(\boldsymbol{x}_j, \theta)}{\pi_i \pi_j} \mathsf{K}_{ij}^2 \tag{35}$$

Proposition 2.5 of [34] states:

$$\forall (i,j) \qquad \mathsf{K}_{ij}^2 = \left(\boldsymbol{v}_i^\top \boldsymbol{v}_j\right)^2 = \frac{1}{r} \left\|\boldsymbol{v}_i\right\|^2 \left\|\boldsymbol{v}_j\right\|^2 + \frac{r-1}{r} \tilde{\boldsymbol{v}}_i^\top \tilde{\boldsymbol{v}}_j \tag{36}$$

$$= \frac{1}{r} \pi_i \pi_j + \frac{r-1}{r} \tilde{\boldsymbol{v}}_i^\top \tilde{\boldsymbol{v}}_j. \tag{37}$$

Replacing this in Eq. (35) yields the desired result. □

Remark 3.6. *The variance of the DPP estimator is partly due to the fact that the number of samples is random, which is not the case with the iid scheme we compare it to. The following corollary compares variances when the number of samples is fixed, i.e., in the case where the DPP is projective.*

Corollary 3.7. *The marginal kernel of a projective DPP with a (fixed) number of samples μ is, by definition, of rank $r = \mu$. In this case:*

$$\forall \theta \in \Theta \qquad \mathrm{Var}(\hat{L}) = \mathrm{Var}(\hat{L}_{\mathrm{iid}}) - \frac{\mu - 1}{\mu} \left\| \sum_i \frac{f(\boldsymbol{x}_i, \theta)}{\pi_i} \tilde{\boldsymbol{v}}_i \right\|^2. \tag{38}$$

The variance is thus necessarily improved when using a projective DPP compared to its iid counterpart. This result is remarkable: the variance reduction is independent of the sign of f (supposed positive in the coreset context). This opens interesting generalizing perspectives to a more general class of cost functions L.

3.2.3. A link with tight frames

The elements we currently have to design the ideal marginal kernel K are as follows:

- The previous corollary suggests to design a projective DPP, that is: $\mathsf{K} = \mathsf{V}^\top\mathsf{V}$ with $\mathsf{V}\mathsf{V}^\top = \mathsf{I}_m$.
- Theorem 3.1 suggests to set $\pi_i = \mathsf{K}_{ii} = \frac{m\sigma_i}{\mathfrak{S}}$.

Finding such a marginal kernel boils down to finding $\mathsf{V} = (\boldsymbol{v}_1|\dots|\boldsymbol{v}_n)$ a set of n vectors $\boldsymbol{v}_i$ in dimension m with specified norms $\|\boldsymbol{v}_i\|^2 = \pi_i$, such that $\sum_i \pi_i = m$ and $\mathsf{V}\mathsf{V}^\top = \mathsf{I}_m$. This is exactly the problem of finding a tight frame of n vectors in dimension m, with specified norms [35].

Lemma 3.8. *Such a tight frame exists.*

Proof. Let us denote by $\pi_{(i)}$ the non-decreasing ordered sequence of π_i: $\pi_{(1)} \leqslant \pi_{(2)} \leqslant \ldots \leqslant \pi_{(n)}$. The Schur-Horn theorem states that a hermitian matrix K of size $n \times n$ with diagonal entries π_i and eigenvalues $(0,\dots,0,1,\dots,1)$ with $n-m$ zeros and m ones, exists if $\pi_{(i)}$ majorizes $(0,\dots,0,1,\dots,1)$, that is, if all the following inequalities are simultaneously verified:

$$\pi_{(1)} \geqslant 0, \quad \pi_{(1)} + \pi_{(2)} \geqslant 0, \quad \cdots, \quad \sum_{i=1}^{n-m} \pi_{(i)} \geqslant 0$$

$$\sum_{i=1}^{n-m+1} \pi_{(i)} \geqslant 1, \quad \cdots, \quad \sum_{i=1}^{n-1} \pi_{(i)} \geqslant m-1, \quad \sum_{i=1}^{n} \pi_{(i)} \geqslant m.$$

The first $n-m$ inequalities are trivially verified as all π_i are supposed positive. Now, $\sum_{i=1}^{n-m+1} \pi_{(i)} \geqslant 1$ is also verified. Indeed, if it was not case, *i.e.*, if $\sum_{i=1}^{n-m+1} \pi_{(i)} < 1$, then $\sum_{i=1}^{n} \pi_{(i)} < m$ as the largest $m-1$ values of π_i are by hypothesis upper bounded by 1. This would contradict $\sum_i \pi_i = m$. A similar argument can be applied to the remaining inequalities. □

Also, a tight frame not only exists, but several solutions exist in general, and efficient algorithms have been designed to build one (see for instance [36]). Out of all these possibilities, the ideal would be to find the tight frame that minimizes the variance of Eq.(38). Up to our knowledge, this is an open and difficult question, rooted in frame theory.

Let us recap the above variance results. We showed that a DPP sampling scheme has necessarily a lower variance than its Poisson uncorrelated counterpart, *regardless of the choice of off-diagonal elements of* K, provided that K stays PSD with eigenvalues between 0 and 1. We also showed that a projective DPP sampling scheme has necessarily a lower variance than its iid counterpart *regardless of the choice of off-diagonal elements of* K, provided that K stays projective. We finally showed that finding the projective DPP that minimizes the variance is equivalent to a difficult problem in frame theory. In other words: finding the optimal DPP for a given problem and dataset may be very hard, but on the other hand *any* DPP is guaranteed to do at least as well as iid sampling, in the sense discussed above. Further, we can easily design DPPs which are not optimal, but still have valuable properties, as the next section shows.

3.3. DPPs provide balanced sampling: a new type of guarantee

An important insight of coreset theory is that the datapoints which are different from the rest should be kept in the sample. We show in this section that one can construct a DPP which asymptotically guarantees a rebalancing of the datapoints $\mathcal{X}$, meaning that points which are relatively isolated have a high chance of being retained. For instance, in the k-means setting, this property implies that, asymptotically, one can construct a DPP that provably produces a balanced sample across clusters,

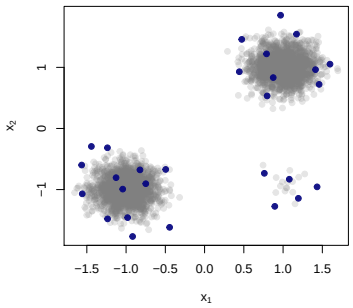

Figure 1. *It is possible to construct a DPP with a (asymptotic) rebalancing property, meaning that it will sample several points from each cluster even when clusters are severely imbalanced. Here, we show three imbalanced clusters: two have size 2,000 and one has size 20. In blue, a sample from a polynomial DPP (see text for definition): it samples from each cluster despite their very different sizes. The formal rebalancing property is illustrated in Fig. 2.*

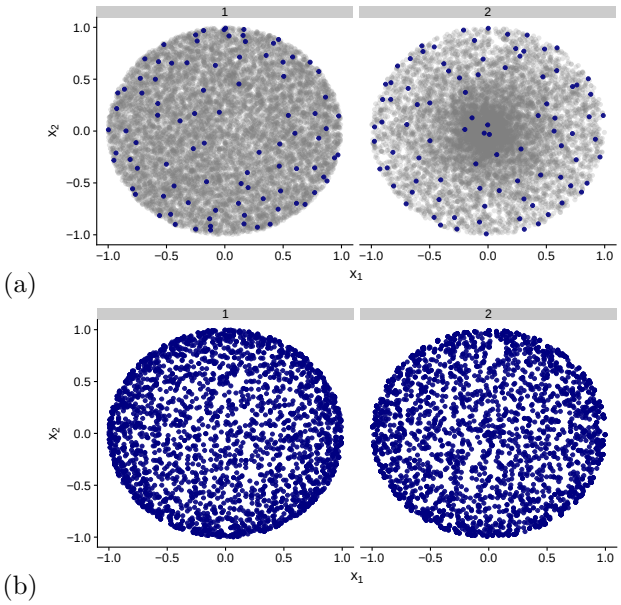

Figure 2. *An illustration of the rebalancing result. (a) We sample ground sets (grey points) from two different distributions on the disc. In blue, two realisations of a polynomial DPP constructed from the ground set. Note how similar the two realisations are (density-wise), despite the very different ground sets they are drawn from. (b) We overlay 30 realisations of each DPP: the two resulting densities are very similar, again despite the different ground sets. Our result states that they should indeed converge in large n and m, and that the limiting density depends only on the shape of the domain. Here points close to the boundary are oversampled relative to points in the center, as predicted.*

even in datasets where some clusters are much smaller than others. The result is illustrated in Figs. 1 and 2.

In a nutshell, the result is as follows. Suppose that the data $\mathcal{X}$ is a set of n elements drawn iid from a continuous distribution μ defined on $\Omega \subset \mathbb{R}^d$. Build a projective DPP $\mathcal{S}$ of size m based on the monomials of the x_i's (see Section 3.3.1 for a precise definition). Under mild regularity assumptions

on μ, we show that the intensity measure of $\mathcal{S}$, marginalized over $\mathcal{X}$ is *independent* of μ. Our proof is based on a powerful theorem from [37].

Note that this rebalancing property also occurs for iid sampling with sensitivities (that provide a sort of density estimation: the lower the density of points around x_i, the larger σ_i, the higher the chance of sampling it). What is noteworthy here is that the rebalancing property occurs "naturally": without any sort of prior density-like estimation. We will emphasize this important point at the end of this Section.

In Section 3.3.1, we present the specific type of polynomial DPPs for which our proof holds, that are similar to those used in [38]. Our result is then formally stated in Section 3.3.2.

3.3.1. Projective polynomial DPPs

The L-ensemble we shall build is based on the first m monomials. In dimension one this is easy to define, so we start there and generalize later to dimension $d \geqslant 2$. For $d = 1$, we denote by $\mathcal{X} = \{x_1, \dots, x_n\}$ the original set (supposed to be drawn iid from μ defined on Ω), and form the $n \times m$ Vandermonde matrix

$$\mathsf{V}(\mathcal{X}) = [x_i^{j-1}]_{i,j=1}^{n,m} \in \mathbb{R}^{n \times m}. \tag{39}$$

Note that this matrix has rank m a.s. (as μ is supposed regular enough) and contains all monomials up to degree $m-1$. The L-ensemble we consider equals:

$$\mathsf{L} = \mathsf{V}\mathsf{V}^\top \in \mathbb{R}^{n \times n}. \tag{40}$$

The orthogonal polynomials (defined on Ω) under the empirical measure $d\mu_n = (1/n)\sum \delta_{x_i}$ associated to $\mathcal{X}$, are defined in the usual manner, i.e. $q_0(x)$ of degree 0, $q_1(x)$ of degree 1, ... such that: $\int q_i(x) q_j(x) d\mu_n = \delta_{ij}$ and $\int x^i q_j(x) d\mu_n = 0$ if $i < j$. In other words, the sequence is constructed from Gram-Schmidt orthogonalisation under $d\mu_n$. Let us write $\boldsymbol{q}_j(\mathcal{X}) = (q_j(x_1), \dots, q_j(x_n))^\top \in \mathbb{R}^n$ the vector consisting of the polynomial $q_j(x)$ taken at values in $\mathcal{X}$. It is well-known (and easily verified) that the QR decomposition of V verifies:

$$\mathsf{V} = \mathsf{Q}\mathsf{R} \tag{41}$$

with $\mathsf{Q} = (\boldsymbol{q}_0(\mathcal{X}) | \dots | \boldsymbol{q}_{m-1}(\mathcal{X})) \in \mathbb{R}^{n \times m}$ and $\mathsf{R} \in \mathbb{R}^{m \times m}$ an upper triangular matrix.

Now, consider the m-DPP $\mathcal{S}$ with L-ensemble $\mathsf{L} = \mathsf{V}\mathsf{V}^\top$. Using the fact that $\det(AB) = \det(A)\det(B)$ if A and B are square, we have:

$$\begin{aligned}
\mathbb{P}(\mathcal{S}) &= Z^{-1} \det(\mathsf{L}_\mathcal{S}) \\
&= Z^{-1} \det((\mathsf{Q}\mathsf{R}\mathsf{R}^\top\mathsf{Q}^\top)_\mathcal{S}) \\
&= Z^{-1} \det((\mathsf{Q}\mathsf{Q}^\top)_\mathcal{S}) \det(\mathsf{R})^2 \\
&= Z'^{-1} \det((\mathsf{Q}\mathsf{Q}^\top)_\mathcal{S})
\end{aligned}$$

such that $\mathcal{S}$ is also a m-DPP with L-ensemble $\mathsf{Q}\mathsf{Q}^\top$. As $\mathsf{Q}^\top\mathsf{Q} = \mathsf{I}_m$, *i.e.*, $\mathsf{Q}\mathsf{Q}^\top$ is projective, $\mathcal{S}$ is in fact a projective DPP. As a result, its associated marginal kernel is also $\mathsf{Q}\mathsf{Q}^\top$ (see Lemma. 2.7) and, for instance:

$$\mathbb{P}(x_i \in \mathcal{S}) = \sum_{j=0}^{m-1} q_j^2(x_i). \tag{42}$$

The extension to $d > 1$ is mostly straightforward, but there are a few differences to keep in mind when defining the Vandermonde matrix of monomials. Monomials $\boldsymbol{x}^{\boldsymbol{\alpha}}$ are now defined as:

$$\boldsymbol{x}^{\boldsymbol{\alpha}} = \prod_{j=1}^{d} x(j)^{\alpha_j} \tag{43}$$

The total degree of a monomial equals the sum of the degrees in each variable, i.e. $\sum \alpha_i = ||\boldsymbol{\alpha}||_1$. The most significant difference between the one-dimensional case and the general case is that there

is more than one monomial of total degree ϕ. For example, in dimension 2, $\boldsymbol{x} = (x(1), x(2))^\top$ and the monomials of degree 2 are given by the powers $(2,0)$, $(0,2)$ and $(1,1)$: $x(1)^2$, $x(2)^2$, $x(1)x(2)$. A good way of thinking about the construction of a polynomial DPP in the multidimensional case is to pick first a maximum order (e.g. $\phi = 3$), meaning that all monomials with total degree up to 3 are included. Then the natural sample size m for the DPP equals the total number of features, giving $m = \binom{d+\phi}{\phi}$. Again, for $d = 2$ and $\phi = 3$, this gives $m = 1 + 2 + 3 + 4 = 10$. In fact, there is one monomial of order 0: 1, two monomials of order 1: $x(1)$ and $x(2)$, three monomials of order 2 (the ones stated above), and four monomials of order 3: $x(1)^3$, $x(2)^3$, $x(1)^2x(2)$ and $x(1)x(2)^2$. This implies that in dimension d, the m-DPP detailed earlier is well-defined only for specific values of m: $m = \binom{d+1}{1} = d+1$, or $m = \binom{d+2}{2} = \frac{1}{2}(d+1)(d+2)$, or $m = \binom{d+3}{3} = \frac{1}{6}(d+1)(d+2)(d+3)$, etc.

A slight technical difficulty arises in defining the orthogonal polynomials of a multivariate measure: in dimension 1, the fact that there is a single monomial of a given degree leads to a natural order in which to perform the Gram-Schmidt procedure. In higher dimensions the order is only a partial order, so that we can introduce the monomials by blocks of equal degree, but within a block the ordering is arbitrary. So we may pick any arbitrary order (e.g. lexicographic) and run Gram-Schmidt in that order (for more see [39]). Given this choice the QR decomposition remains well-defined and all properties given above in the 1D case carry over to the general case. In particular, the link with the orthogonal polynomials on the discrete measure μ_n stays valid.

3.3.2. The rebalancing theorem

Formally, the result is as follows. The intensity function $\iota(\boldsymbol{x})$ of a point process quantifies the expected number of points to be found around $\boldsymbol{x}$. We characterize the asymptotics of the intensity function of a DPP $\mathcal{S}$ when both $\mathcal{S}$ and the ground set $\mathcal{X}$ are large, and show that, in that limit, the intensity is *independent* of the measure μ from which $\mathcal{X}$ is sampled from.

The result is stated formally as a double limit, letting first n go to infinity (an easy discrete-to-continuous limit), and then increasing the order ϕ of the polynomial DPP, which implies m going to infinity too. We emphasize that, empirically speaking, rebalancing occurs for reasonable values of n and m but the rate of convergence is hard to quantify.

Certain regularity assumptions are inherited from [37], to which we refer for more thorough details. The formal assumptions are as follows:

1) The initial dataset $\mathcal{X} = \{x_1, \dots, x_n\}$ is drawn i.i.d. from a measure μ over a compact, convex[4] domain $\Omega \subset \mathbb{R}^d$.
2) μ and the Lebesgue measure ν are mutually absolutely continuous on Ω, so that μ', the density, is well-defined everywhere on the domain (we use the Lebesgue measure for simplicity, another measure may be substituted)
3) We are interested in convergence "in the bulk", ie. inside the domain. Formally, the results hold for $D \subset D_1 \subset \Omega$, where D is compact and D_1 is open
4) μ' is bounded above and below on D_1
5) We form a m-DPP $\mathcal{S}$ on the set $\mathcal{X}$, with a polynomial kernel of degree ϕ (defined in the previous Section), such that $m = \binom{\phi+d}{\phi}$.
6) (technical) μ is regular in the sense of Stahl, Totik, and Ullman, and the Christoffel function with respect to μ verifies condition (1.7) in [37].

The intensity measure of $\mathcal{S}$, marginalizing over $\mathcal{X}$, which we denote by $I_{n,\phi}(\mathcal{A})$ equals the expected number of points of $\mathcal{S}$ in set $\mathcal{A}$, i.e.:

$$I_{n,\phi}(\mathcal{A}) = \mathbb{E}_{\mathcal{X},\mathcal{S}}\left(|\mathcal{S} \cap \mathcal{A}|\right) = \mathbb{E}_{\mathcal{X},\mathcal{S}}\left\{\sum_{s \in \mathcal{S}} \mathbb{I}\left(s \in \mathcal{A}\right)\right\} \tag{44}$$

Note that the expectation is over *both* $\mathcal{X}$ and $\mathcal{S}$. Furthermore, $I_{n,\phi}(\Omega)$ equals m, the total number of points in $\mathcal{S}$.

Our result may be stated as follows.

[4]The convexity assumption can probably be relaxed

Theorem 3.9. *Under the assumptions above, for all $\mathcal{A} \subset D_1$,*

$$\lim_{\phi\to\infty} \frac{1}{\binom{\phi+d}{\phi}} \lim_{n\to\infty} I_{n,\phi}(\mathcal{A}) = \int_{\mathcal{A}} \kappa(\boldsymbol{y}) d\boldsymbol{y} \tag{45}$$

where κ is a density independent *of μ*

The proof is in Appendix C.

Lemma 3.10. *κ is mostly dependent on the distance to the boundaries of Ω. For example, if Ω is the unit ball in $\mathbb{R}^d$, $\kappa(\boldsymbol{y}) = \left(1 - ||\boldsymbol{y}||^2\right)^{-1/2}$*

See [37] for a proof.
Several important remarks are in order:

- unlike iid sampling with sensitivities or other density-related measure for which such rebalancing property will also occur, there is here no prior density estimation: the L-ensemble is defined via the Vandermonde matrix that is trivial to compute. Thus, this rebalancing is a property that "naturally" arises from the DPP.
- this is only an asymptotic result as n and m go to infinity. Finding minimal values of m for which rebalancing is highly probable, or even rates of convergence is likely a difficult endeavour. We emphasize nevertheless that, empirically speaking, rebalancing occurs for reasonable values of n and m, as visible in Figs. 1 and 2.

This ends the theoretical results of this paper. We now move on to applications. In the next Section, we apply the results to two problems: k-means and linear regression. In Section 5, implementation details are provided. Finally, experimental validation on artificial and real-world datasets is provided in Section 6.

4. Application to two problems: k-means and linear regression

We focus on two problems: k-means and linear regression. Admittedly, these are not the best problems to exhibit the usefulness of coresets: there already exists very efficient algorithms to solve them and the need for a controlled summary is in fact rare. We nevertheless focus on these two problems as they have been well studied in the iid setting, which it is our goal to improve on. Moreover, we derived analytical formulas for the sensitivity in the 1-means and the linear regression settings: we will thus be able to compare, in those two cases, DPP sampling vs the ideal iid setting (later in the experimental Section 6).

4.1. Application to k-means

The theoretical results of Section 3 are valid for any learning problem of the form detailed in Section 2.1. We now specifically consider the k-means problem on a set $\mathcal{X}$ comprised of n datapoints in $\mathbb{R}^d$. This problem boils down to finding k centroids $\theta = (c_1, \ldots, c_k)$, all in $\mathbb{R}^d$, such that the following cost is minimized:

$$L(\mathcal{X}, \theta) = \sum_{x\in\mathcal{X}} f(x, \theta) \quad \text{with} \quad f(x, \theta) = \min_{c\in\theta} \|x - c\|^2 .$$

Let ρ be the diameter of the minimum enclosing ball of $\mathcal{X}$ (the smallest ball enclosing all points in $\mathcal{X}$). Theorem 3.1 and its corollaries are applicable to the k-means problem, such that:

Corollary 4.1 (m-DPP for k-means)**.** *Let $\mathcal{S}$ be a sample from an m-DPP with L-ensemble L. Let $\epsilon, \delta \in (0,1)^2$. With probability at least $1-\delta$, $\mathcal{S}$ is a ϵ-coreset provided that:*

$$m \geqslant m^* = \frac{32}{\epsilon^2} \left(\max_i \frac{\sigma_i}{\bar{\pi}_i}\right)^2 \left(kd \log\left(\frac{24\rho^2}{\epsilon \langle f \rangle_{opt}} + 1\right) + \log\frac{4}{\delta}\right), \tag{46}$$

with $\forall i, \bar{\pi}_i = \pi_i/m$.

Setting the marginal probabilities to their optimal values $\pi_i = m\sigma_i/\mathfrak{S}$, $\mathcal{S}$ is a ϵ-coreset with probability larger than $1-\delta$ provided that:

$$m \geqslant \frac{32}{\epsilon^2}\mathfrak{S}^2 \left(kd \log\left(\frac{24\rho^2}{\epsilon \langle f \rangle_{opt}} + 1 \right) + \log\frac{4}{\delta} \right).$$

Proof. Let us write $\mathcal{B}$ the minimum enclosing ball of $\mathcal{X}$, of diameter ρ. The potentially interesting centroids are necessarily included in $\mathcal{B}$ such that the space of parameters Θ in the k-means setting is the set of all possible k centroids in $\mathcal{B}$: $\Theta = \mathcal{B}^k$. The metric d_Θ we consider is the Hausdorff metric associated with the Euclidean distance:

$$\forall \theta, \theta', \qquad d_\Theta(\theta, \theta') = \max\left\{ \max_{c\in\theta} \min_{c'\in\theta'} \|c - c'\|_2 \,,\ \max_{c'\in\theta'} \min_{c\in\theta} \|c - c'\|_2 \right\}.$$

An ϵ'-net of Θ. Consider $\Gamma_\mathcal{B}$ an ϵ'-net of $\mathcal{B}$ consisting of $(\frac{2\rho}{\epsilon'} + 1)^d$ small balls of radius ϵ'. Such a covering indeed exists: see, e.g., Lemma 2.5 in [40]. Consider $\Gamma = \Gamma_\mathcal{B}^k$ of cardinality $|\Gamma| = (\frac{2\rho}{\epsilon'} + 1)^{kd}$. Let us show that Γ is an ϵ'-net of Θ, that is:

$$\forall \theta \in \mathcal{B}^k, \quad \exists \theta^* \in \Gamma \quad \text{s.t.} \quad d_\Theta(\theta, \theta^*) \leqslant \epsilon'.$$

In fact, consider $\theta = (c_1, \ldots, c_k) \in \mathcal{B}^k$. By construction, as $\Gamma_\mathcal{B}$ is an ϵ'-net of $\mathcal{B}$, we have:

$$\forall i = 1, \ldots, k \qquad \exists c_i^* \in \Gamma_\mathcal{B} \quad \text{s.t.} \quad \|c_i - c_i^*\| \leqslant \epsilon'.$$

Writing $\theta^* = (c_1^*, \ldots, c_k^*) \in \Gamma$, one has:

$$d_\Theta(\theta, \theta^*) \leqslant \epsilon',$$

which proves that Γ is an ϵ'-net of Θ. The number of balls of radius $\epsilon' = \epsilon\langle f \rangle_{\text{opt}}/6\gamma$ sufficient to cover Θ is thus $\eta = (\frac{12\rho\gamma}{\epsilon\langle f \rangle_{\text{opt}}} + 1)^{kd}$.

$f(x, \theta)$ is γ-Lipschitz with $\gamma = 2\rho$. Consider any θ, θ' and $x \in \mathcal{X}$. We want to show that:

$$-\gamma\ d_\Theta(\theta, \theta') \leqslant f(x,\theta) - f(x, \theta') \leqslant \gamma\ d_\Theta(\theta, \theta').$$

Let us write $c = \text{argmin}_{t\in\theta} \|x - t\|^2$ the centroid in θ closest to x and $c' = \text{argmin}_{t'\in\theta'} \|x - t'\|^2$ the centroid in θ' closest to x. Moreover, let us write $\tilde{c}' = \text{argmin}_{t'\in\theta'} \|c - t'\|^2$ the centroid in θ' closest to c. Note that c' and $\tilde{c}'$ are not necessarily equal. By definition of c', one has:

$$\|x - c'\| \leqslant \|x - \tilde{c}'\|,$$

such that:

$$\|x - c'\|_2 - \|x - c\|_2 \leqslant \|x - \tilde{c}'\|_2 - \|x - c\|_2 \leqslant \|\tilde{c}' - c\|_2 \leqslant d_\Theta(\theta, \theta').$$

Thus:

$$\begin{aligned} f(x, \theta') - f(x, \theta) = \|x - c'(x)\|^2 - \|x - c\|^2 &= (\|x - c'\| - \|x - c\|)(\|x - c'\| + \|x - c\|) \\ &\leqslant (\|x - c'\| + \|x - c\|)\ d_\Theta(\theta, \theta') \leqslant 2\rho\ d_\Theta(\theta, \theta'). \end{aligned}$$

Finally, $n\sigma_{\min} \geqslant 1$, as shown by the second lemma of Appendix D.

Given all these elements, Theorem 3.1 and its subsequent corollaries are thus applicable to the k-means setting and one obtains the desired result. □

Note that, in the case of DPPs, one could apply Theorem B.1 to the k-means problem, and obtain similar results.

4.2. Application to linear regression

We now consider the linear regression problem: find $\theta \in \mathbb{R}^d$ such that a measured vector $y \in \mathbb{R}^n$ is closest to $\mathsf{X}\theta$ where $\mathsf{X}^\top = (x_1|\dots|x_n) \in \mathbb{R}^{d\times n}$ are n data points in $\mathbb{R}^d$. Let us write $X_i = (y_i, x_i)$ and $\mathcal{X} = \{X_1, \dots, X_n\}$. The least squares estimator minimizes:

$$L(\mathcal{X}, \theta) = \|y - \mathsf{X}\theta\|_2^2 .$$

By denoting

$$f(X_i, \theta) = (y_i - x_i^\top \theta)^2, \tag{47}$$

one can thus write the least squares solution to the linear regression problem in the form of the problems investigated in this paper: the objective is to minimize the cost L with f a positive function:

$$L(\mathcal{X}, \theta) = \|y - \mathsf{X}\theta\|_2^2 = \sum_{i=1}^n f(X_i, \theta). \tag{48}$$

We suppose that all $\boldsymbol{x}_i$ are enclosed in the unit ball in dimension d and that $y_i \in [0, 1]$. Moreover, we suppose that the space Θ is bounded and enclosed in a d-dimensional ball $\mathcal{B}$ centered in 0 of diameter ρ.

Even though we derived the analytical formulation of the sensitivity for linear regression (Lemma D.3), we were not able to show that $n\sigma_{\min} \geqslant 1$ in general. We thus have the following slightly more complicated result:

Corollary 4.2 (m-DPP for linear regression)**.** *Let $\mathcal{S}$ be a sample from an m-DPP with L-ensemble L. Let $\epsilon, \delta \in (0, 1)^2$. With probability at least $1 - \delta$, $\mathcal{S}$ is a ϵ-coreset provided that:*

$$m \geqslant \max(m_1^*, m_2^*) \tag{49}$$

with

$$m_1^* = \frac{32}{\epsilon^2}\left(\max_i \frac{\sigma_i}{\bar{\pi}_i}\right)^2 \left(d \log\left(\frac{12\rho(4\rho+2)}{\epsilon\langle f\rangle_{opt}} + 1\right) + \log\frac{4}{\delta}\right), \tag{50}$$

$$m_2^* = \frac{32}{\epsilon^2}\left(\frac{1}{n\bar{\pi}_{min}}\right)^2 \left(d \log\left(\frac{12\rho(4\rho+2)}{\epsilon\langle f\rangle_{opt}} + 1\right) + \log\frac{4}{\delta}\right) \tag{51}$$

with $\forall i, \bar{\pi}_i = \pi_i/m$.

Setting the marginal probabilities to their optimal values $\pi_i = m\sigma_i/\mathfrak{S}$, $\mathcal{S}$ is a ϵ-coreset with probability larger than $1 - \delta$ provided that:

$$m \geqslant \frac{32}{\epsilon^2}\max\left(\mathfrak{S}^2, \frac{\mathfrak{S}^2}{n^2\sigma_{min}^2}\right)\left(d \log\left(\frac{12\rho(4\rho+2)}{\epsilon\langle f\rangle_{opt}} + 1\right) + \log\frac{4}{\delta}\right).$$

Proof. The metric d_Θ we consider is the Euclidean distance in dimension d.

- An ϵ'-net of Θ. Consider $\Gamma_\mathcal{B}$ an ϵ'-net of $\mathcal{B}$ consisting of $(\frac{2\rho}{\epsilon'} + 1)^d$ small balls of radius ϵ'. Such a covering indeed exists: see, e.g., Lemma 2.5 in [40]. The number of balls of radius $\epsilon' = \epsilon\langle f\rangle_{\text{opt}}/6\gamma$ sufficient to cover Θ is thus $\eta = (\frac{12\rho\gamma}{\epsilon\langle f\rangle_{\text{opt}}} + 1)^d$.

- $f(X, \theta)$ is γ-Lipschitz with $\gamma = 4\rho + 2$. Consider any θ, θ', $\boldsymbol{x}_i$ and y_i. We want to show that:

$$\left((y_i - \boldsymbol{x}_i^\top\theta)^2 - (y_i - \boldsymbol{x}_i^\top\theta')^2\right)^2 \leqslant \gamma^2 \ \|\theta - \theta'\|^2 .$$

In fact:

$$\begin{aligned}
\left((y_i - \boldsymbol{x}_i^\top\theta)^2 - (y_i - \boldsymbol{x}_i^\top\theta')^2\right)^2 &= \left(\theta^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\theta - \theta'^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\theta' - 2y_i\boldsymbol{x}_i^\top(\theta - \theta')\right)^2 \\
&= \left[\left(2\theta'^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top + (\theta - \theta')^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top - 2y_i\boldsymbol{x}_i^\top\right)(\theta - \theta')\right]^2 \\
&\leqslant \left[\left\|2\theta'^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top + (\theta - \theta')^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top - 2y_i\boldsymbol{x}_i^\top\right\| \|\theta - \theta'\|\right]^2 \\
&\leqslant \left[2\left\|\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\| \|\theta'\| + \left\|\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\| \|\theta - \theta'\| + 2y_i\|\boldsymbol{x}_i\|\right]^2 \|\theta - \theta'\|^2
\end{aligned} \tag{52}$$

by triangular inequality and writing $\left\|\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\|$ the 2-norm of the matrix $\boldsymbol{x}_i\boldsymbol{x}_i^\top$, which is equal to $\|\boldsymbol{x}_i\|^2$ and bounded by one by hypothesis. As Θ is supposed to be enclosed in a ball of radius ρ, we further have:

$$\left((y_i - \boldsymbol{x}_i^\top\theta)^2 - (y_i - \boldsymbol{x}_i^\top\theta')^2\right)^2 \leqslant (4\rho+2)^2\left\|\theta-\theta'\right\|^2 \tag{53}$$

Given these elements, Theorem 3.1 is applicable to the linear regression setting and one obtains the desired result. □

5. Implementation

5.1. The DPP's ideal marginal kernel

Following the theoretical results, the ideal strategy (although unrealistic) to build the marginal kernel K of the ideal DPP sampling scheme would be as follows. 1/ Deal with outliers as explained in Appendix E until $\sigma_{\max}$ is not too large. 2/ Compute all σ_i. 3/ Set all π_i to $m\sigma_i/\mathfrak{S}$ with m sufficiently large as detailed in the theorems. 4/ Find all non-diagonal elements of K in order to minimize for all θ the estimator's variance, as derived in Eq. (28):

$$\mathrm{Var}(\hat{L}) = \sum_i \left(\frac{1}{\pi_i} - 1\right) f(x_i,\theta)^2 - \sum_{i\neq j} \frac{\mathsf{K}_{ij}^2}{\pi_i\pi_j} f(\boldsymbol{x}_i,\theta)f(\boldsymbol{x}_j,\theta). \tag{54}$$

while constraining K to be a valid marginal kernel, *i.e.*: SDP with $0 \preceq \mathsf{K} \preceq 1$, 5/ sample a DPP with kernel K. On our way to derive a practical algorithm with a linear complexity in n, many obstacles stand before us: there is no known polynomial algorithm to compute all σ_i in the general setting, solving exactly the minimization problem of step 4 under eigenvalue constraint remains open, and sampling from this engineered ideal K costs $\mathcal{O}(n^3)$ number of operations (see Alg. 1 of [8]: it necessitates a full diagonalization of K). Designing a linear-time algorithm that provably verifies under a controlled error the conditions of our previous theorems is out-of-scope of this paper. In the following, we prefer to first recall the intuitions behind the construction of a good kernel, and then discuss two choices of kernel we advocate: a Gaussian kernel and a Vandermonde-based kernel.

5.2. A first choice: the Gaussian kernel

In order for K to be a good candidate for coresets, it needs to verify the following two properties:

- As indicated by the theorems, the diagonal entries K_{ii} should increase as the associated σ_i increases.
- As indicated by the variance equation, off-diagonal elements should be as large as possible (in absolute value) given the eigenvalue constraints. In fact, we cannot set all non-diagonal entries of K to large values as the matrix's 2-norm would rapidly be larger than 1. We thus need to choose the best pairs (i,j) for which it is worth setting a large value of K_{ij}. A first glance at the variance equation indicates that the larger $f(x_i,\theta)f(x_j,\theta)$ is, the larger K_{ij} should be, in order to decrease the variance as much as possible. Recall nevertheless that in the coreset setting, all sampling parameters should be independent of θ. The off-diagonal elements should thus verify the following property: the larger is the correlation between x_i and x_j (the more similar are $f(x_i,\theta)$ and $f(x_j,\theta)$ for all θ), the larger K_{ij} should be.

We show in the following in what ways the choice of marginal kernel

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

with L the Gaussian kernel matrix with parameter τ:

$$\forall(i,j) \quad \mathsf{L}_{ij} = \exp^{-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\tau^2}},$$

is a good candidate to build coresets for k-means (the linear regression case is discussed later). Let us write $\mathsf{U} = (\boldsymbol{u}_1|\ldots|\boldsymbol{u}_n)$ the orthonormal eigenvector basis of L and $\Lambda = \mathrm{diag}(\lambda_1|\ldots|\lambda_n)$ its diagonal

matrix of sorted eigenvalues, $0 \leqslant \lambda_1 \leqslant \ldots \leqslant \lambda_n$. U and Λ naturally depend on τ. One shows for instance that, with respect to τ, λ_n is a monotonically increasing function between 1 and n.

Concerning the off-diagonal elements of K, let us first note that if x_i and x_j are correlated (that is, in the k-means setting, if they are close to each other), then

$$\mathsf{K}_{ij} = \sum_k \frac{\lambda_k}{1+\lambda_k} u_k(i) u_k(j)$$

should be large in absolute value. In fact, in the limit where $x_i = x_j$, then $\forall k, u_k(i) = u_k(j)$ and $\mathsf{K}_{ij} = \mathsf{K}_{ii} = \mathsf{K}_{jj}$. The determinant of the 2×2 submatrix of K indexed by i and j is therefore null: sampling both will never occur. Thus, the closer are x_i and x_j, the lower is the chance of sampling both jointly. Moreover, if x_i and x_j are far from each other (for instance, in different clusters), then the entries i and j of $\mathsf{L}'s$ eigenvectors will be very different. For instance, say the dataset contains two well separated clusters of similar size. If the Gaussian parameter τ is set to the size of these clusters, then the kernel matrix L will be quasi-block diagonal, with each block corresponding to the entries of each cluster. Also, each eigenvector $\boldsymbol{u}_k$ will have energy either in one cluster or the other such that K_{ij} is necessarily small if i and j belong to different clusters, and the event of sampling both jointly is probable.

Concerning the probability of inclusion of i, we have:

$$\mathsf{K}_{ii} = \sum_k \frac{\lambda_k}{1+\lambda_k} v_i(k)^2,$$

where $\boldsymbol{v}_i$ is the vector of size n verifying $\forall k, v_i(k) = u_k(i)$. For all i, $\|\boldsymbol{v}_i\|^2 = 1$. The probability of inclusion is thus directly linked to the values of k that contain the energy of $\boldsymbol{v}_i$: the more the energy of $\boldsymbol{v}_i$ is contained on high values of k, the larger is the probability of inclusion. Say we are again in a situation where the clusters and the choice of Gaussian parameter τ are such that L is quasi block diagonal. Within each block, the eigenvector associated with the highest eigenvalue corresponds approximately to the constant vector. These eigenvectors being normalized, the associated entry of $v_i(k)$ is thus approximately equal to $1/\sqrt{\#C_i}$ where $\#C_i$ is the size of the cluster containing data x_i. Typically, if the cluster is small, that is, if $\#C_i$ tends to 1, the associated entry $v_i(k)$ tends to 1 as well, such that all the energy of v_i is drawn towards high values of k, thus increasing the probability of inclusion of i. In other words, the more isolated, the higher the chance of being sampled. This corresponds to the intuition one may obtain for the sensitivity σ_i. It has indeed been shown that the sensitivity may be interpreted as a measure of outlierness [26].

In the linear regression case, a similar argumentation is possible, up to the fact that point i can be an outlier from the point of view of $\boldsymbol{x}_i$ and/or y_i, such that the kernel should take both into account: we suggest the Gaussian kernel in dimension $d+1$:

$$\forall (i,j) \quad \mathsf{L}_{ij} = \exp^{-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2}{2\tau^2}},$$

with $\boldsymbol{z}_i = [\boldsymbol{x}_i^\top, y_i]^\top \in \mathbb{R}^{d+1}$.

In both contexts, we thus advocate to sample DPPs via a Gaussian kernel L-ensemble. We now move on to detailing an efficient sampling implementation.

5.2.1. Efficient implementation

Sampling exactly a DPP from the Gaussian L-ensemble verifying

$$\forall (i,j) \qquad \mathsf{L}_{ij} = \exp^{-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\tau^2}}$$

consists in the following steps:

1) Compute L.
2) Diagonalize L in its set of eigenvectors $\{\boldsymbol{u}_k\}$ and eigenvalues $\{\lambda_k\}$.
3) Sample a DPP given $\{\boldsymbol{u}_k\}$ and $\{\lambda_k\}$ via Alg. 1 of [8].

Algorithm 1 The Gaussian kernel coreset sampling heuristics

Input: $\mathcal{X} = \{x_i\}$ a set of n points in $\mathbb{R}^d$, a Gaussian kernel parameter τ, a number of samples m
- Draw $r \geqslant \mathcal{O}(m)$ random Fourier vectors associated to the Gaussian kernel with parameter τ
- Compute the associated RFF matrix $\Psi \in \mathbb{R}^{2r \times n}$ as explained in Appendix F.1
- Compute $\mathsf{C} = \Psi\Psi^\top \in \mathbb{R}^{2r \times 2r}$ the dual representation
- Compute the eigendecomposition of C: obtain eigenvectors $\{\boldsymbol{v}_k\}$ and eigenvalues $\{\nu_k\}$
- Draw a sample $\mathcal{S}$ from a m-DPP with L-ensemble $\mathsf{L} = \Psi^\top\Psi$ as explained in Appendix F.3.
- Compute the marginal probabilities π_s for all $\boldsymbol{x}_s \in \mathcal{S}$ as explained in Appendix F.3, and set weights $\omega(\boldsymbol{x}_s) = 1/\pi_s$.

Output: $\{\mathcal{S}, \omega\}$ a weighted sample of size m.

Step 1 costs $\mathcal{O}(n^2 d)$, step 2 costs $\mathcal{O}(n^3)$, step 3 costs $\mathcal{O}(n\mu^3)$, where we recall that μ is the expected number of samples of the DPP. This naive approach is thus not practical. We detail in Appendix F how to reduce the overall complexity to $\mathcal{O}(n\mu^2)$, by 1/ taking advantage of Random Fourier Features (RFF) [41] to estimate a low dimensional representation $\Psi \in \mathbb{R}^{2r \times n}$ of the L-ensemble $\mathsf{L} \simeq \Psi^\top\Psi$, where r is the chosen number of features; and 2/ running a DPP sampling algorithm adapted to such a low rank representation.

In the experimental section, we will concentrate on m-DPPs as they are simpler to compare with state of the art methods that all have a fixed known-in-advance number of samples. The overall m-DPP sampling algorithm adapted to the k-means problem that we will consider is summarized in Alg. 1: given the data $\mathcal{X}$, the number of desired samples m, and the Gaussian parameter τ, it outputs a weighted set of m samples $\mathcal{S}$ that is a good candidate to be a coreset if m is large enough. The runtime to build Ψ is $\mathcal{O}(ndr)$; to compute C and diagonalize it is $\mathcal{O}(nr^2)$; to sample a m-DPP given this dual eigendecomposition is $\mathcal{O}(nm^2)$. Given that r is set to a few times m, the overall runtime of Alg. 1 is $\mathcal{O}(ndm + nm^2)$.

Given a number of samples m to draw, how should one set the Gaussian parameter τ? The larger is τ, the more repulsive is the m-DPP, and the smaller is the numerical rank of Ψ (the number of eigenvalues ν such that $n\nu$ is larger than the machine's precision). Now, numerical instabilities arise while sampling an m-DPP if the numerical rank of Ψ decreases below m: τ should not be set too large. Also, the smaller is τ, the closer is L to the identity matrix, such that the closer is the m-DPP to uniform sampling without replacement: τ should not be set too small. We will see in the following experimental section how the choice of τ affects results.

5.3. A second choice: a projective DPP based on the Vandermonde matrix

A second choice of DPP sampling, that derives from our analysis, is the projective DPP with m samples from the L-ensemble $\mathsf{L} = \mathsf{V}\mathsf{V}^\top$ where V is the Vandermonde matrix (discussed in Section 3.3.1). This choice has several advantages over the Gaussian kernel:

- V takes $\mathcal{O}(nm)$ operations to compute: the overall m-DPP sampling cost is thus naturally $\mathcal{O}(nm^2)$, with no need for any approximation technique.
- no particular scale τ is introduced.

This choice however has the drawback that in dimension higher than 1, not all values of m are allowed (only values of m for which there exists $\phi \in \mathbb{N}$ s.t. $m = \binom{\phi+d}{\phi}$), as explained at the end of Section 3.3.1.

6. Experiments

6.1. Different strategies to compare...

We will empirically compare results obtained with the five following approaches:

1) `m-DPP` : The strategy summarized in Alg. 1.
2) `PolyProj-DPP` : The strategy summarized in Alg. 2.

Algorithm 2 The Vandermonde-based coreset sampling heuristics

Input: $\mathcal{X} = \{x_i\}$ a set of n points in $\mathbb{R}^d$, a number of samples m
- m should verify: $\exists \phi \in \mathbb{N}$ such that $m = \binom{\phi+d}{\phi}$.
- Compute the Vandermonde matrix $\mathsf{V} \in \mathbb{R}^{n \times m}$.
- Compute the $(\mathsf{Q} \in \mathbb{R}^{n \times m}, \mathsf{R} \in \mathbb{R}^{m \times m})$ decomposition of V: $\mathsf{V} = \mathsf{QR}$ with $\mathsf{Q}^\top \mathsf{Q} = \mathsf{I}_m$ and R an upper triangular matrix.
- Draw a sample $\mathcal{S}$ from a projective DPP with L-ensemble $\mathsf{L} = \mathsf{Q}\mathsf{Q}^\top$ as explained in Algorithm 3.
- Compute the marginal probabilities π_s for all $\boldsymbol{x}_s \in \mathcal{S}$ with $\pi_s = \|\mathsf{Q}(s,:)\|^2$ the energy of the s-th line of Q; and set weights $\omega(\boldsymbol{x}_s) = 1/\pi_s$.

Output: $\{\mathcal{S}, \omega\}$ a weighted sample of size m.

3) `matched iid` : An iid sampling strategy with replacement, matched to either `m-DPP` or `PolyProj-DPP` (depending on the context). More precisely, m samples are drawn iid with replacement, the probability of selecting x_i at each draw being set to $p_i = \pi_i/m$, where π_i is the marginal probability of drawing x_i in `m-DPP` (or `PolyProj-DPP`).
4) `uniform iid` : Uniform iid sampling with replacement.
5) `sensitivity iid` : The current state of the art iid sampling based on a bi-criteria approximation to upper bound the sensitivity (Alg. 2 of [16]), or, if available (for instance in the case of 1-means and linear regression), an analytical formula of the sensitivity.

For the three iid methods (methods 3, 4 and 5), we will use the importance sampling estimator adapted to iid sampling of Eq. (9). For methods 1 and 2, we will use the importance sampling estimator adapted to correlated sampling of Eq. (16).

Empirically, when the ambient dimension d is small, performance of all methods is enhanced if the weights in $\hat{\mathsf{L}}$ are set via Voronoi cells rather than set to inverse probabilities: given the sample $\mathcal{S}$ of size m, compute its Voronoi tessellation in m cells, and associate to each sample $\boldsymbol{x}_s$ a weight $\omega(\boldsymbol{x}_s)$ equal to the number of datapoints in its associated Voronoi cell. We will call the associated cost estimators $\hat{L}$ the Voronoi estimators.

For completeness, we compare all these methods with another negatively correlated sampling method called D^2-sampling (commonly used for k-means++ seeding [42]):

6) `D`^{2} : sample the first element of $\mathcal{S}$ uniformly at random. Each subsequent element of $\mathcal{S}$ is drawn according to a probability proportional to the squared distance to the closest of the already sampled elements. The marginal probabilities are not known in this algorithm, so we will only be able to build the associated Voronoi cost estimator.

To measure the performance of each method, we will empirically estimate the probability that, given the method's sampled weighted subset, it verifies the coreset property of Eq. 6 for a given randomly chosen θ (setting ϵ to 0.1). On the artificial data models we investigate, we estimate this probability via 50 randomly chosen θ on 1000 realizations of the data. On the real-world datasets, we estimate this probability via 5000 randomly chosen θ. We will in general plot this probability versus the number of samples: the closer it is to 1, the better the sampling method for coresets.

In Sections 6.2.2 and 6.2.3, we will not only compare the coreset property of the samples obtained by each method, we will also compare the result of Lloyd's classical k-means heuristics [29] performed on the entire data versus the result obtained on the weighted samples of each method. To be precise, once the k-means heuristics on the weighted subset outputs k centroids, we classify all nodes (sampled or not) according to their closest distance to the centroids: this gives us a partition that we then compare using the Adjusted Rand (AR) similarity index [43] to the ground truth associated to the dataset. The AR index is a number between -1 and 1: the closer it is to 1, the closer are the partitions, the better the sampling method.

6.2. ...on different datasets

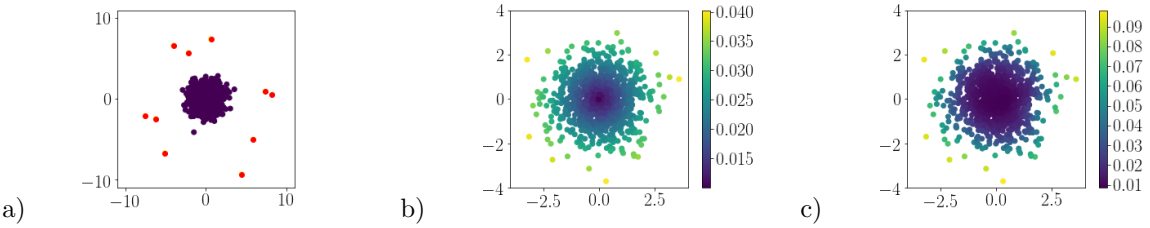

Figure 3. *a) A realization of an artificial dataset of $n = 1000$ data points; blue points are drawn from an isotropic Gaussian, a proportion $q = 0.01$ of the points are drawn as outliers (displayed in red). b) In a case without outliers, and for $m = 20$, we represent the inverse importance sampling weights of* `sensitivity iid`*, i.e., $m\sigma_i/\mathfrak{S}$, using the exact analytical formulation of the sensitivity in the 1-means case (see Lemma D.1). c) On the same data realization, and also setting $m = 20$, we represent the inverse importance sampling weights of* `m-DPP`*: the inclusion probability π_i.*

6.2.1. To start with: two well controlled cases

We start with two perfectly controlled cases (for which we derived the sensitivities analytically – see the first and third lemmas of Appendix D)):

- the 1-means case, for which we show that, supposing without loss of generality that the data is centered ($\sum_j x_j = 0$), the sensitivity verifies the following analytic form:

$$\sigma_i = \frac{1}{n}\left(1 + \frac{\|x_i\|^2}{v}\right), \tag{55}$$

  where $v = \frac{1}{n}\sum_{x\in\mathcal{X}} \|x\|^2$.
- the linear regression case, for which we show that:

$$\forall i \qquad \sigma_i = x_i^\top \mathsf{H}^{-1} x_i + \frac{(y_i - y_i^*)^2}{\|y - y^*\|^2} \tag{56}$$

  where $\mathsf{H} = \mathsf{X}^\top\mathsf{X}$ and y^* reads $y^* = \mathsf{X}\theta^* = \mathsf{X}\mathsf{H}^{-1}\mathsf{X}^\top y$.

We are thus able to compare our method versus the ideal iid sampling scheme for which we set p_i, the probability of drawing x_i, exactly to its ideal value given in Theorem 2.3: $p_i = \sigma_i/\mathfrak{S}$ ($= \sigma_i/2$ for 1-means, $= \sigma_i/(d+1)$ for linear regression).

Experiments with 1-means. We will work on a simple isotropic Gaussian dataset of $n = 1000$ points in dimension $d = 2, 20$ or 100. A percentage q of the n points are drawn as outliers (uniformly in the ambient space and far from the Gaussian mean). An instance of such a dataset in $d = 2$ dimensions, and with $q = 0.01$ is shown in Fig. 3a.

We start by showing in Fig. 4 the results of `m-DPP` versus the number of dimensions and the choice of parameter τ for the Gaussian kernel. All shown results are with $q = 0$ (no outlier) and with a number of random Fourier features $r = 200$. Several comments are in order. Firstly, compared to the importance sampling estimator, the Voronoi estimator produces good results in low dimensions, and fails as the dimension increases. Secondly, the performance of all methods increase and uniformize as the dimension increases. This is due to the fact that in large dimensions, interpoint distances tend to uniformize such that any pair of points tend to be representative of all interpoint distances, thus simplifying the problem of finding good coresets. This may also explain why the choice of τ is less crucial in higher dimension. In low dimensions, however, the choice of τ has a strong impact on performance. The best choice for τ depends in fact on the number of samples m one requires: as m increases, τ should be set smaller. This is in fact natural: if one desires a very short summary of the dataset (small m), the repulsion of the DPP has to be strong in order to sample a diverse subset. Whereas if the length of the summary is less constrained, τ should be decreased to allow for a less coarse-grained description. This observation leads to the natural question of the optimal τ given the data and m.

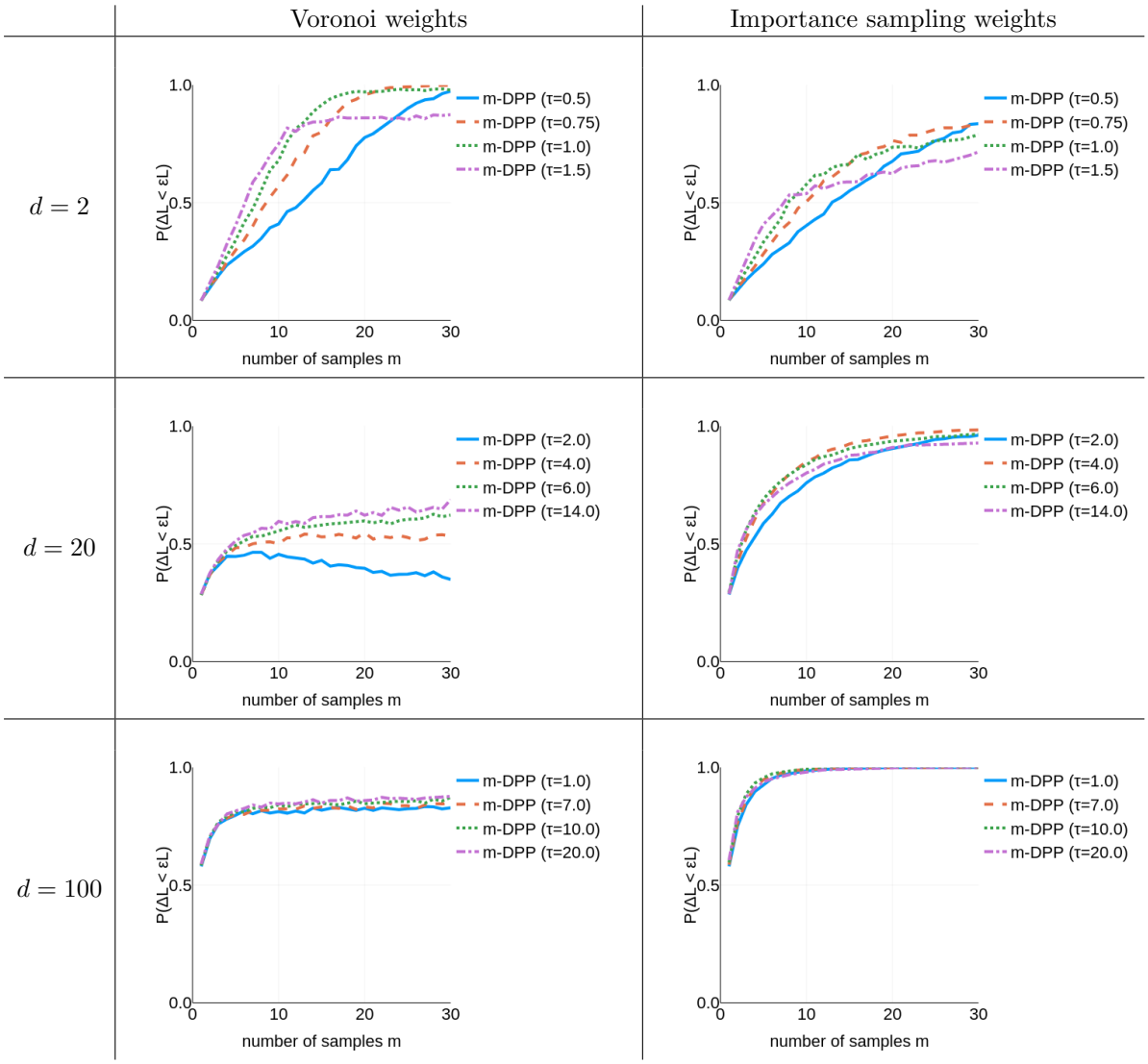

Figure 4. *Performance of* `m-DPP` *on the* 1*-means problem, versus the dimension* d*, the parameter* τ *of the Gaussian kernel and the choice of weights (Voronoi or importance sampling weights) in the cost estimator.*

We currently lack of a satisfying answer to this question, both theoretically and empirically. A usual heuristics in kernel methods is to set τ to the average (or median) interdistance of the points in the dataset. In the experiments of Fig. 4, the average interdistance corresponds to $\tau \simeq 1.7, 6.2$ and 14.0 for $d = 2, 20$ and 100 respectively, which give in fact a good order of magnitude for the choice of τ. In the following, to simplify the discussion, we will sometimes set τ to be the average interdistance, that we will denote by $\bar{\tau}$.

We compare next the performance of several methods in Fig. 5. One observes that the superior performance of the Voronoi estimator over the importance sampling estimator in low dimension d is verified for all methods. Moreover, as the dimension increases, all methods converge to the performance of the uniform iid sampling method. Finally, `m-DPP` associated with Voronoi weights is competitive with $\texttt{D}^2$ in low d; and, regardless of how one chooses the weights, m-`DPP` has a clear edge over the sensitivity-based iid random sampling (the lower the dimension, the clearer the edge). Finally, `PolyProj-DPP` matches the performance of `sensitivity iid` for importance sampling weights. For

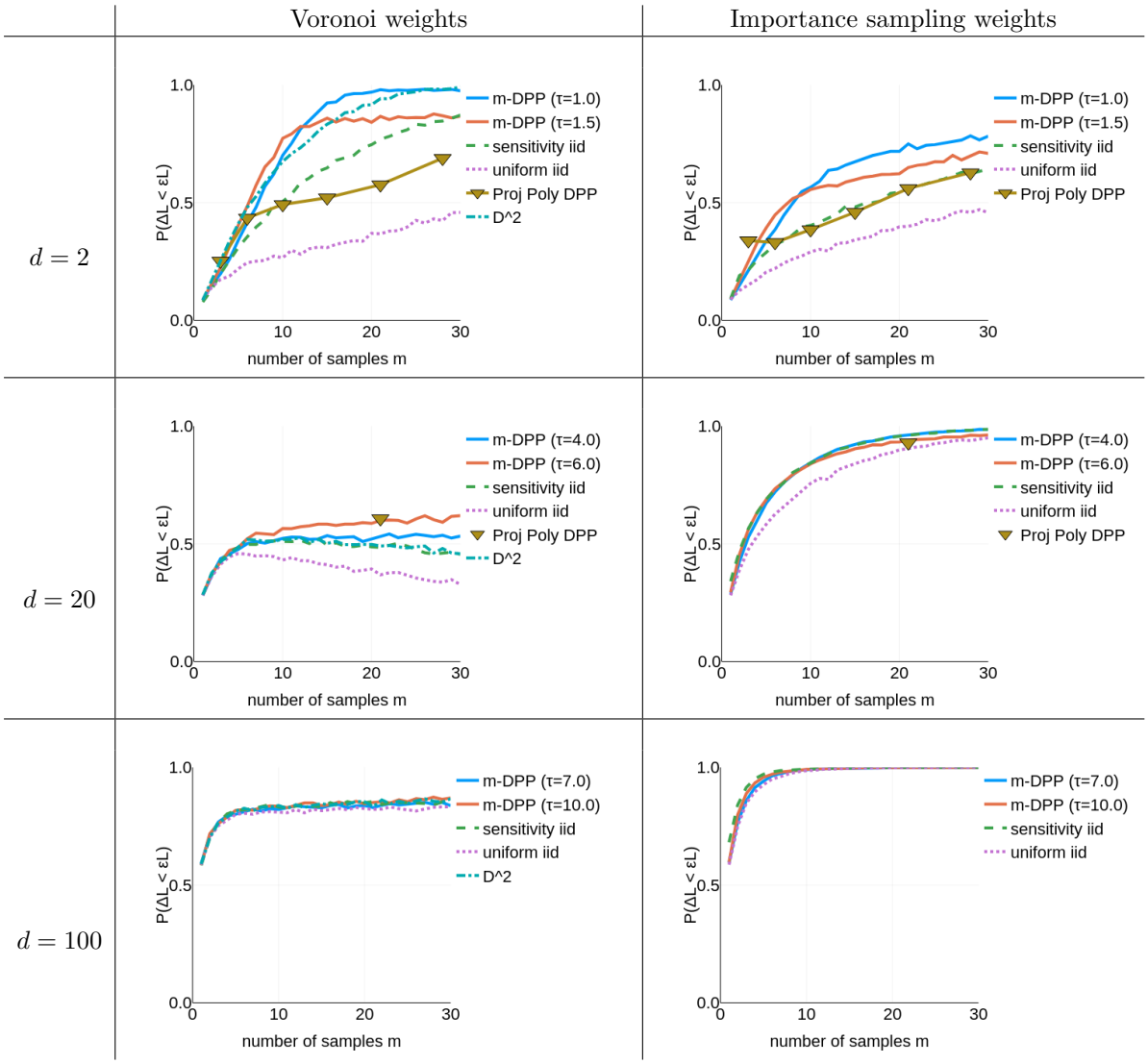

Figure 5. *Performance comparison of different sampling methods on the* 1*-means problem, versus the dimension* d *and the choice of weights (Voronoi or importance sampling weights) in the cost estimator.*

Voronoi weights, the results for $d = 2$ and $d = 20$ are contradictory and we cannot conclude.

In order to clarify further discussion, we will now focus on the importance sampling estimated cost. One should keep in mind that in low dimensions, Voronoi-based estimated costs usually perform well, but fail (sometimes drastically) as the dimension increases.

A natural question arises at this point: is the observed edge of `m-DPP` over `sensitivity iid` due to a better probability of inclusion of the point process? Or is it truly due to the negative correlations induced by the determinantal nature of our method? In fact, we compare in Fig. 3 the probability of inclusion for `sensitivity iid` versus `m-DPP`: they have a similar general behavior but are nevertheless quantitatively different. In Fig. 6 (left), we compare `m-DPP` and `PolyProj-DPP` versus `matched iid` and `sensitivity iid`: the observed edge is clearly due to the negative correlations induced by the determinantal nature of our method. As expected from Corollary 4.1, the best inclusion probability is based on the sensitivity. Nevertheless, the figure shows that even if it is not set to its ideal value (as

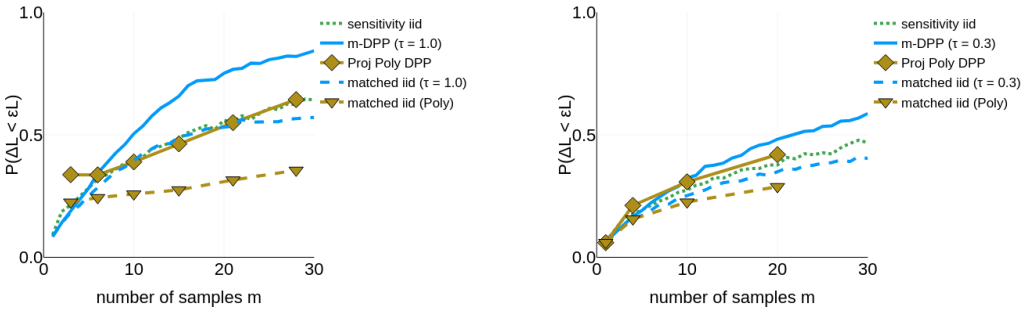

Figure 6. *Comparison between* `m-DPP` *and* `PolyProj-DPP` *versus* `matched iid` *and* `sensitivity iid` *on the 1-means problem (left) and the linear regression problem (right).*

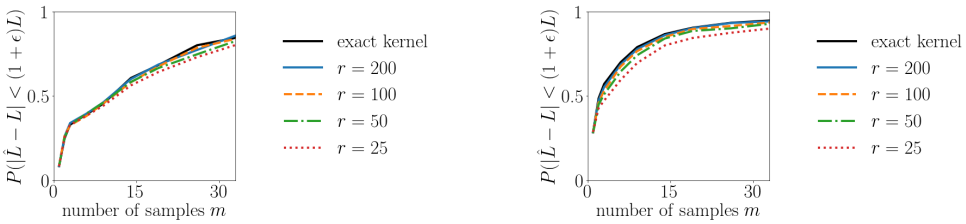

Figure 7. *Performance comparison of* `m-DPP` *on the* 1*-means problem versus the number of RFF* r*, for* $d = 2$ *(left) and* $d = 20$ *(right).*

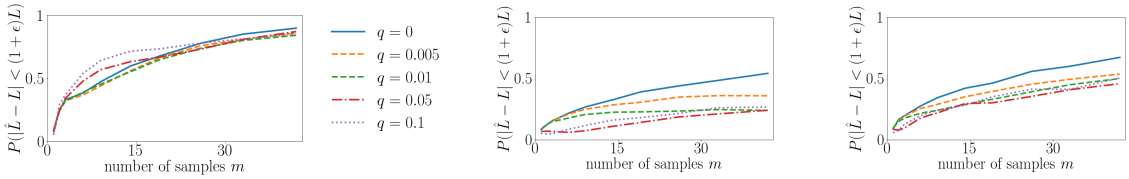

Figure 8. *Performance comparison of* `m-DPP` *(left),* `uniform iid` *(middle) and* `senisitivity iid` *(right) on the* 1*-means problem versus the percentage of outliers* q*.*

in `m-DPP` and `PolyProj-DPP`), one can still improve the performance by inducing negative correlations.

For completeness, we still need to discuss the impact of two variables: the number of random Fourier features r used in `m-DPP`, and the percentage of outliers q in the data. In the following, we set τ to $\bar{\tau}$, the average interdistance. Fig. 7 shows the impact of the choice of r on performances: as expected, as r increases, performance increases, and as d increases, performances become more sensitive to the choice of r. The impact of the choice of r is nevertheless very limited: setting r to a multiple of m has been a safe choice in all our experiments. Finally, Fig. 8 shows the impact of the percentage of outliers q on performances. Empirically, we see here that outliers have a smaller impact on DPP sampling than on uniform or sensitivity-based iid sampling.

Experiments with linear regression. The data $\mathcal{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are generated by sampling n points in the hypercube $[0,1]^d$. Each entry of the vector y is also sampled uniformly from $[0,1]$. The outlier percentage q is set to zero. We show the equivalent of Figs 4 and 5 in, respectively, Figs 9 and 10. Results for $d = 20$ and $d = 100$ are very similar so the case $d = 100$ is not shown.

We observe similar results: `m-DPP` matches $\mathtt{D}^2$ in the Voronoi estimator, and outperforms `sensitivity iid` in all cases; `PolyProj-DPP` at least matches `sensitivity iid` in all cases.

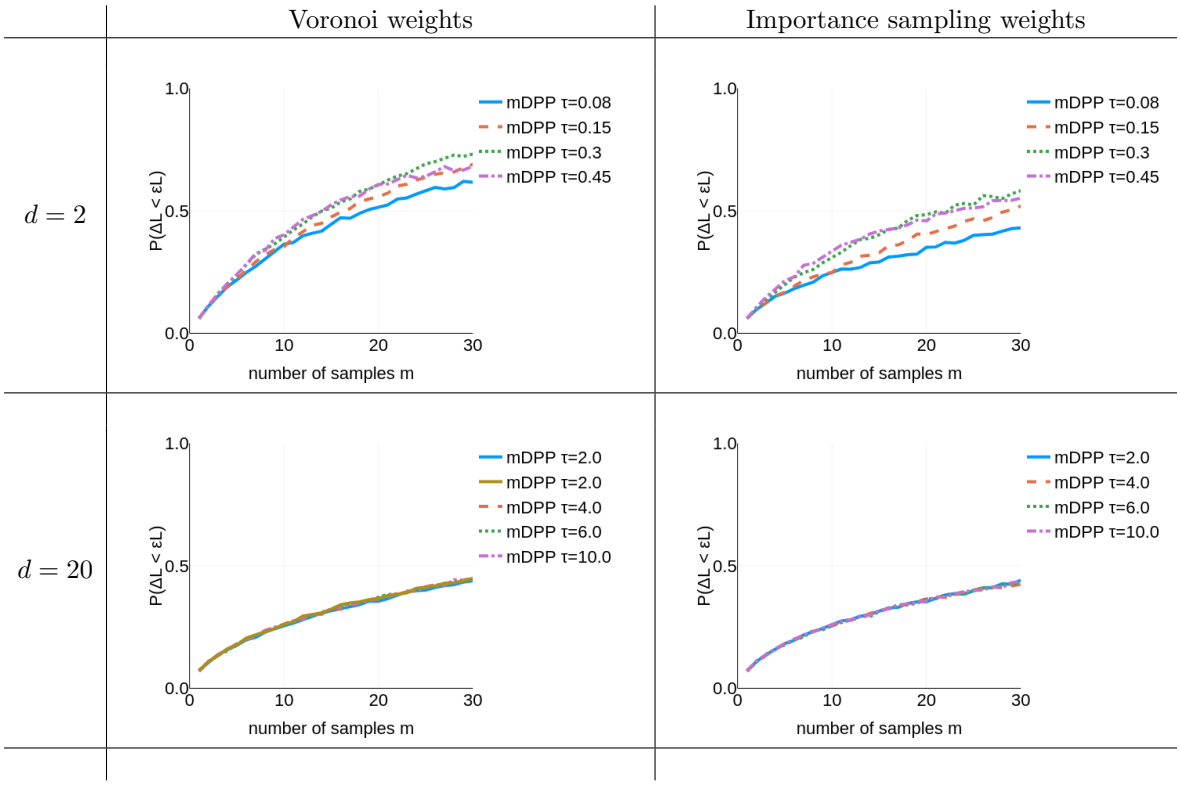

Figure 9. *Performance of* `m-DPP` *on the linear regression problem, versus the dimension d, the parameter* τ *of the Gaussian kernel and the choice of weights (Voronoi or importance sampling weights) in the cost estimator.*

Finally, in Fig. 6 (right), we compare `m-DPP` and `PolyProj-DPP` versus `matched iid` and `sensitivity iid` for the linear regression problem: once again, the observed edge is clearly due to the negative correlations induced by the determinantal nature of our method.

We conclude these first well-controlled experimental results by summarizing the observed behaviors:

- `m-DPP` outperforms the current state of the art `sensitivity iid`, even in the 1-means and the linear regression cases, where sensitivities do not need to be estimated but may be computed exactly.
- `PolyProj-DPP` matches and in some cases outperforms `sensitivity iid`, at least for the importance sampling estimator.
- As the dimension increases, the edge over iid sampling decreases. In fact, all performances tend to `uniform iid`.
- The best choice of parameter τ of the Gaussian kernel in `m-DPP` is still an open problem. Empirically, a good order of magnitude is the average interdistance of the datapoints. Ideally, nevertheless, τ should increase as m, the number of wanted samples, decreases.
- Regarding the number of RFFs r, setting r to $\mathcal{O}(m)$ is sufficient.
- Regarding the impact of outliers. Our theorems are not well suited to outliers (due to the proof techniques used); nevertheless, in practice, we see that outliers are not an issue in our method: they even have a smaller impact on our method's performances than on other methods.
- Replacing weights by Voronoi weights yields in general better results, but only in low dimension. As the dimension increases, the Voronoi cost estimator fails (sometimes drastically).

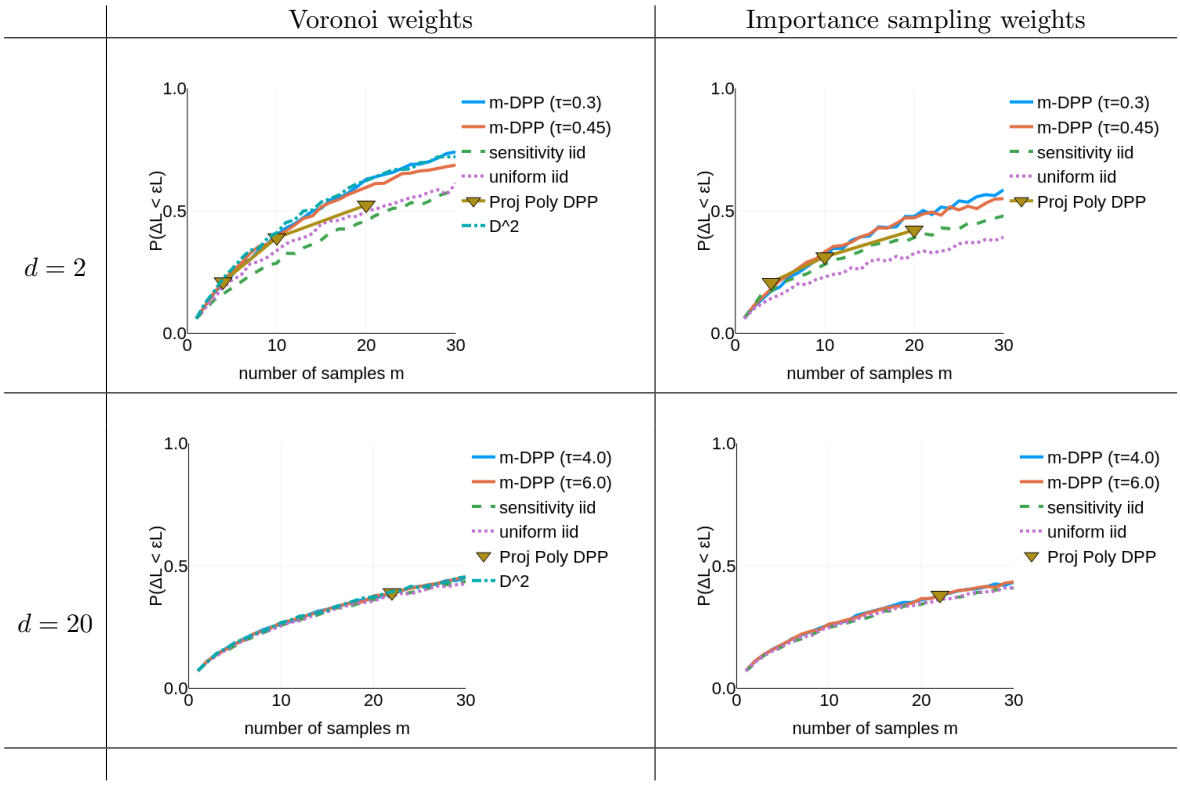

Figure 10. *Performance comparison of different sampling methods on the linear regression problem, versus the dimension d and the choice of weights (Voronoi or importance sampling weights) in the cost estimator.*

6.2.2. Experiments on non-Gaussian data: the case of spectral features

Spectral features. Given a graph of n nodes where $\mathsf{W} \in \mathbb{R}^{n\times n}$ is the adjacency matrix (*i.e.*, $\mathsf{W}_{ij} = 1$ if nodes i and j are connected, and 0 ortherwise), a standard problem consists in partitioning the nodes in k communities, *i.e.*, sets of nodes more connected to themselves than to other nodes of the graph [44]. A classical algorithm to solve efficiently this problem is the so-called spectral clustering algorithm [45]:

- Define the normalized Laplacian matrix $\mathsf{L} = \mathsf{I} - \mathsf{D}^{-\frac{1}{2}}\mathsf{W}\mathsf{D}^{-\frac{1}{2}} \in \mathbb{R}^{n\times n}$ where I is here the identity matrix in dimension n, and $\mathsf{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with $\mathsf{D}_{ii} = d_i = \sum_j \mathsf{W}_{ij}$ the degree of node i.
- Compute via Arnoldi iterations or a similar algorithm the k first eigenvectors of L: $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)$.
- Associate to each node i a (spectral) feature vector $\boldsymbol{x}_i \in \mathbb{R}^k$: $\forall l = 1, \ldots, k \quad x_i(l) = u_l(i)$.
- Normalize all feature vectors: $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i / \|\boldsymbol{x}_i\|_2$.
- Run k-means on all such normalized spectral features.

An extensive literature exists on spectral clustering and it has shown to be a very successful unsupervised classification algorithm in many situations [46, 47].

The Stochastic Block Model (SBM). We consider random community-structured graphs drawn from the SBM, a classical class of structured random graphs (see for instance [48]). We first look at graphs with k communities of same size n/k. In the SBM, the probability of connection between any two nodes i and j is q_1 if they are in the same community, and q_2 otherwise. One can show that the average degree reads $c = q_1\left(\frac{n}{k} - 1\right) + q_2\left(n - \frac{n}{k}\right)$. Thus, instead of providing the probabilities (q_1, q_2), one may characterize a SBM by considering $(\zeta = \frac{q_2}{q_1}, c)$. The larger ζ, the fuzzier the community structure, the harder the classification task. In fact, authors in [49] show that above the critical value

Figure 11. *Examples of SBM spectral features $\boldsymbol{x}_i$, here with $k = 2$. Each colour corresponds to one block of the SBM. On the left, for an "easy" classification task ($\zeta = \zeta_c/4$), and on the right, for a harder setting ($\zeta = \zeta_c/2$).*

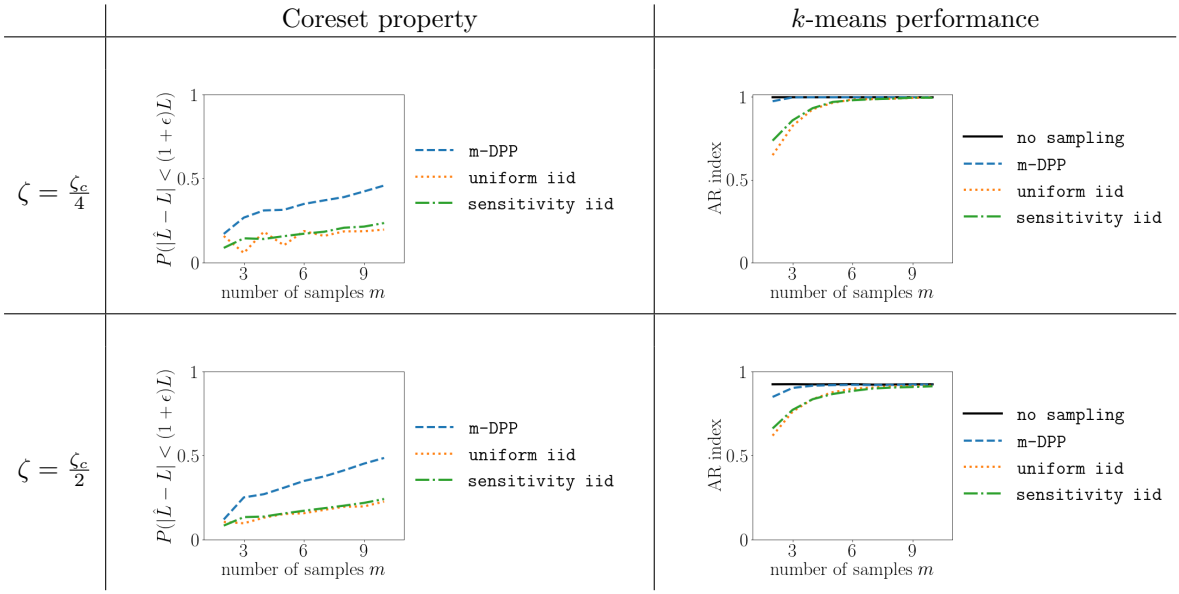

Figure 12. *Performance comparison of different methods on the k-means problem for spectral features of balanced SBM graphs (here, $k = 2$). Left: testing the coreset property. Right: the Adjusted Rand index between the partition recovered by k-means on the weighted subsets and the ground truth partition of the SBM. ζ quantifies the difficulty of the classification task (see text): the lower it is, the easier the classification task.*

$\zeta_c = (c - \sqrt{c})/(c + \sqrt{c}(k-1))$, community structure becomes undetectable in the large n limit. In the following, we set $n = 1000$ and $c = 16$; k and ζ will vary. Note that spectral features $\boldsymbol{x}_i$ are not Gaussian and, in fact, do not fall into any classical data model (see Fig. 11 to visualize instances of SBM spectral features with $k = 2$). They are thus interesting candidates to test k-means algorithms.

Results. For different values of ζ and k, we generate 1000 such SBM graphs from which we sample subsets according to different methods. We test both the coreset property (as before) and the k-means performance on the weighted subset compared to the k-means performed on all data. We plot in Fig. 12 (resp. Fig. 13) the results obtained for $k = 2$ (resp. $k = 10$). Note that in this case, we have no explicit formula for the sensitivity such that for `sensitivity iid`, we use the bi-criteria approximation scheme provided in [16] (Alg. 2). Here again, we see how our method outperforms iid sampling schemes, even in difficult classification contexts (for instance when $\zeta = \zeta_c/2$: even with all the data, k-means' performance saturates at an AR index of 0.9). Moreover, as the dimension increases (here $d = k$), performances of all methods tend to uniformize. Surprisingly, `uniform iid` performs as well ($k = 2$) and even outperforms ($k = 10$) `sensitivity iid`. We believe this is due to approximation errors of the bi-criteria scheme used to find upper bounds of the sensitivity. Also, in

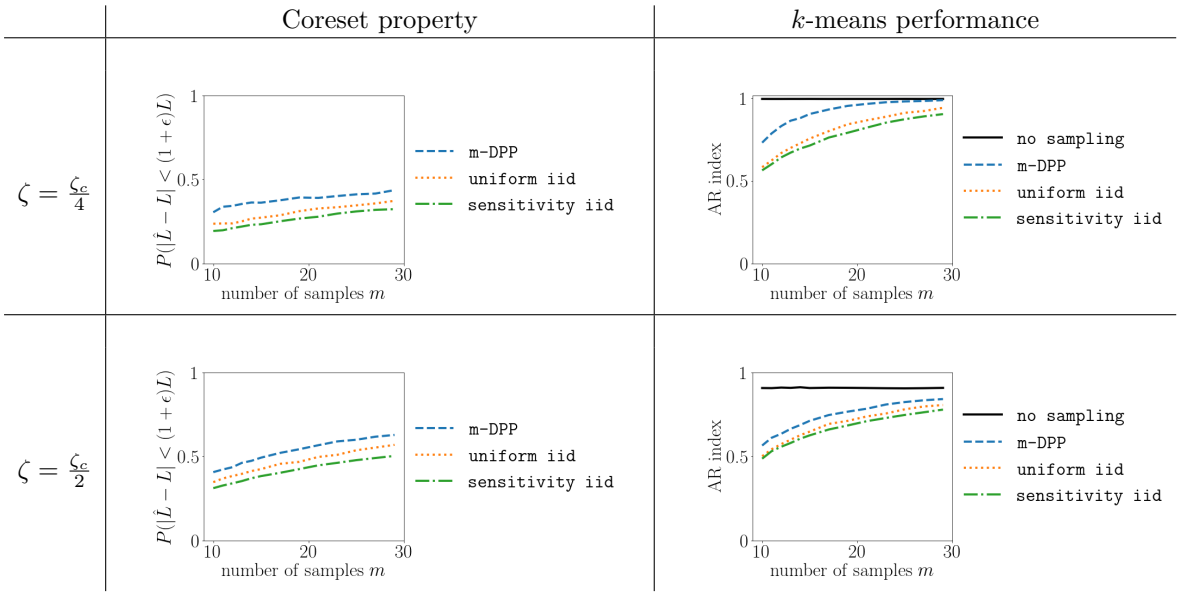

Figure 13. *Same as Fig. 12 but with* $k = 10$.

this balanced case (communities have the same number of nodes), uniform sampling is in fact a good option. We will now see how this changes in the unbalanced case.

The unbalanced case. In the unbalanced case, ζ_c is no longer a recovery threshold, but we may still use ζ as a marker of difficulty of the recovery task. We set ζ to $\zeta_c/4$ and perform the same experiments as previously with $k = 2$ blocks of unbalanced size. Results are shown in Fig. 14. For a fixed ζ, the more unbalanced, the more difficult the recovery task. Also, the more unbalanced, the better is `sensitivity iid` compared to `uniform iid`. Nevertheless, `m-DPP` shows an edge over all iid methods in all tested configurations.

6.2.3. Experiments on two real world datasets

The MNIST dataset. We perform a first experiment on the MNIST dataset [50] that consists in $7 \cdot 10^4$ images of handwritten digits (from 0 to 9) for which the ground truth is known. The classical associated machine learning goal is to classify them in 10 classes (one for each digit). To do so, we pre-process the data in the following unsupervised way. We consider all images and extract SIFT descriptors [51] for each image. We then use FLANN [52] to compute a κ-nearest neighbor graph (with κ=10) based on these descriptors. We finally run the spectral clustering algorithm with $k = 10$ to find the 10 classes corresponding to each digit, as explained in Section 6.2.2. The k-means step is thus the last step of the overall processing. We compare results obtained with different sampling methods versus the results obtained without sampling in Fig. 15 (bottom). For `m-DPP`, several values of τ were tried, and we show here the result obtained for $\tau = 1.5$. Also, a number $r = 200$ of Fourier features were used. We see that, in the Voronoi weight setting, `m-DPP` is competitive with `D`^{2}. Moreover, `uniform iid` outperforms `sensitivity iid` certainly due to approximation errors of the bi-criteria procedure and to the fact that the data is balanced (there are more or less $7 \cdot 10^3$ instances of each digit in the dataset), thus favoring uniform sampling. Finally, `m-DPP` outperforms once again the iid random sampling techniques. Note that the methods' classification performance is remarkable. Without sampling, the overall classification performance in terms of AR index with the ground truth is 0.95. With only ~ 20 samples, `m-DPP` reaches a performance of ~ 0.9!

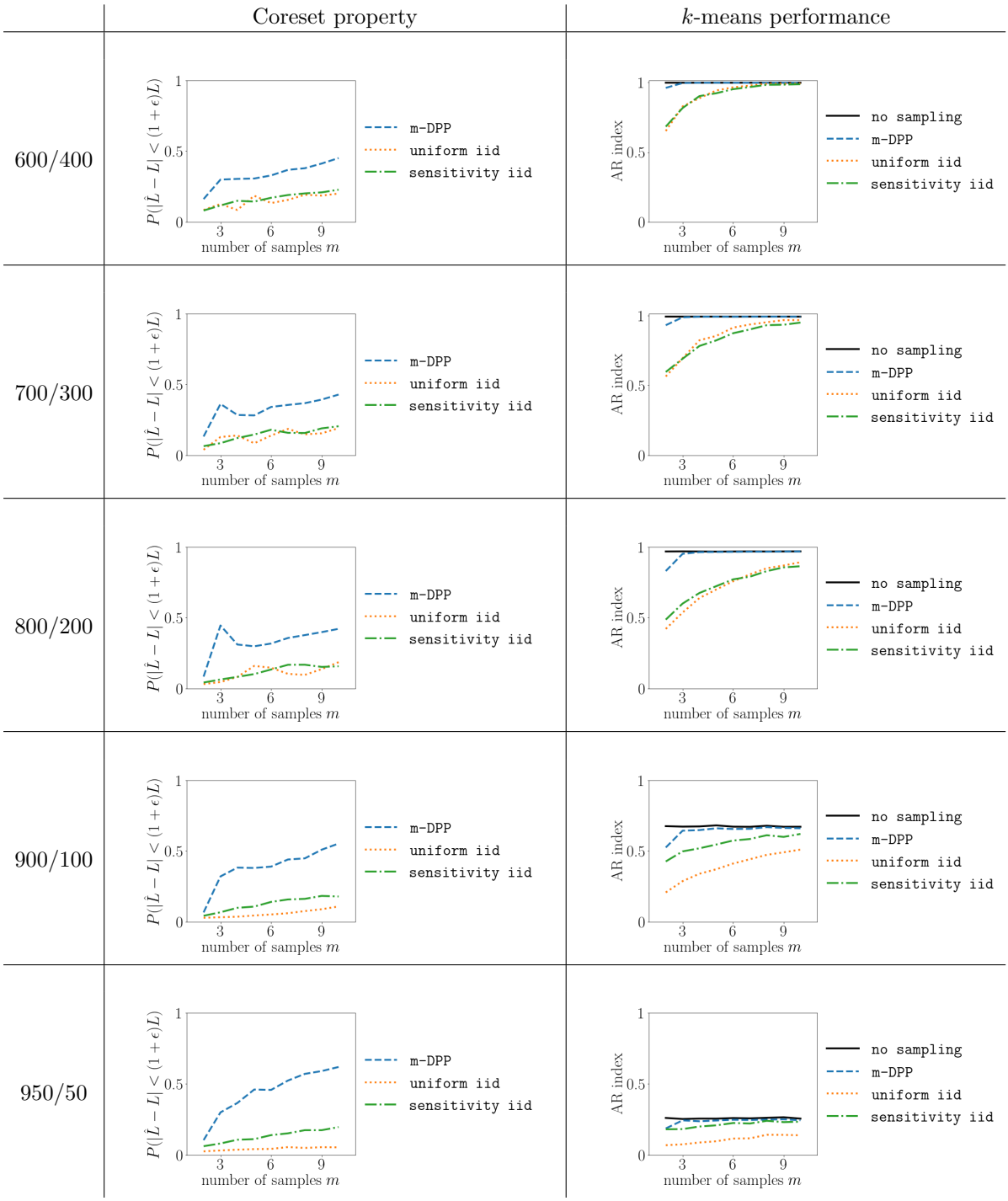

Figure 14. *Same as Fig. 12 but with a fixed* $\zeta = \zeta_c/4$ *and a varying level of balance within the sizes of the* $k = 2$ *communities.* n_1/n_2 *means one community with* n_1 *nodes and the other with* n_2 *nodes.*

The US Census dataset. We also perform experiments on the 1990 US Census dataset[5], that consists in $n = 2458285$ surveyed persons, and $d = 68$ categorical attributes such as age, income, etc.

[5]downloaded from `https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990)`

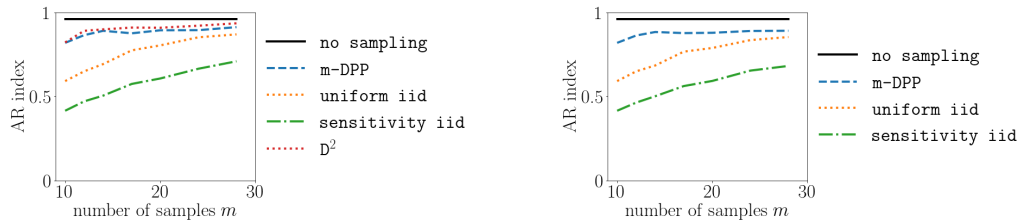

Figure 15. *Classification performance on the MNIST dataset obtained with different sampling methods versus the result obtained without sampling, using Voronoi weights (left), or importance sampling weights (right).*

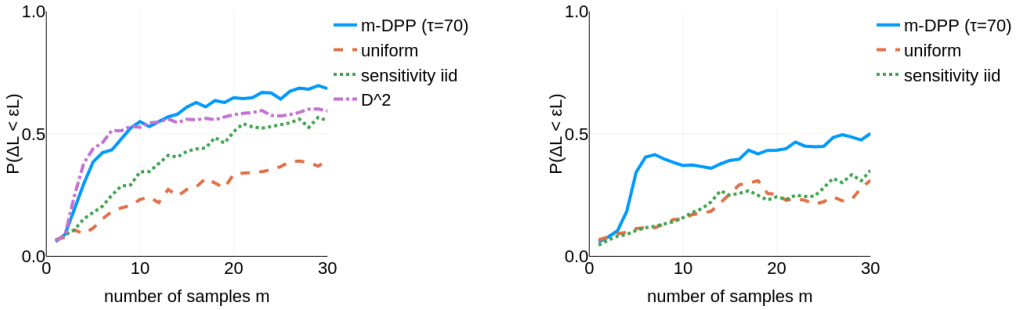

Figure 16. *Performance of different sampling methods on the US Census dataset. Left: performance using Voronoi weights. Right: performance using importance sampling weights.*

The data was pre-processed by a series of operation detailed on its download webpage. As there is no ground truth in this dataset to compare to, we arbitrarily decide $k = 15$ classes, and show solely the coreset property of the samples obtained via different methods. For `m-DPP`, τ was set to 70 (the mean interdistance estimated on 1000 randomly chosen pairs of datapoints), and a number $r = 30$ of Fourier features was chosen. Experiments were done with τ ranging from $\tau = 30$ to $\tau = 140$ with no qualitative change in performance (not shown). Fig. 16 shows the results of the experiments. We see that `m-DPP` outperforms all other methods, in both Voronoi and importance sampling settings. In this example, note that `sensitivity iid` outperforms `uniform iid` probably due to the fact that the 15 potential classes are unbalanced.

6.3. Computation time

Comparing computation times is always a tricky endeavour: they heavily depend on the quality of implementation, choice of language, choice of experiments, choice of parameters, hardware, etc. Giving the full picture is out-of-scope of this paper. We hope here to give some insight in the computational complexity of the proposed methods. We suggest to first recap theoretical times before looking at times observed in a few experiments.

In theory. The theoretical time for `m-DPP` is $\mathcal{O}(nrd)$ for the RFFs, $\mathcal{O}(nr^2)$ for the SVD of the RFFs, $\mathcal{O}(nm^2)$ for the sampling, and $\mathcal{O}(nm + r)$ for the computation of the inclusion probabilities π. As r is set to $\mathcal{O}(m)$, this sums up to $\mathcal{O}(nm^2 + nmd)$. The theoretical time for `PolyProj-DPP` is $\mathcal{O}(nm)$ to compute the Vandermonde matrix, $\mathcal{O}(nm^2)$ for the QR decomposition, $\mathcal{O}(nm^2)$ for the sampling, and $\mathcal{O}(nm)$ to compute the inclusion probabilities π (which are simply the sum of squares of each line of Q); which sums up to $\mathcal{O}(nm^2)$. The theoretical time for Alg. 2 of [16] used here for the bi-criteria approximation of the sensitivities in the k-means context, is $\mathcal{O}(Indk)$ where I is the number of times Alg. 1 of [16] is run to find the best initialization (we set it to $I = 10$ in our experiments). Once upper bounds of the sensitivities have been computed, the iid sampling time is negligeable. The theoretical

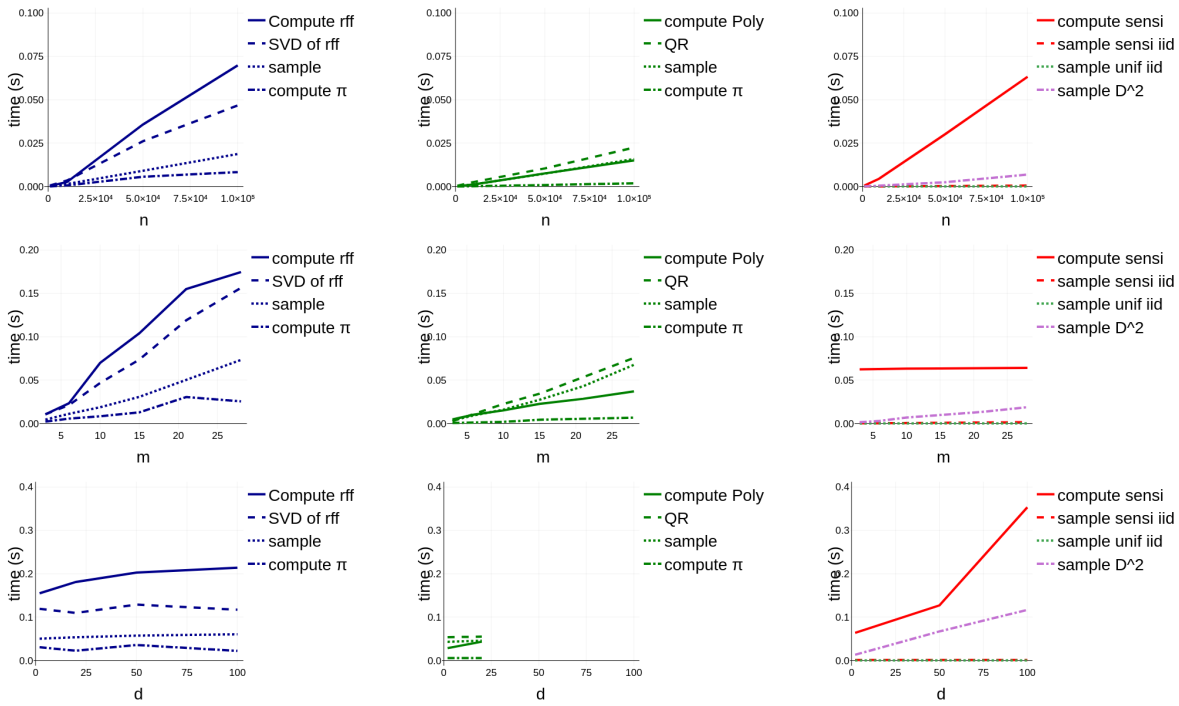

Figure 17. *Computation times for all methods. Top line:* $d = 2$, $m = 10$, *versus* n. *Middle line:* $n = 10^5$, $d = 2$, *versus* m. *Bottom line:* $n = 10^5$, $m = 21$, *versus* d. *Left column: the four fundamental operations associated to* `m-DPP`*: i/ compute the RFFs with* r *set to* m *here, ii/ compute the SVD of the RFFs, iii/ sample from the* m*-DPP, iv/ compute the marginal probabilities* π *useful for the importance sampling estimator. Middle column: the four equivalent operations for* `PolyProj-DPP`*: i/ compute the Vandermonde matrix ("compute Poly"), ii/ compute the QR decomposition of that matrix, iii/ sample from the projective DPP, iv/ compute the marginal probabilities* π*. Right column: the two fundamental operations for* `sensitivity iid`*: i/ the computation of the bi-criteria approximation ("compute sensi"), ii/ the weighted iid sampling with replacement itself, along with* `uniform iid` *and* $\mathtt{D}^2$.

time of $\mathtt{D}^2$ is $\mathcal{O}(ndm)$.

In practice. Experiments were made on a laptop with 8 cores and 16 GB of memory, with the Julia toolbox available on the authors's website[6]. Fig. 17 shows some computation times versus m, d and n for the 1-means problem. We observe a linear progression in n of the determinantal methods, as well as a superlinear progression in m for some parts of the computations. Comparing with `sensitivity iid`, one observes: the lower m and the larger d, the more our methods are comparable in terms of computation time.

Note that this comparison is in fact in favor of `sensitivity iid`: the computation time of the bi-criteria approximation increases in fact with the number of expected classes k. The figures represented here are for $k = 1$. We reproduce in the following table the computation times for the US-Census dataset, with $k = 15$, $m = 30$, $r = 30$:

`m-DPP`: rff	`m-DPP`: svd	`m-DPP`: sample	`m-DPP`: π	`sensi iid`: bi-criteria	`sensi iid`: sample	$\mathtt{D}^2$
2.6s	2.5s	2.7s	1.3s	18.7s	0.04s	3.1s

The total time for m-DPP sampling is thus half the time necessary for the bi-criteria approximation; for a substantial gain in performance, as seen in Fig. 16.

[6] or at `https://gricad-gitlab.univ-grenoble-alpes.fr/tremblan/dpp4coresets.jl`

To conclude, the determinantal methods proposed in this paper are in many situations, especially as m increases, and d decreases, heavier in computation time than iid sampling. Due to the observed gain in coreset performance, however, we believe that the additional cost is worth the effort.

7. Conclusion

In this work, we introduced a new random sampling method based on DPPs to build coresets. Different from sensitivity-based iid random sampling, our method introduces negative correlations between samples due to its determinantal nature. Also, different from D^2 sampling, also known to be repulsive, the proposed method is tractable in the sense that marginal probabilities are known and importance sampling schemes can be used. Our theoretical results may be summarized in three points. Firstly, Thms 3.1 and B.1 provide coreset guarantees in function of the point process' probabilities of inclusion. These guarantees are not stronger than the iid case and are in fact similar: they both show that the ideal marginal probabilities are proportional to the sensitivity. Nevertheless, these results do not take into account higher order inclusion probabilities coding for the repulsion within the sampled subsets and are in fact verified for any choice of such high-order marginals (provided K stays SDP with eigenvalues between 0 and 1). This leads to the second point: given that these higher-order inclusion probabilities offer extra degrees of freedom and due to simple variance arguments (theorems of Section 3.2), we show that DPP-based random sampling necessarily yields better performance than its independent counterparts. On the theoretical side, additional work is required to specify precisely the minimum number of required samples guaranteeing the coreset property. We expect that further research on concentration properties of strongly Rayleigh measures, involving not only first order marginals, but higher-order ones, should enable to move forward in this direction. The third and final point is the rebalancing property of polynomial DPPs: without any prior density-like estimation, polynomial DPPs such as the ones described in Section 3.3.1 provide samples that are asymptotically independent of the underlying distribution of the data. Even though this result is only asymptotic, it is yet another argument in favor of DPP sampling for coresets.

From an application point-of-view, the coreset theorems were applied to the ubiquitous k-means and linear regression problems. Given a dataset, the ideal L-ensemble L adapted to these problems is untractable and we thus propose two heuristics, one via random Fourier features of the Gaussian kernel, and one based on the Vandermonde matrix, in order to efficiently sample a DPP that has the desirable properties to sample coresets (if not provably, at least quantitatively). To sample a subset of size m, our heuristics run resp. in $\mathcal{O}(nm^2 + nmd)$ and $\mathcal{O}(nm^2)$. This is more expensive than the sensitivity-based iid strategy, especially as the number of samples m increases; but empirically provides better coresets on different artificial and real-world datasets.

Finally, this work calls for several extensions. First of all, two likely difficult theoretical questions: how to improve the concentration inequalities for DPPs? (such improvements would directly benefit the coreset theorem's bounds) How to find the optimal DPP kernel given a dataset? (which asks in fact difficult questions in frame theory) Also, these DPP sampling schemes should be extended to the streaming and/or distributed settings.

. Acknowledgments

This work was partly funded by the ANR GenGP (ANR-16-CE23-0008), the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01), the CNRS PEPS I3A (Project RW4SPEC), the Grenoble Data Institute (ANR-15-IDEX-02) and the LIA CNRS/Melbourne Univ Geodesic.

Appendix A - Proof of Theorem 3.1

Proof. The theorem consists in proving that Eq. (6) is true under the assumptions of Thm. 3.1. We follow a classical proof scheme from compressed sensing [53], in four steps:

1) we first use concentration arguments for a given $\theta \in \Theta$.
2) we then build an ϵ-net paving the space of parameters Θ.

3) via the union bound, we obtain the result for all θ in the ϵ-net.
4) via the Lipschitz property of f, we obtain the desired result for all $\theta \in \Theta$.

Step 1 (Concentration around $\theta \in \Theta$) DPPs are instances of sampling schemes that are strongly Rayleigh. Since strongly Rayleigh distributions are closed under truncation, any m-DPP is also strongly Rayleigh. We can thus apply the concentration results of [33]. For a given $\theta \in \Theta$, we have : $\forall \epsilon \in (0,1), \forall \delta \in (0,1)$:

$$\mathbb{P}\left(\left|\frac{\hat{L}}{L} - 1\right| \geqslant \epsilon\right) = \mathbb{P}\left(\left|\hat{L} - L\right| \geqslant \epsilon L\right) \leqslant \delta, \tag{57}$$

provided that:

$$m \geqslant \frac{8}{\epsilon^2} C^2 \log \frac{2}{\delta}, \tag{58}$$

with $C = \max_i \frac{f(x_i, \theta)}{L \bar{\pi}_i}$, where $\bar{\pi}_i$ is a shorthand for π_i / m, and π_i is the marginal probability of sampling element i.

Using the same concentration results, we also have:

$$\forall (\epsilon, \delta) \in (0,1)^2, \ \mathbb{P}\left(\left|\frac{\sum_i \frac{\epsilon_i}{\pi_i}}{n} - 1\right| \geqslant \epsilon\right) \leqslant \delta, \tag{59}$$

provided that:

$$m \geqslant \frac{8}{\epsilon^2} \frac{1}{n^2 \bar{\pi}_{\min}^2} \log \frac{2}{\delta}, \tag{60}$$

where $\bar{\pi}_{\min} = \min_i \bar{\pi}_i$.

Step 2 (ϵ'-net of Θ) Consider $\Gamma_{\epsilon'} = (\theta_1^*, \ldots, \theta_\eta^*)$ the smallest subset of Θ such that balls of radius ϵ' centered around the elements in $\Gamma_{\epsilon'}$ cover Θ. $\Gamma_{\epsilon'}$ is called an ϵ'-net of Θ and $\eta = |\Gamma_{\epsilon'}|$ its covering number. The covering property entails that:

$$\forall \theta \in \Theta \quad \exists \theta^* \in \Gamma_{\epsilon'} \ \text{ s.t. } \ d_\Theta(\theta, \theta^*) \leqslant \epsilon'. \tag{61}$$

Step 3. (Union bound) Write $\delta' = \delta / 2\eta$. From step 1, we know that, $\forall \theta^* \in \Gamma_{\epsilon'}$:

$$\mathbb{P}\left(\left|\frac{\hat{L}}{L} - 1\right| \geqslant \epsilon\right) \leqslant \delta' \tag{62}$$

provided that:

$$m \geqslant \frac{8}{\epsilon^2} C^2 \log \frac{2}{\delta'}. \tag{63}$$

From the union bound, we have:

$$\mathbb{P}\left(\forall \theta^* \in \Gamma_{\epsilon'}, \quad \left|\frac{\hat{L}}{L} - 1\right| \leqslant \epsilon\right) \geqslant 1 - \sum_{\theta^* \in \Gamma} \delta' = 1 - \frac{\delta}{2}, \tag{64}$$

provided that:

$$m \geqslant \frac{8}{\epsilon^2} \left(\max_{\theta^* \in \Gamma_{\epsilon'}} C\right)^2 \log \frac{4\eta}{\delta} \tag{65}$$

Given that $\bar{\pi}_i$ will *in fine* be independent of θ (as we want the coreset property to be true for all $\theta \in \Theta$),

$$\max_{\theta^* \in \Gamma_{\epsilon'}} C = \max_{\theta^* \in \Gamma_{\epsilon'}} \max_i \frac{f(x_i, \theta)}{L \bar{\pi}_i} \tag{66}$$

$$= \max_i \frac{1}{\bar{\pi}_i} \max_{\theta^* \in \Gamma_{\epsilon'}} \frac{f(x_i, \theta)}{L} \tag{67}$$

$$\leqslant \max_i \frac{1}{\bar{\pi}_i} \max_{\theta \in \Theta} \frac{f(x_i, \theta)}{L} = \max_i \frac{\sigma_i}{\bar{\pi}_i}, \tag{68}$$

where we see how the sensitivity σ_i naturally arises in the proof. Eq. (68) entails that Eq. (65) is verified if $m \geqslant m_1$ with

$$m_1 \overset{=}{} \frac{8}{\epsilon^2} \left(\max_i \frac{\sigma_i}{\bar{\pi}_i} \right)^2 \log \frac{4\eta}{\delta}. \tag{69}$$

Write $\delta'' = \delta/2$. From Eq. (59), we have:

$$\mathbb{P}\left(\left| \frac{\sum_i \frac{\epsilon_i}{\pi_i}}{n} - 1 \right| \geqslant \epsilon \right) \leqslant \delta'', \tag{70}$$

provided that $m \geqslant m_2$ with

$$m_2 = \frac{8}{\epsilon^2 n^2 \bar{\pi}^2_{\min}} \log \frac{4}{\delta}. \tag{71}$$

We have (with the union bound again):

$$\mathbb{P}\left(\left| \frac{\sum_i \frac{\epsilon_i}{\pi_i}}{n} - 1 \right| \leqslant \epsilon \quad \text{AND} \quad \forall \theta^* \in \Gamma_{\epsilon'}, \quad \left| \frac{\hat{L}}{L} - 1 \right| \leqslant \epsilon \right)$$

$$\geqslant 1 - \delta/2 - \delta'' = 1 - \delta, \tag{72}$$

provided that:

$$m \geqslant \max(m_1, m_2). \tag{73}$$

Step 4 (Continuity argument) Suppose that $m \geqslant \max(m_1^*, m_2^*)$ with m_1^*, m_2^* as defined in the theorem. The result of step 3 with $\epsilon \leftarrow \epsilon/2$ states that, with probability at least $1 - \delta$, one has:

$$\left| \frac{\sum_i \frac{\epsilon_i}{\pi_i}}{n} - 1 \right| \leqslant \frac{\epsilon}{2} \quad \text{AND} \quad \forall \theta^* \in \Gamma_{\epsilon'}, \ \left| \frac{\hat{L}}{L} - 1 \right| \leqslant \frac{\epsilon}{2}. \tag{74}$$

We now look for the maximum value of ϵ' such that Eq. (74) implies the following desired result:

$$\forall \theta \in \Theta, \qquad \left| \frac{\hat{L}}{L} - 1 \right| \leqslant \epsilon. \tag{75}$$

Consider $\theta \in \Theta$. By the covering property of $\Gamma_{\epsilon'}$, we have:

$$\exists \theta^* \in \Gamma_{\epsilon'} \ \text{ s.t. } \ d_\Theta(\theta, \theta^*) \leqslant \epsilon'. \tag{76}$$

Moreover, as f is γ-Lipschitz, $\forall x_i \in \mathcal{X}$:

$$|f(x_i, \theta) - f(x_i, \theta^*)| \leqslant \gamma \ d_\Theta(\theta, \theta^*) \leqslant \gamma \epsilon'. \tag{77}$$

Thus, using Eqs. (77) and then (74):

$$\hat{L}(\mathcal{X}, \theta) \leqslant \hat{L}(\mathcal{X}, \theta^*) + \gamma \epsilon' \sum_i \frac{\epsilon_i}{\pi_i} \tag{78}$$

$$\leqslant (1 + \frac{\epsilon}{2})(L(\mathcal{X}, \theta^*) + n\gamma\epsilon'). \tag{79}$$

Also, using Eq. (77) again:

$$L(\mathcal{X},\theta^*) \leqslant L(\mathcal{X},\theta) + n\gamma\epsilon'. \tag{80}$$

Thus:

$$\hat{L}(\mathcal{X},\theta) \leqslant (1+\frac{\epsilon}{2})L(\mathcal{X},\theta) + 2n\gamma\epsilon'(1+\frac{\epsilon}{2}). \tag{81}$$

Similarly, for the lower bound, one obtains:

$$(1-\frac{\epsilon}{2})(L(\mathcal{X},\theta) - 2n\gamma\epsilon') \leqslant \hat{L}(\mathcal{X},\theta) \tag{82}$$

In order for Eqs (81) and (82) to imply Eq.(75), we need:

$$2n\gamma\epsilon'(1+\frac{\epsilon}{2}) \leqslant \frac{\epsilon}{2}L(\mathcal{X},\theta), \tag{83}$$

i.e.:

$$\epsilon' \leqslant \frac{\epsilon L(\mathcal{X},\theta)}{4n\gamma(1+\frac{\epsilon}{2})} \leqslant \frac{\epsilon L(\mathcal{X},\theta)}{6n\gamma}. \tag{84}$$

In order for this condition to be true for all θ, we choose:

$$\epsilon' = \frac{\epsilon \min_{\theta\in\Theta} L(\mathcal{X},\theta)}{6n\gamma} = \frac{\epsilon L^{\text{opt}}}{6n\gamma} = \frac{\epsilon\langle f\rangle_{\text{opt}}}{6\gamma}. \tag{85}$$

Concluding the proof. Consider $\mathcal{S}$ a sample from a DPP with L-ensemble L, with marginal probabilities π_i and normalized marginal probabilities $\bar{\pi}_i = \pi_i/m$. Consider $\epsilon \in (0,1)$ and $\delta \in (0,1)$. Define ϵ' as in Eq. (85) and Γ the set of centers of the η balls of radius ϵ' covering the parameter space. We showed that if $m \geqslant \max(m_1^*, m_2^*)$, then $\mathcal{S}$ is an ϵ-coreset with probability at least $1-\delta$. □

Appendix B - Coreset results for DPPs

Theorem B.1 (DPP for coresets)**.** *Consider $\mathcal{S}$ a sample from a DPP with L-ensemble L, and an average number of sample $\mu = \sum_i \frac{\lambda_i}{1+\lambda_i}$. Let $\epsilon \in (0,1)$, $\delta \in (0,1)$. Denote by η the minimum number of balls of radius $\epsilon\langle f\rangle_{opt}/6\gamma$ necessary to cover Θ. With probability higher than $1-\delta$, $\mathcal{S}$ is a ϵ-coreset provided that*

$$\mu \geqslant \mu^* = \max(\mu_1^*, \mu_2^*) \tag{86}$$

with:

$$\mu_1^* = \frac{32}{\epsilon^2}\left(\epsilon \max_i \frac{\sigma_i}{\bar{\pi}_i} + 4\left(\max_i \frac{\sigma_i}{\bar{\pi}_i}\right)^2\right)\log\frac{10\eta}{\delta}, \tag{87}$$

$$\mu_2^* = \frac{32}{\epsilon^2}\left(\frac{\epsilon}{n\bar{\pi}_{min}} + \frac{4}{n^2\bar{\pi}_{min}^2}\right)\log\frac{10}{\delta}, \tag{88}$$

and $\forall i,\ \bar{\pi}_i = \pi_i/\mu$.

Proof. According to [33], replace Eq. (58) by:

$$\mu \geqslant \frac{16}{\epsilon^2}\left(\epsilon C + 2C^2\right)\log\frac{5}{\delta}, \tag{89}$$

with $C = \max_i \frac{f(x_i,\theta)}{L\bar{\pi}_i}$, where $\bar{\pi}_i$ is a shorthand for π_i/μ; and Eq. (60) by:

$$\mu \geqslant \frac{16}{\epsilon^2 n\bar{\pi}_{\min}}\left(\epsilon + \frac{2}{n\bar{\pi}_{\min}}\right)\log\frac{5}{\delta}, \tag{90}$$

and change accordingly the rest of the proof. □

For the same reasons as the m-DPP case, we have:

Corollary B.2. *If $n\sigma_{min} \geqslant 1$, then $\mu_1^* \geqslant \mu_2^*$ and the coreset property of Theorem B.1 is verified if:*

$$\mu \geqslant \mu^* = \frac{32}{\epsilon^2}\left(\epsilon \max_i \frac{\sigma_i}{\bar{\pi}_i} + 4\left(\max_i \frac{\sigma_i}{\bar{\pi}_i}\right)^2\right)\log\frac{10\eta}{\delta}. \tag{91}$$

Corollary B.3. *If there exists $\alpha > 0$ and $\beta \geqslant 1$ such that:*

$$\forall i \qquad \alpha\sigma_i \leqslant \pi_i \leqslant \alpha\beta\sigma_i, \tag{92}$$

$$\text{and} \qquad \frac{\alpha}{\beta} \geqslant \frac{32}{\epsilon^2}(\epsilon + 4\mathfrak{S})\log\frac{10n}{\delta}, \tag{93}$$

then $\mathcal{S}$ is a ϵ-coreset with probability at least $1-\delta$. In this case, the expected number of samples verifies:

$$\mu \geqslant \frac{32}{\epsilon^2}\beta\mathfrak{S}(\epsilon + 4\mathfrak{S})\log\frac{10n}{\delta}.$$

Appendix C - Proof of Theorem 3.9

We split the proof into two parts. We first show the convergence of the discrete intensity function to its continuous limit (as n goes to infinity). We then deal with the outer limit (as the degree ϕ goes to infinity) to prove the theorem.

C.1. Discrete-to-continuous limit

The discrete DPP defined in Section 3.3.1 has a natural continuous counterpart: namely, instead of sampling $\mathcal{S}$ from $\mathcal{X}$, we directly sample $\mathcal{S}$ from Ω. The corresponding orthogonal polynomials are now orthogonal w.r.t. the measure μ, and the inclusion probabilities turn into intensity functions (in the continous limit, any given point in Ω has probability 0 of being selected, which is why we need to integrate over an ϵ ball). The counterpart of the discrete marginal kernel $\mathsf{K} = \mathsf{Q}\mathsf{Q}^\top$ is now a positive-definite kernel

$$k_\mu(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m} q_{\mu,i}(\boldsymbol{x}) q_{\mu,i}(\boldsymbol{y}) \tag{94}$$

where $q_{\mu,i}$ is the i'th orthogonal polynomial under μ. As in the discrete case, the intensity function for the continuous DPP simply equals the diagonal values of the kernel, i.e:

$$\iota(\boldsymbol{y}) = k_\mu(\boldsymbol{y}, \boldsymbol{y}) \tag{95}$$

We need to introduce the Christoffel functions of a measure. The Christoffel function of μ is defined as:

$$\lambda_{\mu,\phi}(\boldsymbol{y}) = \min_{f \in \Pi_\phi^d} \frac{\int f(\boldsymbol{x})^2 d\mu}{f(\boldsymbol{y})^2} \tag{96}$$

where Π_ϕ^d is the set of polynomials in $\mathbb{R}^d$ of degree less than or equal to ϕ. Re-expressing f in the orthonormal basis for μ, and solving for the argmin in (96), we find:

$$\lambda_{\mu,\phi}(\boldsymbol{y}) = \frac{1}{k_\mu(\boldsymbol{y}, \boldsymbol{y})} \tag{97}$$

so that the intensity function of the (continuous) polynomial DPP is just one over the Christoffel function. The argument is also valid for the discrete case, replacing μ with the empirical distribution $\mu_n = (1/n)\sum \delta_{\boldsymbol{x}_i}$. We may rewrite the empirical Christoffel function as:

$$\lambda_{\mu_n,\phi}(\boldsymbol{y}) = \min_{f \in \Pi_\phi^d, f(\boldsymbol{y})=1} \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)^2 \tag{98}$$

Convergence of the empirical Christoffel function to its continuous limit is proved formally in [54] (th. 3.11), and is easy to see from the formula above ($\lambda_{\mu_n,\phi}(\boldsymbol{y})$ is just the minimum of a quadratic empirical functional). In the large n limit, we have:

$$\lim_{n\to\infty} \lambda_{\mu_n,\phi}(\boldsymbol{y}) = \lambda_{\mu,\phi}(\boldsymbol{y}) \tag{99}$$

a.s., uniformly in $\boldsymbol{y} \in D$. $\lambda_{\mu,\phi}(\boldsymbol{y})$ for $\boldsymbol{y} \in D$ is bounded below in convex domains [55], so that convergence of the inclusion probabilities to the intensity function follows:

$$\lim_{n\to\infty} \mathbb{P}(\boldsymbol{x}_i \in \mathcal{S}|\mathcal{X}) = \iota_\phi(\boldsymbol{x}_i) \tag{100}$$

For the unconditional intensity measure, we need to average the left-hand side over $\mathcal{X}$:

$$\begin{aligned}\lim_{n\to\infty} I_{n,\phi}(\mathcal{A}) &= \lim_{n\to\infty} \mathbb{E}_{\mathcal{X},\mathcal{S}}\left\{\sum_{s_i\in\mathcal{S}} \mathbb{I}\left(s_i \in \mathcal{A}\right)\right\} \\ &= \lim_{n\to\infty} \mathbb{E}_{\mathcal{X}}\left\{\sum_{x_i\in\mathcal{X}} \mathbb{P}(\boldsymbol{x}_i \in \mathcal{S}|\mathcal{X})\mathbb{I}\left(x_i \in \mathcal{A}\right)\right\}\end{aligned}$$

The quantities in the expectation are bounded ($\leqslant 1$), and by dominated convergence we may interchange the limit and the expectation, so that:

$$\lim_{n\to\infty} I_{n,\phi}(\boldsymbol{y}) = \int_{\mathcal{A}} \iota_\phi(\boldsymbol{y}) d\mu(\boldsymbol{y}) \tag{101}$$

uniformly in $\boldsymbol{y} \in D$.

C.2. Large-m asymptotics

For the next step, we use the fact that asymptotics of Christoffel functions are well-studied. We let $\phi \to \infty$, in which case Th. 1.5 in [37] gives us:

$$\lim_{\phi\to\infty} \frac{\iota_\phi(\boldsymbol{y})}{k_{\nu,\phi}(\boldsymbol{y},\boldsymbol{y})} = \frac{1}{\mu'(\boldsymbol{y})} \tag{102}$$

Here $k_{\nu,\phi}$ is the projection kernel for orthogonal polynomials of degree l under the Lebesgue measure. Note that $\iota_\phi(\boldsymbol{y})$ and $k_{\nu,\phi}(\boldsymbol{y},\boldsymbol{y})$ both integrate to m over Ω, and $\frac{1}{k_{\nu,\phi}(\boldsymbol{y},\boldsymbol{y})}$ is the Christoffel function for ν, which tends to a well-defined limit independent of μ. Injecting (102) into (101) proves our result.

Appendix D - Proof of three Lemmas

Lemma D.1. *In the 1-means problem (the k-means problem with $k = 1$), and supposing without loss of generality that the data is centered (i.e.: $\sum_j x_j = 0$), we have:*

$$\sigma_i = \frac{1}{n}\left(1 + \frac{\|x_i\|^2}{v}\right), \tag{103}$$

where $v = \frac{1}{n}\sum_{x\in\mathcal{X}} \|x\|^2$. Thus:

$$\mathfrak{S} = \sum_i \sigma_i = 2. \tag{104}$$

Proof. By definition:

$$\frac{1}{\sigma_i} = \min_c \frac{\sum_x \|x - c\|^2}{\|x_i - c\|^2}.$$

Consider $\mathcal{S}(x_i, R)$ the sphere centered on x_i and radius $R \geqslant 0$. We have that:

$$\min_c = \min_{R\geqslant 0} \min_{c\in\mathcal{S}}$$

We thus have:

$$\frac{1}{\sigma_i} = \min_{R \geqslant 0} \frac{1}{R^2} \min_{c \in \mathcal{S}} \sum_x \|x - c\|^2 .$$

Writing $x - c = x - x_i - (c - x_i)$, we may write

$$\sum_x \|x - c\|^2 = nR^2 + \sum_x \|x - x_i\|^2 - 2R \left\| \sum_x x - x_i \right\| \cos\theta,$$

with θ the angle formed by $\sum_x x - x_i$ and $c - x_i$. As the minimum is sought for c on the sphere, the angle θ may take any value, such that the minimum is always attained with θ s.t. $\cos\theta = 1$. We finally obtain:

$$\frac{1}{\sigma_i} = n + \min_{R \geqslant 0} \frac{1}{R^2} \left(\sum_x \|x - x_i\|^2 - 2R \left\| \sum_x x - x_i \right\| \right).$$

Studying analytically the function $f(R) = \frac{a - 2bR}{R^2}$, its minimum is attained for $R^* = \frac{a}{b}$ and $f(R^*) = -\frac{b^2}{a}$, such that:

$$\frac{1}{\sigma_i} = n - \frac{\| \sum_x x - x_i \|^2}{\sum_x \|x - x_i\|^2}.$$

Supposing without loss of generality that the data is centered, *i.e.*: $\sum_x x = 0$ and denoting $v = \frac{1}{n} \sum_x \|x\|^2$, we have:

$$\frac{1}{\sigma_i} = n - \frac{n^2 \|x_i\|^2}{nv + n \|x_i\|^2}.$$

Inverting this equation yields:

$$\sigma_i = \frac{v + \|x_i\|^2}{nv + n \|x_i\|^2 - n \|x_i\|^2} = \frac{1}{n} \left(1 + \frac{\|x_i\|^2}{v} \right)$$

□

Lemma D.2. *In the k-means problem, $n\sigma_{min} \geqslant 1$.*

Proof. Consider $\theta^{\text{opt}} = (c_1^{\text{opt}}, \ldots, c_k^{\text{opt}})$ the optimal solution of k-means and $\{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k\}$ their associated Voronoi sets. Consider $x_i \in \mathcal{X}$ and suppose, without loss of generality that $x_i \in \mathcal{V}_1$. Also, for any $x \in \mathcal{X}$, we denote by $c(x) = \text{argmin}_{c \in \theta} \|x - c\|^2$. We have:

$$\begin{aligned}
\frac{1}{\sigma_i} &= \min_{c_1, \ldots, c_k} \frac{\sum_{x \in \mathcal{X}} \|x - c(x)\|^2}{\|x_i - c(x_i)\|^2} \\
&= \min_{c_1, \ldots, c_k} \frac{\sum_{x \in \mathcal{V}_1} \|x - c(x)\|^2}{\|x_i - c(x_i)\|^2} + \sum_{j=2}^{k} \frac{\sum_{x \in \mathcal{V}_j} \|x - c(x)\|^2}{\|x_i - c(x_i)\|^2}
\end{aligned}$$

Given that, by definition of $c(x)$, $\forall j,\ \|x - c(x)\|^2 \leqslant \|x - c_j\|^2$, we have:

$$\frac{1}{\sigma_i} \leqslant \min_{c_1, \ldots, c_k} \frac{\sum_{x \in \mathcal{V}_1} \|x - c_1\|^2}{\|x_i - c(x_i)\|^2} + \sum_{j=2}^{k} \frac{\sum_{x \in \mathcal{V}_j} \|x - c_j\|^2}{\|x_i - c(x_i)\|^2}$$

To further bound this quantity, let us constrain the domain over which the minimum is sought. Consider $\mathcal{B}(x_i, R)$ the ball centered on x_i and radius $R \geqslant 0$. Consider $\mathcal{S}(x_i, R)$ its surface (*i.e.*, the associated sphere). We have that:

$$\min_{c_1, \ldots, c_k} \leqslant \min_{R \geqslant 0} \min_{c_1 \in \mathcal{S}, (c_2, \ldots, c_k) \notin \mathcal{B}}$$

Given this restricted search space, we have: $c(x_i) = c_1$ and $\|x_i - c_1\|^2 = R^2$, and thus:

$$\frac{1}{\sigma_i} \leqslant \min_{R \geqslant 0} \frac{1}{R^2} \min_{c_1 \in \mathcal{S}} \left(\sum_{x \in \mathcal{V}_1} \|x - c_1\|^2 + \min_{(c_2,\dots,c_k) \notin \mathcal{B}} \sum_{j=2}^{k} \sum_{x \in \mathcal{V}_j} \|x - c_j\|^2 \right)$$

Now, one may show, for all $j = 2, \dots, k$, that:

$$\sum_{x \in \mathcal{V}_j} \|x - c_j\|^2 = \sum_{x \in \mathcal{V}_j} \left\|x - c_j^{\text{opt}}\right\|^2 + \#\mathcal{V}_j \left\|c_j - c_j^{\text{opt}}\right\|^2,$$

due to the fact that $c_j^{\text{opt}} = \frac{1}{\#\mathcal{V}_j} \sum_{x \in \mathcal{V}_j} x$. Given that the minimum of $\left\|c_j - c_j^{\text{opt}}\right\|^2$ is necessarily smaller than R^2:

$$\min_{c_j \notin \mathcal{B}} \sum_{x \in \mathcal{V}_j} \|x - c_j\|^2 \leqslant \sum_{x \in \mathcal{V}_j} \left\|x - c_j^{\text{opt}}\right\|^2 + \#\mathcal{V}_j R^2,$$

such that:

$$\begin{aligned} \frac{1}{\sigma_i} &\leqslant \min_{R \geqslant 0} \frac{1}{R^2} \min_{c_1 \in \mathcal{S}} \left(\sum_{x \in \mathcal{V}_1} \|x - c_1\|^2 + \alpha + (n - \#\mathcal{V}_1) R^2 \right) \\ &= n - \#\mathcal{V}_1 + \min_{R \geqslant 0} \frac{1}{R^2} \min_{c_1 \in \mathcal{S}} \left(\sum_{x \in \mathcal{V}_1} \|x - c_1\|^2 + \alpha \right) \end{aligned}$$

with $\alpha = L^{\text{opt}\backslash\mathcal{V}}$ the optimal $(k-1)$-means cost on $\mathcal{X}\backslash\mathcal{V}$. Writing $x - c_1 = x - x_i - (c_1 - x_i)$, we may decompose $\sum_{x \in \mathcal{V}_1} \|x - c_1\|^2$ in $R^2 \#\mathcal{V}_1 + \sum_{x \in \mathcal{V}_1} \|x - x_i\|^2 - 2R \left\|\sum_{x \in \mathcal{V}_1} x - x_i\right\| \cos\theta$, with θ the angle formed by $\sum_{x \in \mathcal{V}_1} x - x_i$ and $c_1 - x_i$. As the minimum is sought for c_1 on the sphere, the angle θ may take any value, such that the minimum is always attained with θ s.t. $\cos\theta = 1$. We finally obtain, denoting $\forall x \in \mathcal{V}_1,\ y = x - x_i$:

$$\frac{1}{\sigma_i} \leqslant n + \min_{R \geqslant 0} \frac{1}{R^2} \left(\sum_{x \in \mathcal{V}_1} \|y\|^2 - 2R \left\| \sum_{x \in \mathcal{V}_1} y \right\| + \alpha \right).$$

Studying analytically the function $f(R) = \frac{a - 2bR + \alpha}{R^2}$, its minimum is attained for $R^* = \frac{a+\alpha}{b}$ and $f(R^*) = -\frac{b^2}{a+\alpha}$, such that:

$$\frac{1}{\sigma_i} \leqslant n - \frac{|| \sum_{x \in \mathcal{V}_1} y||^2}{\sum_{x \in \mathcal{V}_1} \|y\|^2 + \alpha} \leqslant n.$$

This is true for all i, and in particular for $\sigma_{\min}$. □

Lemma D.3. *In the linear regression problem:*

$$\forall i \qquad \sigma_i = x_i^\top \mathsf{H}^{-1} x_i + \frac{(y_i - y_i^*)^2}{\|y - y^*\|^2} \tag{105}$$

where $\mathsf{H} = \mathsf{X}^\top \mathsf{X}$ *and* y^* *reads* $y^* = \mathsf{X}\theta^* = \mathsf{X}\mathsf{H}^{-1}\mathsf{X}^\top y$. *Also:*

$$\mathfrak{S} = \sum_i \sigma_i = d + 1. \tag{106}$$

Proof. We have:

$$\frac{1}{\sigma_i} = \min_{\theta \in \Theta} \frac{\sum_j (y_j - x_j^\top \theta)^2}{(y_i - x_i^\top \theta)^2} \tag{107}$$

Let us write $\theta = u + v$ where u is colinear to x_i and v is orthogonal to x_i. We obtain:

$$\frac{1}{\sigma_i} = \min_{u,v} \frac{\sum_j (y_j - x_j^\top (u+v))^2}{(y_i - x_i^\top u)^2} \tag{108}$$

$$= \min_u \frac{1}{(y_i - x_i^\top u)^2} \left(\|y\|^2 + \min_v \left[(u+v)^\top \mathsf{H} (u+v) - 2(u+v)^\top \mathsf{X}^\top y \right] \right) \tag{109}$$

$$= \min_u \frac{1}{(y_i - x_i^\top u)^2} \left(\|y\|^2 + u^\top \mathsf{H} u - 2u^\top \mathsf{X}^\top y + \min_v \left[v^\top \mathsf{H} v + 2u^\top \mathsf{H} v - 2v^\top \mathsf{X}^\top y \right] \right) \tag{110}$$

where $\mathsf{H} = \sum_j x_j x_j^\top = \mathsf{X}^\top \mathsf{X}$. Let us first concentrate on solving:

$$\min_v v^\top \mathsf{H} v + 2u^\top \mathsf{H} v - 2v^\top \mathsf{X}^\top y = \min_v v^\top \mathsf{H} v + 2z^\top v \tag{111}$$

with $z = \mathsf{H}^\top u - \mathsf{X}^\top y$. The minimum is to be found for v orthogonal to x_i. We write the Lagrangian:

$$L(v, \lambda) = v^\top \mathsf{H} v + 2z^\top v - \lambda x_i^\top v. \tag{112}$$

We solve it wrt v:

$$2\mathsf{H} v + 2z - \lambda x_i = 0 \tag{113}$$

i.e.:

$$v = \mathsf{H}^{-1} \left(\frac{\lambda}{2} x_i - z \right). \tag{114}$$

We know that v should be orthogonal to x_i such that:

$$0 = x_i^\top \mathsf{H}^{-1} \left(\frac{\lambda}{2} x_i - z \right) \tag{115}$$

i.e.:

$$\frac{\lambda}{2} = \frac{x_i^\top \mathsf{H}^{-1} z}{x_i^\top \mathsf{H}^{-1} x_i}. \tag{116}$$

We finally have:

$$v^* = \mathsf{H}^{-1} \left(\frac{x_i^\top \mathsf{H}^{-1} z}{x_i^\top \mathsf{H}^{-1} x_i} x_i - z \right) \tag{117}$$

and thus:

$$\min_v v^\top \mathsf{H} v + 2z^\top v = v^{*\top} \mathsf{H} v^* + 2z^\top v^* = \frac{(x_i^\top \mathsf{H}^{-1} z)^2}{x_i^\top \mathsf{H}^{-1} x_i} - z^\top \mathsf{H}^{-1} z. \tag{118}$$

i.e.:

$$\frac{1}{\sigma_i} = \min_u \frac{1}{(y_i - x_i^\top u)^2} \left(\|y\|^2 + \frac{(x_i^\top u - x_i^\top \mathsf{H}^{-1} \mathsf{X}^\top y)^2}{x_i^\top \mathsf{H}^{-1} x_i} - y^\top \mathsf{X} \mathsf{H}^{-1} \mathsf{X}^\top y \right) \tag{119}$$

$$= \frac{1}{x_i^\top \mathsf{H}^{-1} x_i} \min_u \frac{\left(\|y\|^2 - y^\top \mathsf{X} \mathsf{H}^{-1} \mathsf{X}^\top y \right) x_i^\top \mathsf{H}^{-1} x_i + \left(x_i^\top u - x_i^\top \mathsf{H}^{-1} \mathsf{X}^\top y \right)^2}{\left(y_i - x_i^\top u \right)^2} \tag{120}$$

Let us write $u = \alpha \frac{x_i}{\|x_i\|^2}$. We have:

$$\frac{1}{\sigma_i} = \frac{1}{x_i^\top \mathsf{H}^{-1} x_i} \min_\alpha \frac{a + (\alpha - b)^2}{(\alpha - c)^2} \tag{121}$$

with: $a = \left(\|y\|^2 - y^\top \mathsf{X} \mathsf{H}^{-1} \mathsf{X}^\top y \right) x_i^\top \mathsf{H}^{-1} x_i$, $b = x_i^\top \mathsf{H}^{-1} \mathsf{X}^\top y$ and $c = y_i$. The minimum of $f(\alpha) = \frac{a + (\alpha - b)^2}{(\alpha - c)^2}$ is attained for $\alpha^* = \frac{a}{b-c} + b$ which entails:

$$f(\alpha^*) = \frac{a}{a + (b-c)^2}. \tag{122}$$

And thus:

$$\frac{1}{\sigma_i} = \frac{\|y\|^2 - y^\top \mathsf{X}\mathsf{H}^{-1}\mathsf{X}^\top y}{\left(\|y\|^2 - y^\top \mathsf{X}\mathsf{H}^{-1}\mathsf{X}^\top y\right) x_i^\top \mathsf{H}^{-1} x_i + \left(x_i^\top \mathsf{H}^{-1}\mathsf{X}^\top y - y_i\right)^2} \tag{123}$$

i.e.:

$$\sigma_i = x_i^\top \mathsf{H}^{-1} x_i + \frac{\left(x_i^\top \mathsf{H}^{-1}\mathsf{X}^\top y - y_i\right)^2}{\|y\|^2 - y^\top \mathsf{X}\mathsf{H}^{-1}\mathsf{X}^\top y}. \tag{124}$$

Writing $\theta^* = \mathsf{H}^{-1}\mathsf{X}^\top y$ the least-square solution to the problem, this is re-written:

$$\sigma_i = x_i^\top \mathsf{H}^{-1} x_i + \frac{\left(x_i^\top \theta^* - y_i\right)^2}{\|y\|^2 - \theta^{*\top} \mathsf{H} \theta^*}. \tag{125}$$

Finally, denoting $y^* = \mathsf{X}\theta^*$:

$$\sigma_i = x_i^\top \mathsf{H}^{-1} x_i + \frac{(y_i - y_i^*)^2}{\|y\|^2 - \|y^*\|^2} \tag{126}$$

$$= x_i^\top \mathsf{H}^{-1} x_i + \frac{(y_i - y_i^*)^2}{\|y - y^*\|^2} \tag{127}$$

Thus:

$$\mathfrak{S} = \sum_i \sigma_i = \mathrm{Tr}(\mathsf{X}^\top \mathsf{H}^{-1}\mathsf{X}) + \sum_i \frac{(y_i - y_i^*)^2}{\|y - y^*\|^2} \tag{128}$$

$$= d + 1. \tag{129}$$

□

Appendix E - The issue of outliers

Corollary 3.3 is applicable to cases where $\sigma_{\max}$ is not too large. In fact, in order for $\alpha\sigma_i$ to be smaller than π_i, and thus smaller than 1 as π_i is a probability, α should always be set inferior to $\frac{1}{\sigma_{\max}}$. Now, if $\sigma_{\max}$ is so large that $\frac{1}{\sigma_{\max}} \leqslant \frac{32}{\epsilon^2}\mathfrak{S}\log\frac{4\eta}{\delta}$, then, even by setting β to its minimum value 1, there is no admissible α verifying both conditions (23) and (24). Large values of σ_i means strong outliers[7]. A simple workaround in this case is to separate the data in two: $\mathcal{X}_o = \{x_i \text{ s.t. } \sigma_i > \sigma^*\}$ the set of outliers and $\bar{\mathcal{X}} = \{x_i \text{ s.t. } \sigma_i \leqslant \sigma^*\}$ the others, where σ^* is the threshold sensitivity over which a data point is considered as an outlier (it is discussed in the following). The initial cost L may also be separated in two: $L = L_o + \bar{L}$ where

$$L_o = \sum_{x\in\mathcal{X}_o} f(x,\theta) \qquad \text{and} \qquad \bar{L} = \sum_{x\in\bar{\mathcal{X}}} f(x,\theta). \tag{130}$$

Let us write $\bar{\sigma}_i$ the sensitivity of data point i in $\bar{\mathcal{X}}$ and $\bar{\mathfrak{S}} = \sum_{x\in\bar{\mathcal{X}}} \bar{\sigma}_i$. Let us choose σ^* to be the largest value in $[0,1]$ for which $\frac{1}{\bar{\sigma}_{\max}} \geqslant \frac{32}{\epsilon^2}\bar{\mathfrak{S}}\log\frac{4\eta}{\delta}$ is verified. One can thus apply the corollary to $\bar{\mathcal{X}}$ to obtain $\bar{\mathcal{S}}$ such that:

$$\forall\theta\in\Theta \qquad (1-\epsilon)\bar{L}(\bar{\mathcal{X}},\theta) \leqslant \hat{\bar{L}}(\bar{\mathcal{S}},\theta) \leqslant (1+\epsilon)\bar{L}(\bar{\mathcal{X}},\theta).$$

Trivially, one may add to $\bar{\mathcal{S}}$ all outliers in $\mathcal{X}_o$ and associate to each of them a weight 1 in the estimated cost. The resulting set $\mathcal{S}$ is thus necessarily a coreset for all datapoints:

$$\forall\theta\in\Theta \qquad (1-\epsilon)L \leqslant (1-\epsilon)\bar{L} + L_o \leqslant \hat{L} = \hat{\bar{L}} + L_o \leqslant (1+\epsilon)\bar{L} + L_o \leqslant (1+\epsilon)L.$$

The number of required samples is thus the number required for $\bar{\mathcal{S}}$ to be a coreset for $\bar{\mathcal{X}}$ plus the number of outliers in $\mathcal{X}_o$: $\mathcal{O}(|\mathcal{X}_o| + \frac{\bar{\mathfrak{S}}^2}{\epsilon^2}\log\frac{\eta}{\delta})$. The exact value of σ^* is application and data dependent.

[7] sensitivities have indeed been shown to be good outlierness indicators [26]

In general, we expect it to be $\mathcal{O}(1)$, such that the number of outliers $|\mathcal{X}_o|$ may be considered as a constant and the number of required samples is of the order $\mathcal{O}(\frac{\mathfrak{S}^2}{\epsilon^2}\log\frac{\eta}{\delta})$.

Appendix F - Implementation

F.1. Approximating the kernel via Random Fourier Features

In order to approximate L in time linear in n, we rely on random Fourier features (RFF) [41]. We briefly recall the RFF framework in the following.

Let us write κ the Gaussian kernel that we use: $\kappa(\boldsymbol{t}) = \exp(-\boldsymbol{t}^2/2\tau^2)$. Its Fourier transform is:

$$\hat{\kappa}(\boldsymbol{\omega}) = \int_{\mathbb{R}^d} \kappa(\boldsymbol{t}) \exp^{-i\boldsymbol{\omega}^\intercal \boldsymbol{t}} \mathrm{d}\boldsymbol{t}. \tag{131}$$

It has real values as κ is symmetrical. One may write:

$$\kappa(\boldsymbol{x},\boldsymbol{y}) = \kappa(\boldsymbol{x}-\boldsymbol{y}) = \frac{1}{Z}\int_{\mathbb{R}^d} \hat{\kappa}(\boldsymbol{\omega}) \exp^{i\boldsymbol{\omega}^\intercal(\boldsymbol{x}-\boldsymbol{y})} \mathrm{d}\boldsymbol{\omega}, \tag{132}$$

where, in order to ensure that $\kappa(\boldsymbol{x},\boldsymbol{x}) = 1$:

$$Z = \int_{\mathbb{R}^d} \hat{\kappa}(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}. \tag{133}$$

According to Bochner's theorem, and due to the fact that κ is positive-definite, $\hat{\kappa}/Z$ is a valid probability density function. $\kappa(\boldsymbol{x},\boldsymbol{y})$ may thus be interpreted as the expected value of $\exp^{i\boldsymbol{\omega}^\intercal(\boldsymbol{x}-\boldsymbol{y})}$ provided that $\boldsymbol{\omega}$ is drawn from $\hat{\kappa}/Z$:

$$\kappa(\boldsymbol{x},\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{\omega}}\left(\exp^{i\boldsymbol{\omega}^\intercal(\boldsymbol{x}-\boldsymbol{y})}\right) \tag{134}$$

The distribution $\hat{\kappa}/Z$ from which ω should be drawn from may be shown to be $\mathcal{N}(\omega; 0, 1/\tau^2)$, where $\mathcal{N}(x;\mu,v)$ is the normal law:

$$\mathcal{N}(x;\mu,v) = \frac{1}{\sqrt{2v\pi}}\exp^{-\frac{(x-\mu)^2}{2v}}. \tag{135}$$

In practice, we draw r random Fourier vectors from $\hat{\kappa}/Z$:

$$\Omega_r = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r).$$

For each data point $\boldsymbol{x}_j$, we define a column feature vector associated to Ω_r:

$$\boldsymbol{\psi}_j = \frac{1}{\sqrt{r}}[\cos(\boldsymbol{\omega}_1^\intercal\boldsymbol{x}_j)|\cdots|\cos(\boldsymbol{\omega}_r^\intercal\boldsymbol{x}_j)|\sin(\boldsymbol{\omega}_1^\intercal\boldsymbol{x}_j)|\cdots|\sin(\boldsymbol{\omega}_r^\intercal\boldsymbol{x}_j)]^\intercal \in \mathbb{R}^{2r}, \tag{136}$$

and call $\Psi = (\boldsymbol{\psi}_1|\cdots|\boldsymbol{\psi}_n) \in \mathbb{R}^{2r\times n}$ the RFF matrix. Other embeddings are possible in the RFF framework, but this one was shown to be the most appropriate to the Gaussian kernel [56]. As r increases, $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$ concentrates around its expected value: $\boldsymbol{\psi}_i^\intercal\boldsymbol{\psi}_j \simeq \kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$. The Gaussian kernel matrix is thus approximated via:

$$\mathsf{L} \simeq \Psi^\intercal\Psi. \tag{137}$$

Computing the RFF matrix requires $\mathcal{O}(nrd)$ operations.

Remark F.1. *How many random features r should we choose? Firstly, note that the entry-wise concentration of $\Psi^\intercal\Psi$ around its expected value L is controlled by a multiplicative error ϵ provided that $r \geqslant \mathcal{O}(d/\epsilon^2)$ [41]. Thus, r should at least be of the order of the dimension d if one wants a proper approximation of the Gaussian kernel. However, this is not our goal here. In fact, what is needed is to obtain in average μ samples from a DPP with L-ensemble $\Psi^\intercal\Psi$. The maximum number of samples of such a DPP is the rank of Ψ, such that r should necessarily be chosen larger than μ. In the following, we thus set r to simply be a few times μ.*

F.2. Fast sampling of DPPs

In order to sample a DPP from a L-ensemble given its eigenvectors $\{\boldsymbol{u}_k\}$ and eigenvalues λ_k, one may follow Alg. 1 of [8], originally from [57]. This algorithm runs in $\mathcal{O}(n\mu^3)$ in average. The limiting step of the overall sampling algorithm is the $\mathcal{O}(n^3)$ cost of the diagonalisation of L. Thankfully, the RFFs not only provide us with an approximation of L in linear time, it also provides us with a dual representation, *i.e.*, a representation of L in the form

$$\mathsf{L} = \Psi^\intercal \Psi. \tag{138}$$

Thus, we may circumvent the prohibitive diagonalization cost of L and only diagonalize its dual form:

$$\mathsf{C} = \Psi \Psi^\intercal \in \mathbb{R}^{2r \times 2r}, \tag{139}$$

costing only $\mathcal{O}(nr^2) = \mathcal{O}(n\mu^2)$ (time to compute C from Ψ and to compute the low-dimensional diagonalization). C's eigendecomposition yields:

$$\mathsf{C} = \mathsf{V}\mathsf{D}\mathsf{V}^\intercal, \tag{140}$$

with $\mathsf{V} = (\boldsymbol{v}_1 | \dots | \boldsymbol{v}_{2r})$ the orthonormal basis of eigenvectors and D the diagonal matrix of eigenvalues such that $0 \leqslant \nu_1 \leqslant \ldots \leqslant \nu_{2r}$.

Note that all eigenvectors associated to non-zero eigenvalues of L can be recovered from C's eigendecomposition (see, e.g., Proposition 3.1 in [8]). More precisely, if $\boldsymbol{v}_k$ is an eigenvector of C associated to eigenvalue ν_k, then:

$$\boldsymbol{u}_k = \frac{1}{\sqrt{\nu_k}} \Psi^\intercal \boldsymbol{v}_k \tag{141}$$

is a normalized eigenvector of L associated to the same eigenvalue.

In the case of such a dual representation, two standard approaches are used in the literature: 1) either follow Alg. 1 of [8] with the reconstructed eigenvectors $\mathsf{U} = \Psi^\intercal \mathsf{V} \mathsf{D}^{-1/2}$ as inputs, running in $\mathcal{O}(n\mu^3)$; 2) or follow Alg. 3 of [8] with the dual eigendecomposition $\{\boldsymbol{v}_k\}$ and $\{\nu_k\}$ as inputs, running in $\mathcal{O}(nr\mu^2 + r^2\mu^3)$. Both approaches are nevertheless suboptimal and we show in [58] that the first (resp. second) one has an equivalent formulation running in $\mathcal{O}(n\mu^2)$ (resp. $\mathcal{O}(n\mu r)$). In this paper, we work with the following sampling strategy, given the dual eigendecomposition $\{\boldsymbol{v}_k\}$ and $\{\nu_k\}$:

- i/ Sample eigenvectors. Draw n Bernoulli variables with parameters $\nu_k/(1+\nu_k)$: for $k = 1, \dots, 2r$, add k to the set of sampled indices $\mathcal{J}$ with probability $\nu_k/(1+\nu_k)$. We generically denote by J the number of elements in $\mathcal{J}$. Note that the expected value of J is μ.
- ii/ Run Alg. 3 to sample a J-DPP with projective L-ensemble $\mathsf{P} = \mathsf{W}\mathsf{W}^\intercal$ where $\mathsf{W} \in \mathbb{R}^{n \times J}$ concatenates all the reconstructed eigenvectors $\boldsymbol{u}_k = \frac{1}{\sqrt{\nu_k}} \Psi^\intercal \boldsymbol{v}_k$ such that $k \in \mathcal{J}$.

Algorithm 3 Efficient J-DPP sampling algorithm with projective L-ensemble $\mathsf{P} = \mathsf{W}\mathsf{W}^\intercal$

Input: $\mathsf{W} \in \mathbb{R}^{n \times J}$ such that $\mathsf{W}^\intercal \mathsf{W} = \mathsf{I}_J$
Write $\forall i,\ \boldsymbol{y}_i = \mathsf{W}^\intercal \boldsymbol{\delta_i} \in \mathbb{R}^J$.
$\mathcal{S} \leftarrow \emptyset$
Define $\boldsymbol{p} \in \mathbb{R}^n : \forall i, \quad p(i) = \|\boldsymbol{y}_i\|^2$
for $n = 1, \dots, J$ **do**:
- Draw s_n with proba $\mathbb{P}(s) = p(s)/\sum_i p(i)$
- $\mathcal{S} \leftarrow \mathcal{S} \cup \{s_n\}$
- Compute $\boldsymbol{f}_n = \boldsymbol{y}_{s_n} - \sum_{l=1}^{n-1} \boldsymbol{f}_l (\boldsymbol{f}_l^\intercal \boldsymbol{y}_{s_n}) \in \mathbb{R}^J$
- Normalize $\boldsymbol{f}_n \leftarrow \boldsymbol{f}_n / \sqrt{\boldsymbol{f}_n^\intercal \boldsymbol{y}_{s_n}}$
- Update $\boldsymbol{p} : \forall i \quad p(i) \leftarrow p(i) - (\boldsymbol{f}_n^\intercal \boldsymbol{y}_i)^2$

end for
Output: $\mathcal{S}$ of size J.

The runtime of this strategy given the dual eigendecomposition is $\mathcal{O}(n\mu^2)$. Also, for a proof that this strategy does sample from a DPP with L-ensemble $\mathsf{L} = \mathsf{\Psi}^\intercal\mathsf{\Psi}$ we refer the reader to our technical report [58].

F.3. Fast sampling of m-DPPs

In the experiments, we will only provide results for m-DPP sampling. In fact, results are easier to compare with classical i.i.d. coreset methods when the number of samples is fixed and not random. Given the eigendecomposition of the dual representation C, one samples a m-DPP via the following two steps (we refer once again to [58] for a proof):

i/ Sample m eigenvectors. Draw $2r$ Bernoulli variables with parameters $\nu_k/(1+\nu_k)$ under the constraint that exactly m variables should be equal to one. Call $\mathcal{J}$ the set of indices thus drawn: $|\mathcal{J}| = J = m$.

ii/ Run Alg. 3 to sample a J-DPP with projective L-ensemble $\mathsf{P} = \mathsf{W}\mathsf{W}^\intercal$ where $\mathsf{W} \in \mathbb{R}^{n\times J}$ concatenates all the reconstructed eigenvectors $\boldsymbol{u}_k = \frac{1}{\sqrt{\nu_k}}\mathsf{\Psi}^\intercal\boldsymbol{v}_k$ such that $k \in \mathcal{J}$.

The only difference with a usual DPP is in the first step, where the n Bernoulli variables are not drawn independently anymore, but under constraint that exactly m of them should be equal to one. To do so, one may follow Alg. 8 of [8] which runs in $\mathcal{O}(nm)$. Step ii/ runs in $\mathcal{O}(nm^2)$, such that the overall cost of sampling a m-DPP given the dual eigendecomposition is also $\mathcal{O}(nm^2)$. Alg. 8 of [8] makes use of elementary polynomials. Given the eigenvalues of C, $\{\nu_i\}$, the p-th order associated elementary polynomial reads:

$$e_p(\nu_1,\ldots,\nu_{2r}) = \sum_{\mathcal{J}\subseteq\{1,2,\ldots,2r\} \text{ s.t. } |\mathcal{J}|=p} \prod_{j\in\mathcal{J}} \nu_j \qquad \in \mathbb{R}. \tag{142}$$

As r increases, these polynomials become less and less stable to compute and Alg. 8 of [8] fails in many practical situations due to numerical precision errors as m becomes too large. In order to avoid these errors, we follow the saddle-point approximation method detailed in [59]. This method has the additional advantage of providing very accurate approximations of the probabilities of inclusion of the m-DPP (that are exactly written as a ratio of elementary polynomials and thus also vulnerable to numerical instability). We in fact need these marginals for the importance sampling estimator.

References

[1] P. Agarwal, S. Har-Peled, and K. Varadarajan, "Geometric approximation via coresets. Combinatorial and Computational Geometry (JE Goodman, J. Pach, and E. Welzl, eds.)," *Math. Sci. Research Inst. Pub., Cambridge*, vol. 6, p. 42, 2005.

[2] S. Har-Peled, *Geometric approximation algorithms*. American Mathematical Soc., 2011, no. 173.

[3] K. L. Clarkson, "Coresets, sparse greedy approximation and the Frank-Wolfe algorithm," in *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 922–931.

[4] M. Langberg and L. J. Schulman, "Universal ϵ-approximators for integrals," in *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 2010, pp. 598–607.

[5] O. Bachem, M. Lucic, and S. Lattanzi, "One-Shot Coresets: The Case of k-Clustering," *arXiv:1711.09649 [stat]*, Nov. 2017, arXiv: 1711.09649. [Online]. Available: http://arxiv.org/abs/1711.09649

[6] K. Chen, "On Coresets for k -Median and k -Means Clustering in Metric and Euclidean Spaces and Their Applications," *SIAM Journal on Computing*, vol. 39, no. 3, pp. 923–947, Jan. 2009. [Online]. Available: http://epubs.siam.org/doi/10.1137/070699007

[7] V. Braverman, D. Feldman, and H. Lang, "New Frameworks for Offline and Streaming Coreset Constructions," *arXiv preprint arXiv:1612.00889*, 2016. [Online]. Available: https://arxiv.org/abs/1612.00889

[8] A. Kulesza and B. Taskar, "Determinantal Point Processes for Machine Learning," *Foundations and Trends® in Machine Learning*, vol. 5, no. 2–3, pp. 123–286, 2012. [Online]. Available: http://dx.doi.org/10.1561/2200000044

[9] A. Munteanu and C. Schwiegelshohn, "Coresets-Methods and History: A Theoreticians Design Pattern for Approximation and Streaming Algorithms," *KI - Künstliche Intelligenz*, Dec. 2017. [Online]. Available: http://link.springer.com/10.1007/s13218-017-0519-3

[10] S. Har-Peled and S. Mazumdar, "On coresets for k-means and k-median clustering," in *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*. ACM, 2004, pp. 291–300.

[11] S. Har-Peled and A. Kushal, "Smaller coresets for k-median and k-means clustering," in *Proceedings of the twenty-first annual symposium on Computational geometry*. ACM, 2005, pp. 126–134. [Online]. Available: http://dl.acm.org/citation.cfm?id=1064114

[12] M. Badoiu and K. L. Clarkson, "Optimal core-sets for balls," *Computational Geometry*, vol. 40, no. 1, pp. 14–22, May 2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0925772107000454

[13] W. F. de la Vega, M. Karpinski, C. Kenyon, and Y. Rabani, "Approximation Schemes for Clustering Problems," in *Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing*, ser. STOC '03. New York, NY, USA: ACM, 2003, pp. 50–58. [Online]. Available: http://doi.acm.org/10.1145/780542.780550

[14] A. Kumar, Y. Sabharwal, and S. Sen, "Linear-time approximation schemes for clustering problems in any dimensions," *Journal of the ACM (JACM)*, vol. 57, no. 2, p. 5, 2010.

[15] D. Feldman and M. Langberg, "A unified framework for approximating and clustering data," in *Proceedings of the forty-third annual ACM symposium on Theory of computing*. ACM, 2011, pp. 569–578. [Online]. Available: http://dl.acm.org/citation.cfm?id=1993712

[16] O. Bachem, M. Lucic, and A. Krause, "Practical Coreset Constructions for Machine Learning," *arXiv:1703.06476 [stat]*, Mar. 2017, arXiv: 1703.06476. [Online]. Available: http://arxiv.org/abs/1703.06476

[17] J. M. Phillips, "Coresets and Sketches," *arXiv:1601.00617 [cs]*, Jan. 2016, arXiv: 1601.00617. [Online]. Available: http://arxiv.org/abs/1601.00617

[18] D. P. Woodruff, "Sketching as a tool for numerical linear algebra," *Foundations and Trends® in Theoretical Computer Science*, vol. 10, no. 1–2, pp. 1–157, 2014.

[19] M. W. Mahoney, "Randomized Algorithms for Matrices and Data," *Foundations and Trends® in Machine Learning*, vol. 3, no. 2, pp. 123–224, 2011. [Online]. Available: http://dx.doi.org/10.1561/2200000035

[20] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas, "Randomized Dimensionality Reduction for k -Means Clustering," *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 1045–1062, Feb. 2015. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6967844

[21] C. Boutsidis and A. Gittens, "Improved matrix algorithms via the Subsampled Randomized Hadamard Transform," *arXiv:1204.0062 [cs, math]*, Mar. 2012, arXiv: 1204.0062. [Online]. Available: http://arxiv.org/abs/1204.0062

[22] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu, "Dimensionality Reduction for k-Means Clustering and Low Rank Approximation," *arXiv:1410.6801 [cs]*, Oct. 2014, arXiv: 1410.6801. [Online]. Available: http://arxiv.org/abs/1410.6801

[23] N. Keriven, N. Tremblay, Y. Traonmilin, and R. Gribonval, "Compressive K-means," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 6369–6373.

[24] K. L. Clarkson and D. P. Woodruff, "Low Rank Approximation and Regression in Input Sparsity Time," *arXiv:1207.6365 [cs]*, Jul. 2012, arXiv: 1207.6365. [Online]. Available: http://arxiv.org/abs/1207.6365

[25] P. Drineas and M. W. Mahoney, "Lectures on Randomized Numerical Linear Algebra," *arXiv:1712.08880 [cs, stat]*, Dec. 2017, arXiv: 1712.08880. [Online]. Available: http://arxiv.org/abs/1712.08880

[26] M. Lucic, O. Bachem, and A. Krause, "Linear-time Outlier Detection via Sensitivity," *arXiv:1605.00519 [cs, stat]*, May 2016, arXiv: 1605.00519. [Online]. Available: http://arxiv.org/abs/1605.00519

[27] M.-F. F. Balcan, S. Ehrlich, and Y. Liang, "Distributed k-means and k-median Clustering on General Topologies," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 1995–2003. [Online]. Available: http://papers.nips.cc/paper/5096-distributed-k-means-and-k-median-clustering-on-general-topologies.pdf

[28] K. Makarychev, Y. Makarychev, M. Sviridenko, and J. Ward, "A bi-criteria approximation algorithm for k Means," *arXiv:1507.04227 [cs]*, Jul. 2015, arXiv: 1507.04227. [Online]. Available: http://arxiv.org/abs/1507.04227

[29] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[30] S. Barthelmé, P.-O. Amblard, and N. Tremblay, "Asymptotic equivalence of fixed-size and varying-size determinantal point processes," *Bernoulli*, 2019.

[31] Y. Li, P. M. Long, and A. Srinivasan, "Improved Bounds on the Sample Complexity of Learning," *Journal of Computer and System Sciences*, vol. 62, no. 3, pp. 516 – 527, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022000000917410

[32] Y. Baraud, "A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression," *Bernoulli*, vol. 16, no. 4, pp. 1064–1085, 2010. [Online]. Available: https://doi.org/10.3150/09-BEJ245

[33] R. Pemantle and Y. Peres, "Concentration of Lipschitz functionals of determinantal and other strong Rayleigh measures," *Combinatorics, Probability and Computing*, vol. 23, no. 1, p. 140–160, 2014.

[34] M. S. Copenhaver, Y. H. Kim, C. Logan, K. Mayfield, S. K. Narayan, M. J. Petro, and J. Sheperd, "Diagram vectors and tight frame scaling in finite dimensions," *Operators and Matrices*, no. 1, pp. 73–88, 2014. [Online]. Available: http://oam.ele-math.com/08-02

[35] P. G. Casazza and G. Kutyniok, *Finite frames: Theory and applications*. Springer, 2012.

[36] J. Tropp, I. Dhillon, and R. Heath, "Finite-Step Algorithms for Constructing Optimal CDMA Signature Sequences," *IEEE Transactions on Information Theory*, vol. 50, no. 11, pp. 2916–2921, Nov. 2004. [Online]. Available: http://ieeexplore.ieee.org/document/1347383/

[37] A. Kroó and D. S. Lubinsky, "Christoffel Functions and Universality in the Bulk for Multivariate Orthogonal Polynomials," *Canadian Journal of Mathematics*, vol. 65, no. 3, pp. 600–620, Jun. 2013. [Online]. Available: https://www.cambridge.org/core/product/identifier/S0008414X00002236/type/journal_article
[38] R. Bardenet and A. Hardy, "Monte Carlo with determinantal point processes," *arXiv preprint arXiv:1605.00361*, 2016.
[39] C. F. Dunkl and Y. Xu, *Orthogonal polynomials of several variables*. Cambridge University Press, 2014, vol. 155.
[40] S. A. van de Geer, *Empirical Processes in M-estimation*. Cambridge university press, 2000, vol. 6.
[41] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2008, pp. 1177–1184.
[42] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.
[43] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.
[44] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.
[45] A. Ng, M. Jordan, Y. Weiss, and others, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.
[46] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.
[47] N. Tremblay and A. Loukas, "Approximating Spectral Clustering via Sampling: a Review," *arXiv:1901.10204 [cs, stat]*, Jan. 2019, arXiv: 1901.10204. [Online]. Available: http://arxiv.org/abs/1901.10204
[48] E. Abbe and C. Sandon, "Community Detection in General Stochastic Block models: Fundamental Limits and Efficient Algorithms for Recovery," in *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*. IEEE, 2015, pp. 670–688.
[49] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," *Phys. Rev. E*, vol. 84, no. 6, p. 066106, Dec. 2011.
[50] Y. LeCun, "The MNIST database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.
[51] A. Vedaldi and B. Fulkerson, "Vlfeat: an open and portable library of computer vision algorithms." ACM Press, 2010, p. 1469. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1873951.1874249
[52] M. Muja and D. G. Lowe, "Scalable Nearest Neighbor Algorithms for High Dimensional Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2227–2240, Nov. 2014.
[53] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A Simple Proof of the Restricted Isometry Property for Random Matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, Dec. 2008. [Online]. Available: http://link.springer.com/10.1007/s00365-007-9003-x
[54] J. B. Lasserre and E. Pauwels, "The empirical christoffel function in statistics and machine learning," *arXiv preprint arXiv:1701.02886*, 2017.
[55] A. Prymak, "Upper estimates of christoffel function on convex domains," *Journal of Mathematical Analysis and Applications*, vol. 455, no. 2, pp. 1984–2000, 2017.
[56] D. J. Sutherland and J. Schneider, "On the Error of Random Fourier Features," *arXiv:1506.02785 [cs, stat]*, Jun. 2015, arXiv: 1506.02785. [Online]. Available: http://arxiv.org/abs/1506.02785
[57] J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág, "Determinantal Processes and Independence," *Probability Surveys*, vol. 3, no. 0, pp. 206–229, 2006. [Online]. Available: http://projecteuclid.org/euclid.ps/1146832696
[58] N. Tremblay, S. Barthelme, and P.-O. Amblard, "Optimized Algorithms to Sample Determinantal Point Processes," *arXiv:1802.08471 [cs, stat]*, Feb. 2018, arXiv: 1802.08471. [Online]. Available: http://arxiv.org/abs/1802.08471
[59] S. Barthelme, P.-O. Amblard, and N. Tremblay, "Asymptotic Equivalence of Fixed-size and Varying-size Determinantal Point Processes," *arXiv:1803.01576 [math, stat]*, Mar. 2018, arXiv: 1803.01576. [Online]. Available: http://arxiv.org/abs/1803.01576