Subspace sampling using determinantal point processes.

Ayoub Belhadji

Numéro d'ordre: 428

Centrale Lille

THESE

Présentée en vue d'obtenir le grade de

DOCTEUR

En

Spécialité :
Automatique, Génie informatique, Traitement du signal et des images

Par

Ayoub Belhadji

Doctorat délivré par Centrale Lille

Titre de la thèse :

Subspace sampling using determinantal point processes

Échantillonnage des sous-espaces à l'aide des processus ponctuels déterminantaux

Soutenue le 03 novembre 2020 devant le jury d'examen:

Rapporteurs	Agnès Desolneux	Directrice de Recherche CNRS, ENS Paris-Saclay
	Francis Bach	Directeur de Recherche Inria, ENS Paris
Examinateurs	Gersende Fort	Directrice de Recherche CNRS, IMT Toulouse
(Président)	Rémi Gribonval	Directeur de Recherche Inria, ENS Lyon
Invité	Chris Oates	Professor, Newcastle University
Directeurs de thèse	Pierre Chainais	Professeur des universités, Centrale Lille
	Rémi Bardenet	Chargé de Recherche CNRS, Université de Lille

Thèse préparée dans le Laboratoire
Centre de Recherche en Informatique Signal et Automatique de Lille
Université de Lille, Centrale Lille, CNRS, UMR 9189 - CRIStAL
École Doctorale SPI 072

To my parents, Mohamed and Zohra.
To my sisters, Wafae and Hiba.

REMERCIEMENTS

Avant tout, j'aimerais remercier mes encadrants Pierre Chainais et Rémi Bardenet qui m'ont accompagné tout au long de ces trois années. Je vous remercie pour votre disponibilité, vos conseils, votre soutien, votre patience et votre confiance. J'étais chanceux d'être admis pour travailler sur ce projet : je me suis retrouvé au centre d'une collaboration scientifique qui croise des champs de recherche très variés. J'ai beaucoup appris à vos côtés et je vous en serai éternellement reconnaissant.

J'aimerais remercier Agnès Desolneux et Francis Bach pour avoir accepté de rapporter mon manuscrit et pour avoir partagé leurs commentaires sur ce travail. J'aimerais remercier également Gersende Fort, Rémi Gribonval et Chris Oates pour avoir accepté de participer à mon jury de thèse. Je suis profondément honoré de vous avoir au sein de mon jury.

Durant mes trois années de thèse, j'ai eu la chance d'avoir rencontré des chercheurs de différents endroits et avec qui j'ai eu des discussions scientifiques stimulantes. J'aimerais remercier pour cela Arnaud Poinas, Slim Kammoun et Simon Barthelmé. Finalement, j'aimerais remercier Jean-François Coeurjolly pour l'accueil qu'il m'a accordé lors de mon séjour à Montréal et pour l'intérêt qu'il a porté à mes travaux de recherche. Discuter de la science en observant la chute de neige est une expérience inoubliable.

Un grand merci à l'équipe SigMA pour m'avoir accueilli durant ces trois années et pour m'avoir fourni un environnement propice à la recherche scientifique. Ça fait plaisir de voir que cette équipe grandit et s'enrichit de nouvelles personnes. Je remercie tous ses membres passés et présents : Patrick, Rémy, John, Jérémie, Noha, Amine, Clément, Julien, Guillaume, Maxime, Quentin, Solène, Ouafa, Rui, Arnaud, Barbara et Yoann... Mon expérience au sein de l'équipe SigMA a été marquée par la présence de Théo, son humour et son charisme... Je te remercie pour cette agréable compagnie ...

Je remercie mes amis Aimad, Saleh, Wissam, Wajih, Rajae, Souhail et Filippo pour avoir gardé le contact durant ces années de thèse. Je remercie également les tovarichs de Lille Loïc, Shao, Lydia, Mira, Lilia et Yulia, Alexis et Yohann pour les bons moments qu'on a passés ensemble à cuisiner, à camper ou à regarder des navets et des nanars...

Le goût de la science, la curiosité, l'esprit critique et la persévérance, je les dois à ma famille et à mes enseignants. J'aimerais remercier mes enseignants au lycée Abdelkrim el Khattabi à Nador : Mohamed Agouram, Mohamed Bensaïd et Mohamed Mattich pour m'avoir équipé du bagage scientifique et pour m'avoir préparé mentalement pour l'aventure d'après le bac. Je remercie également mes cousins Mohamed Taamouti et Abderrahim Taamouti, qui sont ma source d'inspiration, pour leur soutien et leurs conseils aux moments clé de ma formation universitaire.

Enfin, à mes parents et à mes sœurs, je ne pourrai jamais assez vous remercier...

CONTENTS

1

INTRODUCTION

The real world is much smaller than the imaginary

Friedrich Nietzsche

Subsampling is a recurrent task in applied mathematics. This paradigm has many applications in data analysis, signal processing, machine learning and statistics: continuous signal discretization, numerical integration, dimension reduction, learning on budget, preconditioning... Seemingly unrelated, these problems can be tackled using the same strategy: looking for the most representative elements in a set. For example, the aim of subset selection is to select the most representative elements of a large set to approximate an object that can be scalar, a vector, or a matrix. On the other hand, quadrature and interpolation aim to select the most representative elements of a continuous domain to approximate a function or an integral. These approximations are conducted, mainly, for three purposes: improving the numerical complexity of an existing algorithm (*numerical issue*), reducing the cost of labelling (*sensing issue*), or making an algorithm more interpretable (*interpretation issue*).

For some problems with linear structure, the set can be embedded in a vector space. A good subset of representatives would have a geometric characterization that can be expressed in a simple way using the spectrum of some linear operator. In other words, the sub-sampling problem boils down to a geometric problem. One way to set up a linear representation on a set is to define a kernel. Indeed, kernels are universal tools that can be defined on discrete sets or continuous domains. Moreover, they combine well with linearisation.

Once an embedding is defined, one can make use of the abundant tools of linear operators such as linear algebra or functional analysis to solve the geometrical problem that corresponds to the sub-sampling problem. Under this perspective, we can explain the recourse to random sub-sampling in many problems with linear structure: non-asymptotic random matrix theory offers strong tools to tackle these geometric problems in a universal way. A universality that lacks to deterministic sub-sampling approaches.

Until now, sampling vectors independently, with a probability proportional to their (squared) length was the gold standard of randomized subsampling. Indeed, lengths are amenable to fast evaluation, approximation, and sampling. Moreover, these simple sub-sampling schemes come with relatively strong guarantees. Still, lengths contain only first-order information about the set to be sampled. One may expect to improve the quality of sampling by considering high-order information such as volumes.

Indeed, a good subset of representatives would capture the substance of the information avoiding any unnecessary redundancy. This redundancy may be measured by the volume spanned by these elements. Intuitively, a subset of vectors would be redundant if defining a polytope with small volume, and it would be non-redundant if defining a polytope with large volume. It turns out that there exists a family of probabilistic models that define random subsets with a repulsion property: informally,

the probability of appearance of a subset is proportional to the squared volume it defines in the vector space. These models are called determinantal point processes and they were the topic of intense research in various fields: random matrices, quantum optics, spatial statistics, image processing, machine learning and recently numerical integration. These probabilistic models seem to furnish the plausible framework for the study of sub-sampling using high-order information. Indeed, these probabilistic models are expressed using the language of linear operators and kernels: they naturally extend first-order sub-sampling schemes.

This thesis is devoted to investigate the suitability and the pertinence of using determinantal point processes as models for randomized sub-sampling for problems with a linear structure.

1.1 SUBSAMPLING PROBLEMS WITH LINEAR STRUCTURE

In order to situate the contributions of this thesis, a brief review of randomized subsampling techniques for problems with a linear structure is given in the following.

We start by the problem of linear regression under constraint on the number of observations. Consider a matrix $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ that contains the d features of N observations, with $N \geq d$, and assume the rank of $\boldsymbol{X}$ is equal to d. A recurrent task in data analysis and statistics is linear regression, which consists in looking for the minimum-length vector $\boldsymbol{w} \in \mathbb{R}^d$ among the vectors minimizing the residual

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2, \tag{1.1}$$

for a given vector of labels $\boldsymbol{y} \in \mathbb{R}^N$. It is well known that the solution is given by $\boldsymbol{w}^* = \boldsymbol{X}^+\boldsymbol{y}$, where $\boldsymbol{X}^+$ is the Moore-Penrose pseudo-inverse of $\boldsymbol{X}$ and the optimal residual writes

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{X}^+\boldsymbol{y}\|^2 = \|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{U}^\intercal\boldsymbol{y}\|^2, \tag{1.2}$$

with $\boldsymbol{U} \in \mathbb{R}^{N\times d}$ is a matrix containing the left eigenvectors of $\boldsymbol{X}$.

In some situations, calculating $\boldsymbol{w}^*$ using a pseudo-inverse formula is not affordable, either because N is very large and the computation resources are not sufficient (the numerical issue), or because obtaining the N labels $(y_n)_{n\in[N]}$ is expensive (the sensing issue). In such a situation, looking for representatives can be beneficial.

Formally, a family of representatives corresponds to a subset $S = \{i_1, \dots, i_c\} \subset [N]$ with $c \in \mathbb{N}^*$, $i_1 < i_2 < \cdots < i_c$, and a matrix $\boldsymbol{S} \in \mathbb{R}^{c\times N}$ such that for $j \in [c]$, $\boldsymbol{S}_{j,i_j} = s_j > 0$, and the other elements of $\boldsymbol{S}$ are equal to 0. The constant c represents the *budget* of sub-sampling, and $\boldsymbol{S}^\intercal\boldsymbol{X} \in \mathbb{R}^{c\times d}$ is the sub-matrix of $\boldsymbol{X}$ that corresponds to the subset S, after rescaling the rows by the factors s_i.

Now, one can define an approximation of $\boldsymbol{w}^*$ based on the subset S. For example, Drineas, Mahoney, and Muthukrishnan, 2006 proposed

$$\boldsymbol{w}_S^* = (\boldsymbol{S}^\intercal\boldsymbol{X})^+(\boldsymbol{S}^\intercal\boldsymbol{y}). \tag{1.3}$$

In this approximation, the pair $(\boldsymbol{X}, \boldsymbol{y})$ was replaced by the reduced pair $(\boldsymbol{S}^\intercal\boldsymbol{X}, \boldsymbol{S}^\intercal\boldsymbol{y})$. A good choice of the representatives would define a sampling matrix $\boldsymbol{S}$ that satisfies

$$\forall \boldsymbol{y} \in \mathbb{R}^N,\ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_S^*\|^2 \leq (1+\epsilon)\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}^*\|^2, \tag{1.4}$$

for some tolerable error $\epsilon > 0$. This condition can be written as

$$(1-\epsilon)\mathbb{I}_d \prec \sum_{j\in[c]} s_j \tilde{x}_{i_j}\tilde{x}_{i_j}^\intercal \prec (1+\epsilon)\mathbb{I}_d \iff \left\|\mathbb{I}_d - \sum_{j\in[c]} s_j \tilde{x}_{i_j}\tilde{x}_{i_j}^\intercal\right\|_2 \leq \epsilon, \tag{1.5}$$

where $\mathbb{I}_d$ is the identity matrix of dimension d, $\tilde{x}_n = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1/2} x_n$ and $\|.\|_2$ is the spectral norm.

Therefore, the sub-sampling problem boils down to looking for a matrix $\boldsymbol{S}$ that makes the reduced matrix $\sum_{j\in[c]} s_j \tilde{x}_{i_j}\tilde{x}_{i_j}^\intercal$ close to the identity matrix $\mathbb{I}_d$, which itself writes as the sum of rank one matrices

$$\mathbb{I}_d = \sum_{n\in[N]} \tilde{x}_n \tilde{x}_n^\intercal. \tag{1.6}$$

One way to design a sampling matrix $\boldsymbol{S}$ that satisfies condition (1.5) is to consider a randomized sub-sum of (1.6), and to make use of a matrix concentration inequality to prove that this sub-sum concentrates around $\mathbb{I}_d$. For example, consider the following random sum

$$I_c = \frac{1}{c}\sum_{i\in[c]} \frac{1}{p_{n_i}} \tilde{x}_{n_i}\tilde{x}_{n_i}^\intercal, \tag{1.7}$$

where the p_n are positive and sum to one, and the n_i are sampled independently from the multinomial random variable of parameter $\boldsymbol{p} = (p_n)_{n\in[N]}$. We can easily prove that $\mathbb{E}\, I_c = \mathbb{I}_d$. Moreover, the fluctuations of I_c around $\mathbb{I}_d$ can be controlled using a concentration inequality.

Theorem 1.1 (Matrix Bernstein concentration inequality). *(Tropp, 2015) Consider a finite sequence (X_j) of independent, random, self-adjoint $d \times d$ matrices. Assume that each matrix X_j satisfies*

$$\mathbb{E}\, X_j = \mathbf{0}_{d\times d}, \tag{1.8}$$

$$\|X_j\|_2 \leq \tau, \tag{1.9}$$

where $\tau > 0$, and define $\sigma^2 = \|\mathbb{E}\sum_j X_j^2\|$. Then for $\epsilon > 0$,

$$\mathbb{P}\left(\left\|\sum_j X_j\right\|_2 \geq \epsilon\right) \leq 2de^{-\frac{3}{2}\frac{\epsilon^2}{3\sigma^2+\tau\epsilon}}. \tag{1.10}$$

Theorem 1.1 can be applied to the matrix I_c in (1.7) by considering the random matrices $X_i = (\tilde{x}_{n_i}\tilde{x}_{n_i}^\intercal/p_{n_i} - \mathbb{I}_d)/c$. In particular, when $\boldsymbol{p} = (1/N)$, the bounds of Theorem 1.1 are significant if $c \geq CN\log(d)\max_{n\in[N]}\|\tilde{x}_n\|^2$, where C is a universal constant that does not depend on d and N: if there exists n such that $\|\tilde{x}_n\| = 1$, then one needs to take $c \geq CN\log(d)$ and the subsampling does not allow to reduce the complexity of the problem; however, if $\max_{n\in[N]}\|\tilde{x}_n\|^2$ is equal to its minimal value d/N, then the budget $c = Cd\log(d)$ is enough to have good guarantees. In other words, uniform sub-sampling reduces the complexity of the problem if the *coherence* $\max_{n\in[N]}\|\tilde{x}_n\|^2$ approaches its lowest possible value d/N, otherwise the minimal budget has to be larger than N, which is at odds with the idea of sub-sampling. Intuitively, the variance of such a randomized sub-sum is governed by the vectors $\tilde{x}_n$ that have large norms, and including them in this sum with a probability larger than $1/N$ would

make this variance low. This intuition motivates the introduction of the *leverage scores* $(\ell_n)_{n\in[N]}$ defined by

$$\ell_n = \|\tilde{x}_n\|^2. \tag{1.11}$$

Indeed, using again Theorem 1.1, if $\boldsymbol{p} = (\ell_n/d)_{n\in[N]}$, the budget $c = Cd\log(d)$ is enough to have good guarantees independently of the value of the coherence, and more importantly, independently of N. This constitutes a significant reduction of the complexity of the problem in situations where N is very large compared to d.

Beside vanilla linear regression, this reduction can be beneficial in other problems with a linear structure. For example, for the problem of graph sparsification, i.e. sub-sampling the edges of a graph while maintaining its spectral properties. For this problem, the dimension d is the number of nodes, which can be very small compared to the number of edges N. Leverage score sampling was studied for this problem under the name of *effective resistance sampling* (Spielman and Teng, 2004; Spielman and Srivastava, 2011). Sub-sampling nodes in a graph can also be beneficial, e.g. for the recovery of band-limited signals (Puy, Tremblay, Gribonval, and Vandergheynst, 2018). In this setting, d is the size of the band of the signal and N is the number of nodes. Finally, the notion of leverage score sampling was used for low-rank approximations, of general matrices (Drineas, Mahoney, and Muthukrishnan, 2007), as we shall see more in detail in Chapter 3.

Another field of application of matrix sub-sampling is low-rank approximation of kernel matrices. These approximations are widely used to scale up kernel methods (Schölkopf and Smola, 2018; Shawe-Taylor and Cristianini, 2004). To give an example, consider the kernel ridge regression problem, defined for some elements $x_1, \dots, x_N$ of a metric space $\mathcal{X}$ equipped with a p.s.d. kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{K}(\boldsymbol{x})\boldsymbol{\alpha}\|^2 + \lambda\boldsymbol{\alpha}^\intercal\boldsymbol{K}(\boldsymbol{x})\boldsymbol{\alpha}, \tag{1.12}$$

where $\boldsymbol{y} \in \mathbb{R}^N$ and $\boldsymbol{K}(\boldsymbol{x}) = (k(x_n, x_{n'}))_{n,n'\in[N]} \in \mathbb{R}^{N\times N}$. The solution of (1.12) has a tractable form

$$\boldsymbol{\alpha}^* = (\boldsymbol{K}(\boldsymbol{x}) + N\lambda\mathbb{I}_N)^{-1}\boldsymbol{y}, \tag{1.13}$$

yet it is impractical to compute for large values of N, because the required number of operations to invert $\boldsymbol{K}(\boldsymbol{x}) + N\lambda\mathbb{I}_N$ scales as $\mathcal{O}(N^3)$. One way to scale up this problem is to consider a low rank approximation of the kernel matrix $\boldsymbol{K}(\boldsymbol{x})$. Indeed, if a low rank approximation of $\boldsymbol{K}(\boldsymbol{x})$ is known, then approximating of $\boldsymbol{w}^*$ can be done in $\mathcal{O}(Nr + r^3)$, with r the rank of the approximation (Smola and Schölkopf, 2000; Williams and Seeger, 2001). Nyström approximation is a widely used low rank approximation; it is simple to implement and amenable to theoretical analysis. This approximation is based on a subset $S \subset [N]$, and writes

$$\boldsymbol{K}(\boldsymbol{x}) \approx \boldsymbol{K}(\boldsymbol{x})_{:,S}\boldsymbol{K}(\boldsymbol{x})^{+}_{S,S}\boldsymbol{K}(\boldsymbol{x})^\intercal_{:,S}, \tag{1.14}$$

where $\boldsymbol{K}(\boldsymbol{x})_{:,S}$ and $\boldsymbol{K}(\boldsymbol{x})_{S,S}$ are respectively the columns and the sub-matrix of $\boldsymbol{K}(\boldsymbol{x})$ corresponding to the indices in the set S.

The theoretical analysis of Nyström approximation for kernel ridge regression was carried out in (Bach, 2013) under uniform sub-sampling, and in (Alaoui and Mahoney, 2015) under the *ridge leverage score* distribution defined by

$$\ell_n^\lambda = k(x_n, .)^\intercal\big(\boldsymbol{K}(\boldsymbol{x}) + N\lambda\mathbb{I}_N\big)^{-1}k(x_n, .). \tag{1.15}$$

If a spectral decomposition of $\boldsymbol{K}(\boldsymbol{x}) = \boldsymbol{U}\,\mathrm{Diag}((\sigma_\ell)_{\ell\in[N]})\boldsymbol{U}^\intercal$ is known, then

$$\ell_n^\lambda = \sum_{\ell\in[N]} \frac{\sigma_\ell}{\sigma_\ell + N\lambda} \boldsymbol{U}_{n,\ell}^2. \tag{1.16}$$

The analysis conducted in these works highlighted the importance of a quantity called *the effective dimension* defined by

$$d_{\mathrm{eff}}(\lambda) = \mathrm{Tr}\,\boldsymbol{K}(\boldsymbol{x})(\boldsymbol{K}(\boldsymbol{x}) + N\lambda\mathbb{I}_N)^{-1}. \tag{1.17}$$

The effective dimension plays the same role as the dimension d for the sparsification of the identity matrix $\mathbb{I}_d$ reviewed around Theorem 1.1: the minimal budget $|S|$ has to scale as $\mathcal{O}(d_{\mathrm{eff}}(\lambda)\log d_{\mathrm{eff}}(\lambda))$ to obtain a statistical risk within a factor $(1+\epsilon)$ of the statistical risk obtained when the full matrix $\boldsymbol{K}(\boldsymbol{x})$ is used. This suggests that kernel ridge regression under ridge leverage score sampling would scale as $\mathcal{O}(Nd_{\mathrm{eff}}(\lambda)^2)$. Nevertheless, the calculation of the ridge leverage scores through the definition (1.15), or the spectral representation (1.16) is as expensive as the initial problem. To close the loop, many approximation schemes of the ridge leverage scores were proposed and analysed. These methods scale better with N; see (Calandriello, Lazaric, and Valko, 2017; Calandriello, 2017) for an algorithm that runs in $\mathcal{O}(Nd_{\mathrm{eff}}(\lambda)^3)$ operations.

As one would observe, all the problems we reviewed so far are *fixed design* problems, in the sense that it is required to sub-sample from a given set $\{x_1,\dots,x_N\}$. This setting is typical in numerical linear algebra and machine learning applications, where $\{x_1,\dots,x_N\}$ corresponds to some fixed matrix or dataset. Yet, as it was shown by Bach, 2017, the analysis of some *random design* problems can be conducted using the same spectral techniques used for fixed design problems. Indeed, the author considered the problem of approximating a function μ defined on some metric space $\mathcal{X}$ by a finite mixture of kernel translates $k(x_n,.)$

$$\mu \approx \sum_{n\in[N]} w_n k(x_n,.). \tag{1.18}$$

More precisely, $\mathcal{X}$ is assumed to be a measurable space equipped with a measure ω, and μ is assumed to write as

$$\mu = \mu_g := \int_{\mathcal{X}} k(x,.)g(x)\mathrm{d}\omega(x), \tag{1.19}$$

with $g \in \mathbb{L}_2(\mathrm{d}\omega)$. By construction, μ belongs to the RKHS $\mathcal{F}$ associated to the kernel k, which we assume to be embedded in $\mathbb{L}_2(\mathrm{d}\omega)$. Ensuring a "good" approximation in (1.18) is particulary useful for quadrature. Indeed, we have

$$\forall f\in\mathcal{F},\ \left|\int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x) - \sum_{n\in N} w_n f(x_n)\right| \le \|f\|_{\mathcal{F}}\left\|\mu_g - \sum_{n\in[N]} w_n k(x_n,.)\right\|_{\mathcal{F}}, \tag{1.20}$$

so that controlling the approximation error in r.h.s. of (1.20) yields a control on the integration error in the l.h.s. Moreover, $\|\mu_g - \sum_{n\in[N]} w_n k(x_n,.)\|_{\mathcal{F}}$ is actually the worst case integration error on the unit ball of $\mathcal{F}$.

Bach, 2017 proposed to choose the *nodes* x_n to be independent random variables that follow a density q with respect to the measure ω, and to choose the weights $(w_n)_{n\in N}$ as the solution of the optimization problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^N} \left\|\mu_g - \sum_{n\in[N]} \frac{w_n}{q(x_n)^{1/2}} k(x_n,.)\right\|_{\mathcal{F}}^2 + N\lambda\|\boldsymbol{w}\|^2. \tag{1.21}$$

This optimization problem admits a unique solution

$$\boldsymbol{w}^* = (\overline{\boldsymbol{K}}(\boldsymbol{x}) + \lambda \mathbb{I}_N)^{-1} \mu_g(\boldsymbol{x}), \tag{1.22}$$

where $\mu_g(\boldsymbol{x}) = (\mu_g(x_n))_{n\in[N]} \in \mathbb{R}^N$, and $\overline{\boldsymbol{K}}(\boldsymbol{x}) = (q(x_n)^{-1/2} k(x_n, x_{n'}) q(x_{n'})^{-1/2})_{n,n'\in[N]}$. In particular, there exists a density q^*_λ called the *continuous ridge leverage score density* for which

$$\sup_{\substack{g\in\mathbb{L}_2(\mathrm{d}\omega)\\ \|g\|_{\mathrm{d}\omega}\leq 1}} \left\| \mu_g - \sum_{n\in[N]} \frac{w^*_n}{q^*_\lambda(x_n)^{1/2}} k(x_n, .) \right\|^2_{\mathcal{F}} \leq \lambda, \tag{1.23}$$

with high probability, whenever $N \geq C d_{\mathrm{eff}}(\lambda) \log d_{\mathrm{eff}}(\lambda)$, with $d_{\mathrm{eff}}(\lambda)$ is the effective degree of freedom defined by

$$d_{\mathrm{eff}}(\lambda) = \operatorname{Tr} \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda \mathbb{I}_{\mathcal{F}})^{-1}, \tag{1.24}$$

and $\boldsymbol{\Sigma}$ is the integration operator associated to the kernel k and the measure ω, defined on $\mathbb{L}_2(\mathrm{d}\omega)$ by $\boldsymbol{\Sigma} g = \int_{\mathcal{X}} g(.) k(x, .) \mathrm{d}\omega(x)$. Similarly to the discrete ridge leverage score distribution, the continuous ridge leverage score density has a spectral expression

$$q^*_\lambda(x) = \sum_{n\in\mathbb{N}^*} \frac{\sigma_n}{\sigma_n + \lambda} e_n(x)^2, \tag{1.25}$$

where the (σ_n, e_n) are the eigenpairs of the integration operator $\boldsymbol{\Sigma}$.

In a nutshell, this result implies that the resulting quadrature $(\boldsymbol{x}, \boldsymbol{w}^*)$ has a worst case integration error that goes below λ, with high probability, if $N \geq C d_{\mathrm{eff}}(\lambda) \log d_{\mathrm{eff}}(\lambda)$. Yet, for a fixed level of regularization λ, sampling more nodes from q^*_λ does not improve the worst case integration error. Nevertheless, one can build up an infinite sequence of quadratures $(\boldsymbol{x}_t, \boldsymbol{w}^*_t)_{t\in\mathbb{N}^*}$, each associated to a regularization parameter λ_t, such that $\lim_{t\to+\infty} \lambda_t \to 0$. This way, Bach, 2017 was able to derive the rates of convergence of these sequence of quadratures in the limiting asymptotic $\lambda \to 0$ in different RKHSs. While intricate, this method allows to prove that the rates of the resulting quadratures are almost optimal, up to logarithmic terms. Moreover, these rates depend on the eigenvalues σ_n of the integration operator: the smoother the kernel, the faster the convergence to 0. However, and unlike the discrete case, the ridge leverage score density 1.25 may not have a tractable expression, in general; even worse the spectral decomposition of $\boldsymbol{\Sigma}$ may not be available.

Remarkably, the proof of this result relies on an extension of Bernstein matrix concentration inequality to self-adjoint operators (Minsker, 2017). This suggests that even random design problems can be analysed using spectral techniques.

We continue on this direction on Chapter 4 and Chapter 5 where we try to solve the problem of the intractability of the continuous ridge leverage score density using DPPs and one of its variants, continuous volume sampling. We mention that the contributions of these two chapters are extensions of the work initiated in Chapter 3 on the column subset selection problem. In particular, we investigated in this chapter volume sampling through the lens of DPPs. Before giving a detailed account of these contributions, we review, in the following section, the existing work on the volume sampling distribution and some of its applications to subsampling problems with a linear structure.

1.2 DETERMINANTAL SAMPLING

As we have seen in the previous section, independent sampling was used in a variety of approximation tasks that have linear structure. These sub-sampling methods rely on first-order information such as the leverage score distribution.

Now we will review another approach of sampling that relies on high order information. This is achieved by determinantal point processes and their variants. Indeed, this class of distributions defines the most natural extension of leverage score sampling with negative correlation property.

We start again with the problem of linear regression under budget constraint. Remember that, for this problem, it is required to estimate $\boldsymbol{X}^+\boldsymbol{y}$ using a subset of $[N]$. Now, if $\boldsymbol{X}$ is of full rank, the Moore-Penrose pseudo inverse has a *determinantal representation* (Ben-Tal and Teboulle, 1990)

$$\boldsymbol{X}^+\boldsymbol{y} = \sum_{I \in \mathcal{I}_{N,d}} \frac{\Delta_I}{\Delta} \boldsymbol{X}^+_{I,:}\boldsymbol{y}_I, \tag{1.26}$$

where $\mathcal{I}_{N,d} = \{I \subset [N], |I| = d\}$, $\Delta_I = \mathrm{Det}^2 \boldsymbol{X}_{I,:}$ and $\Delta = \sum_{I \in \mathcal{I}_{N,d}} \Delta_I$. In other words, the solution of the regression problem defined for the pair $(\boldsymbol{X}, \boldsymbol{y})$ is a mixture of the solutions of the smaller regression problems defined by the pairs $(\boldsymbol{X}_{I,:}, \boldsymbol{y}_I)$, and the weight is proportional to $\mathrm{Det}^2 \boldsymbol{X}_{I,:}$ that is, the volume spanned by the rows of $\boldsymbol{X}_{I,:}$ in the space $\mathbb{R}^d$. In particular, singular sub-matrices $\boldsymbol{X}_{I,:}$ do not participate in this mixture. Obviously, this determinantal representation cannot be used in this raw form. Indeed, an exhaustive enumeration of $\mathcal{I}_{N,d}$ would cost much more than calculating $\boldsymbol{X}^+\boldsymbol{y}$. Yet, its structure is amenable to a probabilistic interpretation. Indeed, if we consider a random element S of $\mathcal{I}_{N,d}$ such that

$$\forall I \in \mathcal{I}_{N,d}, \ \mathbb{P}(S = I) = \Delta_I / \Delta, \tag{1.27}$$

then

$$\boldsymbol{X}^+\boldsymbol{y} = \mathbb{E}\, \boldsymbol{X}^+_{S,:}\boldsymbol{y}_S\,. \tag{1.28}$$

In other words, under the distribution (1.27), also called *volume sampling*, $\boldsymbol{w}^*_S = \boldsymbol{X}^+_{S,:}\boldsymbol{y}_S$ is an unbiased estimator of $\boldsymbol{w}^* = \boldsymbol{X}^+\boldsymbol{y}$. This idea was exploited in (Derezinski and Warmuth, 2017), where the authors studied the mean square error of this estimator and proved that

$$\mathbb{E}\, \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}^*_S\|^2 \leq (1+d)\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}^*\|^2. \tag{1.29}$$

This bound holds for a budget of subsampling that is equal to the dimension d, that is, the smallest budget that we can expect, otherwise the rank of the submatrix $\boldsymbol{X}_{S,:}$ is smaller than d and and the estimator $\boldsymbol{w}^*_S$ is ill-defined. This "sparsity level" has no equivalent under independent sampling for which the minimal budget scales as $\mathcal{O}(d \log d)$. Indeed, under leverage score sampling (with or without replacement) and in the "minimalist sampling" regime i.e. when c is very close to d, the matrix $\boldsymbol{X}_{S,:}$ is rank-defficient with high probability; see (Ipsen and Wentworth, 2014) for details on this phenomenon and Section 6.1.2 in (Tropp, 2015) for a discussion on the optimality of the Matrix Bernstein concentration inequality. On the other hand, the bound (1.29) does not tell much about the high-order moments of $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}^*_S\|^2$, which would be

useful to derive probabilistic bounds with exponential tails as it is the case for leverage score sampling. Therefore, the two distributions are not comparable.

However, if we must compare the two distributions, we may recall that volume sampling belongs to a larger class of distributions called *dual volume sampling*. These distributions charge subsets $I \subset [N]$ of a fixed cardinality $c \geq d$ as follows

$$\mathbb{P}(S = I) \propto \operatorname{Det} \boldsymbol{X}_{I,:}^{\intercal} \boldsymbol{X}_{I,:} \mathbb{1}_{\{|I|=c\}}. \tag{1.30}$$

In particular the case $c = d$ corresponds to the volume sampling distribution. These distributions satisfy the unbiasedness property (1.28), as shown by Derezinski and Warmuth, 2017. However, the empirical investigation conducted in (Derezinski et al., 2018) showed that dual volume sampling in the regime $c = \Omega(d \log d)$ can lead to an inferior approximation compared to leverage score sampling. In other words, dual volume sampling is better used in the "minimalist sampling" regime $c \approx d$, and leverage score sampling is better used in "Bernstein" regime $c = \Omega(d \log d)$.

This suggests that volume sampling should be used for problems where we need to keep the budget of sub-sampling c to its lowest possible value. As an example, volume sampling can be used to tackle the column subset selection problem, which we discuss in detail in Chapter 3. In a nutshell, the subset selection, in this setting, is meant to resolve the interpretation issue of principal component analysis. The latter outputs features that do not have, in general, a physical meaning. Hence, the idea to sub-sample a set of features S of the matrix $\boldsymbol{X}$, that are represented by the matrix $\boldsymbol{C} = \boldsymbol{X}_{:,S}$, with the constraint to control the residual $\boldsymbol{X} - \boldsymbol{\Pi}_{\operatorname{Span} \boldsymbol{C}} \boldsymbol{X}$, where $\boldsymbol{\Pi}_{\operatorname{Span} \boldsymbol{C}}$ the projection onto the subspace of $\mathbb{R}^N$ spanned by the columns $\boldsymbol{C}$. In this setting, volume sampling is defined on the set of subsets of $[d]$ by

$$\mathbb{P}_{\mathrm{VS}}(S = I) \propto \operatorname{Det}(\boldsymbol{X}_{:,I}^{\intercal} \boldsymbol{X}_{:,I}) \mathbb{1}_{\{|I|=k\}}\,. \tag{1.31}$$

It satisfies (Deshpande, Rademacher, Vempala, and Wang, 2006)

$$\mathbb{E}_{\mathrm{VS}} \|\boldsymbol{X} - \boldsymbol{\Pi}_{\operatorname{Span} \boldsymbol{C}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq (1 + k) \|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_{\mathrm{Fr}}^2, \tag{1.32}$$

and

$$\mathbb{E}_{\mathrm{VS}} \|\boldsymbol{X} - \boldsymbol{\Pi}_{\operatorname{Span} \boldsymbol{C}} \boldsymbol{X}\|_2^2 \leq (d - k)(1 + k) \|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_2^2, \tag{1.33}$$

where $\boldsymbol{\Pi}_k$ is the projection onto the first left eigenvectors of $\boldsymbol{X}$. As we shall see in Chapter 3, the constant $1 + k$ in (1.32) is optimal in the sense that there exists a matrix $\boldsymbol{X}$ for which this bound is almost matched.

The starting point of this thesis is to challenge the optimality of this bound. The first step was to see the volume sampling distribution through the lens of determinantal point processes. This perspective allows us to propose a new sub-sampling algorithm based on a DPP, which has a better geometric interpretation than volume sampling, and which leads to better theoretical guarantees and empirical results under some conditions. Intuitively, we can see that this specific choice of DPP defines a distribution on the set of subspaces of $\mathbb{R}^d$ of dimension k, the *Grassmannian* $\mathbb{G}(\mathbb{R}^d, k)$, that favours the subspaces that are close to the (right) principal subspace of $\boldsymbol{X}$ of dimension k. This is to be compared to the uniform distribution on the Grassmannian $\mathbb{G}(\mathbb{R}^d, k)$ usually used in the compressed sensing literature (Chikuse, 2012)[Theorem 2.2.1.]. We push this geometric interpretation further in Chapter 4 to tackle the problem of kernel quadrature

in an RKHS. In particular we connect the previous work on numerical integration using DPPs (Bardenet and Hardy, 2020) and the work of (Bach, 2017) to propose a new class of quadrature, based on DPP nodes, and weights that solve the unregularized version of the optimization problem (1.21). We prove convergence bounds of these quadratures for smooth kernels. This contribution partially solves the problem of intractability of the continuous ridge leverage density of (Bach, 2017). Indeed, under the condition that the spectral decomposition of $\mathbf{\Sigma}$ is known, the numerical implementation of this DPP is tractable as it relies only on the first eigenfunctions of $\mathbf{\Sigma}$. Finally, by reconsidering the links between DPPs and volume sampling that we have discussed in Chapter 3, we define the continuous volume sampling distribution in Chapter 5 to tackle the more general problem of kernel interpolation. We prove sharp bounds for kernel interpolation under this distribution along with other properties related to the unbiasedness of volume sampling in the finite setting. The advantage of continuous volume sampling is that it is amenable to approximate sampling via a *fully kernelized* MCMC algorithm: this approximate sampler can be implemented using only evaluations of the kernel k, without needing to the spectral decomposition of $\mathbf{\Sigma}$.

LAYOUT OF THE THESIS AND THE CONTRIBUTIONS

We give in the following a detailed layout of the manuscript, and mention the associated publications.

Determinantal point processes

Chapter 2 is dedicated to give a formal introduction to determinantal point processes (DPPs). We give an abstract definition that includes the two settings: discrete DPPs and continuous DPPs. Moreover, we give several examples of DPPs, and we recall some elements on the numerical simulation of these probabilistic models.

Column subset selection using Projection DPPs

In Chapter 3, we propose and analyze a new column subset selection algorithm for low rank matrix factorization based on a projection DPP. In particular, we compare a particular projection DPP with volume sampling, and we prove that this projection DPP can lead to better theoretical guarantees and empirical results under a condition called *the sparsity of the k-leverage scores*. Moreover, we prove theoretical guarantees under a more realistic condition called *the relaxed sparsity of the k-leverage scores*. Interestingly, the introduction of this notion of sparsity is inspired from an analysis that we conducted of a worst-case example in the literature (Deshpande and Rademacher, 2010). Our novel analysis proves that it is possible to bypass this lower bound if we assume a sparsity condition. These sparsity conditions are shown to be satisfied by several datasets. A major contribution of this chapter is to highlight the importance of geometric parameters called the principal angles between subspaces. Besides their geometric intuition, the introduction of these parameters enables closed form calculation of some upper bounds of the expected approximation error under the projection DPP distribution. This new parametrization also allows to give a simultaneous analysis under the Frobenius norm and the spectral norm. Moreover, the conducted analysis allows the study of the regression bias. Figure 1.1 illustrates the bound of the projection DPP, that improves

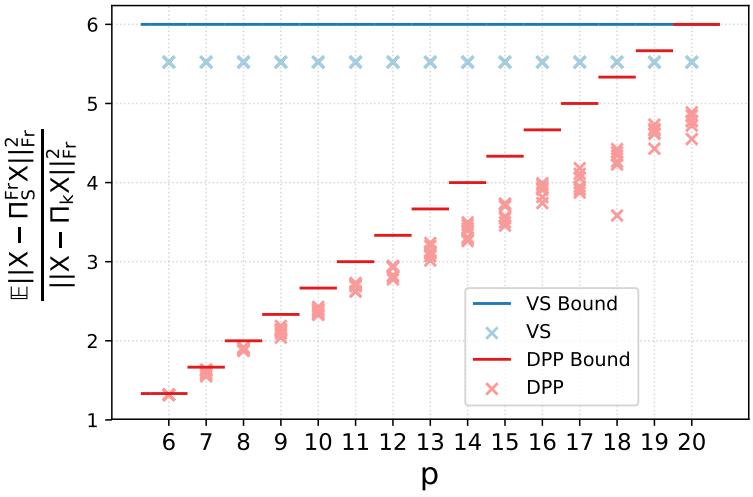


Figure 1.1 – The projection DPP improves on volume sampling under the condition of sparsity of the k-leverage scores quantified by parameter p on the x-axis.

when the sparsity of the k-leverage scores is small, compared to the bound of volume sampling on toy datasets. The material of this chapter is based on the following article

- A. Belhadji, R. Bardenet, and P. Chainais (2020a). "A determinantal point process for column subset selection". In: *Journal of Machine Learning Research* 21.197, pp. 1–62.

Projection DPPs for kernel quadrature

In Chapter 4, we introduce a new class of quadratures based on projection DPPs. These quadratures are suited for functions living in a reproducing kernel Hilbert space. Compared to the work of Bach, 2017, our quadrature is ridgeless ($\lambda = 0$), i.e. we solve the optimization problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^N} \|\mu_g - \sum_{n\in[N]} w_n k(x_n, .)\|_{\mathcal{F}}^2. \tag{1.34}$$

The solution of (1.34) writes $\hat{\boldsymbol{w}} = \boldsymbol{K}(\boldsymbol{x})^{-1}\mu_g(\boldsymbol{x})$, and the optimal mixture $\sum_{n\in[N]} \hat{w}_n k(x_n, .)$ is the projection of μ_g on the subspace $\mathcal{T}(\boldsymbol{x}) = \mathrm{Span}(k(x_n, .)_{n\in[N]})$ denoted $\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g$, and the optimal value of (1.34) is called the interpolation error of μ_g. As for the nodes, we take a random set $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ that follows the distribution of a projection DPP that depends only on the first eigenfunctions of $\boldsymbol{\Sigma}$: this projection DPP provides an implementable alternative to the continuous ridge leverage scores distribution when the spectral decomposition of $\boldsymbol{\Sigma}$ is known.

We give a theoretical analysis of this class of quadratures and we show that the convergence rate depends on the eigenvalues $(\sigma_n)_{n\in\mathbb{N}^*}$ of the integration operator as $\mathcal{O}(N\sum_{n\geq N+1}\sigma_n)$, where N is the number of quadrature nodes. Numerical simulations hint that the rate of convergence actually scales as fast as $\mathcal{O}(\sigma_N)$. This is illustrated in Figure 1.2, where we compare the worst-case interpolation error $\sup_{\|g\|_{d\omega}\leq 1} \|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2$ of our quadrature (DPPKQ) compared the rate $\mathcal{O}(\sigma_{N+1})$ and the quadrature based on continuous ridge leverage score sampling (LVSQ) and the

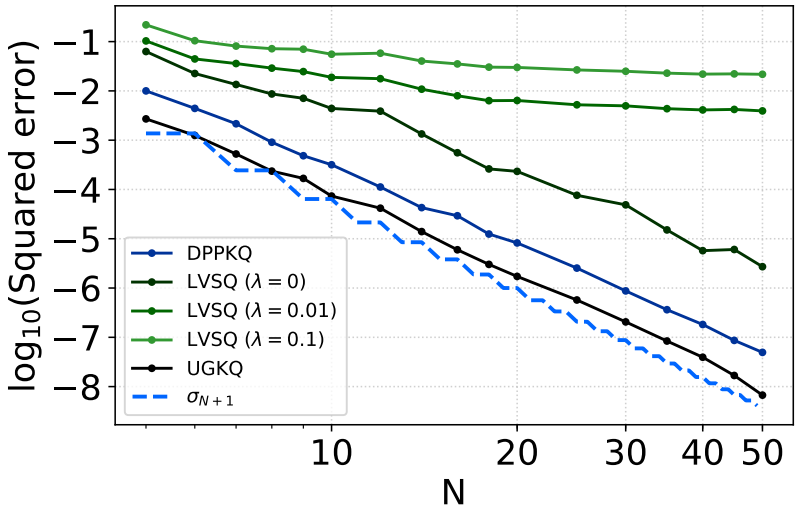


Figure 1.2 – The worst-case interpolation error on the unit ball of $\mathbb{L}_2(\mathrm{d}\omega)$ under projection DPP (DPPKQ), continuous ridge leverage score density (LVSQ) and the uniform grid (UGKQ) compared to the eigenvalue of order σ_{N+1}, in the periodic Sobolev space of order $s = 3$.

quadrature based on uniform grid (UGKQ) and the weights that solve the optimization problem (1.34) in the case of the periodic Sobolev space of order $s = 3$.

From a technical point of view, we harness the geometric intuitions developed in Chapter 3 to work out the theoretical analysis of these quadratures. In particular, we make use of the notion of principal angles between subspaces to derive a tractable upper bound of the expected value of the squared interpolation error

$$\mathbb{E}_{\mathrm{DPP}} \|\mu_g - \mathbf{\Pi}_{\mathcal{T}(x)} \mu_g\|_{\mathcal{F}}^2, \tag{1.35}$$

which has no tractable formula. The material of this chapter is based on the following article

- A. Belhadji, R. Bardenet, and P. Chainais (2019a). "Kernel quadrature with DPPs". In: *Advances in Neural Information Processing Systems* 32, pp. 12907–12917.

Continuous volume sampling for kernel interpolation

In Chapter 5, we continue the line of research initiated in Chapter 3 and Chapter 4, and we study kernel quadrature and kernel interpolation based on nodes that follow the continuous volume sampling distribution, a generalization of discrete volume sampling. Contrary to the projection DPP of Chapter 4, we show that the expected value of the squared interpolation error has a closed formula under the distribution of continuous volume sampling

$$\mathbb{E}_{\mathrm{CVS}} \|\mu_g - \mathbf{\Pi}_{\mathcal{T}(x)} \mu_g\|_{\mathcal{F}}^2 = \sum_{m \in \mathbb{N}^*} \langle g, e_m \rangle_{\mathrm{d}\omega}^2 \epsilon_m(N), \tag{1.36}$$

where

$$\epsilon_m(N) = \sum_{\substack{|T|=N \\ m \notin T}} \prod_{t \in T} \sigma_t \Big/ \sum_{|T|=N} \prod_{t \in T} \sigma_t. \tag{1.37}$$

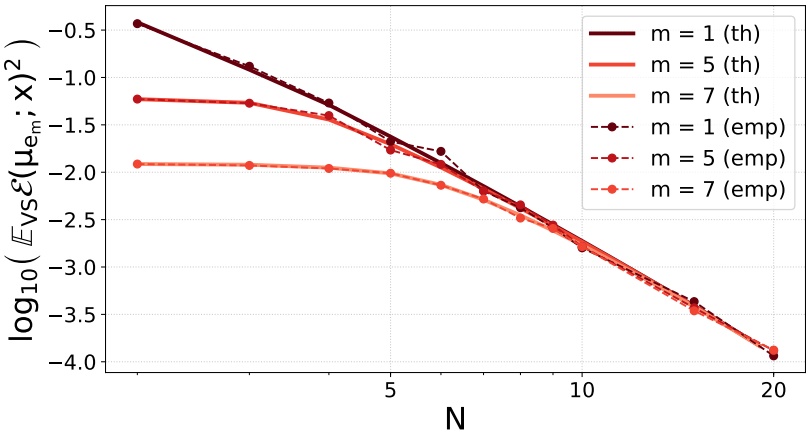


Figure 1.3 – The theoretical values of the $\epsilon_m(N)$ compared to the empirical estimation using a fully kernelized MCMC. The RKHS is the periodic Sobolev space of order $s = 2$.

Moreover, we show that the $\epsilon_m(N)$ scales as σ_{N+1} for several RKHSs, and we give convergence guarantees for interpolation outside the scope of kernel quadrature. Finally, we prove an extension of the unbiasedness property (1.28) of the discrete volume sampling distribution, and we give an interpretation of this result.

The advantage of this distribution is that it is amenable to approximation using an MCMC algorithm that does not require the spectral decomposition of $\mathbf{\Sigma}$, unlike projection DPP of Chapter 4 and the continuous ridge leverage score density of (Bach, 2017).

Figure 1.3 illustrates the comparison between the theoretical values of the $\epsilon_m(N))$ and their estimation using the MCMC sampler of (Rezaei and Gharan, 2019) in the case of the periodic Sobolev space of order $s = 2$. The material of this chapter is based on the following article

- A. Belhadji, R. Bardenet, and P. Chainais (2020b). "Kernel interpolation with continuous volume sampling". In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 725–735.

Table 1.1 situates the contributions of this thesis regarding the related work.

Distribution/ Setting	Discrete	Continuous
Projection DPP	Belhadji et al., 2020a	Belhadji et al., 2019a
Volume sampling	Beck and Teboulle, 2003 Deshpande et al., 2006 Derezinski and Warmuth, 2017	Belhadji et al., 2020b
Ridge leverage scores	Bach, 2013 Alaoui and Mahoney, 2015	Bach, 2017

Table 1.1 – A summary of the contributions of this thesis compared to existing of work.

PUBLICATIONS

JOURNAL ARTICLES

- A. Belhadji, R. Bardenet, and P. Chainais (2020a). "A determinantal point process for column subset selection". In: *Journal of Machine Learning Research* 21.197, pp. 1–62

PEER-REVIEWED CONFERENCES

- A. Belhadji, R. Bardenet, and P. Chainais (2019a). "Kernel quadrature with DPPs". In: *Advances in Neural Information Processing Systems 32*, pp. 12907–12917.
- A. Belhadji, R. Bardenet, and P. Chainais (2020b). "Kernel interpolation with continuous volume sampling". In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 725–735.
- A. Belhadji, R. Bardenet, and P. Chainais (2019b). "Un processus ponctuel déterminantal pour la sélection d'attributs". In: *GRETSI 2019*.

2

DETERMINANTAL POINT PROCESSES

Determinantal point processes (DPPs) were introduced by Macchi, 1975 as probabilistic models for beams of fermions in quantum optics and their use widely spread after the 2000s in random matrix theory (Johansson, 2005), machine learning (Kulesza and Taskar, 2012), spatial statistics (Lavancier et al., 2015), image processing (Launay, Desolneux, and Galerne, 2020a), and Monte Carlo methods (Bardenet and Hardy, 2020), among others.

This chapter is dedicated to give a formal introduction on DPPs. Intuitively, a DPP defines a random discrete subset of negatively correlated particles. These probabilistic models appear mainly in two forms: discrete DPPs and continuous DPPs. We shall give an abstract definition that includes the two settings to emphasize on the universality of this mathematical object.

We start by recalling some elements of the theory of point processes in Section 2.1. The definition of a DPP along with existence results and some basic properties of DPPs are given in Section 2.2. In particular, we give several examples of DPPs in Section 2.2.4 and we recall some elements on the numerical simulation of a DPP in Section 2.2.5.

We adopted a different notation in this chapter that will change in the next chapters. The rationale behind this choice is to keep the notation as close as possible to the usual ones in the corresponding literatures.

2.1 POINT PROCESSES

Intuitively, a point process is a random discrete subset of points in some measurable space $(\mathcal{D}, \mathcal{B})$. In order to define this object rigorously, it is more convenient to see a discrete subset of $\mathcal{D}$ as an atomic measure defined on $\mathcal{D}$ as illustrated in Figure 2.1. Indeed, as we shall see, under some conditions on $\mathcal{D}$, the set of measures on $\mathcal{D}$ have nice properties that are compatible with the standard setting of probability theory.

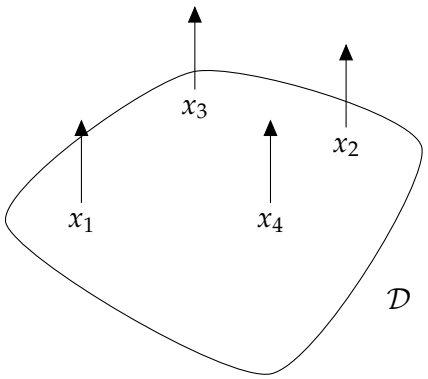


Figure 2.1 – A set $\{x_1, x_2, x_3, x_4\}$ can be identified with the atomic measure $\sum_{n=1}^{4} \delta_{x_n}$.

Polish spaces are recurrently used in probability theory. They define a large class of topological spaces that includes any countable set equipped with the discrete topology, closed subspaces of $\mathbb{R}^d$ endowed with the usual topology induced by Euclidean norm and many more. This abstract class seems to be convenient to give a unified definition of DPPs.

Definition 2.1. *Let $(\mathcal{D}, \mathcal{T})$ be a topological space. $(\mathcal{D}, \mathcal{T})$ is said to be a Polish space, if there exists a metric δ on $\mathcal{D}$ such that*

- *$(\mathcal{D}, \delta)$ is a* complete *metric space,*
- *$(\mathcal{D}, \delta)$ is* separable*: there exists a countable subset $\{x_n; n \in \mathbb{N}\} \subset \mathcal{D}$ dense in $\mathcal{D}$,*

and the topology induced by δ is equal to $\mathcal{T}$.

We say that a Polish space is a *separable completely metrizable* topological space. The requirements of a Polish space allow to work with the classical concepts of probability theory: the existence of conditional distributions, the definition of weak convergence, the regularity of probability measures...

2.1.1 *Discrete subsets as counting measures*

Once we have defined the topological space, we move to defining the corresponding Borel σ-algebra $\mathcal{B}$: the smallest σ-algebra in $\mathcal{D}$ that contains all open subsets of $\mathcal{D}$. Its elements are called Borel sets. A measure μ is a function from $\mathcal{B}$ to $\mathbb{R}_+ \cup \{\infty\}$ that is *non-negative*, *σ-additive* and vanishes at the empty set. A locally finite measure is a measure μ on $(\mathcal{D}, \mathcal{B})$ such that for every $B \in \mathcal{B}$ that is *relatively compact* (its closure is compact) we have

$$\mu(B) < +\infty. \tag{2.1}$$

Denote by $\mathbf{M}(\mathcal{D})$ the set of locally finite measures on $\mathcal{D}$. We supply $\mathbf{M}(\mathcal{D})$ with the σ-algebra $\mathcal{M}(\mathcal{D})$ generated by the evaluation maps defined for every Borel set $B \in \mathcal{B}$

$$\begin{aligned} \Phi_B : \mathcal{M}(\mathcal{D}) &\longrightarrow \mathbb{R}_+ \cup \{\infty\}. \\ \mu &\longmapsto \mu(B) \end{aligned}$$

In other words, $\mathcal{M}(\mathcal{D})$ is the smallest σ-algebra for which all the evaluations maps Φ_B are measurables. This σ-algebra is generated by the cylinder sets

$$\Big\{\mu \in \mathbf{M}(\mathcal{D}),\ \mu(B_1) \in [0, r_1], \dots, \mu(B_m) \in [0, r_m]\Big\}, \tag{2.2}$$

where $m \in \mathbb{N}^*$, $B_1, \dots, B_m$ are relatively compact open subsets of $\mathcal{D}$ and $r_1, \dots, r_m \in \mathbb{R}_+$; see Lemma 3.1.1 in (Schneider and Weil, 2008).

The set of counting measures defined by

$$\mathbf{N}(\mathcal{D}) = \Big\{\gamma \in \mathbf{M}(\mathcal{D});\ \forall B \in \mathcal{B}, \gamma(B) \in \mathbb{N} \cup \{+\infty\}\Big\} \tag{2.3}$$

is an example of an element of $\mathcal{M}(\mathcal{D})$; see Lemma 3.1.2 in (Schneider and Weil, 2008). Another example is given by the set of simple counting measures Lemma 3.1.4 in (Schneider and Weil, 2008)

$$\mathbf{N}_s(\mathcal{D}) = \Big\{\gamma \in \mathbf{M}(\mathcal{D});\ \forall x \in \mathcal{D}, \gamma(\{x\}) \in \{0, 1\}\Big\}. \tag{2.4}$$

γ as a discrete subset	γ as a simple counting measure
$x \in \gamma$	$\gamma(\{x\}) = 1$
$\{x_1, \dots, x_N\} \subset \gamma$	$\gamma(\{x_1, \dots, x_N\}) = N$
$\lvert\gamma\rvert = M$	$\gamma(\mathcal{D}) = M$
$\gamma \cap C = \emptyset$	$\gamma(C) = 0$
$\lvert\gamma \cap C\rvert = M$	$\gamma(C) = M$

Table 2.1 – A dictionary of notations between discrete subsets and simple counting measures.

Starting from a collection of points $x_1, \dots, x_N$ in $\mathcal{D}$, one can define a counting measure

$$\gamma = \sum_{n \in [N]} \delta_{x_n}, \tag{2.5}$$

where

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

In this case, $\gamma \in \mathbf{N}_s(\mathcal{D})$ if and only if the x_n are pairwise distinct; and γ can be identified to the subset $\{x_1, \dots, x_N\}$. In particular, many operations on discrete subsets can be expressed using operations on simple counting measures. Table 2.1 contains some of these operations expressed using both notations. These two notations will be used interchangeably in this manuscript. Now, if an element x_n is repeated more than once in the collection, γ is no longer simple and it can be seen as a discrete subset of $\mathcal{D}$ with multiplicities. This intuition is based on the following result.

Theorem 2.1 ((Daley and Vere-Jones, 2007)[Proposition 9.1.III]**.** *] Let γ be a counting measure on $\mathcal{D}$. Then γ can be written as*

$$\gamma = \sum_{i \in I} m_i \delta_{x_i}, \tag{2.6}$$

where I is a countable set and $(x_i)_{i \in I}$ is a sequence of points in $\mathcal{D}$ without accumulation point and $(m_i)_{i \in I}$ is a sequence of positive integers.

In the case of a simple counting measure, the integers m_i belong to $\{0, 1\}$.

2.1.2 *Point processes as random measures*

Now, we give the definition of a point process. It is a special case of a random measure.

Definition 2.2. *A random measure γ on $\mathcal{D}$ is a measurable map from some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ into the measurable space $(\mathbf{M}(\mathcal{D}), \mathcal{M}(\mathcal{D}))$. The distribution of γ is given by the probability measure $\mathbb{P}_\gamma$ defined on $\mathcal{M}(\mathcal{D})$ by*

$$\forall M \in \mathcal{M}(\mathcal{D}),\ \mathbb{P}_\gamma(M) = \mathbb{P}(\gamma \in M). \tag{2.7}$$

Intuitively, a random measure γ on $\mathcal{D}$ defines a stochastic process with values in $\mathbb{R}_+$ indexed by $\mathcal{B}$: $\{\gamma(B)\}_{B \in \mathcal{B}}$; see Proposition 1.1.7. in (Baccelli, Błaszczyszyn, and Karray, 2020).

Definition 2.3. *A point process γ on $\mathcal{D}$ is a random measure on $\mathcal{D}$ such that*

$$\mathbb{P}\big(\gamma \in \mathbf{N}(\mathcal{D})\big) = 1. \tag{2.8}$$

The point process γ is said to be a simple if

$$\mathbb{P}\big(\gamma \in \mathbf{N}_s(\mathcal{D})\big) = 1. \tag{2.9}$$

Consider the following example. Let ω be a probability measure on $\mathcal{D}$ and let ν be a probability measure on $\mathbb{N}$. Let $x_0, x_1 \ldots,$ be independent random variables that follow the distribution ω and N a random variable that follows the distribution ν. Then the random measure γ defined by

$$\gamma = \sum_{n \in [N]} \delta_{x_n}, \tag{2.10}$$

is a point process on $\mathcal{D}$. This point process is called the *mixed binomial process* with mixing distribution ν and sampling distribution ω. Now, if $\nu = \delta_{N_0}$ for some $N_0 \in \mathbb{N}^*$, we recover the binomial process defined by

$$\forall B \in \mathcal{B},\ \forall n \in \{0, \ldots, N_0\},\ \mathbb{P}(\gamma(B) = n) = \binom{N_0}{n} \omega(B)^n (1 - \omega(B))^{N_0 - n}. \tag{2.11}$$

On the other hand, if ν follows the distribution of a Poisson random variable of parameter 1, then we recover a Poisson point process of intensity ω that we define in the following.

Definition 2.4. *Let μ be a locally finite measure on $\mathcal{D}$. A Poisson process with intensity measure $\mathrm{d}\omega$ is a point process γ such that for all pairwise disjoint Borel sets $B_1, \ldots, B_m \in \mathcal{B}$, the random variables $\gamma(B_1), \ldots, \gamma(B_m)$ are independent Poisson random variables with respective parameters $\omega(B_1), \ldots, \omega(B_m)$.*

In other words, under a Poisson point process of intensity ω, for every $B_1, \ldots, B_m \in \mathcal{B}$ pairwise disjoint Borel sets, and $n_1, \ldots, n_m \in \mathbb{N}$

$$\mathbb{P}\Big(\gamma(B_1) = n_1, \ldots, \gamma(B_m) = n_m\Big) = e^{-\omega(B_1)} \frac{\omega(B_1)^{n_1}}{n_1!} \times \cdots \times e^{-\omega(B_m)} \frac{\omega(B_m)^{n_m}}{n_m!}. \tag{2.12}$$

A Poisson point process is simple if and only if its intensity measure is *diffuse*, i.e. it has no atoms (Lemma 3.2.1 in (Schneider and Weil, 2008))

$$\forall x \in \mathcal{D},\ \omega(\{x\}) = 0. \tag{2.13}$$

2.1.3 *The mean measure of a point process*

The description of a point process through the cylinder sets

$$\big\{\gamma(B_1) \ = \ n_1, \ \ldots, \ \gamma(B_m) \ = \ n_m \ \big\}$$

is convenient in some cases such as the binomial process and the Poisson point process. However, in general the probability

$$\mathbb{P}\Big(\gamma(B_1) = n_1, \ldots, \gamma(B_m) = n_m\Big) \tag{2.14}$$

have no tractable formula and it is more convenient to work with an alternative description offered by the moment measures that are defined in the following. We start with the mean measure.

Definition 2.5. *Let γ be a point process. The mean measure of γ is the measure γ_1 defined by*

$$\forall B \in \mathcal{B},\ \gamma_1(B) = \mathbb{E}\gamma(B). \tag{2.15}$$

The mean measure is well defined but may take infinite values.

Example 2.1. *Let γ be a Poisson point process associated to an intensity ω. It is immediate from (2.12) that*

$$\forall B \in \mathcal{B},\ \mathbb{E}\gamma(B) = \omega(B). \tag{2.16}$$

Therefore, the mean measure of γ is equal to ω.

Now, consider ω to be a measure on $(\mathcal{D}, \mathcal{B})$ that we call the *reference measure*. If γ_1 is absolutely continuous with respect to ω, then the Radon Nikodym derivative ρ_1 is called the *first intensity function* and it satisfies

$$\forall B \in \mathcal{B},\ \mathbb{E}\gamma(B) = \gamma_1(B) = \int_B \rho_1(x)\mathrm{d}\omega(x). \tag{2.17}$$

Example 2.2. *Let $\mathcal{D}$ be a finite set and define the counting measure $\omega = \sum_{x\in\mathcal{D}} \delta_x$. Let γ be a simple point process defined on $\mathcal{D}$. Then, by definition of γ_1*

$$\begin{aligned}
\gamma_1(B) &= \sum_{n\in\mathbb{N}^*} \mathbb{P}(\gamma(B) = n)n \\
&= \sum_{n\in\mathbb{N}^*} \mathbb{P}(\gamma(B) = n) \sum_{b\in B} \mathbb{P}\left(\gamma(\{b\}) = 1 | \gamma(B) = n\right) \\
&= \sum_{b\in B} \sum_{n\in\mathbb{N}^*} \mathbb{P}(\gamma(B) = n)\, \mathbb{P}\left(\gamma(\{b\}) = 1 | \gamma(B) = n\right) \\
&= \sum_{b\in B} \mathbb{P}\left(\gamma(\{b\}) = 1\right) \\
&= \sum_{b\in B} \mathbb{P}\left(\gamma(\{b\}) = 1\right)\omega(\{b\}) \\
&= \int_B \rho_1(x)\mathrm{d}\omega(x),
\end{aligned} \tag{2.18}$$

where

$$\rho_1(x) = \mathbb{P}\big(\gamma(\{x\}) = 1\big) = \mathbb{P}\big(x \in \gamma\big). \tag{2.19}$$

In other words, the first intensity function of a simple point process is the inclusion probability $\mathbb{P}\big(x \in \gamma\big)$.

We shall see later a similar characterization of the first intensity function of simple point processes in the case of a continuous domain.

Now, given a simple point process γ, and by Theorem 2.1, the identity (2.15) can be rewritten

$$\forall B \in \mathcal{B},\ \int_{\mathcal{D}} \mathbb{1}_B \mathrm{d}\gamma_1(x) = \mathbb{E}\sum_{x\in\gamma} \mathbb{1}_B(x). \tag{2.20}$$

The measurable function $\mathbb{1}_B$ in the identity (2.20) can be replaced by any nonnegative measurable function as it is shown in the following result.

Theorem 2.2 (Campbell–Hardy theorem). *Let γ be simple point process in $\mathcal{D}$, and let $f : \mathcal{D} \to \mathbb{R}$ be a nonnegative measurable function. Then $\sum_{x\in\gamma} f(x)$ is measurable and*

$$\int_{\mathcal{D}} f \mathrm{d}\gamma_1(x) = \mathbb{E} \sum_{x\in\gamma} f(x). \tag{2.21}$$

In particular, if γ_1 is absolutely continuous with respect to ω, then

$$\int_{\mathcal{D}} f(x)\rho_1(x)\mathrm{d}\omega(x) = \mathbb{E} \sum_{x\in\gamma} f(x). \tag{2.22}$$

Note that the Campbell theorem can be applied for general point processes: the term $\sum_{x\in\gamma} f(x)$ should be replaced by $\int_{\mathcal{D}} f \mathrm{d}\gamma$. Moreover, it can be extended to functions $f : \mathcal{D} \to \mathbb{R}$ that are integrable with respect to γ_1. See Theorem 1.2.5 in (Baccelli et al., 2020) for a general statement and its proof.

2.1.4 *High-order moment measures of a point process*

The mean measure gives an idea of the random measure γ evaluated on one Borel set $B \in \mathcal{B}$ yet it does not capture the correlations between the evaluation of γ on a family of Borel sets $B_1, \dots, B_L \in \mathcal{B}$. There are many ways to estimate this interaction. For example, we can define the L-th power of γ that is a point process on the measurable space $\mathcal{D}^L$, equipped with the product σ-algebra, defined by

$$\forall \prod_{\ell\in[L]} B_\ell \in \mathcal{B}^L,\ \gamma^{\otimes_L}(\prod_{\ell\in[L]} B_\ell) = \prod_{\ell\in[L]} \gamma(B_\ell), \tag{2.23}$$

and we consider the mean measure of $\gamma^{\otimes_L}$ defined by

$$\gamma_1^{\otimes_L}(\prod_{\ell\in[L]} B_\ell) = \mathbb{E}\,\gamma^{\otimes_L}(\prod_{\ell\in[L]} B_\ell) = \mathbb{E} \prod_{\ell\in[L]} \gamma(B_\ell). \tag{2.24}$$

The measure $\gamma_1^{\otimes_L}$ is also called the L-th moment measure of γ. The definition of $\gamma_1^{\otimes_L}$ is straightforward, yet it is not convenient in the study of DPPs that will be presented later. An alternative definition that is more compatible with the structure of DPPs is slightly more technical, and it is given in the following.

Define

$$\mathcal{D}^L_{\neq} = \{(x_1, \dots, x_L) \in \mathcal{D}^L;\ \forall \ell, \ell' \in [L],\ x_\ell \neq x_{\ell'}\}. \tag{2.25}$$

The set $\mathcal{D}^L_{\neq}$ is an open set of $\mathcal{D}^L$ (equipped with the product topology), therefore $\mathcal{D}^L_{\neq} \in \mathcal{B}^L$; and we can define the restriction of the point process $\gamma^{\otimes_L}$ to $\mathcal{D}^L_{\neq}$, denoted $\gamma^{\otimes_L}\lfloor\mathcal{D}^L_{\neq}$, by

$$\gamma^{\otimes_L}\lfloor\mathcal{D}^L_{\neq}(\prod_{\ell\in[L]} B_\ell) = \gamma^{\otimes_L}(\prod_{\ell\in[L]} B_\ell \cap \mathcal{D}^L_{\neq}). \tag{2.26}$$

$\gamma^{\otimes_L}\lfloor\mathcal{D}^L_{\neq}$ is a point process on $\mathcal{D}^L$ and we can define its mean measure that we denote γ_L

$$\gamma_L(\prod_{\ell\in[L]} B_\ell) = \mathbb{E}\,\gamma^{\otimes_L}\lfloor\mathcal{D}^L_{\neq}(\prod_{\ell\in[L]} B_\ell) = \mathbb{E}\,\gamma^{\otimes_L}(\prod_{\ell\in[L]} B_\ell \cap \mathcal{D}^L_{\neq}). \tag{2.27}$$

Now, by definition of $\mathcal{D}^L_{\neq}$, if the $B_1, \dots, B_L$ are pairwise distinct, then

$$\gamma_L\big(\prod_{\ell\in[L]} B_\ell\big) = \mathbb{E}\,\gamma^{\otimes_L}\big(\prod_{\ell\in[L]} B_\ell \cap \mathcal{D}^L_{\neq}\big) = \mathbb{E}\,\gamma^{\otimes_L}\big(\prod_{\ell\in[L]} B_\ell\big) = \mathbb{E}\prod_{\ell\in[L]}\gamma(B_\ell). \tag{2.28}$$

Yet, the equality between $\gamma_L\big(\prod_{\ell\in[L]} B_\ell\big)$ and $\mathbb{E}\prod_{\ell\in[L]}\gamma(B_\ell)$ is not valid in general, and this is the main difference between the point process $\gamma^{\otimes_L}$ and the point process $\gamma^{\otimes_L}\lfloor\mathcal{D}^L_{\neq}$. This difference is better illustrated for simple point processes.

Let γ be a simple point process, the two point processes $\gamma^{\otimes_L}$ and $\gamma^{\otimes_L}\lfloor\mathcal{D}^L_{\neq}$ are simple. Indeed, let $\prod_{\ell\in[L]}\{x_\ell\} \subset \mathcal{D}^L$, we have by the simplicity of γ

$$\gamma^{\otimes_L}\big(\prod_{\ell\in[L]}\{x_\ell\}\big) = \prod_{\ell\in[L]}\gamma\big(\{x_\ell\}\big) \in \{0,1\},\ \text{a.s.} \tag{2.29}$$

As for $\gamma^{\otimes_L}\lfloor\mathcal{D}^L_{\neq}$, we have two cases: either $\prod_{\ell\in[L]}\{x_\ell\} \subset \mathcal{D}^L_{\neq}$ or $\prod_{\ell\in[L]}\{x_\ell\} \cap \mathcal{D}^L_{\neq} = \emptyset$. In the first case

$$\gamma^{\otimes_L}\lfloor\mathcal{D}^L_{\neq}\big(\prod_{\ell\in[L]}\{x_\ell\}\big) = \gamma^{\otimes_L}\big(\prod_{\ell\in[L]}\{x_\ell\}\big) = \prod_{\ell\in[L]}\gamma(\{x_\ell\}) \in \{0,1\}. \tag{2.30}$$

In the second case,

$$\gamma^{\otimes_L}\lfloor\mathcal{D}^L_{\neq}\big(\prod_{\ell\in[L]}\{x_\ell\}\big) = \gamma^{\otimes_L}(\emptyset) = 0 \in \{0,1\}. \tag{2.31}$$

In other words, $\gamma^{\otimes_L}\lfloor\mathcal{D}^L_{\neq}$ excludes any family $(x_1, \dots, x_L)$ that contains the same element more than once; this is not the case of $\gamma^{\otimes_L}$.

Definition 2.6. *Let γ be a simple point process. If γ_L is absolutely continuous with respect to $\omega^{\otimes_L}$, then the corresponding Radon Nikodym derivative ρ_L can be defined and it is called the L-intensity function or the joint intensity function of order L. In particular, for any family of pairwise disjoint Borel subsets $B_1, \dots, B_L \in \mathcal{B}$*

$$\mathbb{E}\prod_{\ell\in[L]}\gamma(B_\ell) = \int_{\prod_{\ell\in[L]} B_\ell} \rho_L(x_1, \dots, x_L)\mathrm{d}\omega(x_1)\dots\mathrm{d}\omega(x_L). \tag{2.32}$$

We give in the following two examples, the intuition behind the joint intensity functions.

Example 2.3. *Consider $\mathcal{D} = \mathbb{R}^d$ equipped with Lebesgue measure ω. Let γ be a simple point process on $\mathcal{D}$. Let $x_1, \dots, x_L \in \mathcal{D}$ pairwise distinct and let $\epsilon > 0$ such that the balls $B_\epsilon(x_\ell)$, centered around the x_ℓ and of radii equal to ϵ, are pairwise distinct. Under some assumptions on the point process γ, we can prove that (see Chapter 1 of (Hough et al., 2009))*

$$\rho_L(x_1, \dots, x_L) = \lim_{\epsilon\to 0} \frac{\mathbb{P}\Big(\forall \ell \in [L],\ \gamma(B_\epsilon(x_\ell)) = 1\Big)}{\omega\Big(B_\epsilon(x_\ell)\Big)}. \tag{2.33}$$

Example 2.4. *Let $\mathcal{D}$ be a finite set and define the counting measure $\omega = \sum_{x\in\mathcal{D}} \delta_x$; and we consider a simple point process γ. Using a similar analysis to the one in Example 2.2, we can prove that*

$$\rho_L(x_1,\dots,x_L) = \mathbb{P}\big(\forall \ell \in [L],\ \gamma(\{x_\ell\}) = 1\big). \tag{2.34}$$

This can be rewritten as

$$\rho_L(x_1,\dots,x_L) = \mathbb{P}\big(\{x_1,\dots,x_L\} \subset \gamma\big). \tag{2.35}$$

Again, $\rho_L(x_1,\dots,x_L)$ can be interpreted as an inclusion probability.

The Campbell-Hardy theorem can be applied to the point process $\gamma^{\otimes_L} \lfloor \mathcal{D}^L_{\neq}$. By observing that, for a nonnegative measurable function $f : \mathcal{D}^L \to \mathbb{R}$, the sum

$$\sum_{(x_1,\dots,x_L)\in\gamma^{\otimes_L}\lfloor\mathcal{D}^L_{\neq}} f(x_1,\dots,x_L) \tag{2.36}$$

can be simplified to

$$\sum_{\substack{(x_1,\dots,x_L)\in\gamma^{\otimes_L}\\ x_\ell\neq x_{\ell'}}} f(x_1,\dots,x_L), \tag{2.37}$$

we obtain the following result.

Theorem 2.3. *Let γ be simple point process in $\mathcal{D}$, and let $f : \mathcal{D}^L \to \mathbb{R}$ be a nonnegative measurable function. Then*

$$\sum_{\substack{(x_1,\dots,x_L)\in\gamma^{\otimes_L}\\ x_\ell\neq x_{\ell'}}} f(x_1,\dots,x_L) \tag{2.38}$$

is measurable and

$$\mathbb{E} \sum_{\substack{(x_1,\dots,x_L)\in\gamma^{\otimes_L}\\ x_\ell\neq x_{\ell'}}} f(x_1,\dots,x_L) = \int_{\mathcal{D}^L} f \mathrm{d}\gamma_L. \tag{2.39}$$

In particular, if the joint intensity ρ_L is defined

$$\mathbb{E} \sum_{\substack{(x_1,\dots,x_L)\in\gamma^{\otimes_L}\\ x_\ell\neq x_{\ell'}}} f(x_1,\dots,x_L) = \int_{\mathcal{D}^L} f(x_1,\dots,x_L)\rho_L(x_1,\dots,x_L)\mathrm{d}\omega(x_1)\times\cdots\times\mathrm{d}\omega(x_L). \tag{2.40}$$

Theorem 2.3 may be called the "fundamental theorem of calculus" using point processes. It is at the heart of many theoretical analysis that concern DPPs.

2.1.5 *The orthogonality of Poisson point processes*

We conclude this section with a property of Poisson point processes that motivates the introduction of DPPs in Section 2.2.

Let γ be a Poisson point process on $\mathcal{D}$ associated to an intensity ω. Let $B_1,\dots,B_L \in \mathcal{B}$ be pairwise disjoint Borel sets. By (2.12), we have

$$\gamma_L\Big(\prod_{\ell\in[L]} B_\ell\Big) = \mathbb{E}\prod_{\ell\in[L]}\gamma(B_\ell) = \prod_{\ell\in[L]}\omega(B_\ell). \tag{2.41}$$

The identity (2.41) reflects the independence of the random variables $\gamma(B_1), \dots, \gamma(B_L)$ when γ follows the distribution of a Poisson point process. In other words, under the law of Poisson point process, there is no interaction between pairwise disjoint Borel sets. This is to be compared to the correlations that appear under the distribution of a determinantal point process.

2.2 DETERMINANTAL POINT PROCESSES

2.2.1 *Definition*

We give now the definition of a DPP.

Definition 2.7. *Let $\kappa : \mathcal{D} \times \mathcal{D} \to \mathbb{C}$ be a measurable function. A simple point process γ on $\mathcal{D}$ is said to be a determinantal point process with kernel κ and reference measure $\mathrm{d}\omega$ if the joint intensities ρ_L with respect to the measure $\mathrm{d}\omega$ are well defined for $L \in \mathbb{N}^*$ and satisfy*

$$\rho_L(x_1, \dots, x_L) = \operatorname{Det} \boldsymbol{\kappa}\big(x_1, \dots, x_L\big), \tag{2.42}$$

where $\boldsymbol{\kappa}\big(x_1, \dots, x_L\big) = \big(\kappa(x_\ell, x_{\ell'})\big)_{\ell,\ell' \in [L]} \in \mathbb{C}^{N \times N}$.

In other words, a DPP defines a point process with the correlations, defined through the joint intensity functions, are described by the kernel κ. From Definition 2.7, we can already define a point process with the *negative correlation* property. Indeed, let γ be a DPP of kernel κ and reference measure $\mathrm{d}\omega$. Assume that the kernel κ is Hermitian

$$\kappa(x, y) = \overline{\kappa(y, x)}. \tag{2.43}$$

Let B_1, B_2 two disjoint Borel sets. By (2.42), we have

$$\begin{aligned}
\mathbb{E}\ \gamma(B_1)\gamma(B_2) &= \int_{B_1 \times B_2} \operatorname{Det} \boldsymbol{\kappa}(x_1, x_2) \mathrm{d}\omega(x_1) \mathrm{d}\omega(x_2) \\
&= \int_{B_1 \times B_2} \Big(\kappa(x_1, x_1)\kappa(x_2, x_2) - \kappa(x_1, x_2)\kappa(x_2, x_1)\Big) \mathrm{d}\omega(x_1) \mathrm{d}\omega(x_2) \\
&= \int_{B_1} \kappa(x_1, x_1) \mathrm{d}\omega(x_1) \int_{B_2} \kappa(x_2, x_2) \mathrm{d}\omega(x_2) - \int_{B_1 \times B_2} |\kappa(x_1, x_2)|^2 \mathrm{d}\omega(x_1) \mathrm{d}\omega(x_2) \\
&= \mathbb{E}\, \gamma(B_1)\, \mathbb{E}\, \gamma(B_2) - \int_{B_1 \times B_2} |\kappa(x_1, x_2)|^2 \mathrm{d}\omega(x_1) \mathrm{d}\omega(x_2).
\end{aligned} \tag{2.44}$$

Therefore

$$\operatorname{Cov}\Big(\gamma(B_1), \gamma(B_2)\Big) = -\int_{B_1 \times B_2} |\kappa(x_1, x_2)|^2 \mathrm{d}\omega(x_1) \mathrm{d}\omega(x_2) \leq 0. \tag{2.45}$$

In other words, the covariance of the random variables $\gamma(B_1)$ and $\gamma(B_1)$ is non-positive. It is negative, when $\int_{B_1 \times B_2} |\kappa(x_1, x_2)|^2 \mathrm{d}\omega(x_1) \mathrm{d}\omega(x_2) > 0$, and the two variables are negatively correlated. This is to be compared with a Poisson point process where the random variables $\gamma(B_1)$ and $\gamma(B_1)$ are independents.

Observe the importance of the condition (2.43) from the second line to the third line in the development of (2.44) where we have used

$$\int_{B_1 \times B_2} \kappa(x_1, x_2)\kappa(x_2, x_1) \mathrm{d}\omega(x_1) \mathrm{d}\omega(x_2) = \int_{B_1 \times B_2} \kappa(x_1, x_2)\overline{\kappa(x_1, x_2)} \mathrm{d}\omega(x_1) \mathrm{d}\omega(x_2). \tag{2.46}$$

Without the condition (2.43), the negative correlation property does not hold. We present in the next section a class of kernels that satisfy this condition and define DPPs with interesting properties.

2.2.2 *DPPs associated to trace class kernels*

The existence of a simple point process that satisfies (2.42) enforces some constraints on κ. First, since the joint intensities functions are non-negative functions κ should satisfy

$$\operatorname{Det} \boldsymbol{\kappa}(x_1, \dots, x_L) \geq 0 \tag{2.47}$$

The constraint (2.47) is satisfied whenever κ is Hermitian (2.43) and positive

$$\forall L \in \mathbb{N}^*,\ \forall a_1, \dots, a_L \in \mathbb{C},\ \forall x_1, \dots, x_L \in \mathcal{D},\ \sum_{\ell, \ell' \in [L]} a_\ell \overline{a_{\ell'}} \kappa(x_\ell, x_{\ell'}) \geq 0. \tag{2.48}$$

In order to define DPPs with interesting properties, we make a further assumption: κ is square integrable on $\mathcal{D}^2$, that is

$$\int_{\mathcal{D}^2} |\kappa(x,y)|^2 \mathrm{d}\omega(x)\mathrm{d}\omega(y) < +\infty, \tag{2.49}$$

so that we can define the corresponding integration operator with respect to $\mathrm{d}\omega$

$$\begin{aligned} \boldsymbol{\Sigma}_\kappa : \mathbb{L}_2(\mathrm{d}\omega) &\longrightarrow \mathbb{L}_2(\mathrm{d}\omega) \\ g &\longmapsto \int_{\mathcal{D}} g(.)\kappa(.,y)\mathrm{d}\omega(y). \end{aligned}$$

Now the condition (2.43) implies that $\boldsymbol{\Sigma}_\kappa$ is a self-adjoint operator on $\mathbb{L}_2(\mathrm{d}\omega)$; the condition (2.48) implies that $\boldsymbol{\Sigma}_\kappa$ is non-negative definite; and (2.49) implies that $\boldsymbol{\Sigma}_\kappa$ is a compact operator. Therefore, by the the spectral theorem for compact self-adjoint operators (Chapter 6 in (Brezis, 2010)), $\mathbb{L}_2(\mathrm{d}\omega)$ have an orthonormal basis $(v_n)_{n \in \mathbb{N}^*}$ of eigenfunctions of $\boldsymbol{\Sigma}_\kappa$, the corresponding eigenvalues σ_n are non-negative, have finite multiplicities and 0 is the only possible accumulation point of the spectrum. Furthermore, $\boldsymbol{\Sigma}_\kappa$ is said to be of *trace class* if

$$\sum_{n \in \mathbb{N}^*} \sigma_n < +\infty. \tag{2.50}$$

In the following we call conditions (2.43), (2.48), (2.49) and (2.50) the *usual conditions*.

Now, we are ready to recall a fundamental characterization of kernels that define DPPs.

Theorem 2.4 (Macchi, 1975; Soshnikov, 2000)**.** *Let κ such that $\boldsymbol{\Sigma}_\kappa$ satisfies the usual conditions. Then κ defines a DPP if and only if the spectrum of $\boldsymbol{\Sigma}_\kappa$ is contained in $[0,1]$.*

Projection kernels form a large class of kernels that satisfy the conditions of Theorem 2.4 and they define what is commonly known as *projection DPPs*. They define DPPs with deterministic cardinalities as it is stated in the following result.

Proposition 2.1 (Lemma 17 in (Hough et al., 2006)). *Let $v_1, \dots, v_M \in \mathbb{L}_2(\mathrm{d}\omega)$, such that $(v_m)_{m\in[M]}$ is an orthonormal family of $\mathbb{L}_2(\mathrm{d}\omega)$, and define the kernel*

$$\kappa(x,y) = \sum_{m\in[M]} v_m(x)\overline{v_m(y)}. \tag{2.51}$$

Let γ be the DPP associated to the kernel κ and reference measure $\mathrm{d}\omega$, then

$$\mathbb{P}\left(\gamma(\mathcal{D}) = M\right) = 1. \tag{2.52}$$

These DPPs are important in the characterization of DPPs that satisfy the conditions of Theorem 2.4. Indeed, we have the following result.

Theorem 2.5 (Theorem 7 in (Hough et al., 2006)). *Let κ be a kernel such that $\mathbf{\Sigma}_\kappa$ satisfies the usual conditions and denote (σ_m, v_m) its eigenpairs. Assume that the spectrum of $\mathbf{\Sigma}_\kappa$ is included in $[0,1]$.*

Let γ be the random measure defined as follows. Let $I_1, I_2, \dots$ be independent Bernoulli random variables such that for every $m \in \mathbb{N}^*$, the parameter of I_m is equal to σ_m; and let γ be a random counting measure that follows the distribution of the projection DPP associated to the reference measure $\mathrm{d}\omega$ and the kernel

$$\kappa_I(x,y) = \sum_{m\in\mathbb{N}^*} I_m v_m(x)\overline{v_m(y)}. \tag{2.53}$$

Then γ follows the distribution of the determinantal point process associated to the kernel κ and the reference measure $\mathrm{d}\omega$.

Theorem 2.5 implies that any DPP defined trough a kernel κ that satisfies the usual conditions is a mixture of projection DPPs. Observe that the intermediate projection kernel κ_I is of finite rank a.s. This is a consequence of the trace class condition that guarantees that

$$\sum_{m\in\mathbb{N}^*} I_m < +\infty \text{ a.s.} \tag{2.54}$$

The distribution of the cardinality of this class of DPPs follows immediately from Theorem 2.5 and Proposition 2.1.

Corollary 2.1. *Let γ be the DPP associated to the kernel κ and reference measure $\mathrm{d}\omega$, then*

$$\gamma(\mathcal{D}) \sim \sum_{m\in\mathbb{N}^*} I_m. \tag{2.55}$$

We shall see in Section 2.2.4 another class of mixtures of projection DPPs that define point processes with determinsitic cardinality.

2.2.3 *The definition of DPPs under weaker assumptions*

The usual assumptions (2.43), (2.49) and (2.50) made previously can be relaxed in many ways. We review two relaxations that are common in the literature.

First, the assumption (2.49) is usually relaxed so that κ is only assumed to be locally square integrable: for any compact set C of $\mathcal{D}$

$$\int_{C^2} |\kappa(x,y)|^2 \mathrm{d}\omega(x)\mathrm{d}\omega(y) < +\infty. \tag{2.56}$$

Under this relaxed assumption, the domain of definition of the integration operator $\mathbf{\Sigma}_\kappa$ should be restrained to elements of $\mathbb{L}_2(\mathrm{d}\omega)$ that vanish *a.e.* outside a compact subset of $\mathcal{D}$. Moreover, Theorem 2.4 remains valid and Theorem 2.5 and its consequences are still valid for the restricted integration operators: if γ follows the distribution of a DPP of kernel κ and reference measure $\mathrm{d}\omega$ then for every compact set $C \subset \mathcal{D}$, $\gamma \cap C$ follows the distribution of a DPP of the kernel κ and the reference measure $\mathrm{d}\omega_C$ defined by

$$\forall B \in \mathcal{B},\ \omega_C(B) = \omega(B \cap C). \tag{2.57}$$

The Hermitianity of κ can be relaxed too: there are examples of DPPs associated to non-Hermitian kernels ; see Section 2.5 in (Soshnikov, 2000) and (Brunel, 2018). Yet, many interesting aspects of DPPs require this condition: negative correlations, tractable numerical simulation ...

2.2.4 *Examples*

We review in this section some examples of DPPs.

DISCRETE DPPS This class of DPPs was the topic of intense research recently for machine learning applications; we refer the reader to (Kulesza and Taskar, 2012) for details. In this setting, the domain is taken to be $\mathcal{D} = [d]$ and usually; the reference measure is the counting measure $\omega = \sum_{i \in [d]} \delta_i$; and the definition 2.7 is equivalent to the following.

Definition 2.8 (Discrete DPP). *Let $\boldsymbol{K} \in \mathbb{R}^{d \times d}$ be a positive semi-definite matrix. A random subset $Y \subseteq [d]$ is drawn from a DPP of marginal kernel $\boldsymbol{K}$ if and only if*

$$\forall S \subseteq [N], \quad \mathbb{P}(S \subseteq Y) = \mathrm{Det}(\boldsymbol{K}_S), \tag{2.58}$$

where $\boldsymbol{K}_S = [\boldsymbol{K}_{i,j}]_{i,j \in S}$. We take as a convention $\mathrm{Det}(\boldsymbol{K}_\emptyset) = 1$.

In this setting, $\kappa(i,j) = \boldsymbol{K}_{i,j}$ for every pair $(i,j) \in \mathcal{D}^2$, and $\mathbb{L}_2(\mathrm{d}\omega)$ coincides with $\mathbb{R}^d$ and the integration operator $\mathbf{\Sigma}_\kappa$ is given by

$$\mathbf{\Sigma}_\kappa g(i) = \int_{\mathcal{D}} g(j)\kappa(i,j)\mathrm{d}\omega(j) = \sum_{j \in [d]} g_j \boldsymbol{K}_{i,j} = (\boldsymbol{K}g)_i, \tag{2.59}$$

therefore

$$\forall g \in \mathbb{R}^d,\ \mathbf{\Sigma}_\kappa g = \boldsymbol{K}g. \tag{2.60}$$

According to Theorem 2.4, a sufficient condition, for a given matrix $\boldsymbol{K}$ to consistently define a DPP, is that $\boldsymbol{K}$ is symmetric and its spectrum is in $[0,1]$. Moreover, for a symmetric matrix $\boldsymbol{K}$, a DPP can be seen as a *repulsive* distribution, in the sense that for all $i, j \in [d]$

$$\begin{aligned}
\mathbb{P}(\{i,j\} \subseteq Y) &= \boldsymbol{K}_{i,i}\boldsymbol{K}_{j,j} - \boldsymbol{K}_{i,j}^2 && (2.61)\\
&= \mathbb{P}(\{i\} \subseteq Y)\,\mathbb{P}(\{j\} \subseteq Y) - \boldsymbol{K}_{i,j}^2 && (2.62)\\
&\leq \mathbb{P}(\{i\} \subseteq Y)\,\mathbb{P}(\{j\} \subseteq Y). && (2.63)
\end{aligned}$$

Besides projection DPPs, there is another natural way of using a kernel matrix to define a random subset of $[d]$ with prespecified cardinality k.

Definition 2.9 (k-DPP). *Let $\boldsymbol{L} \in \mathbb{R}^{d\times d}$ be a positive semidefinite matrix. A random subset $Y \subseteq [d]$ is drawn from a k-DPP of kernel $\boldsymbol{L}$ if and only if*

$$\forall S \subseteq [d], \quad \mathbb{P}(Y = S) \propto \mathbb{1}_{\{|S|=k\}} \operatorname{Det}(\boldsymbol{L}_S) \tag{2.64}$$

where $\boldsymbol{L}_S = [\boldsymbol{L}_{i,j}]_{i,j\in S}$.

DPPs and k-DPPs are closely related but different objects. For starters, k-DPPs are always well-defined provided $\boldsymbol{L}$ has a nonzero minor of size k. The next proposition establishes that k-DPPs also are mixtures of projection DPPs.

Proposition 2.2. *(Kulesza and Taskar (2012, Section 5.2.2)) Let Y be a random subset of $[d]$ sampled from a k-DPP with kernel $\boldsymbol{L}$. We further assume that $\boldsymbol{L}$ is symmetric, we denote its rank by r and its diagonalization by $\boldsymbol{L} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\intercal$. Finally, let $k \leq r$. It holds that*

$$\mathbb{P}(Y = S) = \sum_{\substack{T\subseteq[r]\\|T|=k}} \mu_T \left[\frac{1}{k!} \operatorname{Det}\left(\boldsymbol{V}_{T,S}\boldsymbol{V}_{T,S}^\intercal\right)\right] \tag{2.65}$$

where

$$\mu_T = \frac{\prod_{i\in T}\lambda_i}{\sum_{\substack{U\subseteq[r]\\|U|=k}} \prod_{i\in U}\lambda_i}. \tag{2.66}$$

Each mixture component in square brackets in (2.65) is a projection DPP with cardinality k. The main difference between k-DPPs and DPPs is that all mixture components in (2.65) have the same cardinality k. In particular, projection DPPs are the only DPPs that are also k-DPPs.

THE CIRCULAR UNITARY ENSEMBLE The set of the eigenvalues of a random matrix chosen uniformly (from the Haar measure) on the unitary group $\mathbb{U}_N$, also called the Circular Unitary Ensemble (CUE), was introduced by Dyson, 1962. It is an example of a DPP on a continuous domain: $\mathcal{D}$ is the unit circle in the complex plane and ω is the Lebesgue measure on $\mathcal{D}$. The construction of this DPP goes as follows.

For $N \in \mathbb{N}^*$, denote by $\mathbb{U}_N$ the group of $N \times N$ unitary matrices. $\mathbb{U}_N$ endowed with the induced topology, as a subset of the set of $N \times N$ complex-valued matrices, is compact; therefore, there exists a unique Borel probability measure $\mathrm{d}M$ on $\mathbb{U}_N$, called *Haar measure*, that is invariant under left multiplication by unitary matrices; see Theorem 5.14 in (Rudin, 1991). The following result is fundamental in the statistical description of the CUE.

Theorem 2.6 (Weyl, 1946). *Let $f : \mathbb{U}_N \to \mathbb{C}$ satisfy*

$$\forall\, V, M \in \mathbb{U}_N,\ f(V^{-1}MV) = f(M), \tag{2.67}$$

and denote for $\boldsymbol{\theta} = (\theta_n)_{n\in[N]} \in [0, 2\pi]^N$, the diagonal matrix $D(\boldsymbol{\theta}) = \operatorname{Diag}(e^{\mathbf{i}\theta_n})_{n\in[N]}$. Then

$$\int_{\mathbb{U}_N} f(M)\mathrm{d}M = \frac{1}{N!(2\pi)^N} \int_{[0,2\pi]^N} f(D(\theta))\Delta(\theta)^2 \mathrm{d}^N\theta, \tag{2.68}$$

where

$$\Delta(\theta) = \prod_{1\leq n,n'\leq N} |e^{\mathbf{i}\theta_n} - e^{\mathbf{i}\theta_{n'}}|. \tag{2.69}$$

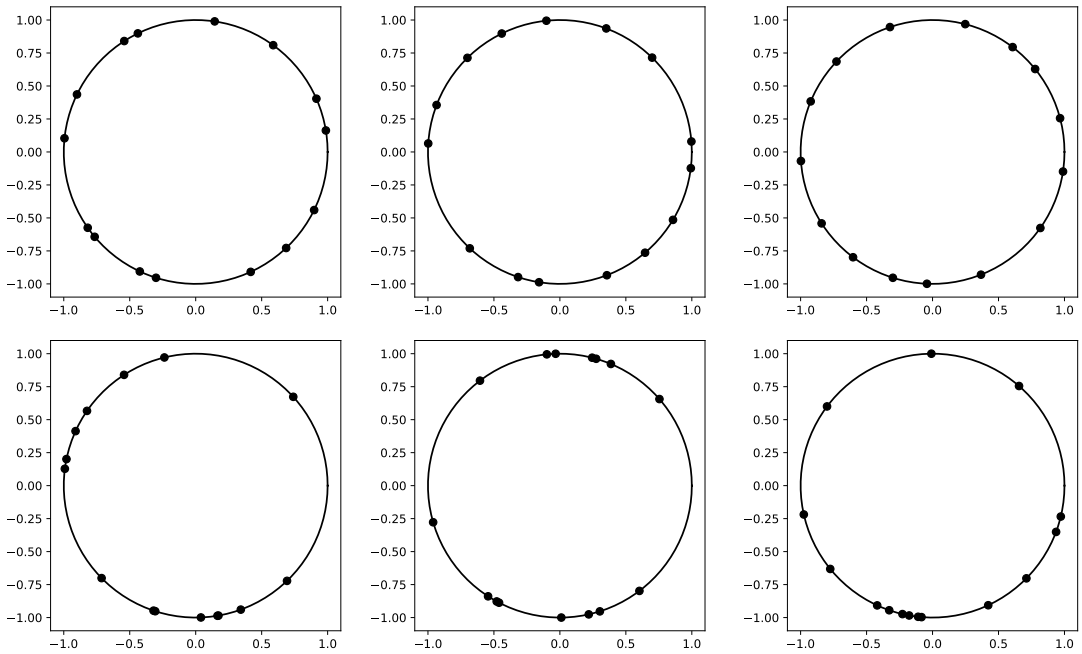


Figure 2.2 – Three realizations from the CUE with $N = 15$ particles (top) compared to 15 particles sampled i.i.d in the unit circle (bottom).

The identity (2.68) is known as the *Weyl integration formula*: it gives an explicit formula to integrate over $\mathbb{U}_N$ a function that only depends on the eigenvalues. This identity highlights the important term $\Delta(\theta)^2$ that has a determinantal representation

$$\Delta(\theta)^2 = \mathrm{Det}(\kappa(e^{\mathbf{i}\theta_n}, e^{\mathbf{i}\theta_{n'}}))_{n,n' \in [N]}, \tag{2.70}$$

where

$$\kappa(e^{\mathbf{i}\theta}, e^{\mathbf{i}\theta'}) = \frac{1}{2\pi} \sum_{n \in [N]} e^{2\pi n \mathbf{i}(\theta - \theta')}. \tag{2.71}$$

This representation implies the following result.

Theorem 2.7 (Dyson, 1962)**.** *The counting measure of the CUE is a determinantal point process on the unit complex circle* $\mathbb{U} = \{u \in \mathbb{C}, |u| = 1\}$ *with kernel* κ *and reference measure the Lebesgue measure on* $\mathbb{U}$*.*

Observe that the cardinality of this DPP is constant a.s. This is an illustration of Proposition 2.1: the kernel κ defines a projection operator onto the N dimensional subspace spanned by the functions $z \mapsto z^n$; $n \in \{0, \dots, N-1\}$.

Figure 2.2 shows three realizations of the CUE ($N = 15$) compared to three realizations from the binomial process ($N = 15$) corresponding to the Lebesgue measure on $\mathbb{U}$. We observe that the CUE tends to cover the unit circle with more regularity than the binomial process. We discuss this regularity later in Section 4.1.3.

GINIBRE ENSEMBLE Ginibre, 1965 studied the statistical properties of the eigenvalues of matrices with independent complex Gaussian entries. This ensemble bears his name and follows the distribution of a DPP defined on $\mathcal{D} = \mathbb{C}$.

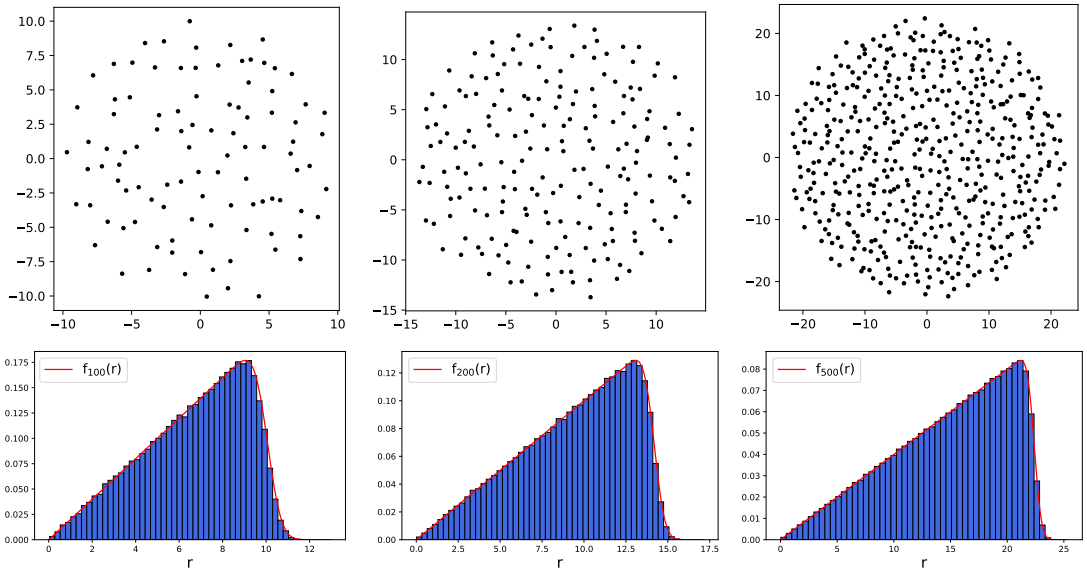


Figure 2.3 – A realization of the Ginibre ensemble with $N = 100$ (left), $N = 200$ (middle) and $N = 500$ (right). The histogram of the radii $|z|$ of 1000 realization of the Ginibre ensemble compared to the theoretical density f_N, with $N = 100$ particles (left) $N = 200$ particles (middle) and $N = 500$ particles (right).

Theorem 2.8 (Ginibre, 1965)**.** *Let M be an $N \times N$ matrix with i.i.d. standard complex Gaussian entries. Then the eigenvalues of M follow the distribution of determinantal point process associated to the kernel defined on the complex plane*

$$\kappa(z, w) = \sum_{\ell=0}^{N-1} \frac{(z\overline{w})^\ell}{\ell!}, \tag{2.72}$$

with respect to the reference measure $\mathrm{d}\omega(z) = \frac{1}{\pi} e^{-|z|^2} \mathrm{d}\lambda(z)$, where λ is the Lebesgue measure on $\mathbb{C}$.

The kernel κ defines a projection operator onto the N dimensional subspace spanned by the functions $z \mapsto z^n$; $n \in \{0, \dots, N-1\}$, and converges, as $N \to +\infty$, to the kernel

$$\kappa_\infty(z, w) = e^{z\overline{w}}, \tag{2.73}$$

that defines the space of all entire functions in $\mathbb{L}_2(\mathrm{d}\omega)$.

The top of Figure 2.3 shows a realization of three Ginibre ensembles corresponding to $N \in \{100, 200, 500\}$. We observe that for the three values of N, the configuration takes a "spherical form" that can be explained by the rotational invariance of the mean measure $\kappa(z, z)\mathrm{d}\omega(z)$. Moreover, the modulus of the elements of the Ginibre ensemble form a point process on $\mathbb{R}_+$, and the corresponding mean measure has a density with respect to the Lebesgue measure that writes

$$f_N(r) = \frac{2}{N} \sum_{\ell=0}^{N-1} \frac{r^{2\ell+1}}{\ell!} e^{-r^2} \mathbb{1}_{\mathbb{R}_+}(r). \tag{2.74}$$

The bottom of the Figure 2.3 shows histograms of the modulus computed on 1000 samples of the Ginibre ensemble for the three values of $N \in \{100, 200, 500\}$ compared to

the respective densities f_{100}, f_{200} and f_{500}. Besides the matching between the theoretical density f_N and the empirical histogram; we observe that density f_N is unimodal and achieves its maximal value at $r \approx \sqrt{N}$, then manifest a sharp drop to very small values.

THE SPHERICAL ENEMBLE The spherical ensemble is another example of a DPP based on the eigenvalues of random matrices. This ensemble is defined on the complex plane; yet, using the inverse of the stereographic projection, it can be seen as a DPP on the sphere $\mathbb{S}^2$, hence the name. This ensemble was already used in theoretical physics for the modeling of repulsive particles in the sphere (Caillol, 1981) (Forrester et al., 1992). The spectral representation of this DPP was discovered by Krishnapur, 2009.

Theorem 2.9 (Theorem 3 in (Krishnapur, 2009))**.** *Let A, B be independent $N \times N$ matrix with i.i.d. standard complex Gaussian entries. The counting measure corresponding to the eigenvalues $\{z_1, \dots, z_N\} \subset \mathbb{C}$ of $A^{-1}B$ follows the distribution of the determinantal point process associated to the kernel defined on the complex plane*

$$\kappa_{\mathbb{C}}(z, w) = (1 + z\overline{w})^{N-1}, \tag{2.75}$$

with respect to the reference measure $\mathrm{d}\omega(z) = \frac{N}{\pi(1+|z|^2)^{N+1}}\mathrm{d}\lambda(z)$, where λ is the Lebesgue measure on $\mathbb{C}$.

Theorem 2.9 have can be reinterpreted as follows. Let $\pi : \mathbb{C} \cup \{\infty\} \to \mathbb{S}^2$ be the inverse of the stereographic projection defined by

$$\pi(z) = \frac{1}{1 + |z|^2}\Big(2\mathfrak{Re}(z), 2\mathfrak{Im}(z), |z|^2 - 1\Big). \tag{2.76}$$

For every $n \in [N]$ let $x_n = \pi(z_n)$. Then, $\{x_1, \dots, x_N\}$ follows the distribution of the determinantal point process associated to the kernel

$$\kappa_{\mathbb{S}^2}(x, y) = \kappa_{\mathbb{C}}(\pi^{-1}(x), \pi^{-1}(y)), \tag{2.77}$$

with respect to the uniform measure on $\mathbb{S}^2$.

2.2.5 *Numerical simulation*

We recall in this section some algorithms for the numerical simulation of DPPs.

Algorithms based on random matrix models

As we have seen in Section 2.2.4, the eigenvalues of some random matrix models correspond to specific projection DPPs. This is the case, for example, of the Circular Unitary Ensemble, the Ginibre Ensemble and the Spherical Ensemble.

Orthogonal Polynomial Ensembles (OPE) on the real line form another class of projection DPPs that can be represented as the eigenvalues of random matrices. For this class of DPPs, the kernel κ writes

$$\kappa(x, y) = \sum_{n=0}^{N-1} P_n(x)P_n(y), \tag{2.78}$$

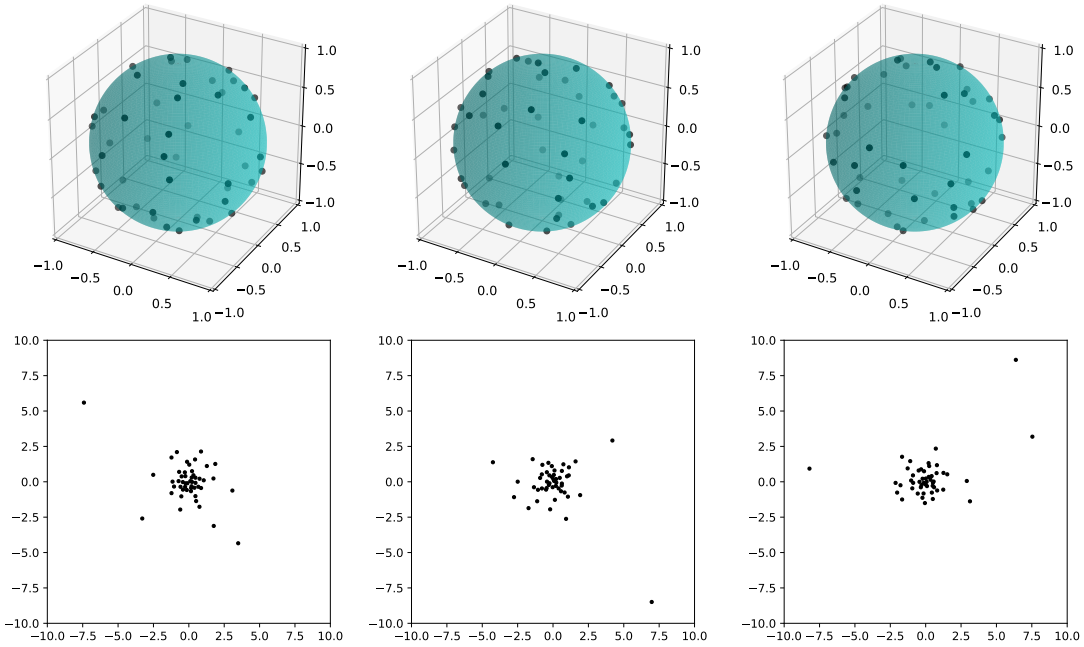


Figure 2.4 – Three realizations from the spherical ensemble with $N = 50$ particles (top) compared to their stereographic projections on the complex plane (bottom).

where (P_n) is a family of orthonormal polynomials with respect to some measure ω.

In particular, when ω is Gaussian, Gamma (Dumitriu and Edelman, 2002), or Beta (Killip and Nenciu, 2004), the corresponding DPP can be sampled by diagonalizing a tridiagonal matrix with independent entries, with a cost that scales as $\mathcal{O}(N^2)$. We illustrate this technique in Section 4.1.1 in Chapter 4. We refer to (Gautier et al., 2020) for more details on this topic.

Now, in general, there exists an algorithm for the numerical simulation of a projection DPP. This algorithm will be recalled in the following section.

The HKPV algorithm for projection DPPs

A universal sampling algorithm of projection DPPs was proposed by Hough et al., 2006. The idea behind the algorithm goes as follow.

Define the projection kernel

$$\kappa(x, y) = \sum_{n \in [N]} \psi_n(x)\psi_n(y), \tag{2.79}$$

where $(\psi_n)_{n\in[N]}$ is an orthonormal family with respect to some reference measure $\mathrm{d}\omega$.

Consider the random variable $(x_1, \dots, x_N)$ defined on $\mathcal{D}^N$ that have the density f_κ with respect to the measure $\mathrm{d}\boldsymbol{\omega}(x_1, \dots, x_N) = \otimes_{n\in[N]}\mathrm{d}\omega(x_n)$

$$f_\kappa(x_1, \dots, x_N) = \frac{1}{N!} \operatorname{Det} \boldsymbol{\kappa}(x_1, \dots, x_n). \tag{2.80}$$

Then the random counting measure $\gamma = \sum_{n\in[N]} \delta_{x_n}$ follows the distribution of the projection DPP $(\kappa, \mathrm{d}\omega)$.

Indeed, $\boldsymbol{\kappa}(x_1, \ldots, x_N)$ is the Gram matrix of the family $\big(\kappa(x_n, .)\big)_{n\in[N]}$, whose i, j entry is given by the product of $\kappa(x_i, .)$ and $\kappa(x_j, .)$ with respect to bilinear form defined by $\mathrm{d}\omega$. In fact, by the orthonormality of $(\psi_n)_{n\in[N]}$

$$\big\langle \kappa(x_i, .), \kappa(x_j, .) \big\rangle_{\mathrm{d}\omega} = \sum_{n\in[N]} \sum_{n'\in[N]} \int_{\mathcal{X}} \psi_n(x_i)\psi_n(x)\psi_{n'}(x_j)\psi_{n'}(x)\mathrm{d}\omega(x) = \kappa(x_i, x_j). \quad (2.81)$$

Hence, $\mathrm{Det}\,\boldsymbol{\kappa}(x_1, \ldots, x_N)$ corresponds to the volume of the parallelepiped defined in $\mathbb{L}_2(\mathrm{d}\omega)$ and spanned by the vectors $(\kappa(x_n, .))_{n\in[N]}$. Using base×height formula, Hough et al., 2006 proved that the density (2.80), with respect to the measure $\mathrm{d}\boldsymbol{\omega}(\boldsymbol{x}) = \otimes_{n\in[N]}\mathrm{d}\omega(x_n)$, factorizes as the product of N densities (with respect to $\mathrm{d}\omega$)

$$f_\kappa(x_1, \ldots, x_N) = \prod_{n\in[N]} f_{\kappa,n}(x_n), \quad (2.82)$$

where

$$f_{\kappa,1}(x) = \frac{1}{N}\kappa(x, x), \quad (2.83)$$

and for $n > 1$

$$f_{\kappa,n}(x) = \frac{1}{N-n+1}\left(\kappa(x, x) - \kappa(x, \boldsymbol{x}_n)^\intercal \boldsymbol{\kappa}(\boldsymbol{x}_n)^{-1}\kappa(x, \boldsymbol{x}_n)\right), \quad (2.84)$$

where $\boldsymbol{x}_n = (x_1, x_2, \ldots, x_{n-1})$ and $\kappa(x, \boldsymbol{x}_n) = (\kappa(x, x_\ell))_{\ell\in[n]}$.

Theorem 2.10 (Proposition 19 in (Hough et al., 2006))**.** *The counting measure associated to the set of points constructed by Algorithm in Figure 2.5 follows the distribution of projection DPP of kernel κ and reference measure $\mathrm{d}\omega$.*

HKPV$(\kappa, \mathrm{d}\omega)$

1 $n \longleftarrow N$
2 $\mathrm{d}\mu(x) \longleftarrow f_{\kappa,1}(x)\mathrm{d}\omega(x)$ ▷ *Initialize the distribution*
3 $\boldsymbol{x} \longleftarrow \emptyset$
4 **while** $n > 0$
5 Pick x_{N-n+1} in $\mathcal{D}$ from μ ▷ *Sample from the distribution*
6 $\boldsymbol{x} \longleftarrow \boldsymbol{x} \cup \{x_{N-n+1}\}$
7 $n \longleftarrow n - 1$
8 $\mathrm{d}\mu(x) \longleftarrow f_{\kappa,n}(x)\mathrm{d}\omega(x)$ ▷ *Update the distribution*
9 **return** $\boldsymbol{x}$

Figure 2.5 – Pseudocode of the HKPV algorithm for sampling from a projection DPP of marginal kernel κ.

In practice, the implementation of this algorithm requires to sample from the conditional distributions $f_{\kappa,n}\mathrm{d}\omega$ (Step 5 in Algorithm in Figure 2.5). In some cases, sampling directly from these distributions is possible. We give two examples.

Example 2.5 (Sampling from a discrete projection DPP). *Let $\mathcal{D} = [d]$ and $\omega = \sum_{x\in\mathcal{D}} \delta_x$. Define the matrix*

$$\boldsymbol{K} = \boldsymbol{V}\boldsymbol{V}^\intercal, \tag{2.85}$$

where $\boldsymbol{V} \in \mathbb{R}^{d\times k}$ such that

$$\boldsymbol{V}^\intercal\boldsymbol{V} = \mathbb{I}_k. \tag{2.86}$$

$\boldsymbol{K}$ is a projection matrix and defines a projection DPP (with respect to ω). The HKPV algorithm can be implemented using a sampler from multinomial distributions. In particular, the first density writes

$$f_{\kappa,1}(x) = \frac{1}{k}\sum_{i\in[k]} \boldsymbol{V}_{x,i}^2, \tag{2.87}$$

which can be calculated in a $\mathcal{O}(kd)$ operations if the matrix $\boldsymbol{V}$ is known. Similarly, the conditionals $f_{\kappa,2}, \dots, f_{\kappa,k}$ can be calculated in a polynomial time in k and d.

The right of Table 2.2 illustrates the densities $f_{\kappa,n}$ as a function of n when $d = 12$ and $k = 8$: once a point x is selected at some step n, $f_{\kappa,m}(x) = 0$ for every $m > n$; in other words, a point in $\mathcal{D}$ is selected only once.

Example 2.6 (Sampling from the projection DPP associated to the Haar wavelets family). *Let $\mathcal{D} = [0,1]$ and ω is the uniform measure $\mathcal{D}$. Define the Haar wavelets family $(\psi_{n,m})_{n\in\mathbb{N},m\in\{0,2^n-1\}}$ by*

$$\psi_{n,m}(x) = 2^{n/2}\psi(2^{n/2}x - m), \tag{2.88}$$

where

$$\psi(x) = \mathbb{1}_{[0,1/2[}(x) - \mathbb{1}_{[1/2,1[}(x). \tag{2.89}$$

This is an orthonormal family with respect to ω. We consider the kernel

$$\kappa(x,y) = \sum_{n=0}^{\nu}\sum_{m=0}^{2^n-1} \psi_{m,n}(x)\psi_{m,n}(y). \tag{2.90}$$

In this case the first density

$$\forall x \in [0,1],\ f_{\kappa,1}(x) = 1, \tag{2.91}$$

is constant, and the conditionals $f_{\kappa,n}$ are step functions that can be sampled from efficiently. Therefore, the HKPV algorithm can be implemented efficiently for this family.

The left of Table 2.2 illustrates the densities $f_{\kappa,n}$ as a function of n when the rank of the projection DPP is equal to 8 ($\nu = 2$): once a point x is selected at some step n, $f_{\kappa,m}(y) = 0$ for every $m > n$ for every y in the interval $[m2^{-\nu}, (m+1)2^{-\nu}[$.

In general, sampling directly from $f_{\kappa,n}$ is not possible. Bardenet and Hardy, 2020 proposed to use rejection sampling based on the following observation

$$f_{\kappa,n}(x) \leq \frac{N}{n} f_{\kappa,1}(x) = \frac{N}{n}\kappa(x,x). \tag{2.92}$$

Therefore, it is enough to sample from the first distribution $f_{\kappa,1}\mathrm{d}\omega$ and use rejection sampling for the other conditionals.

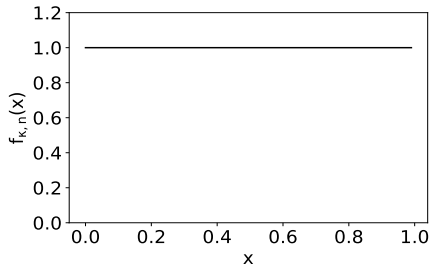
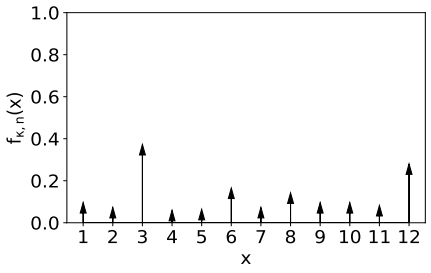
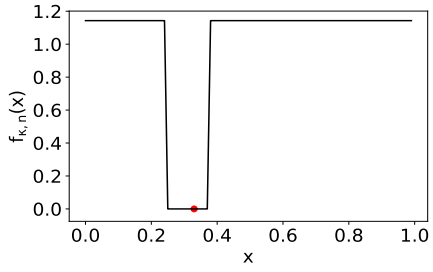
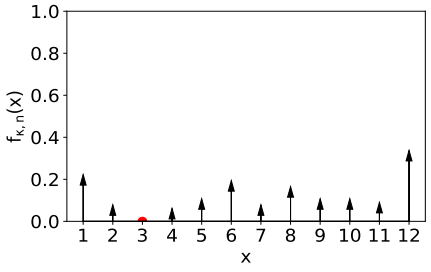
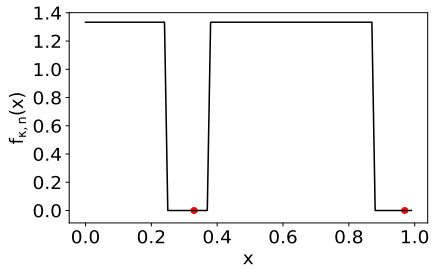
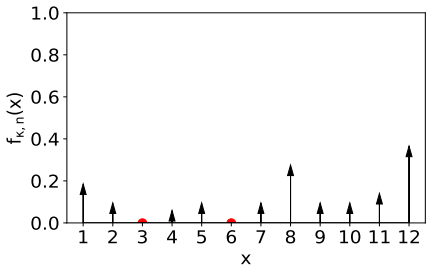
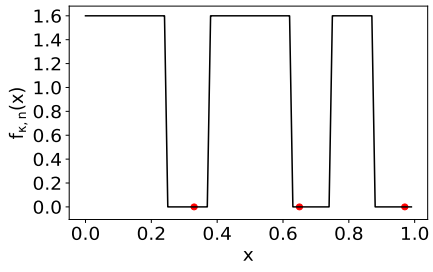
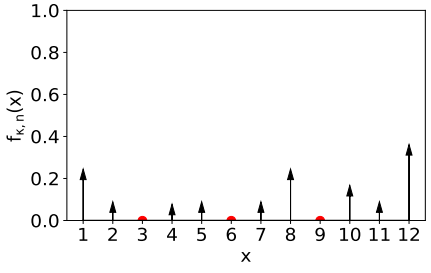
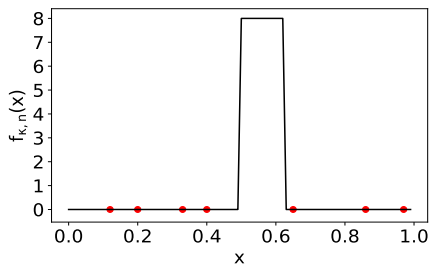
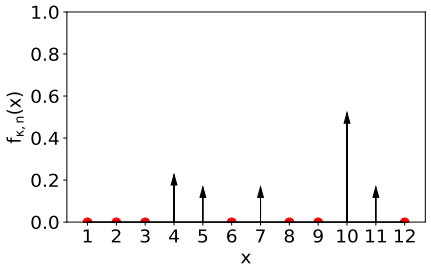
n	The density $f_{\kappa,n}$ (Haar wavelets family)	The density $f_{\kappa,n}$ (Projection matrix)
1		
2		
3		
4		
...	...	...
8		

Table 2.2 – The HKPV algorithm in action: the densities $f_{\kappa,n}$ for the projection DPP defined by the Haar wavelets family and the uniform measure on $[0,1]$(left), the densities $f_{\kappa,n}$ for the projection DPP defined by an orthogonal matrix of rank 8 and the counting measure on the set $\{1,\dots,12\}$ (right).

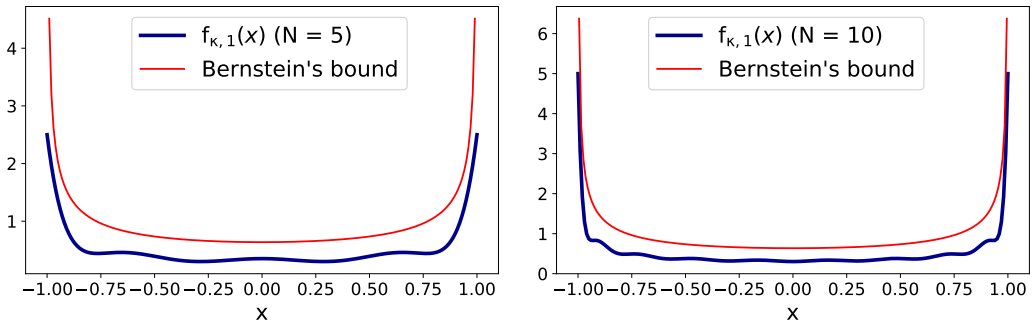


Figure 2.6 – The sharpness of Bernstein's inequality for bounding the density $f_{\kappa,1}$.

Example 2.7. *Let $(L_n)_{n\in\mathbb{N}}$ be the family of normalized Legendre polynomials defined by*

$$\int_{-1}^{1} L_n(x)L_{n'}(x)\mathrm{d}x = \delta_{n,n'}. \tag{2.93}$$

Let κ be the projection kernel

$$\kappa(x,y) = \sum_{n=0}^{N-1} L_n(x)L_n(y). \tag{2.94}$$

We have

$$\forall x \in]-1,1[,\ f_{\kappa,1}(x) \leq \frac{2}{\pi}\frac{1}{\sqrt{1-x^2}}. \tag{2.95}$$

Indeed, by Bernstein's inequality (Lorch, 1983), we have

$$L_n(x)^2 \leq \frac{2}{\pi}\frac{1}{\sqrt{1-x^2}}. \tag{2.96}$$

Figure 2.6 illustrates the sharpness of the bound (2.95) for $N = 5$ and $N = 10$.

Approximate sampling

As we have seen in Theorem 2.5, DPPs corresponding to trace class Hermitian kernels are mixtures of projection DPPs. Therefore, one can use the HKPV algorithm to simulate the intermediate projection DPP κ_I (in Theorem 2.5). However, in some situations, this approach can be non-efficient or even can not be applied at all. Indeed, in many settings both discrete and continuous, the spectral decomposition of $\mathbf{\Sigma}_\kappa$ is not tractable. For this reason, many approximate samplers were proposed for DPPs to circumvent sampling the mixture.

MCMC algorithms were proposed in (Anari et al., 2016) and (Gautier et al., 2017) for DPPs and k-DPPs defined in discrete domains. These algorithms do not require the spectral decomposition of the kernel matrix. An extension of the sampler of (Anari et al., 2016) to continuous domains was proposed in (Rezaei and Gharan, 2019). We will present and discuss this algorithm in Chapter 5. Recently, a fast sampler of intractable OPE was proposed in (Gautier et al., 2020). This algorithm is based on an MCMC in the domain of tri-diagonal matrices; see Section 4.1.1.

Another approximate sampler, suited for translation invariant kernels defined on $\mathbb{R}^d$, was proposed in (Lavancier et al., 2015). This sampler is based on the approximation of the function ϕ in the Fourier domain where ϕ is defined by

$$\kappa(x, y) = \phi(x - y). \tag{2.97}$$

3

COLUMN SUBSET SELECTION USING PROJECTION DPPS

Datasets come in always larger dimensions, and dimension reduction is thus often one the first steps in any machine learning pipeline. Two of the most widespread strategies are principal component analysis (PCA) and feature selection. PCA projects the data in directions of large variance, called principal components. While the initial features (the canonical coordinates) generally have a direct interpretation, principal components are linear combinations of these original variables, which makes them hard to interpret. On the contrary, using a selection of original features will preserve interpretability when it is desirable. Once the data are gathered in an $N \times d$ matrix, of which each row is an observation encoded by d features, feature selection boils down to selecting columns of $\boldsymbol{X}$. Independently of what comes after feature selection in the machine learning pipeline, a common performance criterion for feature selection is the approximation error in some norm, that is, the norm of the residual after projecting $\boldsymbol{X}$ onto the subspace spanned by the selected columns. Optimizing such a criterion over subsets of $\{1, \dots, d\}$ requires exhaustive enumeration of all possible subsets, which is prohibitive in high dimension. One alternative is to use a polynomial-cost, random subset selection strategy that favours small values of the criterion.

This rationale corresponds to a rich literature on randomized algorithms for column subset selection (Deshpande and Vempala, 2006; Drineas et al., 2007; Boutsidis et al., 2011). A prototypal example corresponds to sampling s columns of $\boldsymbol{X}$ i.i.d. from a multinomial distribution of parameter $\boldsymbol{p} \in \mathbb{R}^d$. This parameter $\boldsymbol{p}$ can be the squared norms of each column (Drineas et al., 2004), for instance, or the more subtle k-leverage scores (Drineas et al., 2007). While the former only takes $\mathcal{O}(dN)$ time to evaluate, it comes with loose guarantees; see Section 3.2. The k-leverage scores are more expensive to evaluate, since they call for a truncated SVD of order k, but they come with tight bounds on the ratio of their expected approximation error over that of PCA.

To minimize approximation error, the subspace spanned by the selected columns should be as large as possible. Simultaneously, the number of selected columns should be as small as possible, so that intuitively, diversity among the selected columns is desirable. The column subset selection problem (CSSP) then becomes a question of designing a discrete point process over the column indices $\{1, \dots, d\}$ that favours diversity in terms of directions covered by the corresponding columns of $\boldsymbol{X}$. Beyond the problem of designing such a point process, guarantees on the resulting approximation error are desirable. Since, given a target dimension $k \leq d$ after projection, PCA provides the best approximation in Frobenius or spectral norm, it is often used as a reference: a good CSS algorithm preserves interpretability of the c selected features while guaranteeing an approximation error not much worse than that of rank-k PCA, all of this with c not much larger than k.

In this chapter, we introduce and analyse a new randomized algorithm for selecting k diverse columns. Diversity is ensured using a determinantal point process (DPP). In a sense, the DPP we propose is a nonindependent generalization of the multinomial

sampling with k-leverage scores of Boutsidis et al., 2009. It further naturally connects to volume sampling, the CSS algorithm that has the best error bounds (Deshpande et al., 2006). We give error bounds for DPP sampling that exploit sparsity and decay properties of the k-leverage scores, and outperform volume sampling when these properties hold. Our claim is backed up by experiments on toy and real datasets.

The chapter is organized as follows. Section 3.1 introduces our notations. Section 3.2 is a survey of column subset selection, up to the state of the art to which we later compare. In Section 3.3, we discuss determinantal point processes and their connection to volume sampling. Section 3.4 contains our main results, in the form of both classical bounds on the approximation error and risk bounds when CSS is a prelude to linear regression. In Section 3.5, we numerically illustrate the theoretical results. We conclude with the discussion in Section 3.6. The details of the proofs are gathered in Section 3.7.

The material of this chapter is based on the following article

- A. Belhadji, R. Bardenet, and P. Chainais (2020a). "A determinantal point process for column subset selection". In: *Journal of Machine Learning Research* 21.197, pp. 1–62.

3.1 NOTATION

We use $[n]$ to denote the set $\{1,\dots,n\}$, and $[n:m]$ for $\{n,\dots,m\}$. We use bold capitals $\boldsymbol{A},\boldsymbol{X},\dots$ to denote matrices. For a matrix $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ and subsets of indices $I\subset[m]$ and $J\subset[n]$, we denote by $\boldsymbol{A}_{I,J}$ the submatrix of $\boldsymbol{A}$ obtained by keeping only the rows indexed by I and the columns indexed by J. When we mean to take all rows or $\boldsymbol{A}$, we write $\boldsymbol{A}_{:,J}$, and similarly for all columns. We write $\mathrm{rk}(\boldsymbol{A})$ for the rank of $\boldsymbol{A}$, and $\sigma_i(\boldsymbol{A})$, $i=1,\dots,\mathrm{rk}(\boldsymbol{A})$ for its singular values, ordered decreasingly. Sometimes, we will need the vectors $\Sigma(\boldsymbol{A})$ and $\Sigma(\boldsymbol{A})^2$ with respective entries $\sigma_i(\boldsymbol{A})$ and $\sigma_i^2(\boldsymbol{A})$, $i=1,\dots,\mathrm{rk}(\boldsymbol{A})$. Similarly, when $\boldsymbol{A}$ can be diagonalized, $\Lambda(\boldsymbol{A})$ (and $\Lambda(\boldsymbol{A})^2$) are vectors with the decreasing eigenvalues (squared eigenvalues) of $\boldsymbol{A}$ as entries. If $\boldsymbol{A}$ is a symmetric matrix, $\mathrm{Sp}(\boldsymbol{A})$ denotes the vector of its eigenvalues in decreasing order.

The spectral norm of $\boldsymbol{A}$ is $\|\boldsymbol{A}\|_2=\sigma_1(\boldsymbol{A})$, while the Frobenius norm of $\boldsymbol{A}$ is defined by

$$\|\boldsymbol{A}\|_{\mathrm{Fr}}=\sqrt{\sum_{i=1}^{\mathrm{rk}(\boldsymbol{A})}\sigma_i(\boldsymbol{A})^2}.$$

For $\ell\in\mathbb{N}$, we need to introduce the ℓ-th elementary symmetric polynomial on $L\in\mathbb{N}$ variables, that is

$$e_\ell(X_1,\dots,X_L)=\sum_{\substack{T\subset[L]\\|T|=\ell}}\prod_{j\in T}X_j. \tag{3.1}$$

Finally, we follow Ben-Israel, 1992 and denote spanned volumes by

$$\mathrm{Vol}_q(\boldsymbol{A})=\sqrt{e_q\left(\sigma_1(\boldsymbol{A})^2,\dots,\sigma_{\mathrm{rk}(A)}(\boldsymbol{A})^2\right)},\quad q=1,\dots,\mathrm{rk}(\boldsymbol{A}).$$

Throughout the paper, $\boldsymbol{X}$ will always denote an $N \times d$ matrix that we think of as the original data matrix, of which we want to select $k \leq d$ columns. We do not make any assumption on how N compares to d. Unless otherwise specified, r is the rank of $\boldsymbol{X}$, and matrices $\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}$ are reserved for the SVD of $\boldsymbol{X}$, that is,

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal \tag{3.2}$$

$$= \left[\begin{array}{c|c} \boldsymbol{U}_k & \boldsymbol{U}_{k^\perp} \end{array}\right] \left[\begin{array}{c|c} \boldsymbol{\Sigma}_k & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{\Sigma}_{k^\perp} \end{array}\right] \left[\begin{array}{c} \boldsymbol{V}_k^\intercal \\ \hline \boldsymbol{V}_{k^\perp}^\intercal \end{array}\right], \tag{3.3}$$

where $\boldsymbol{U} \in \mathbb{R}^{N \times d}$ and $\boldsymbol{V} \in \mathbb{R}^{d \times d}$ are orthogonal, and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is diagonal. The diagonal entries of $\boldsymbol{\Sigma}$ are $\sigma_i = \sigma_i(\boldsymbol{X})$, $i = 1, \ldots, r$, and we still assume they are in decreasing order. We will also need the blocks given in (3.3), where we separate blocks of size k corresponding to the largest k singular values. To simplify notation, we abusively write $\boldsymbol{U}_k$ for $\boldsymbol{U}_{:,[k]}$ and $\boldsymbol{V}_k$ for $\boldsymbol{V}_{:,[k]}$ in (3.3), among others. Though they will be introduced and discussed at length in Section 3.2, we also recall here that we denote by $\ell_i^k = \|\boldsymbol{V}_{i,[k]}\|_2^2$ the so-called *k-leverage score* of the i-th column of $\boldsymbol{X}$.

We need some notation for the selection of columns. Let $S \subset [d]$ be such that $|S| = k$, and let $\boldsymbol{S} \in \{0,1\}^{d \times k}$ be the corresponding sampling matrix: $\boldsymbol{S}$ is defined by $\forall \boldsymbol{M} \in \mathbb{R}^{N \times d}, \boldsymbol{M}\boldsymbol{S} = \boldsymbol{M}_{:,S}$. In the context of column selection, it is often referred to $\boldsymbol{X}\boldsymbol{S} = \boldsymbol{X}_{:,S}$ as $\boldsymbol{C}$. We set for convenience $\boldsymbol{Y}_{:,S}^\intercal = (\boldsymbol{Y}_{:,S})^\intercal$.

The result of column subset selection will usually be compared to the result of PCA. We denote by $\boldsymbol{\Pi}_k \boldsymbol{X}$ the best rank-k approximation to $\boldsymbol{X}$. The sense of *best* can be understood either in Frobenius or spectral norm, as both give the same result. On the other side, for a given subset $S \subset [d]$ of size $|S| = s$ and $\nu \in \{2, \mathrm{Fr}\}$, let

$$\boldsymbol{\Pi}_{S,k}^\nu \boldsymbol{X} = \arg\min_A \|\boldsymbol{X} - A\|_\nu$$

where the minimum is taken over all matrices $\boldsymbol{A} = \boldsymbol{X}_{:,S}\boldsymbol{B}$ such that $\boldsymbol{B} \in \mathbb{R}^{s \times d}$ and $\operatorname{rk} \boldsymbol{B} \leq k$; in words, the minimum is taken over matrices of rank at most k that lie in the column space of $\boldsymbol{C} = \boldsymbol{X}_{:,S}$. When $|S| = k$, we simply write $\boldsymbol{\Pi}_S^\nu \boldsymbol{X} = \boldsymbol{\Pi}_{S,k}^\nu \boldsymbol{X}$. In practice, the Frobenius projection can be computed as $\boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X} = \boldsymbol{C}\boldsymbol{C}^+ \boldsymbol{X}$, where $\boldsymbol{C}^+$ is the Moore-Penrose pseudo inverse of $\boldsymbol{C}$, yet there is no simple expression for $\boldsymbol{\Pi}_S^2 \boldsymbol{X}$. However, $\boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}$ can be used as a proxy for $\boldsymbol{\Pi}_S^2 \boldsymbol{X}$ since

$$\|\boldsymbol{X} - \boldsymbol{\Pi}_S^2 \boldsymbol{X}\|_2 \leq \|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_2 \leq \sqrt{2}\|\boldsymbol{X} - \boldsymbol{\Pi}_S^2 \boldsymbol{X}\|_2, \tag{3.4}$$

see Boutsidis et al., 2011, Lemma 2.3. Finally, define

$$\boldsymbol{\Pi}_k \boldsymbol{X} = \arg\min_{\operatorname{rk} \boldsymbol{A} \leq k} \|\boldsymbol{X} - A\|_{\mathrm{Fr}} = \arg\min_{\operatorname{rk} \boldsymbol{A} \leq k} \|\boldsymbol{X} - A\|_2.$$

3.2 RELATED WORK

In this section, we survey existing work about column subset selection.

Rank-revealing QR decompositions

The first k-CSSP algorithm can be traced back to the article of Golub, 1965 on pivoted QR factorization. This work introduced the concept of rank-revealing QR factorization

Algorithm	$p_2(k,d)$	Complexity	References
Pivoted QR	$2^k\sqrt{d-k}$	$\mathcal{O}(dNk)$	(Golub and Van Loan, 1996)
Strong RRQR (Alg. 3)	$\sqrt{(d-k)k+1}$	not polynomial	(Gu and Eisenstat, 1996)
Strong RRQR (Alg. 4)	$\sqrt{f^2(d-k)k+1}$	$\mathcal{O}(dNk\log_f(d))$	(Gu and Eisenstat, 1996)

Table 3.1 – Examples of some RRQR algorithms and their theoretical performances.

(RRQR). The original motivation was to calculate a well-conditioned QR factorization of a matrix $\boldsymbol{X}$ that reveals its numerical rank.

Definition 3.1. *Let $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ and $k \in \mathbb{N}$ $(k \leq d)$. A RRQR factorization of $\boldsymbol{X}$ is a 3-tuple $(\boldsymbol{\Pi}, \boldsymbol{Q}, \boldsymbol{R})$ with $\boldsymbol{\Pi} \in \mathbb{R}^{d\times d}$ a permutation matrix, $\boldsymbol{Q} \in \mathbb{R}^{N\times d}$ an orthogonal matrix, and $\boldsymbol{R} \in \mathbb{R}^{d\times d}$ a triangular matrix, such that $\boldsymbol{X\Pi} = \boldsymbol{QR}$, and*

$$\frac{\sigma_k(\boldsymbol{X})}{p_1(k,d)} \leq \sigma_k(\boldsymbol{R}_{[k],[k]}) \leq \sigma_k(\boldsymbol{X}), \tag{3.5}$$

and

$$\sigma_{k+1}(\boldsymbol{X}) \leq \sigma_1(\boldsymbol{R}_{[k+1:d],[k+1:d]}) \leq p_2(k,d)\sigma_{k+1}(\boldsymbol{X}), \tag{3.6}$$

where $p_1(k,d)$ and $p_2(k,d)$ are controlled.

In practice, a RRQR factorization algorithm interchanges pairs of columns and updates or builds a QR decomposition on the fly. The link between RRQR factorization and k-CSSP was first discussed by Boutsidis, Mahoney, and Drineas, 2009. The structure of a RRQR factorization indeed gives a deterministic selection of a subset of k columns of $\boldsymbol{X}$. More precisely, if we take $\boldsymbol{C}$ to be the first k columns of $\boldsymbol{X\Pi}$, $\boldsymbol{C}$ is a subset of columns of $\boldsymbol{X}$ and $\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_2 = \|\boldsymbol{R}_{[k+1:r],[k+1:r]}\|_2$. By (3.6), any RRQR algorithm thus provides provable guarantees in spectral norm for k-CSSP.

Following (Golub, 1965), many articles gave algorithms that improved on $p_1(k,d)$ and $p_2(k,d)$ in Definition 3.1. Table 3.1 sums up the guarantees of the original algorithm of Golub, 1965 and the state-of-the-art algorithms of Gu and Eisenstat, 1996. Note the dependency of $p_2(k,d)$ on the dimension d through the term $\sqrt{d-k}$; this term is common for guarantees in spectral norm for k-CSSP. We refer to (Boutsidis et al., 2009) for an exhaustive survey on RRQR factorization. A RRQR factorization gives an example of a deterministic column subset selection with a spectral guarantee. We present in Section 3.2 a randomized improvement over strong RRQR, called *double phase*. As we shall see, randomized algorithms can match the bound in the bottom row of Table 3.1 and provide guarantees in Frobenius norm as well.

Length square importance sampling and additive bounds

Drineas, Frieze, Kannan, Vempala, and Vinay, 2004 proposed a randomized CSS algorithm based on independently sampling s indices $S = \{i_1, \ldots, i_s\}$ from a multinomial distribution of parameter $\boldsymbol{p}$, where

$$p_j = \frac{\|\boldsymbol{X}_{:,j}\|_2^2}{\|\boldsymbol{X}\|_{\mathrm{Fr}}^2}, j \in [d]. \tag{3.7}$$

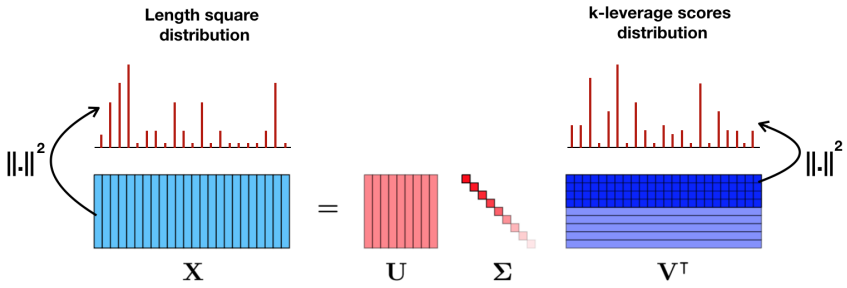


Figure 3.1 – An illustration of the difference between the length squares distributions and the k-leverage scores distribution.

The rationale is that columns with large norms should be kept. Let $\boldsymbol{C} = \boldsymbol{X}_{:,S}$ be the corresponding submatrix. First, we note that some columns of $\boldsymbol{X}$ may appear more than once in $\boldsymbol{C}$. Second, Drineas et al., 2004, Theorem 3 states that

$$\mathbb{P}\left(\|\boldsymbol{X} - \boldsymbol{\Pi}_{S,k}^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq \|\boldsymbol{X} - \boldsymbol{\Pi}_k\boldsymbol{X}\|_{\mathrm{Fr}}^2 + 2\left(1 + \sqrt{8\log\left(\frac{2}{\delta}\right)}\right)\sqrt{\frac{k}{s}}\|\boldsymbol{X}\|_{\mathrm{Fr}}^2\right) \geq 1 - \delta. \tag{3.8}$$

Equation (3.8) is a high-probability, *additive* upper bound for $\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2$. The drawback of such bounds is that they can be very loose if the first k singular values of $\boldsymbol{X}$ are large compared to σ_{k+1}. For this reason, multiplicative approximation bounds have been investigated, using a different distribution that takes into account the geometry of the dataset.

k-leverage scores sampling and multiplicative bounds

Drineas, Mahoney, and Muthukrishnan, 2007 proposed an algorithm with a provable multiplicative upper bound using multinomial sampling, but this time according to k-leverage scores.

Definition 3.2 (k-leverage scores)**.** *Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal \in \mathbb{R}^{N\times d}$ be the SVD of $\boldsymbol{X}$. We denote by $\boldsymbol{V}_k = \boldsymbol{V}_{:,[k]}$ the first k columns of $\boldsymbol{V}$. For $i \in [d]$, the k-leverage score of the i-th column of $\boldsymbol{X}$ is defined by*

$$\ell_i^k = \sum_{j=1}^{k} V_{i,j}^2. \tag{3.9}$$

Intuitively, a large value of ℓ_i^k in (3.9) indicates that the i-th canonical vector is close to the space spanned by the first k eigenvectors. We shall make this intuition more precise in Section 3.3.1. See Figure 3.1 for a graphical depiction of the difference between the length square distribution and the k-leverage scores distribution. For now, we note that

$$\sum_{i\in[d]} \ell_i^k = \sum_{i\in[d]} \|(\boldsymbol{V}_k^\intercal)_{:,i}\|_2^2 = \mathrm{Tr}(\boldsymbol{V}_k\boldsymbol{V}_k^\intercal) = k, \tag{3.10}$$

since $\boldsymbol{V}_k$ is an orthogonal matrix. Therefore, one can consider the multinomial distribution on $[d]$ with parameters

$$p_i = \frac{\ell_i^k}{k}\ , i \in [d]. \tag{3.11}$$

This multinomial is called the *k-leverage scores distribution*.

Theorem 3.1 (Drineas et al., 2007, Theorem 3). *If the number s of sampled columns satisfies*

$$s \geq \frac{3200k^2}{\epsilon^2}, \tag{3.12}$$

then, under i.i.d. sampling from the k-leverage scores distribution,

$$\mathbb{P}\left(\|\boldsymbol{X} - \boldsymbol{\Pi}_{S,k}^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq (1+\epsilon)\|\boldsymbol{X} - \boldsymbol{\Pi}_k\boldsymbol{X}\|_{\mathrm{Fr}}^2 \right) \geq 0.7. \tag{3.13}$$

Drineas et al., 2007 also considered replacing multinomial with Bernoulli sampling, still using the k-leverage scores. The expected number of columns needed for (3.13) to hold is then lowered to $\mathcal{O}(\frac{k\log k}{\epsilon^2})$. A natural question is then to understand how low the number of columns can be, while still guaranteeing a multiplicative bound like (3.13). A partial answer has been given by Deshpande and Vempala, 2006.

Proposition 3.1 (Deshpande and Vempala, 2006, Proposition 4). *Given $\epsilon > 0$, $k, d \in \mathbb{N}$ such that $d\epsilon \geq 2k$, there exists a matrix $\boldsymbol{X}^\epsilon \in \mathbb{R}^{kd \times k(d+1)}$ such that for any $S \subset [d]$,*

$$\|\boldsymbol{X}^\epsilon - \boldsymbol{\Pi}_{S,k}^{\mathrm{Fr}}\boldsymbol{X}^\epsilon\|_{\mathrm{Fr}}^2 \geq (1+\epsilon)\|\boldsymbol{X}^\epsilon - \boldsymbol{X}_k^\epsilon\|_{\mathrm{Fr}}^2. \tag{3.14}$$

This suggests that a lower bound for the number of columns is $2k/\epsilon$, at least in the worst case sense of Proposition 3.1. Interestingly, the k-leverage scores distribution of the matrix $\boldsymbol{X}^\epsilon$ in the proof of Proposition 3.1 is uniform, so that k-leverage score sampling boils down to uniform sampling.

Finally, a deterministic algorithm based on k-leverage score sampling was proposed by Papailiopoulos, Kyrillidis, and Boutsidis, 2014. The algorithm selects the $c(\theta)$ columns of $\boldsymbol{X}$ with the largest k-leverage scores, where

$$c(\theta) \in \arg\min_u \left(\sum_{i=1}^u \ell_i^k > \theta \right), \tag{3.15}$$

and θ is a free parameter that controls the approximation error. To guarantee that there exists a matrix of rank k in the subspace spanned by the selected columns, Papailiopoulos et al., 2014 assume that

$$0 \leq k - \theta < 1. \tag{3.16}$$

Loosely speaking, this condition is satisfied for a low value of $c(\theta)$ if the k-leverage scores (after ordering) are decreasing rapidly enough. The authors give empirical evidence that this condition is satisfied by a large proportion of real datasets.

Theorem 3.2 (Papailiopoulos et al., 2014, Theorem 2). *Let $\epsilon = k - \theta \in [0, 1)$, letting S index the columns with the $c(\theta)$ largest k-leverage scores,*

$$\|\boldsymbol{X} - \boldsymbol{\Pi}_{S,k}^{\nu}\boldsymbol{X}\|_\nu \leq \frac{1}{1-\epsilon}\|\boldsymbol{X} - \boldsymbol{\Pi}_k\boldsymbol{X}\|_\nu, \quad \nu \in \{2, \mathrm{Fr}\}. \tag{3.17}$$

In particular, if $\epsilon \in [0, \frac{1}{2}]$,

$$\|\boldsymbol{X} - \boldsymbol{\Pi}_{S,k}^{\nu}\boldsymbol{X}\|_\nu \leq (1+2\epsilon)\|\boldsymbol{X} - \boldsymbol{\Pi}_k\boldsymbol{X}\|_\nu, \quad \nu \in \{2, \mathrm{Fr}\}. \tag{3.18}$$

Furthermore, they proved that if the k-leverage scores decay like a power law, the number of columns needed to obtain a multiplicative bound can actually be smaller than k/ϵ.

Theorem 3.3 (Papailiopoulos et al., 2014, Theorem 3). *Assume, for $\eta > 0$,*

$$\ell_i^k = \frac{\ell_1^k}{i^{\eta+1}}. \tag{3.19}$$

Let $\epsilon = k - \theta \in [0, 1)$, then

$$c(\theta) = \max\left\{ \left(\frac{4k}{\epsilon}\right)^{\frac{1}{\eta+1}} - 1, \left(\frac{4k}{\eta\epsilon}\right)^{\frac{1}{\eta}}, k \right\}. \tag{3.20}$$

This complements the fact that the worst case example in Proposition 3.1 had uniform k-leverage scores. Loosely speaking, matrices with fast decaying k-leverage scores can be efficiently subsampled.

Negative correlation: volume sampling and the double phase algorithm

In this section, we survey algorithms that randomly sample exactly k columns from $\boldsymbol{X}$, further requiring the columns to be somehow negatively correlated to avoid redundancy. This is to be compared to the multinomial sampling schemes of Sections 3.2 and 3.2, which ignore the joint structure of $\boldsymbol{X}$ and typically require more than k columns.

Deshpande, Rademacher, Vempala, and Wang, 2006 obtained a multiplicative bound on the expected approximation error, with only k columns, using the so-called *volume sampling*.

Theorem 3.4 (Deshpande et al., 2006). *Let S be a random subset of $[d]$, chosen with probability*

$$\mathbb{P}_{\mathrm{VS}}(S) = Z^{-1} \operatorname{Det}(\boldsymbol{X}_{:,S}^{\mathsf{T}} \boldsymbol{X}_{:,S})\, \mathbb{1}_{\{|S|=k\}}, \tag{3.21}$$

where $Z = \sum_{|S|=k} \operatorname{Det}(\boldsymbol{X}_{:,S}^{\mathsf{T}} \boldsymbol{X}_{:,S})$. Then

$$\mathbb{E}_{\mathrm{VS}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq (k+1) \|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_{\mathrm{Fr}}^2 \tag{3.22}$$

and

$$\mathbb{E}_{\mathrm{VS}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^2 \boldsymbol{X}\|_2^2 \leq (d-k)(k+1) \|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_2^2. \tag{3.23}$$

Note that the bound for the spectral norm was proven in Deshpande et al., 2006 for the Frobenius projection, that is, they bound $\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_2$. The bound (3.23) easily follows from (3.4). A more precise description of the approximation error under volume sampling was given in Guruswami and Sinop, 2012.

Theorem 3.5 (Theorem 3.1, Guruswami and Sinop, 2012). *Let $\boldsymbol{X} \in \mathbb{R}^{N \times d}$, and let $\sigma \in \mathbb{R}^d$ be the vector containing the square of the singular values of $\boldsymbol{X}$. The function*

$$\sigma \mapsto \mathbb{E}_{\mathrm{VS}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S \boldsymbol{X}\|_{\mathrm{Fr}}^2 = (k+1) \frac{e_k(\sigma)}{e_{k-1}(\sigma)} \tag{3.24}$$

is Schur-concave.

In other words, the expected approximation error under the distribution of volume sampling for the Frobenius norm is low for flat spectrum and it is large otherwise; see Appendix 3.B for the formal definition of Schur-concavity.

Later, sampling according to (3.21) was shown to be doable in polynomial time (Deshpande and Rademacher, 2010). Using a worst case example, Deshpande et al., 2006 proved that the $k+1$ factor in (3.22) cannot be improved.

Proposition 3.2 (Deshpande et al., 2006)**.** *Let $\epsilon > 0$. There exists a $(k+1) \times (k+1)$ matrix $\boldsymbol{X}^\epsilon$ such that for every subset S of k columns of $\boldsymbol{X}^\epsilon$,*

$$\|\boldsymbol{X}^\epsilon - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}^\epsilon\|_{\mathrm{Fr}}^2 > (1-\epsilon)(k+1)\|\boldsymbol{X}^\epsilon - \boldsymbol{\Pi}_k \boldsymbol{X}^\epsilon\|_{\mathrm{Fr}}^2. \tag{3.25}$$

We note that there has been recent interest in a similar but different distribution called *dual volume sampling* (Li, Jegelka, and Sra, 2017a; Derezinski and Warmuth, 2018), sometimes also confusingly termed *volume sampling*. The main application of dual VS is row subset selection of a matrix $\boldsymbol{X}$ for linear regression on label budget constraints.

Boutsidis et al., 2009 proposed a k-CSSP algorithm, called *double phase*, that combines ideas from multinomial sampling and RRQR factorization. The motivating idea is that the theoretical performance of RRQR factorizations depends on the dimension through a factor $\sqrt{d-k}$; see Table 3.1. To improve on that, the authors propose to first reduce the dimension d to c by preselecting a large number of columns $c > k$ using multinomial sampling from the k-leverage scores distribution, as in Section 3.2. Then only, they perform a RRQR factorization of the reduced matrix $\boldsymbol{V}_k^\intercal \boldsymbol{S}_1 \boldsymbol{D}_1 \in \mathbb{R}^{k \times c}$, where $\boldsymbol{S}_1 \in \mathbb{R}^{d \times c}$ is the sampling matrix of the multinomial phase and $\boldsymbol{D}_1 \in \mathbb{R}^{c \times c}$ is a scaling matrix.

Theorem 3.6 (Boutsidis et al., 2009)**.** *Let S be the output of the double phase algorithm with $c = 1600 c_0^2 k \log(800 c_0^2 k)$. Then*

$$\mathbb{P}_{\mathrm{DPh}}\left(\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}} \leq (1 + 8\sqrt{2k(c-k)+1})\|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_{\mathrm{Fr}} \right) \geq 0.8\,, \tag{3.26}$$

and

$$\mathbb{P}_{\mathrm{DPh}}\left(\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{2} \boldsymbol{X}\|_{2} \leq \big(1 + 2\sqrt{2k(c-k)+1}\big)\|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_{2} + \frac{8\sqrt{2k(c-k)+1}}{c^{1/4}}\|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_{\mathrm{Fr}} \right) \geq 0.8\,. \tag{3.27}$$

We note that c_0 is an unknown constant from (Rudelson and Vershynin, 2007). Although not explicitly stated by Boutsidis et al., 2009, the spectral bound (3.27) easily follows from their result using (3.4). We also note that to obtain their spectral bound, Boutsidis et al., 2009 use a slight modification of the leverage scores in the random phase.

Excess risk in sketched linear regression

So far, we have focused on approximation bounds in spectral or Frobenius norm for the residual $\boldsymbol{X} - \boldsymbol{\Pi}_{S,k}^{\nu} \boldsymbol{X}$. This is a reasonable generic measure of error as long as it is not

known what the practitioner wants to do with the submatrix $\boldsymbol{X}_{:,S}$. In this section, we assume that the ultimate goal is to perform linear regression of some $\mathbf{y} \in \mathbb{R}^N$ onto $\boldsymbol{X}$.

Other measures of performance then become of interest, such as the excess risk incurred by regressing onto $\boldsymbol{X}_{:,S}$ rather than $\boldsymbol{X}$. We use here the framework of Slawski, 2018, further assuming well-specification for simplicity. For every $i \in [N]$, assume $y_i = \boldsymbol{X}_{i,:}\boldsymbol{w}^* + \xi_i$, where the noises ξ_i are i.i.d. real variables with mean 0 and variance v. For a given estimator $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{X}, \boldsymbol{y})$, the excess risk is defined as

$$\mathcal{E}(\boldsymbol{w}) = \mathbb{E}_{\xi}\left[\frac{\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{X}\boldsymbol{w}\|_2^2}{N}\right]. \tag{3.28}$$

In particular, it is easy to show that the ordinary least squares (OLS) estimator $\hat{\boldsymbol{w}} = \boldsymbol{X}^+\boldsymbol{y}$ has excess risk

$$\mathcal{E}(\hat{\boldsymbol{w}}) = v \times \frac{\mathrm{rk}(\boldsymbol{X})}{N}. \tag{3.29}$$

Selecting k columns indexed by S in $\boldsymbol{X}$ prior to performing linear regression yields $\boldsymbol{w}_S = (\boldsymbol{X}\boldsymbol{S})^+\boldsymbol{y} \in \mathbb{R}^k$. We are interested in the excess risk of the corresponding sparse vector

$$\hat{\boldsymbol{w}}_S := \boldsymbol{S}\boldsymbol{w}_S = \boldsymbol{S}(\boldsymbol{X}\boldsymbol{S})^+\boldsymbol{y} \in \mathbb{R}^d$$

which has all coordinates zero, except those indexed by S.

Proposition 3.3 (Theorem 9, Mor-Yosef and Avron, 2019)**.** *Let $S \subset [d]$, such that $|S| = k$. Let $(\theta_i(S))_{i\in[k]}$ be the principal angles between* Span $\boldsymbol{S}$ *and* Span $\boldsymbol{V}_k$*, see Appendix 3.A. Then*

$$\mathcal{E}(\hat{\boldsymbol{w}}_S) \leq \frac{1}{N}\left(1 + \max_{i\in[k]} \tan^2\theta_i(S)\right)\|\boldsymbol{w}^*\|^2\sigma_{k+1}^2 + \frac{vk}{N}. \tag{3.30}$$

Compared to the excess risk (3.29) of the OLS estimator, the second term of the right-hand side of (3.30) replaces $\mathrm{rk}\boldsymbol{X}$ by k. But the price is the first term of the right-hand side of (3.30), which we loosely term *bias*. To interpret this bias term, we first look at the excess risk of the principal component regressor (PCR)

$$\boldsymbol{w}_k^* \in \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathrm{Span}\,\boldsymbol{V}_k} \mathbb{E}_{\xi}\left[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2/N\right]. \tag{3.31}$$

Proposition 3.4 (Corollary 11, Mor-Yosef and Avron, 2019)**.**

$$\mathcal{E}(\boldsymbol{w}_k^*) \leq \frac{\|\boldsymbol{w}^*\|^2\sigma_{k+1}^2}{N} + \frac{vk}{N}. \tag{3.32}$$

The right-hand side of (3.32) is almost that of (3.30), except that the bias term in the CSS risk (3.30) is larger by a factor that measures how well the subspace spanned by S is aligned with the principal eigenspace $\boldsymbol{V}_k$. This makes intuitive sense: the performance of CSS will match PCR if selecting columns yields almost the same eigenspace.

The excess risk (3.30) is yet another motivation to investigate DPPs for column subset selection. We shall see in Section 3.4.2 that the expectation of (3.30) under a well-chosen DPP for S has a particularly simple bias term.

Finally, as mentioned in Section 3.2, probability distributions similar to volume sampling but for *row* subset selection were investigated in the context of regression (Derezinski and Warmuth, 2017; Derezinski et al., 2018), under the name of *dual volume*

sampling[1]. Selecting rows in linear regression is akin to experimental design, and applies to cases where all features are to be used, but only a few labels can be observed due to budget constraints. We emphasize that the two problems are related, but they are not simple transpositions of each other. In particular, the excess risk for the regularized dual volume sampling of Derezinski and Warmuth, 2018 scales as $\mathcal{O}(1/k)$ using all d features and k observations, while the excess risk in the results of Section 3.4.2 rather scales as $\mathcal{O}(1/N)$ using k features and N observations.

3.3 THE PROPOSED ALGORITHM

In this section, we introduce our sub-sampling algorithm based on a projection DPP. We refer to Chapter 2 for the definition of DPPs and k-DPPs. In particular, recall that volume sampling, as defined in Section 3.2, is an example of a k-DPP. Its kernel is the covariance matrix of the data $\boldsymbol{L} = \boldsymbol{X}^\intercal \boldsymbol{X}$. In general, $\boldsymbol{L}$ is not an orthogonal projection matrix, so that volume sampling is not a DPP. In particular, draws from volume sampling have fixed cardinality, and thus cannot be written as a sum of non trivial Bernoulli random variables. However, following (2.65) in Proposition 2.2, volume sampling can be seen as a mixture of projection DPPs indexed by $T \subseteq [d], |T| = k$, with marginal kernels $\boldsymbol{K}_T = \boldsymbol{V}_{:,T}\boldsymbol{V}_{:,T}^\intercal$ and mixture weights $\mu_T \propto \prod_{i \in T} \sigma_i^2$. The component with the highest weight thus corresponds to the k largest singular values, that is, the projection DPP with marginal kernel

$$\boldsymbol{K} := \boldsymbol{V}_k \boldsymbol{V}_k^\intercal. \tag{3.33}$$

This chapter is about column subset selection using precisely this DPP. Alternately, we could motivate the study of this DPP by remarking that its marginals $\mathbb{P}(i \subseteq Y)$ are the k-leverage scores introduced in Section 3.2. Since $\boldsymbol{K}$ is symmetric, this DPP can be seen as a repulsive generalization of leverage score sampling.

3.3.1 *The geometric intuition*

The k-leverage scores can be given a geometric interpretation, the generalization of which serves as a first motivation for our work.

For $i \in [d]$, let $\boldsymbol{e}_i$ be the i-th canonical basis vector of $\mathbb{R}^d$. Let further θ_i be the angle between $\boldsymbol{e}_i$ and the subspace $\mathcal{P}_k = \mathrm{Span}(\boldsymbol{V}_k)$, and denote by $\boldsymbol{\Pi}_{\mathcal{P}_k}\boldsymbol{e}_i$ the orthogonal projection of $\boldsymbol{e}_i$ onto the subspace $\mathcal{P}_k$. Then, by the fact that

$$\langle \boldsymbol{e}_i, \boldsymbol{\Pi}_{\mathcal{P}_k}\boldsymbol{e}_i \rangle = \langle \boldsymbol{\Pi}_{\mathcal{P}_k}\boldsymbol{e}_i, \boldsymbol{\Pi}_{\mathcal{P}_k}\boldsymbol{e}_i \rangle = \|\boldsymbol{\Pi}_{\mathcal{P}_k}\boldsymbol{e}_i\|^2, \tag{3.34}$$

we have

$$\cos^2(\theta_i) := \frac{\langle \boldsymbol{e}_i, \boldsymbol{\Pi}_{\mathcal{P}_k}\boldsymbol{e}_i \rangle^2}{\|\boldsymbol{e}_i\|^2 \|\boldsymbol{\Pi}_{\mathcal{P}_k}\boldsymbol{e}_i\|^2} = \langle \boldsymbol{e}_i, \boldsymbol{\Pi}_{\mathcal{P}_k}\boldsymbol{e}_i \rangle = \langle \boldsymbol{e}_i, \sum_{j=1}^k V_{i,j}\boldsymbol{V}_{:,j} \rangle = \sum_{j=1}^k V_{i,j}^2 = \ell_i^k. \tag{3.35}$$

A large k-leverage score ℓ_i^k thus indicates that $\boldsymbol{e}_i$ is almost aligned with $\mathcal{P}_k$. Selecting columns with large k-leverage scores as in (Drineas et al., 2007) can thus be interpreted

[1] Derezinski and Warmuth, 2017 actually talk of *volume sampling*. To avoid confusion, we rather stick to volume sampling describing the column subset selection algorithm in (Deshpande et al., 2006) and discussed in Section 3.2.

as replacing the principal eigenspace $\mathcal{P}_k$ by a subspace that must contain k of the original coordinate axes. Intuitively, a closer subspace to the original $\mathcal{P}_k$ would be obtained by selecting columns *jointly* rather than independently, considering the angle with $\mathcal{P}_k$ of the subspace spanned by these columns. More precisely, consider $S \subset [d], |S| = k$, and denote $\mathcal{P}_S = \mathrm{Span}(\boldsymbol{e}_j, j \in S)$. A natural definition of the cosine between $\mathcal{P}_k$ and $\mathcal{P}_S$ is in term of the so-called *principal angles* $(\theta_i(\mathcal{P}_S, \mathcal{P}_k))_{i\in[k]}$ that define the relative position of the two subspaces (Golub and Van Loan, 1996, Section 6.4.4); see Figure 3.2 for an illustration and Appendix 3.A for the rigorous definition. In particular, Proposition 3.8 in Appendix 3.A yields

$$\cos^2(\mathcal{P}_k, \mathcal{P}_S) := \prod_{i\in[k]} \cos^2 \theta_i(\mathcal{P}_k, \mathcal{P}_S) = \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2. \tag{3.36}$$

This chapter is precisely about sampling k columns proportionally to (3.36).

In Appendix 3.C, we contribute a different interpretation of k-leverage scores, which relates them to the length-square distribution of Section 3.2.

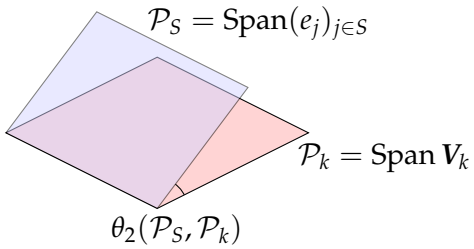


Figure 3.2 – An illustration of the largest principal angle $\theta_2(\mathcal{P}_S, \mathcal{P}_k)$ in the case $d = 3$ and $k = 2$.

3.3.2 *Numerical simulation of the projection DPP and volume sampling*

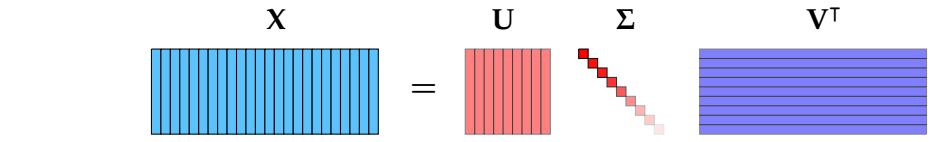
The difference between volume sampling and the DPP with kernel $\boldsymbol{K}$ is also manifested by the sampling procedure. Indeed, as we have mentioned, volume sampling is a mixture of projection DPPs corresponding to the kernels $\boldsymbol{K}_T$. Therefore, given the matrix $\boldsymbol{V}$, volume sampling requires to sample the set of eigenvectors T; while the projection DPP of kernel $\boldsymbol{K}$ does not need this intermediate step. The difference is depicted in Figure 3.3.

As we have seen in Chapter 2, sampling from a projection DPP is possible through the HKPV algorithm. We refer to (Tremblay, Barthelmé, and Amblard, 2018), (Launay, Galerne, and Desolneux, 2020b) and the documentation of the DPPy toolbox[2] (Gautier, Bardenet, and Valko, 2019) for other efficient variants of this algorithm.

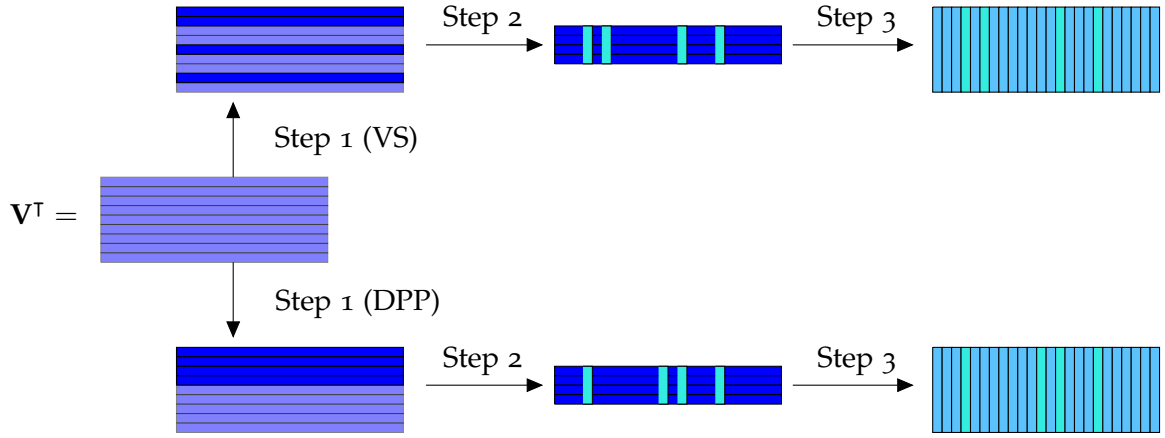
3.4 MAIN RESULTS

In this section, we prove bounds for $\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\nu} \boldsymbol{X}\|_{\nu}^2$ under the projection DPP of marginal kernel $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^{\mathsf{T}}$ presented in Section 3.3. Throughout, we compare our

[2] http://github.com/guilgautier/DPPy



(a) SVD of $\boldsymbol{X}$



(b) Sampling k columns according to VS and our DPP

Figure 3.3 – A graphical depiction of the sampling algorithms for volume sampling (VS) and the DPP with marginal kernel $\boldsymbol{V}_k\boldsymbol{V}_k^\intercal$. (a) Both algorithms start with an SVD. (b) In Step 1, VS randomly selects k rows of $\boldsymbol{V}^\intercal$, while our DPP always picks the first k rows. Step 2 is the same for both algorithms: jointly sample k columns of the subsampled $\boldsymbol{V}^\intercal$, proportionally to their squared volume. Finally, Step 3 is simply the extraction of the corresponding columns of $\boldsymbol{X}$.

bounds to the state-of-the-art bounds of volume sampling obtained by Deshpande et al., 2006; see Theorem 3.4 and Section 3.2. For clarity, we defer the proofs of our results from this section to Section 3.7.

3.4.1 *Multiplicative bounds in spectral and Frobenius norm*

Let S be a random subset of k columns of $\boldsymbol{X}$ chosen with probability:

$$\mathbb{P}_{\mathrm{DPP}}(S) = \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2. \tag{3.37}$$

First, without any further assumption, we have the following result.

Proposition 3.5. *Under the projection DPP of marginal kernel $\boldsymbol{V}_k\boldsymbol{V}_k^\intercal$, it holds that*

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^\nu \boldsymbol{X}\|_\nu^2 \leq k(d+1-k)\|\boldsymbol{X} - \boldsymbol{\Pi}_k\boldsymbol{X}\|_\nu^2, \quad \nu \in \{2, \mathrm{Fr}\}. \tag{3.38}$$

For the spectral norm, the bound is practically the same as that of volume sampling (3.23). However, our bound for the Frobenius norm is worse than (3.22) by a factor $(d-k)$. In the rest of this section, we sharpen our bounds by taking into account the sparsity level of the k-leverage scores and the decay of singular values.

In terms of sparsity, we first replace the dimension d in (3.38) by the number $p \in [d]$ of non zero k-leverage scores

$$p = \left|\{i \in [d], \boldsymbol{V}_{i,[k]} \neq \boldsymbol{0}\}\right|. \tag{3.39}$$

To quantify the decay of the singular values, we define the flatness parameter

$$\beta = \sigma_{k+1}^2 \left(\frac{1}{d-k} \sum_{j \geq k+1} \sigma_j^2 \right)^{-1} . \tag{3.40}$$

In words, $\beta \in [1, d-k]$ measures the flatness of the spectrum of $\boldsymbol{X}$ below the cut-off at $k+1$. Indeed, (3.40) is the ratio of the largest term in a mean to that mean. The closer β is to 1, the more similar the terms in the sum in the denominator of (3.40) to their maximum value σ_{k+1}^2. At the extreme, $\beta = d-k$ when $\sigma_{k+1}^2 > 0$ while $\sigma_j^2 = 0$, $\forall j \geq k+2$. Finally, we also note that β is $(d-k)$ times the inverse of the numerical rank (Rudelson and Vershynin, 2007) of the residual matrix $\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}$.

Proposition 3.6. *Under the projection DPP of marginal kernel $\boldsymbol{V}_k \boldsymbol{V}_k^\mathsf{T}$, it holds that*

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^2 \boldsymbol{X}\|_2^2 \leq (1 + k(p-k)) \|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_2^2 \tag{3.41}$$

and

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq \left(1 + \beta \frac{p-k}{d-k} k \right) \|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_{\mathrm{Fr}}^2 . \tag{3.42}$$

The bound in (3.41) compares favourably with volume sampling (3.23) since the dimension d has been replaced by the sparsity level p. For β close to 1, the bound in (3.42) is better than the bound (3.22) of volume sampling since $(p-k)/(d-k) \leq 1$. Again, the sparser the k-leverage scores, the smaller the bounds. Finally, if needed, bounds in high probability easily follow from Proposition 3.6 using Markov's inequality.

Now, one could argue that, in practice, sparsity is never exact: it can well be that $p = d$ while there still are a lot of small k-leverage scores. We will demonstrate in Section 3.5 that the DPP still performs better than volume sampling in this setting, which Proposition 3.6 doesn't reflect. We introduce two ideas to further tighten the bounds of Proposition 3.6. First, we define an effective sparsity level in the vein of (Papailiopoulos et al., 2014), see Section 3.2. Second, we condition the DPP on a favourable event with controlled probability.

Theorem 3.7. *Let π be a permutation of $[d]$ such that leverage scores are reordered*

$$\ell_{\pi_1}^k \geq \ell_{\pi_2}^k \geq \ldots \geq \ell_{\pi_d}^k . \tag{3.43}$$

For $\delta \in [d]$, let $T_\delta = [\pi_\delta, \ldots, \pi_d]$. Let $\theta \geq 1$ and

$$p_{\mathrm{eff}}(\theta) = \min \left\{ q \in [d] \;\middle|\; \sum_{i \leq q} \ell_{\pi_i}^k \geq k - 1 + \frac{1}{\theta} \right\} . \tag{3.44}$$

Finally, let $\mathcal{A}_\theta$ be the event $\{S \cap T_{p_{\mathrm{eff}}(\theta)} = \emptyset\}$. Then, the probability of $\mathcal{A}_\theta$ is lower bounded

$$\mathbb{P}_{\mathrm{DPP}} (\mathcal{A}_\theta) \geq \frac{1}{\theta} , \tag{3.45}$$

and conditionally on $\mathcal{A}_\theta$,

$$\mathbb{E}_{\mathrm{DPP}} \left[\|\boldsymbol{X} - \boldsymbol{\Pi}_S^2 \boldsymbol{X}\|_2^2 \,\middle|\, \mathcal{A}_\theta \right] \leq (1 + (p_{\mathrm{eff}}(\theta) - k + 1)(k - 1 + \theta)) \|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_2^2 \tag{3.46}$$

and

$$\mathbb{E}_{\mathrm{DPP}} \left[\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \,\middle|\, \mathcal{A}_\theta \right] \leq \left(1 + \beta \frac{(p_{\mathrm{eff}}(\theta) + 1 - k)}{d-k} (k - 1 + \theta) \right) \|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_{\mathrm{Fr}}^2 . \tag{3.47}$$

In Theorem 3.7, the effective sparsity level $p_{\text{eff}}(\theta)$ replaces the sparsity level p of Proposition 3.6. The key is to condition on S not containing any index corresponding to a column with too small a k-leverage score, that is, the event $\mathcal{A}_\theta$. In practice, this is achieved by rejection sampling: we repeatedly and independently sample $S \sim \text{DPP}(\boldsymbol{K})$ until $S \cap T_{p_{\text{eff}}}(\theta) = \emptyset$. The caveat of any rejection sampling procedure is a potentially large number of samples required before acceptance. But in the present case, Equation (3.45) guarantees that the expectation of that number of samples is less than θ. The free parameter θ thus interestingly controls both the "energy" threshold in (3.44), and the complexity of the rejection sampling. The approximation bounds suggest picking θ close to 1, which implies a compromise with the value of $p_{\text{eff}}(\theta)$ that should not be too large either. We have empirically observed that the performance of the DPP is relatively insensitive to the choice of θ.

In order to compare with some of the previous results in Section 3.2, we quickly derive from Theorem 3.7 a bound in probability. We do so for the Frobenius norm, and the proof is similar for the spectral norm. Let $\lambda > 0$. It holds that

$$\mathbb{P}_{\text{DPP}}\left(\|\boldsymbol{X} - \boldsymbol{\Pi}_S^2 \boldsymbol{X}\|_{\text{Fr}} \leq \lambda \|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_{\text{Fr}} \,\middle|\, \mathcal{A}_\theta\right) \tag{3.48}$$

$$\geq 1 - \frac{\left(1 + \beta \frac{(p_{\text{eff}}(\theta)+1-k)}{d-k}(k-1+\theta)\right)}{\lambda^2}, \tag{3.49}$$

where the last inequality follows from Theorem 3.7 and Markov's inequality. Now, for

$$\lambda \;\geq\; \sqrt{5\left(1 + \beta \frac{(p_{\text{eff}}(\theta)+1-k)}{d-k}(k-1+\theta)\right)},$$

it holds that

$$\mathbb{P}_{\text{DPP}}\left(\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\text{Fr}} \boldsymbol{X}\|_{\text{Fr}} \leq \lambda \|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_{\text{Fr}} | \mathcal{A}_\theta\right) \geq 0.8. \tag{3.50}$$

Compare this bound with the result (3.26) of Boutsidis et al., 2009 for the double phase algorithm, namely

$$\mathbb{P}_{\text{DPh}}\left(\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\text{Fr}} \boldsymbol{X}\|_{\text{Fr}} \leq (1 + 8\sqrt{2k(c-k)+1})\|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_{\text{Fr}}\right) \geq 0.8\,, \;\; c = \Theta(k \log k). \tag{3.51}$$

In particular, $(p_{\text{eff}}(\theta) - k + 1)/(d-k) \leq 1 \leq c - k$, so that if

$$\beta(p_{\text{eff}}(\theta) - k + 1)/(d-k) \leq c - k, \tag{3.52}$$

then

$$\sqrt{5\left(1 + \beta \frac{(p_{\text{eff}}(\theta) - k + 1)}{d-k}(k-1+\theta)\right)} \leq 1 + 8\sqrt{2k(c-k)+1}. \tag{3.53}$$

and the DPP with rejection of Theorem 3.7 has a smaller bound than the double phase algorithm. The key condition (3.52) can be expected to hold quite widely as both the decay of the singular values and the leverage scores contribute to make the left-hand side small. In particular, even when β equals its upper bound $d - k$, it is enough to have $p_{\text{eff}}(\theta) = \Theta(k)$.

We can prove a similar bound in probability for the spectral norm, but comparing to double phase becomes trickier, because of the Frobenius norm that appears in the bound (3.27) for double phase.

Finally, we note that using Bayes' theorem, Theorem 3.7 also yields bounds in probability for the projection DPP algorithm used without rejection. For instance, let $\lambda > 0$. It holds that

$$\mathbb{P}_{\mathrm{DPP}}\left(\{\|\boldsymbol{X} - \boldsymbol{\Pi}_S^2 \boldsymbol{X}\|_{\mathrm{Fr}} \leq \lambda \|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_{\mathrm{Fr}}\}\right) \tag{3.54}$$

$$\geq \mathbb{P}_{\mathrm{DPP}}\left(\{\|\boldsymbol{X} - \boldsymbol{\Pi}_S^2 \boldsymbol{X}\|_{\mathrm{Fr}} \leq \lambda \|\boldsymbol{X} - \boldsymbol{\Pi}_k \boldsymbol{X}\|_{\mathrm{Fr}}\} \cap \mathcal{A}_\theta\right) \tag{3.55}$$

$$\geq \frac{1}{\theta}\left(1 - \frac{\left(1 + \beta \frac{(p_{\mathrm{eff}}(\theta)+1-k)}{d-k}(k-1+\theta)\right)}{\lambda^2}\right). \tag{3.56}$$

Such bounds are more flexible than those of double phase, in the sense that we can vary the parameters θ and λ independently, while the bounds of the double phase algorithm are constrained by $c \geq 1600 c_0^2 k \log(800 c_0^2 k)$.

3.4.2 *Bounds for the excess risk in sketched linear regression*

In Section 3.2, we surveyed bounds on the excess risk of ordinary least squares estimators that relied on a subsample of the columns of $\boldsymbol{X}$. Importantly, the generic bound (3.30) of Mor-Yosef and Avron, 2019 has a bias term that depends on the maximum squared tangent of the principal angles between $\mathrm{Span}(\boldsymbol{S})$ and $\mathrm{Span}(\boldsymbol{V}_k)$. When $|S| = k$, this quantity is hard to control without making strong assumptions on the matrix $\boldsymbol{V}_k$. But it turns out that, in expectation under the same DPP as in Section 3.4.1, this bias term drastically simplifies.

Proposition 3.7. *We use the notation of Section 3.2. Under the projection DPP with marginal kernel $\boldsymbol{V}_k \boldsymbol{V}_k^\mathsf{T}$, it holds that*

$$\mathbb{E}_{\mathrm{DPP}}\left[\mathcal{E}(\boldsymbol{w}_S)\right] \leq \left(1 + k(p-k)\right) \frac{\|\boldsymbol{w}^*\|^2 \sigma_{k+1}^2}{N} + \frac{vk}{N}. \tag{3.57}$$

The sparsity level p appears again in the bound (3.57): The sparser the k-leverage scores distribution, the smaller the bias term. The bound (3.57) only features an additional $(1 + k(p-k))$ factor in the bias term, compared to the bound obtained by Mor-Yosef and Avron, 2019 for PCR, see Proposition 3.4. Loosely speaking, this factor is to be seen as the price we accept to pay in order to get more interpretable features than principal components in the linear regression problem. Finally, a natural question is to investigate the choice of k to minimize the bound in (3.57), but this is out of the scope of this paper.

As in Theorem 3.7, for practical purposes, it can be desirable to bypass the need for the exact sparsity level p in Proposition 3.7. We give a bound that replaces p with the effective sparsity level $p_{\mathrm{eff}}(\theta)$ introduced in (3.44).

Algorithm	Pre-processing	Memory	One sample complexity
Our algorithm	$\mathcal{O}(\min(Nd^2, N^2d))$	$\mathcal{O}(dk)$	$\mathcal{O}(dk^2)$
Volume sampling	$\mathcal{O}(\min(Nd^2, N^2d))$	$\mathcal{O}(dr)$	$\mathcal{O}(dk^2)$
Double phase	$\mathcal{O}(\min(Nd^2, N^2d))$	$\mathcal{O}(dk)$	$\mathcal{O}(ck^2 \log_2(k))$

Table 3.2 – Complexity of the three CSSP algorithms.

Theorem 3.8. *Using the notation of Section 3.2 for linear regression, and of Theorem 3.7 for leverage scores and their indices, it holds that*

$$\mathbb{E}_{\mathrm{DPP}}\left[\mathcal{E}(\hat{\boldsymbol{w}}_S) \,\middle|\, \mathcal{A}_\theta\right] \leq \left[1 + (k - 1 + \theta)\left(p_{\mathrm{eff}}(\theta) - k + 1\right)\right] \frac{\|\boldsymbol{w}^*\|^2 \sigma_{k+1}^2}{N} + \frac{vk}{N}. \tag{3.58}$$

In practice, the same rejection sampling routine as in Theorem 3.7 can be used to sample conditionally on $\mathcal{A}_\theta$. Finally, to the best of our knowledge, bounding the excess risk in linear regression has not been investigated under volume sampling.

In summary, we have obtained two sets of results. We have proven a set of multiplicative bounds in spectral and Frobenius norm for $\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^v \boldsymbol{X}\|_v^2$, $v \in \{2, \mathrm{Fr}\}$, under the projection DPP of marginal kernel $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^\mathsf{T}$, see Propositions 3.5 & 3.6 and Theorem 3.7. As far as the linear regression problem is concerned, we have proven bounds for the excess risk in sketched linear regression, see Proposition 3.7 and Theorem 3.8.

3.4.3 *Complexity analysis*

We compare in this section the time and space complexity of our projection DPP, volume sampling and double phase. All three algorithms require the computation of the right eigenvectors of the matrix $\boldsymbol{X}$ as a pre-processing, which can be achieved in $\mathcal{O}(\min(Nd^2, dN^2))$ operations. Our algorithm requires to keep the first k right eigenvectors $\boldsymbol{V}_k$, which means $\mathcal{O}(dk)$ memory cost; every sample costs $\mathcal{O}(dk^2)$ time using the implementation of Tremblay et al., 2018. In comparison, volume sampling requires to keep all the right eigenvectors with non vanishing singular values of $\boldsymbol{X}$: the memory cost is $\mathcal{O}(rd)$, where r is the rank of $\boldsymbol{X}$. Indeed, every sample from VS requires to run 2 steps: 1) sampling the set T of singular values using Algorithm 7 in (Kulesza and Taskar, 2012), which runs in $\mathcal{O}(rk) = \mathcal{O}(dk^2)$ operations, followed by 2) sampling from a projection DPP of marginal kernel $\boldsymbol{V}_{:,T} \boldsymbol{V}_{:,T}^\mathsf{T}$, this time in $\mathcal{O}(dk^2)$. Similarly, for the double phase algorithm, given the singular decomposition of $\boldsymbol{X}$, the complexity of one sample is dominated by the second phase, which runs in $\mathcal{O}(ck^2 \log_2(k))$. The discussion is summarized in Table 3.2.

Volume sampling and projection DPP have comparable time complexities and a slightly lower memory requirement for the DPP. Double phase shares the same pre-processing and space complexity, but the time complexity of obtaining one sample is harder to compare. Remembering the condition on $c = 1600 c_0^2 k \log(800 c_0^2 k)$ for double phase (from Theorem 3.6), the bound on the time complexity can be relatively large, although only cubic in k.

3.5 NUMERICAL EXPERIMENTS

In this section, we empirically compare our algorithm, the projection DPP with kernel $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^\mathsf{T}$, to the state of the art in column subset selection. In Section 3.5.1, the projection DPP with kernel $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^\mathsf{T}$ and volume sampling are compared on toy datasets. In Section 3.5.2, several column subset selection algorithms are compared to the projection DPP on four real datasets from genomics and text processing. In particular, the numerical simulations demonstrate the favourable influence of the sparsity of the k-leverage scores on the performance of our algorithm both on toy datasets and real datasets. Finally, we packaged all CSS algorithms in this section in a publicly available Python toolbox[3].

3.5.1 *Toy datasets*

This section is devoted to comparing the expected approximation error $\mathbb{E}\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2$ for the projection DPP and volume sampling. We focus on the Frobenius norm to avoid effects due to different choices of the projection $\boldsymbol{\Pi}_S^{\nu}$, see (3.4).

In order to be able to evaluate the expected errors *exactly*, we generate matrices of low dimension ($d = 20$) so that the subsets of $[d]$ can be exhaustively enumerated. Furthermore, to investigate the role of leverage scores and singular values on the performance of CSS algorithms, we need to generate datasets $\boldsymbol{X}$ with prescribed spectra and k-leverage scores.

Generating toy datasets

Recall that the SVD of $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ reads $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T}$, where $\boldsymbol{\Sigma}$ is a diagonal matrix and $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices. To sample a matrix $\boldsymbol{X}$, we first let $\boldsymbol{U}$ correspond to the first r columns of an $N \times N$ sample from the Haar measure on $\mathbb{O}_N(\mathbb{R})$. Then, $\boldsymbol{\Sigma}$ is chosen among a few deterministic diagonal matrices that illustrate various spectral properties. Sampling the matrix $\boldsymbol{V}$ is trickier if k-leverage scores are to be prescribed. The first k columns of $\boldsymbol{V}$ are constrained as follows: the number of non vanishing rows of $\boldsymbol{V}_k$ is equal to p and the norms of the nonvanishing rows are prescribed by a vector $\boldsymbol{\ell}$. We thus propose an algorithm that takes as input a leverage scores profile $\boldsymbol{\ell}$ and a spectrum σ^2, and outputs a corresponding random orthogonal matrix $\boldsymbol{V}_k$; see Appendix 3.D. This algorithm is a randomization[4] of the algorithm proposed by Fickus, Mixon, Poteet, and Strawn, 2013. Finally, the matrix $\boldsymbol{V}_k \in \mathbb{R}^{d \times k}$ is completed by applying the Gram-Schmidt procedure to $d - k$ additional i.i.d. unit Gaussian vectors, resulting in a matrix $\boldsymbol{V} \in \mathbb{R}^{d \times d}$. Figure 3.4 summarizes the algorithm we use to generate matrices $\boldsymbol{X}$ with a k-leverage scores profile $\boldsymbol{\ell}$, spectrum $\boldsymbol{\Sigma}$, and a sparsity level p.

Volume sampling vs projection DPP

This section sums up the results of numerical simulations on toy datasets. The number of observations is fixed to $N = 100$, the dimension to $d = 20$, and the number of

[3] http://github.com/AyoubBelhadji/CSSPy
[4] http://github.com/AyoubBelhadji/FrameBuilder

MATRIXGENERATOR$\big(\boldsymbol{\ell} \in \mathbb{R}_+^d, \boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}, p \in [k+1:d]\big)$

1 Sample $\boldsymbol{U}$ from the Haar measure $\mathbb{O}_N(\mathbb{R})$.

2 Generate a matrix $\boldsymbol{V}_k$ with the k-leverage-scores profile $\boldsymbol{\ell}$.

3 Extend the matrix $\boldsymbol{V}_k$ to an orthogonal matrix $\boldsymbol{V}$.

4 **return** $\boldsymbol{X} \longleftarrow \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T}$

Figure 3.4 – The pseudocode of the algorithm generating a matrix $\boldsymbol{X}$ with prescribed profile of k-leverage scores.

selected columns to $k \in \{3,5\}$. Singular values are chosen from the following profiles: a spectrum with a cutoff called the projection spectrum,

$$\boldsymbol{\Sigma}_{k=3,\text{proj}} = 100\sum_{i=1}^{3} \boldsymbol{e}_i\boldsymbol{e}_i^\mathsf{T} + 0.1\sum_{i=4}^{20} \boldsymbol{e}_i\boldsymbol{e}_i^\mathsf{T},$$

$$\boldsymbol{\Sigma}_{k=5,\text{proj}} = 100\sum_{i=1}^{5} \boldsymbol{e}_i\boldsymbol{e}_i^\mathsf{T} + 0.1\sum_{i=6}^{20} \boldsymbol{e}_i\boldsymbol{e}_i^\mathsf{T}.$$

a smooth spectrum

$$\boldsymbol{\Sigma}_{k=3,\text{smooth}} = 100\boldsymbol{e}_1\boldsymbol{e}_1^\mathsf{T} + 10\boldsymbol{e}_2\boldsymbol{e}_2^\mathsf{T} + \boldsymbol{e}_3\boldsymbol{e}_3^\mathsf{T} + 0.1\sum_{i=4}^{20} \boldsymbol{e}_i\boldsymbol{e}_i^\mathsf{T},$$

$$\boldsymbol{\Sigma}_{k=5,\text{smooth}} = 10000\boldsymbol{e}_1\boldsymbol{e}_1^\mathsf{T} + 1000\boldsymbol{e}_2\boldsymbol{e}_2^\mathsf{T} + 100\boldsymbol{e}_3\boldsymbol{e}_3^\mathsf{T} + 10\boldsymbol{e}_4\boldsymbol{e}_4^\mathsf{T} + \boldsymbol{e}_5\boldsymbol{e}_5^\mathsf{T} + 0.1\sum_{i=6}^{20} \boldsymbol{e}_i\boldsymbol{e}_i^\mathsf{T},$$

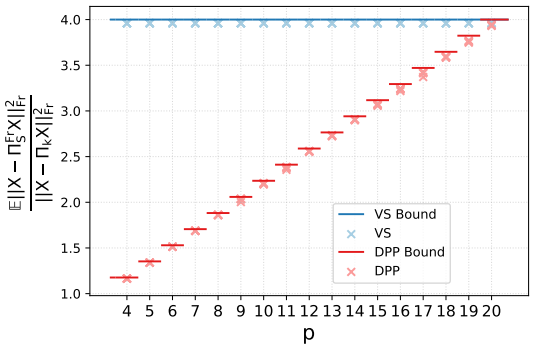
and a flat spectrum with all singular values equal to 1

$$\boldsymbol{\Sigma}_{\text{identity}} = \sum_{i=1}^{20} \boldsymbol{e}_i\boldsymbol{e}_i^\mathsf{T}.$$

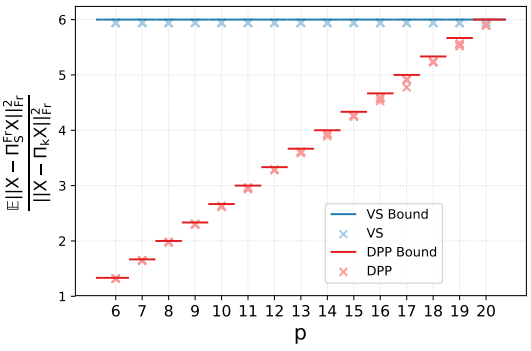
Note that all profiles satisfy $\beta = 1$; see (3.40). We discuss the case $\beta > 1$ at the end of the section. In each experiment, for each spectrum, we sample 200 independent leverage score profiles that satisfy the sparsity constraints $p = \Big|\{i \in [d], \boldsymbol{V}_{i,[k]} \neq \boldsymbol{0}\}\Big|$ from a Dirichlet distribution of dimension p with concentration parameter 1 and equal means. For each leverage score profile, we sample a matrix $\boldsymbol{X}$ from the algorithm in Figure 3.4.

Figure 3.5 compares, on the one hand, the theoretical bounds in Theorem 3.4 for volume sampling and Proposition 3.6 for the projection DPP, to the numerical evaluation of the expected error for sampled toy datasets on the other hand. The x-axis indicates various sparsity levels p. The unit on the y-axis is the error of PCA. There are 400 crosses on each subplot: each of the 200 matrices appears once for both algorithms. The 200 matrices are spread evenly across the values of p.

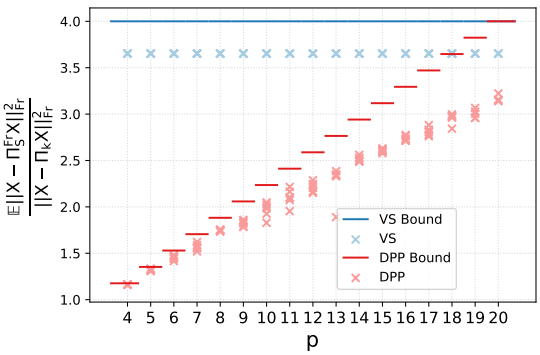
Used as a reference, the VS bounds are proportional to $(k+1)$ and independent of p. In fact, by Theorem 3.5, the expected value of the Frobenius norm of the approximation error only depends on the spectrum of the matrix $\boldsymbol{X}$; in particular, it does not involve



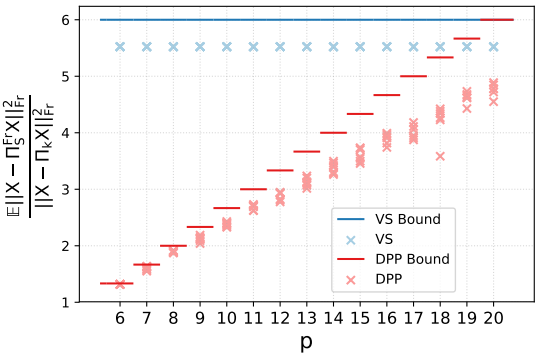
(a) $\mathbf{\Sigma}_{3,\text{proj}}$, $k = 3$



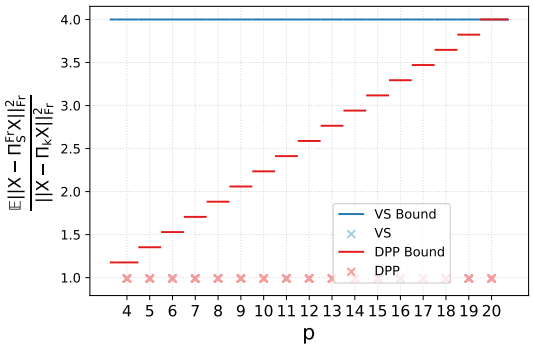
(d) $\mathbf{\Sigma}_{5,\text{proj}}$, $k = 5$



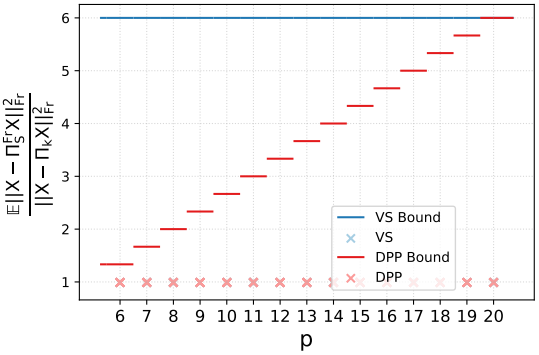
(b) $\mathbf{\Sigma}_{3,\text{smooth}}$, $k = 3$



(e) $\mathbf{\Sigma}_{5,\text{smooth}}$, $k = 5$

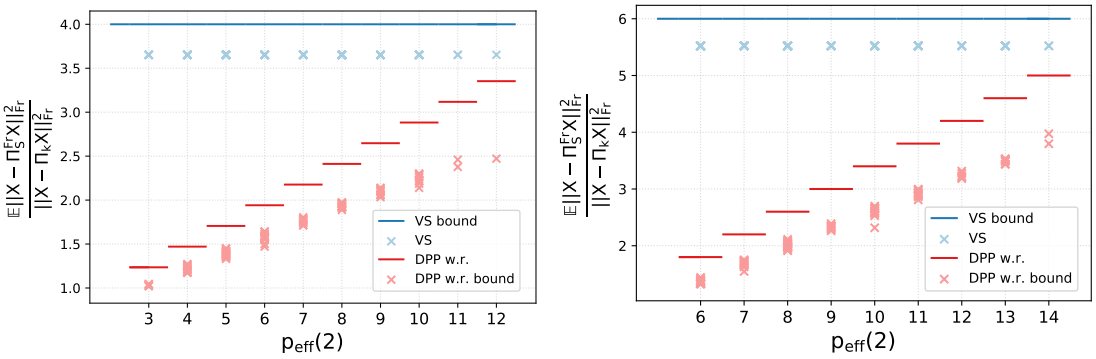


(c) $\mathbf{\Sigma}_{\text{identity}}$, $k = 3$



(f) $\mathbf{\Sigma}_{\text{identity}}$, $k = 5$

Figure 3.5 – Realizations and bounds for $\mathbb{E}\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\text{Fr}}\boldsymbol{X}\|_{\text{Fr}}^2$ as a function of the sparsity level p.



(a) $\boldsymbol{\Sigma}_{3,\text{smooth}}$ (b) $\boldsymbol{\Sigma}_{5,\text{smooth}}$

Figure 3.6 – Realizations and bounds for $\mathbb{E}\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\text{Fr}}\boldsymbol{X}\|_{\text{Fr}}^2$ as a function of the effective sparsity level $p_{\text{eff}}(2)$.

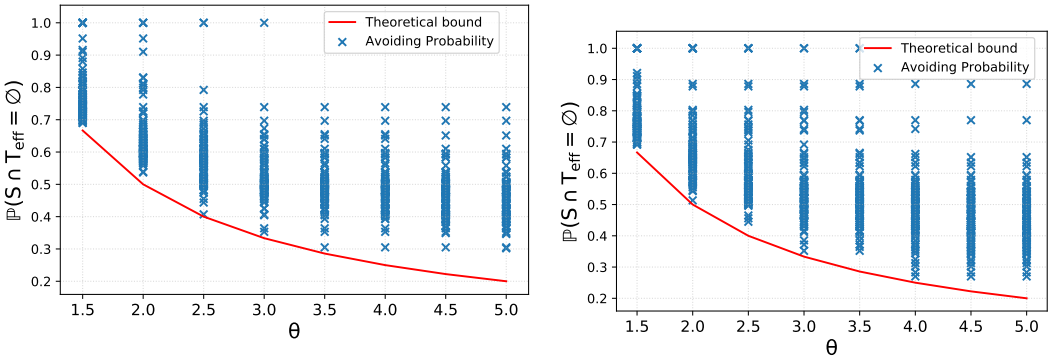
the matrix $\boldsymbol{V}$. These bounds appear to be tight for projection spectra, and looser for smooth spectra.

For the projection DPP, the bound $1 + k\frac{p-k}{d-k}$ is linear in p, and can thus be much lower than the bound of VS. The numerical evaluations of the error also suggest that this DPP bound is tight for a projection spectrum and looser in the smooth case. We emphasize that, in both cases, the bound is representative of the actual behaviour of the algorithm. The bottom row of Figure 3.5 displays the same results for identity spectra, again for $k = 3$ and $k = 5$. This setting is extremely nonsparse and represents an arbitrarily bad scenario where even PCA would not make much practical sense. Then both VS and DPP sampling perform the same as PCA: all crosses superimpose at $y = 1$. In this particular case, our linear bound in p is not representative of the actual behaviour of the error. This observation can be explained for volume sampling using Theorem 3.5, which states that the expected squared error under VS is Schur-concave, and is thus minimized for flat spectra. We have no similar result for the projection DPP.

Figure 3.6 provides a similar comparison for the two smooth spectra $\boldsymbol{\Sigma}_{3,\text{smooth}}$ and $\boldsymbol{\Sigma}_{5,\text{smooth}}$, but this time using the effective sparsity level $p_{\text{eff}}(\theta)$ introduced in Theorem 3.7. Qualitatively, we have observed the results to be robust to the choice of θ: we use $\theta = 2$. The 200 sampled matrices are now unevenly spread across the x-axis, since we do not control $p_{\text{eff}}(\theta)$. Note finally that the DPP here is conditioned on the event $\{S \cap T_{p_{\text{eff}}(\theta)} = \emptyset\}$, and sampled using an additional rejection sampling routine as detailed below Theorem 3.7.

For the DPP, the bound is again linear on the effective sparsity level $p_{\text{eff}}(2)$, and can again be much lower than the VS bound. The behaviours of both VS and the projection DPP are similar to the exact sparsity setting of Figure 3.5: the DPP has uniformly better bounds and actual errors, and the bound reflects the actual behaviour, relatively loosely when $p_{\text{eff}}(2)$ is large.

Figure 3.7 compares the theoretical bound in Theorem 3.7 for the avoiding probability $\mathbb{P}(S \cap T_{p_{\text{eff}}(\theta)} = \emptyset)$ with 200 realizations, as a function of θ. More precisely, we drew 200 matrices $\boldsymbol{X}$, and then for each $\boldsymbol{X}$, we computed exactly – by enumeration – the value $\mathbb{P}(S \cap T_{p_{\text{eff}}(\theta)} = \emptyset)$ for all values of θ. The only randomness is thus in the



(a) $\boldsymbol{\Sigma}_{3,\text{smooth}}$ (b) $\boldsymbol{\Sigma}_{5,\text{smooth}}$

Figure 3.7 – Realizations and bounds for the avoiding probability $\mathbb{P}(S \cap T_{p_{\text{eff}}(\theta)} = \emptyset)$ in Theorem 3.7 as a function of θ.

sampling of $\boldsymbol{X}$, not the evaluation of the probability. Again, the results suggest that the bound is relatively tight.

Finally, we examine relaxing $\beta = 1$. We have observed our results to be robust with respect to β. At the extreme, in Figure 3.8, we compare the errors for two additional spectra $\hat{\boldsymbol{\Sigma}}_{3,\text{proj}}$ and $\hat{\boldsymbol{\Sigma}}_{3,\text{smooth}}$ such that β is close to its maximum value $d - k = 17$:

$$\hat{\boldsymbol{\Sigma}}_{k=3,\text{proj}} = 100 \sum_{i=1}^{3} \boldsymbol{e}_i \boldsymbol{e}_i^\intercal + 0.1 \boldsymbol{e}_4 \boldsymbol{e}_4^\intercal + 10^{-4} \sum_{i=5}^{20} \boldsymbol{e}_i \, \boldsymbol{e}_i^\intercal,$$

and

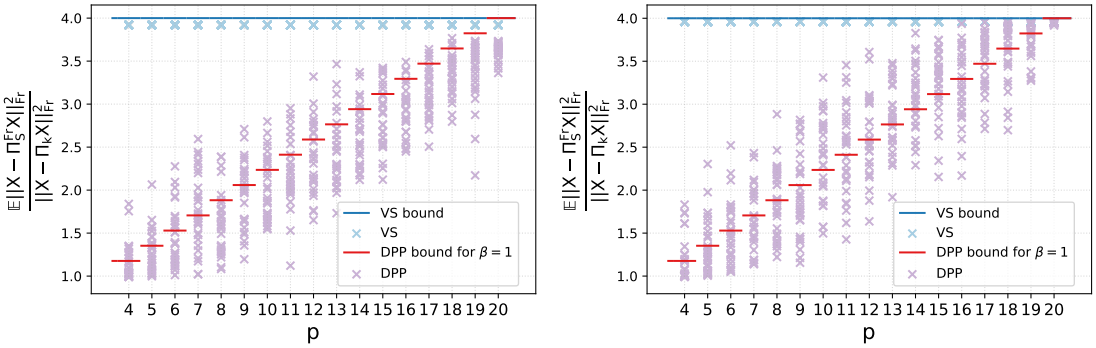
$$\hat{\boldsymbol{\Sigma}}_{k=3,\text{smooth}} = 100 \boldsymbol{e}_1 \boldsymbol{e}_1^\intercal + 10 \boldsymbol{e}_2 \boldsymbol{e}_2^\intercal + \boldsymbol{e}_3 \boldsymbol{e}_3^\intercal + 0.1 \boldsymbol{e}_4 \boldsymbol{e}_4^\intercal + 10^{-4} \sum_{i=5}^{20} \boldsymbol{e}_i \, \boldsymbol{e}_i^\intercal.$$

While the bound for such a large β would be almost vertical and does not reflect anymore the actual behaviour of the algorithm, we observe that the algorithm still performs comparably to the setting where $\beta = 1$, although with more variance, and that the bound with $\beta = 1$ (in red) still represents the behaviour of the algorithm. This is a hint that there is room for improvement in our bounds in the large β regime. The search for a new bound that would be independent of β is nontrivial and a subject of future work.

3.5.2 *Real datasets*

Dataset	Application domain	$N \times d$	References
Colon	genomics	62×2000	(Alon et al., 1999)
Leukemia	genomics	72×7129	(Golub et al., 1999)
Basehock	text processing	1993×4862	(Li et al., 2017b)
Relathe	text processing	1427×4322	(Li et al., 2017b)

Table 3.3 – Datasets used in the experimental section.



(a) $\hat{\mathbf{\Sigma}}_{3,\text{smooth}}$, $k = 3$ (b) $\hat{\mathbf{\Sigma}}_{3,\text{proj}}$, $k = 3$

Figure 3.8 – Realizations and bounds for $\mathbb{E}\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\text{Fr}}\boldsymbol{X}\|_{\text{Fr}}^2$ as a function of the sparsity level p in the case $\beta > 1$.

The datasets described in Table 3.3 are illustrative of two extreme situations regarding the sparsity of the k-leverage scores. For instance, the dataset Basehock has a very sparse profile of k-leverage scores, while the dataset Colon has a quasi-uniform distribution of k-leverage scores, see Figures 3.9a & 3.9b. This section compares the empirical performances of several column subset selection algorithms on these datasets.

We consider the following algorithms presented in Section 3.2: 1) the projection DPP with marginal kernel $\boldsymbol{K} = \boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}$, 2) volume sampling, 3) deterministically picking the largest k-leverage scores, 4) pivoted QR as in (Golub, 1965), although the only known bounds for this algorithm are for the spectral norm, and 5) double phase, with c manually tuned to optimize the performance, usually around $c \approx 10k$.

The rest of Figure 3.9 sums up the empirical results of these algorithms on the Colon and Basehock datasets. Figures 3.9c & 4.15b illustrate the results of the five algorithms in the following setting. An ensemble of 50 subsets are sampled using each algorithm. We give the corresponding boxplots for the Frobenius errors, on Colon and Basehock respectively. Deterministic methods (largest leverage scores and pivoted QR) perform well compared with other algorithms on the Basehock dataset; in contrast, they display very bad performance on the Colon dataset.

Focusing now on the three random sampling methods, we first make sure that the observed differences in Frobenius error are statistically significant at level $\alpha = 0.05$. To that end, we report in Table 3.4 the p-values of the three pairwise Mann-Whitney tests between the three algorithms. More precisely, let F_X denote the CDF of the Frobenius errors for algorithm $\text{X} \in \{\text{DPh}, \text{DPP}, \text{VS}\}$. We test H_0:"$F_\text{X} = F_\text{Y}$" against the so-called *one-sided* alternative H_1 that X is better than Y, in the sense that if you independently run algorithms X and Y, it is more likely that the Frobenius error of X is the smaller of the two. Now, we want to *jointly* test whether all three pairs of algorithms within $\{\text{DPh}, \text{DPP}, \text{VS}\}$ perform differently, so we use a Bonferroni correction (Wasserman, 2013). Looking at Table 3.4 for dataset Colon, all three p-values are smaller than $\alpha/3 = 0.05/3$, so that we simultaneously reject that $F_\text{DPh} = F_\text{DPP}$, $F_\text{DPh} = F_\text{VS}$ and $F_\text{DPP} = F_\text{VS}$, and we declare the differences among algorithms to be statistically significant. The same can be said for dataset Basehock. In particular, we observe that the increase in performance using the projection DPP compared to volume

sampling is more important for the Basehock dataset than for the Colon dataset: this improvement can be explained by the sparsity of the k-leverage scores as predicted by our approximation bounds. The double phase algorithm has the best results on both datasets. However its theoretical guarantees cannot predict such an improvement, as noted in Section 3.2. The performance of the projection DPP is comparable to double phase and makes it a close second, with a slightly larger gap on the Colon dataset. We emphasize that our approximation bounds are sharp compared to numerical observations.

Figures 4.15c & 4.15d show results obtained using a classical boosting technique for randomized algorithms. We repeat 20 times the following procedure: sample 50 subsets $(S_i)_{i\in[50]}$ and take the subset $S_{\min}$ that minimizes the approximation error among the elements of the batch $(S_i)_{i\in[50]}$. Displayed boxplots are for these 20 best results. The same comparisons apply as without boosting, with p-values given in Table 3.5.

Figure 3.10 calls again for similar comments, comparing this time the datasets Relathe (with concentrated profile of k-leverage scores) and Leukemia (with almost uniform profile of k-leverage scores). This time, the same test as for Colon vs. Basehock in Table 3.4 further reveals that we cannot reject the hypothesis that $F_{\mathrm{DPh}} = F_{\mathrm{DPP}}$ on Relathe. In other words, there is no hint that the performance of the double phase is different from that of DPP on that particular dataset (at level $\alpha = 0.05$). The same is true for the boosted version of the algorithms; see Table 3.5.

Dataset \ X vs. Y	DPP vs. VS	DPh vs. VS	DPh vs. DPP
Colon	6.10^{-6}	9.10^{-18}	2.10^{-16}
Leukemia	5.10^{-5}	4.10^{-13}	2.10^{-5}
Basehock	10^{-17}	10^{-17}	3.10^{-5}
Relathe	9.10^{-18}	10^{-17}	**0.15**

Table 3.4 – p-values for Mann–Whitney U-test comparisons.

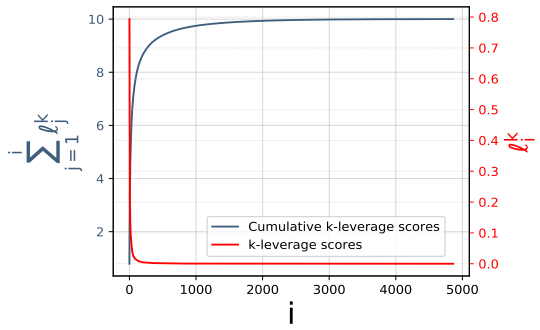
Dataset \ X vs. Y	DPP vs. VS	DPh vs. VS	DPh vs. DPP
Colon	4.10^{-8}	10^{-4}	4.10^{-8}
Leukemia	3.10^{-6}	3.10^{-8}	3.10^{-6}
Basehock	3.10^{-8}	3.10^{-8}	7.10^{-7}
Relathe	3.10^{-8}	3.10^{-8}	**0.053**

Table 3.5 – p-values for Mann–Whitney U-test comparisons, for the boosted algorithms.

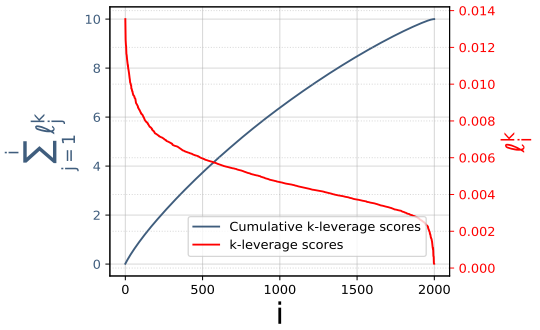
3.5.3 *Regression with column subset selection*

This section compares the empirical performance of several column subset selection algorithms for regression tasks on the datasets in Table 3.3. We compare column subset selection algorithms on synthetic regression vectors.

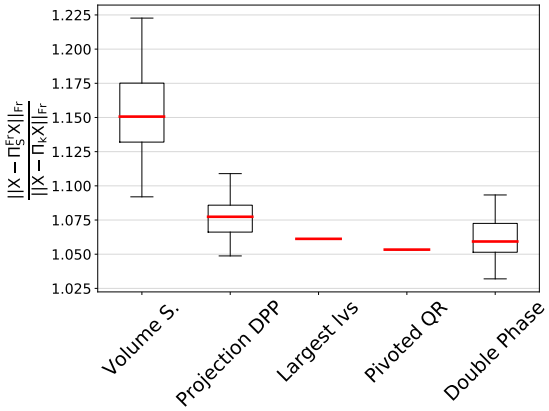
We consider the following algorithms: 1) the projection DPP with marginal kernel $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^\intercal$, 2) volume sampling, 3) double phase with $c = 10k$ and 4) principal component regression (PCR).



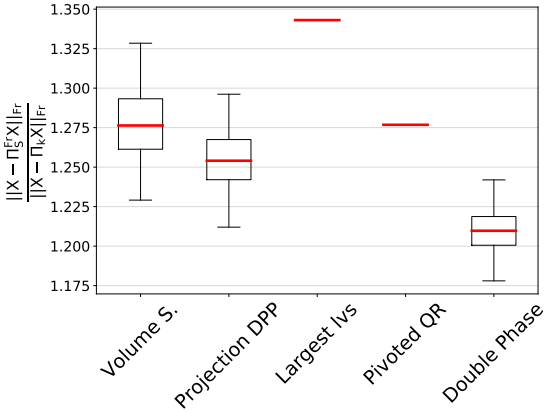
(a) k-leverage scores profile and cumulative profile for the dataset Basehock (k=10).



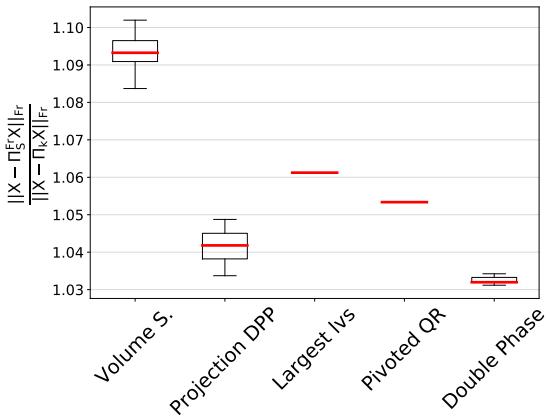
(b) k-leverage scores profile and cumulative profile for the dataset Colon (k=10).



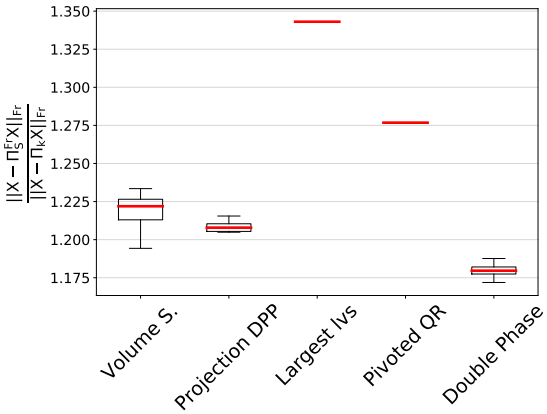
(c) Boxplots of $\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the five algorithms on the dataset Basehock (k=10).



(d) Boxplots of $\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the five algorithms on the dataset Colon (k=10).

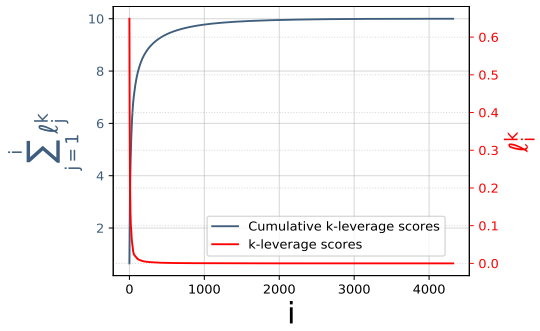


(e) Boxplots of $\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the boosting of randomized algorithms on the dataset Basehock (k=10).

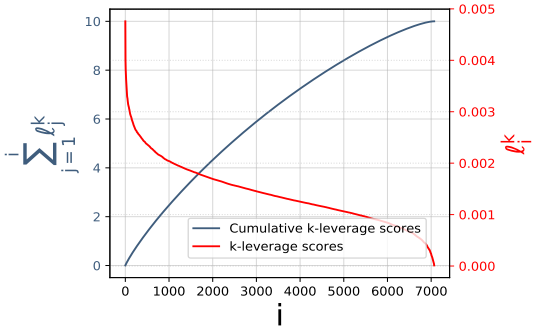


(f) Boxplots of $\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the boosting of randomized algorithms on the dataset Colon (k=10).

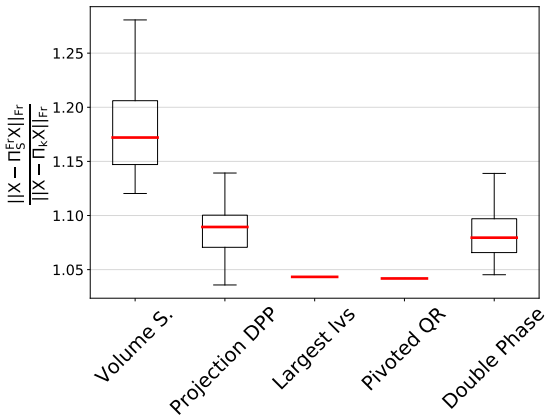
Figure 3.9 – Comparison of several column subset selection algorithms for two datasets with different leverage score profiles: Basehock and Colon.



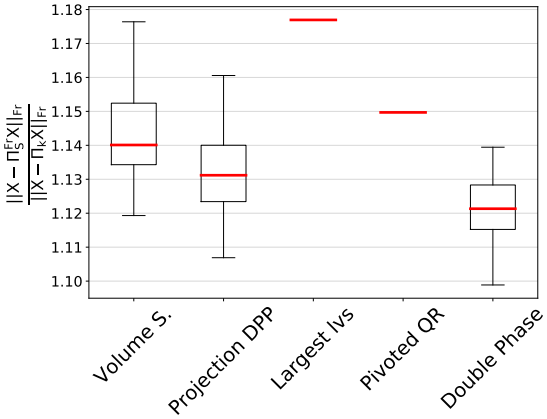
(a) k-leverage scores profile and cumulative profile for the dataset Relathe (k=10).



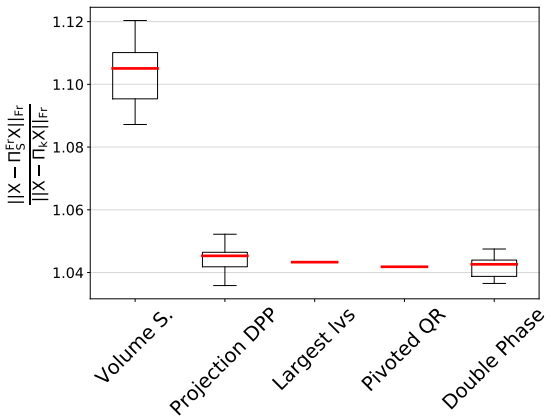
(b) k-leverage scores profile and cumulative profile for the dataset Leukemia (k=10).



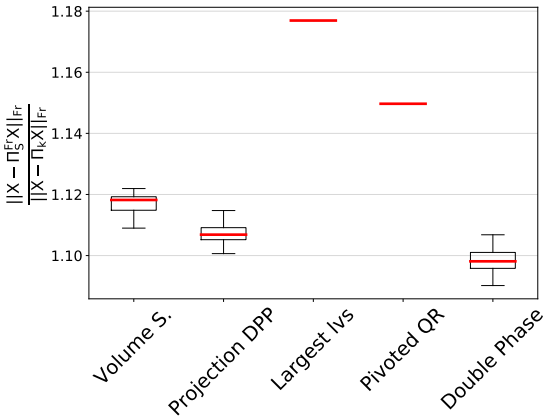
(c) Boxplots of $\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the five algorithms on the dataset Relathe (k=10).



(d) Boxplots of $\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the five algorithms on the dataset Leukemia (k=10).



(e) Boxplots of $\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the boosting of randomized algorithms on the dataset Relathe (k=10).



(f) Boxplots of $\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the boosting of randomized algorithms on the dataset Leukemia (k=10).

Figure 3.10 – Comparison of several column subset selection algorithms for two datasets with different leverage score profiles: Relathe and Leukemia.

To investigate the effect of the alignment of $\mathbf{y}$ with the principal subspaces of $\boldsymbol{X}$, we use two different label vectors $\mathbf{y}$. More precisely, we define a principal subpsace of dimension $k_0 = 20$ and define two directions

$$\mathbf{y}_1 \propto \frac{1}{k_0} \sum_{i \in [k_0]} \boldsymbol{U}_{:,i}, \tag{3.59}$$

and

$$\mathbf{y}_2 \propto \frac{1}{d - k_0} \sum_{i \in [k_0+1:d]} \boldsymbol{U}_{:,i}. \tag{3.60}$$

that are respectively aligned with or orthogonal to the principal subspace of dimension k_0. We take $\mathbf{y}_1$ and $\mathbf{y}_2$ to be normed vectors, and we note that $\mathbf{y}_1 \in \mathrm{Span}(\boldsymbol{U}_{:,i})_{i \in [k_0]}$, while $\mathbf{y}_2 \in \mathrm{Span}(\boldsymbol{U}_{:,i})_{i \in [k_0+1:d]}$. Adapted PCR with $k = k_0$ is expected to perform perfectly well for $\mathbf{y}_1$ and badly for $\mathbf{y}_2$.

Figure 3.11 illustrates the results of the four algorithms in the following setting. An ensemble of 50 subsets are sampled from each randomized algorithm. We give the corresponding approximation errors $\|\mathbf{y}_i - \boldsymbol{X}\hat{\boldsymbol{w}}_S\|_2$, on Colon and Basehock respectively, for every value of $k \in \{10, 15, 20, 25, 30\}$.

First, we observe that the relative performance of the column selection algorithms compared to PCR depends on the regressed vector $\mathbf{y}_i$. As expected, for $\mathbf{y}_1$, PCR has the best approximation error. In particular, the approximation error for PCR is 0 for $k \geq k_0$, while, for the column subset selection algorithms, the approximation error decreases with k without vanishing. On the other hand, PCR has the worst error for $\mathbf{y}_2$.

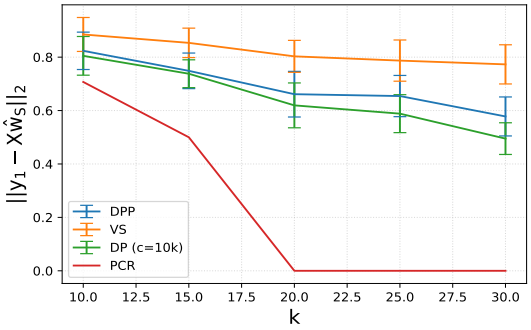
Now, comparing column subset selection algorithms, we observe that the relative performances depend on $\mathbf{y}_i$ and the leverage score profile. Double phase and the projection DPP perform similarly in all cases. Volume sampling displays minimal error for $\mathbf{y}_2$ but has the worst performance for $\mathbf{y}_1$. Similarly to previous observations, the differences between VS and the rest are amplified on the dataset with concentrated leverage score profile (Basehock).

3.5.4 *Conclusion*

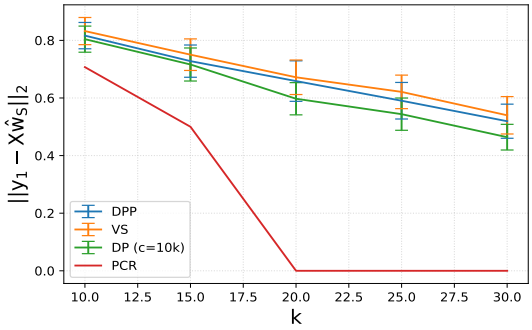
The performance of our projection DPP algorithm has been compared to state-of-the-art column subset selection algorithms. We emphasize that the theoretical performance of the proposed approach takes advantage from the sparsity of the k-leverage scores, as in Proposition 3.6, or their fast decrease, as in Proposition 3.7. The actual behaviour of the algorithm is in very good agreement with our theoretical bounds when the spectrum is flat above k (i.e., β is close to 1). In contrast, state-of-the-art algorithms like volume sampling come with both looser bounds and worse performance; double phase displays great performance but has overly pessimistic theoretical bounds. When β is large, our bounds become pessimistic even though the behaviour of the DPP selection remains very competitive for low-rank approximation.

3.6 DISCUSSION

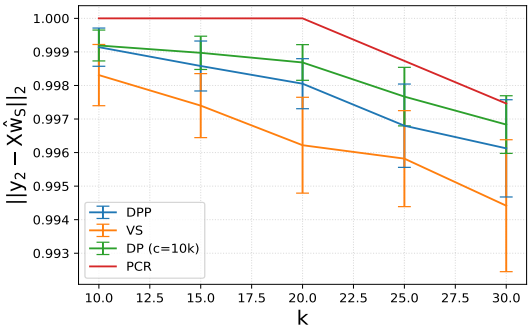
We have proposed, analysed, and empirically investigated a new randomized column subset selection (CSS) algorithm. The crux of our algorithm is a discrete determinantal



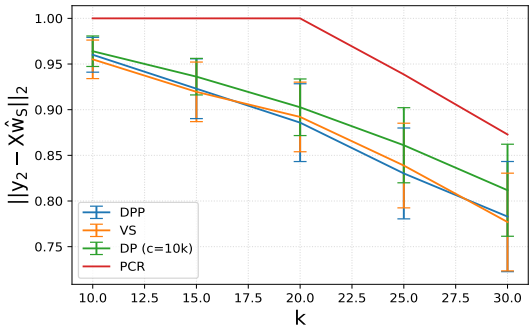
(a) The value of $\|\mathbf{y}_1 - \boldsymbol{X}\hat{\boldsymbol{w}}_S\|_2$ as a function of k on a batch of 50 samples for the algorithms: DPP, VS, DP and PCR on the dataset Basehock.



(b) The value of $\|\mathbf{y}_1 - \boldsymbol{X}\hat{\boldsymbol{w}}_S\|_2$ as a function of k on a batch of 50 samples for the algorithms: DPP, VS, DP and PCR on the dataset Colon.



(c) The value of $\|\mathbf{y}_2 - \boldsymbol{X}\hat{\boldsymbol{w}}_S\|_2$ as a function of k on a batch of 50 samples for the algorithms: DPP, VS, DP and PCR on the dataset Basehock.



(d) The value of $\|\mathbf{y}_2 - \boldsymbol{X}\hat{\boldsymbol{w}}_S\|_2$ as a function of k on a batch of 50 samples for the algorithms: DPP, VS, DP and PCR on the dataset Colon.

Figure 3.11 – Comparison of several column subset selection algorithms for the datasets Basehock and Colon on a regression task.

point process (DPP) that selects a diverse set of k columns of a matrix $\boldsymbol{X}$. This DPP is tailored to CSS through its parametrization by the marginal kernel $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^\mathsf{T}$, where $\boldsymbol{V}_k$ are the first k right singular vectors of the matrix $\boldsymbol{X}$. This specific kernel is related to volume sampling, the state-of-the-art for CSS guarantees in Frobenius and spectral norm.

We have identified generic conditions on the matrix $\boldsymbol{X}$ under which our algorithm has bounds that improve on volume sampling. In particular, our bounds highlight the importance of the sparsity and the decay of the k-leverage scores on the approximation performance of our algorithm. We have further numerically illustrated this relation to the sparsity and decay of the k-leverage scores using toy and real datasets. In these experiments, our algorithm performs comparably well to the so-called double phase algorithm, which is the empirical state-of-the-art for CSS despite more conservative theoretical guarantees than volume sampling. Thus, our DPP sampling inherits both favourable theoretical bounds and increased empirical performance under sparsity or fast decay of the k-leverage scores. Both are common features of real datasets.

As detected in the experimental section, our bounds are sharp except in the large β regime. Surprisingly, the actual behaviour of the algorithm remains very close to the case $\beta = 1$, which further speaks in favour for the DPP approach. This is a hint that our bounds can probably be refined to more sharply account for large βs.

Although generally studied as an independent task, in practice CSS is often a prelude to a learning algorithm. We have considered linear regression and we have given a bound on the excess risk of a regression performed on the selected columns only. In particular, the sparsity and decay of the k-leverage scores are again involved: the more localized the k-leverage scores, the smaller the excess risk bounds. Such an analysis of the excess risk in regression further highlights the interest of the DPP: it would be difficult to conduct for either volume sampling or double phase. Future work in this direction includes investigating the importance of the sparsity of the k-leverage scores on the performance of other learning algorithms such as spectral clustering or support vector machines. Moreover, our theoretical analysis may found an application in the field of graph signal reconstruction (Tremblay et al., 2017; Puy et al., 2018).

In terms of computational cost, our algorithms scale with the cost of finding the k first right singular vectors, which is currently the main bottleneck. In line with (Drineas et al., 2012) and (Boutsidis et al., 2011), where the authors estimate the k-leverage scores using random projections, we plan to investigate the impact of random projections to estimate the full matrix $\boldsymbol{K}$ on the approximation guarantees of our algorithms (Magen and Zouzias, 2008).

3.7 PROOFS

3.7.1 *Technical lemmas*

We start with two useful lemmas borrowed from the literature.

Lemma 3.1 (Lemma 3.1, Boutsidis et al., 2011)**.** *Let $S \subset [d]$, then*

$$\|\boldsymbol{X} - \boldsymbol{\Pi}^{\nu}_{S,k}\boldsymbol{X}\|^2_{\nu} \leq \|\boldsymbol{E}(\boldsymbol{I} - \boldsymbol{P}_S)\|^2_{\nu}, \quad \nu \in \{2, \mathrm{Fr}\}, \tag{3.61}$$

where $\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{\Pi}_k\boldsymbol{X}$ and $\boldsymbol{P}_S = \boldsymbol{S}(\boldsymbol{V}_k^\intercal \boldsymbol{S})^{-1}\boldsymbol{V}_k^\intercal$. Furthermore,

$$\|\boldsymbol{X} - \boldsymbol{\Pi}^{\nu}_{S,k}\boldsymbol{X}\|^2_{\nu} \leq \frac{1}{\sigma_k^2(\boldsymbol{V}_{S,[k]})}\|\boldsymbol{X} - \boldsymbol{\Pi}_k\boldsymbol{X}\|^2_{\nu}, \quad \nu \in \{2, \mathrm{Fr}\}. \tag{3.62}$$

The following lemma was first proven by Deshpande et al., 2006, and later rephrased in (Deshpande and Rademacher, 2010).

Lemma 3.2 (Lemma 11, Deshpande and Rademacher, 2010)**.** *Let $\boldsymbol{V} \in \mathbb{R}^{k\times d}$, $r = \mathrm{rk}(\boldsymbol{V})$ and $\ell \in [1:r]$. Then*

$$\sum_{S\subset[d],|S|=\ell} e_\ell(\Sigma(\boldsymbol{V}_{:,S})^2) = e_\ell(\Sigma(\boldsymbol{V})^2) \tag{3.63}$$

where e_ℓ is the ℓ-th elementary symmetric polynomial on r variables, see Section 3.1.

Elementary symmetric polynomials play an important role in the proof of Proposition 3.7, in particular their interplay with the Schur order; see Appendix 3.B for definitions.

Lemma 3.3. *Let $\phi, \psi : \mathbb{R}_+^{*k} \to \mathbb{R}_+^*$ be defined by*

$$\phi : \boldsymbol{\sigma} \mapsto \frac{e_{k-1}(\boldsymbol{\sigma})}{e_k(\boldsymbol{\sigma})} \tag{3.64}$$

and

$$\psi : \boldsymbol{\sigma} \mapsto e_k(\boldsymbol{\sigma}). \tag{3.65}$$

Then both functions are symmetric, ϕ is Schur-convex, and ψ is Schur-concave.

of Lemma 3.3. Let $i, j \in [k], i \neq j$. Let $\sigma_i, \sigma_j \in \mathbb{R}_+^*$, it holds

$$\begin{aligned}(\sigma_i - \sigma_j)(\partial_i\phi(\boldsymbol{\sigma}) - \partial_j\phi(\boldsymbol{\sigma})) &= (\sigma_i - \sigma_j)(-\frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2}) \\ &= \frac{(\sigma_i - \sigma_j)^2(\sigma_i + \sigma_j)}{\sigma_i^2\sigma_j^2} \geq 0,\end{aligned}$$

so that ϕ is Schur-convex by Proposition 3.11. Similarly,

$$\begin{aligned}(\sigma_i - \sigma_j)(\partial_i\psi(\sigma) - \partial_j\psi(\sigma)) &= (\sigma_i - \sigma_j)(\prod_{\ell\neq i}\sigma_\ell - \prod_{\ell\neq j}\sigma_\ell)\\ &= -(\sigma_i - \sigma_j)^2 \prod_{\ell\neq i,j}\sigma_\ell \geq 0,\end{aligned}$$

so that ψ is Schur-concave by Proposition 3.11. □

Elementary symmetric polynomials also interact nicely with "marginalizing" sums.

Lemma 3.4. *Let $\boldsymbol{V}$ be a real $k \times d$ matrix and let $r = \mathrm{rk}(\boldsymbol{V})$. Denote by p the number of non zero columns of $\boldsymbol{V}$. Then for all $k \leq r+1$,*

$$\sum_{\substack{S\subset[d],|S|=k\\ \mathrm{Vol}_k(\boldsymbol{V}_{:,S})^2>0}} \sum_{\substack{T\subset S\\ |T|=k-1}} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2) \leq (p-k+1)e_{k-1}(\Sigma(\boldsymbol{V})^2). \tag{3.66}$$

A fortiori,

$$\sum_{\substack{S\subset[d],|S|=k\\ \mathrm{Vol}_k(\boldsymbol{V}_{:,S})^2>0}} \sum_{\substack{T\subset S\\ |T|=k-1}} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2) \leq (d-k+1)e_{k-1}(\Sigma(\boldsymbol{V})^2). \tag{3.67}$$

of Lemma 3.4. For $T \subset [d]$, $|T| = k-1$,

$$\begin{aligned}\Omega_1(T) &= \{S \subset [d] : |S| = k, T \subset S, \forall i \in S, \boldsymbol{V}_{:,i} \neq \boldsymbol{0}\}\\ \Omega_2(T) &= \left\{S \subset [d] : |S| = k, T \subset S, \mathrm{Vol}_k(\boldsymbol{V}_{:,S})^2 > 0\right\}.\end{aligned}$$

Note that $\Omega_2(T) \subset \Omega_1(T)$ so that

$$\begin{aligned}\sum_{\substack{S\subset[d],|S|=k\\ \mathrm{Vol}_k(\boldsymbol{V}_{:,S})^2>0}} \sum_{\substack{T\subset S\\ |T|=k-1}} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2) &= \sum_{\substack{T\subset[d]\\ |T|=k-1}} \sum_{S\in\Omega_2(T)} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2)\\ &\leq \sum_{\substack{T\subset[d]\\ |T|=k-1}} \sum_{S\in\Omega_1(T)} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2).\end{aligned}$$

The set $\Omega_1(T)$ has at most $(p-k+1)$ elements so that

$$\sum_{\substack{T\subset[d]\\ |T|=k-1}} \sum_{S\in\Omega_1(T)} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2) \leq (p-k+1) \sum_{\substack{T\subset[d]\\ |T|=k-1}} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2). \tag{3.68}$$

Lemma 3.2 for $\ell = k-1$ further yields

$$(p-k+1) \sum_{\substack{T\subset[d]\\ |T|=k-1}} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2) \leq (p-k+1)\, e_{k-1}(\Sigma(\boldsymbol{V})^2). \tag{3.69}$$

□

3.7.2 *Proof of Proposition 3.5*

First, Lemma 3.1 yields

$$\sum_{S\subset[d],|S|=k} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 \|\boldsymbol{X}-\boldsymbol{\Pi}_S^{\nu}\boldsymbol{X}\|_{\nu}^2 \leq \sum_{S\subset[d],|S|=k} \frac{1}{\sigma_k^2(\boldsymbol{V}_{S,[k]})} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 \, \|\boldsymbol{X}-\boldsymbol{\Pi}_k\boldsymbol{X}\|_{\nu}^2$$
$$= \|\boldsymbol{X}-\boldsymbol{\Pi}_k\boldsymbol{X}\|_{\nu}^2 \sum_{S\subset[d],|S|=k} \prod_{\ell=1}^{k-1} \sigma_\ell^2(\boldsymbol{V}_{S,[k]}), \tag{3.70}$$

where the last equality follows from

$$\mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 = \prod_{\ell=1}^{k} \sigma_\ell^2(\boldsymbol{V}_{S,[k]}). \tag{3.71}$$

By definition of the polynomial e_{k-1}, it further holds

$$\prod_{\ell=1}^{k-1} \sigma_\ell^2(\boldsymbol{V}_{S,[k]}) \leq e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2), \tag{3.72}$$

so that (3.70) leads to

$$\sum_{S\subset[d],|S|=k} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 \|\boldsymbol{X}-\boldsymbol{\Pi}_S^{\nu}\boldsymbol{X}\|_{\nu}^2 \leq \|\boldsymbol{X}-\boldsymbol{\Pi}_k\boldsymbol{X}\|_{\nu}^2 \sum_{S\subset[d],|S|=k} e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2). \tag{3.73}$$

Now, Lemma 3.2 applied to the matrix $\boldsymbol{V}_{S,[k]}^{\mathsf{T}}$ gives

$$e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) = \sum_{T\subset S,|T|=k-1} e_{k-1}(\Sigma(\boldsymbol{V}_{T,[k]})^2), \tag{3.74}$$

Therefore, Lemma 3.4 yields

$$\sum_{S\subset[d],|S|=k} e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) \leq (d-k+1) \sum_{T\subset[d],|T|=k-1} e_{k-1}(\Sigma(\boldsymbol{V}_{T,[k]})^2). \tag{3.75}$$

Using Lemma 3.2 and the fact that $\boldsymbol{V}_k$ is orthogonal, we finally write

$$\sum_{T\subset[d],|T|=k-1} e_{k-1}(\Sigma(\boldsymbol{V}_{T,[k]})^2) = e_{k-1}(\Sigma(\boldsymbol{V}_k)^2) = k. \tag{3.76}$$

Plugging (3.76) into (3.75), and then into (3.73) concludes the proof of Proposition 3.5.

3.7.3 *Proof of Proposition 3.6*

We first prove the Frobenius norm bound, which requires more work. The spectral bound is easier and uses a subset of the arguments for the Frobenius norm.

Frobenius norm bound

Recall that $\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{\Pi}_k\boldsymbol{X}$. We start with Lemma 3.1:

$$\begin{aligned} \|\boldsymbol{X}-\boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 &\leq \|\boldsymbol{E}(\boldsymbol{I}-\boldsymbol{P}_S)\|_{\mathrm{Fr}}^2 \\ &= \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \mathrm{Tr}(\boldsymbol{E}^{\mathsf{T}}\boldsymbol{E}\boldsymbol{P}_S\boldsymbol{P}_S^{\mathsf{T}}) - 2\,\mathrm{Tr}(\boldsymbol{P}_S^{\mathsf{T}}\boldsymbol{E}^{\mathsf{T}}\boldsymbol{E}). \end{aligned} \tag{3.77}$$

Since $\boldsymbol{E}^\intercal \boldsymbol{E} = \boldsymbol{V}_{k^\perp} \boldsymbol{\Sigma}_{k^\perp}^2 \boldsymbol{V}_{k^\perp}^\intercal$ and $\boldsymbol{P}_S = \boldsymbol{S}(\boldsymbol{V}_k^\intercal \boldsymbol{S})^{-1} \boldsymbol{V}_k^\intercal$,

$$\begin{aligned} \operatorname{Tr}(\boldsymbol{P}_S^\intercal \boldsymbol{E}^\intercal \boldsymbol{E}) &= \operatorname{Tr}\left(\boldsymbol{V}_k \left((\boldsymbol{V}_k^\intercal \boldsymbol{S})^\intercal \right)^{-1} \boldsymbol{S}^\intercal \boldsymbol{V}_{k^\perp} \boldsymbol{\Sigma}_{k^\perp} \boldsymbol{V}_{k^\perp}^\intercal \right) \\ &= \operatorname{Tr}\left(\boldsymbol{V}_{k^\perp}^\intercal \boldsymbol{V}_k \left((\boldsymbol{V}_k^\intercal \boldsymbol{S})^\intercal \right)^{-1} \boldsymbol{S}^\intercal \boldsymbol{V}_{k^\perp} \boldsymbol{\Sigma}_{k^\perp} \right) \\ &= 0, \end{aligned} \tag{3.78}$$

where the last equality follows from $\boldsymbol{V}_{k^\perp}^\intercal \boldsymbol{V}_k = \boldsymbol{0}$. Therefore, (3.77) becomes

$$\|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \operatorname{Tr}(\boldsymbol{E}^\intercal \boldsymbol{E} \boldsymbol{P}_S \boldsymbol{P}_S^\intercal). \tag{3.79}$$

Taking expectations,

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \sum_{S \subset [d], |S| = k} \operatorname{Det}(\boldsymbol{V}_{S,[k]})^2 \operatorname{Tr}(\boldsymbol{E}^\intercal \boldsymbol{E} \boldsymbol{P}_S \boldsymbol{P}_S^\intercal). \tag{3.80}$$

Proposition 3.8 expresses $\operatorname{Det}(\boldsymbol{V}_{S,[k]})^2$ as a function of the principal angles $(\theta_i(S))$ between $\operatorname{Span}(\boldsymbol{V}_k)$ and $\operatorname{Span}(\boldsymbol{S})$, namely

$$\operatorname{Det}(\boldsymbol{V}_{S,[k]})^2 = \prod_{i \in [k]} \cos^2(\theta_i(S)). \tag{3.81}$$

The remainder of the proof is in two steps. First, we bound the second factor in the sum in the right-hand side of (3.80) with a similar geometric expression. This allows trigonometric manipulations. Second, we work our way back to elementary symmetric polynomials of spectra, and we conclude after some simple algebra.

First, for $S \subset [d]$, $|S| = k$, let

$$\boldsymbol{Z}_S = \boldsymbol{V}_{k^\perp}^\intercal \boldsymbol{S} (\boldsymbol{V}_k^\intercal \boldsymbol{S})^{-1} = \boldsymbol{V}_{k^\perp}^\intercal \boldsymbol{P}_S \boldsymbol{V}_k.$$

It allows us to write

$$\operatorname{Tr}(\boldsymbol{E}^\intercal \boldsymbol{E} \boldsymbol{P}_S \boldsymbol{P}_S^\intercal) = \operatorname{Tr}(\boldsymbol{V}_{k^\perp} \boldsymbol{\Sigma}_{k^\perp}^2 \boldsymbol{V}_{k^\perp}^\intercal \boldsymbol{P}_S \boldsymbol{P}_S^\intercal) = \operatorname{Tr}(\boldsymbol{\Sigma}_{k^\perp}^2 \boldsymbol{Z}_S \boldsymbol{V}_k \boldsymbol{V}_k^\intercal \boldsymbol{Z}_S^\intercal). \tag{3.82}$$

However, for real symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with the same size, a simple diagonalization argument yields

$$\operatorname{Tr}(\boldsymbol{A}\boldsymbol{B}) \leq \|\boldsymbol{A}\|_2 \operatorname{Tr}(\boldsymbol{B}), \tag{3.83}$$

so that

$$\begin{aligned} \operatorname{Tr}(\boldsymbol{E}^\intercal \boldsymbol{E} \boldsymbol{P}_S \boldsymbol{P}_S^\intercal) &= \operatorname{Tr}(\boldsymbol{\Sigma}_{k^\perp}^2 \boldsymbol{Z}_S \boldsymbol{V}_k \boldsymbol{V}_k^\intercal \boldsymbol{Z}_S^\intercal) \\ &= \operatorname{Tr}(\boldsymbol{Z}_S^\intercal \boldsymbol{\Sigma}_{k^\perp}^2 \boldsymbol{Z}_S \boldsymbol{V}_k \boldsymbol{V}_k^\intercal) \\ &\leq \operatorname{Tr}(\boldsymbol{Z}_S^\intercal \boldsymbol{\Sigma}_{k^\perp}^2 \boldsymbol{Z}_S) \|\boldsymbol{V}_k \boldsymbol{V}_k^\intercal\|_2 \\ &\leq \operatorname{Tr}(\boldsymbol{Z}_S^\intercal \boldsymbol{\Sigma}_{k^\perp}^2 \boldsymbol{Z}_S) \\ &\leq \|\boldsymbol{\Sigma}_{k^\perp}^2\|_2 \operatorname{Tr}(\boldsymbol{Z}_S \boldsymbol{Z}_S^\intercal) \\ &\leq \sigma_{k+1}^2 \operatorname{Tr}(\boldsymbol{Z}_S \boldsymbol{Z}_S^\intercal). \end{aligned} \tag{3.84}$$

In Appendix 3.A, we characterize $\operatorname{Tr}(\boldsymbol{Z}_S \boldsymbol{Z}_S^\intercal)$ using principal angles, see (3.135). This reads

$$\operatorname{Tr}(\boldsymbol{Z}_S \boldsymbol{Z}_S^\intercal) = \sum_{j \in [k]} \tan^2(\theta_j(S)). \tag{3.85}$$

Combining (3.80), (3.84), (3.81), and (3.85), we obtain the following intermediate bound

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \sigma_{k+1}^2 \sum_{S \subset [d], |S|=k} \left[\prod_{i \in [k]} \cos^2(\theta_i(S)) \right] \left[\sum_{j \in [k]} \tan^2(\theta_j(S)) \right]. \tag{3.86}$$

Distributing the sum and using trigonometric identities, the general term of the sum in (3.86) becomes

$$\begin{aligned} \left[\prod_{i \in [k]} \cos^2(\theta_i(S)) \right] \left[\sum_{j \in [k]} \tan^2(\theta_j(S)) \right] &= \sum_{i \in [k]} (1 - \cos^2(\theta_i(S))) \prod_{j \in [k], j \neq i} \cos^2(\theta_j(S)) \\ &= \sum_{i \in [k]} \prod_{j \in [k], j \neq i} \cos^2(\theta_j(S)) - \sum_{i \in [k]} \prod_{j \in [k]} \cos^2(\theta_j(S)). \end{aligned} \tag{3.87}$$

The $(\cos(\theta_j(S)))_{j \in [k]}$ are the singular values of the matrix $\boldsymbol{V}_{S,[k]}$ so that

$$\sum_{i \in [k]} \prod_{j \in [k], j \neq i} \cos^2(\theta_j(S)) = e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2), \tag{3.88}$$

and

$$\prod_{j \in [k]} \cos^2(\theta_j(S)) = e_k(\Sigma(\boldsymbol{V}_{S,[k]})^2). \tag{3.89}$$

Back to (3.87), one gets

$$\begin{aligned} \left[\prod_{i \in [k]} \cos^2(\theta_i(S)) \right] \left[\sum_{j \in [k]} \tan^2(\theta_j(S)) \right] &= e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) - \sum_{i \in [k]} e_k(\Sigma(\boldsymbol{V}_{S,[k]})^2) \\ &= e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) - k e_k(\Sigma(\boldsymbol{V}_{S,[k]})^2). \end{aligned} \tag{3.90}$$

Thus, plugging (3.90) back into the intermediate bound (3.86), it comes

$$\begin{aligned} &\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \\ &\qquad \leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \sigma_{k+1}^2 \left[\sum_{\substack{S \subset [d] \\ |S|=k}} e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) - k \sum_{\substack{S \subset [d] \\ |S|=k}} e_k(\Sigma(\boldsymbol{V}_{S,[k]})^2) \right]. \end{aligned} \tag{3.91}$$

Using Lemma 3.2 twice, it comes

$$\begin{aligned} &\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \\ &\qquad \leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \sigma_{k+1}^2 \left[\sum_{\substack{S \subset [d] \\ |S|=k}} \sum_{\substack{T \subset S \\ |T|=k-1}} e_{k-1}(\Sigma(\boldsymbol{V}_{T,[k]})^2) - k e_k(\Sigma(\boldsymbol{V}_{:,[k]})^2) \right]. \end{aligned} \tag{3.92}$$

Lemmas 3.4 and the identities $e_{k-1}(\Sigma(\boldsymbol{V}_{:,[k]})^2) = k$ and $e_k(\Sigma(\boldsymbol{V}_{:,[k]})^2) = 1$ allow us to conclude

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \sigma_{k+1}^2 \left[(p-k+1)e_{k-1}(\Sigma(\boldsymbol{V}_{:,[k]})^2) - k\right] \tag{3.93}$$

$$= \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \sigma_{k+1}^2 (p-k)k. \tag{3.94}$$

By definition of β (3.40), we have proven (3.42), i.e.,

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 \left(1 + \beta \frac{p-k}{d-k} k\right).$$

Spectral norm bound

The bound in spectral norm is easier to derive. We start with Lemma 3.1:

$$\begin{aligned} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^2 \boldsymbol{X}\|_2^2 &\leq \|\boldsymbol{E}(\boldsymbol{I} - \boldsymbol{P}_S)\|_2^2 \\ &\leq \|\boldsymbol{E}\|_2^2 + \|\boldsymbol{E}\boldsymbol{P}_S\|_2^2 \\ &\leq \|\boldsymbol{E}\|_2^2 + \|\boldsymbol{E}\|_2^2 \|\boldsymbol{V}_{k^\perp}^\mathsf{T} \boldsymbol{S}(\boldsymbol{V}_k^\mathsf{T} \boldsymbol{S})^{-1} \boldsymbol{V}_k^\mathsf{T}\|_2^2 \\ &\leq \|\boldsymbol{E}\|_2^2 (1 + \|\boldsymbol{Z}_S\|_2^2), \end{aligned} \tag{3.95}$$

where the notation is the same as in Section 3.7.3. Now

$$\|\boldsymbol{Z}_S\|_2^2 \leq \|\boldsymbol{Z}_S\|_{\mathrm{Fr}}^2 = \sum_{i \in [k]} \tan^2(\theta_i(S)), \tag{3.96}$$

thus by (3.95), (3.96) and (3.81)

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^2 \boldsymbol{X}\|_2^2 = \sum_{S \subset [d], |S|=k} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 \|\boldsymbol{X} - \boldsymbol{\Pi}_S \boldsymbol{X}\|_2^2 \tag{3.97}$$

$$\leq \|\boldsymbol{E}\|_2^2 \left(1 + \sum_{\substack{S \subset [d], |S|=k \\ \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 > 0}} \prod_{i=1}^{k} \cos^2(\theta_i(S)) \sum_{i \in [k]} \tan^2(\theta_i(S))\right). \tag{3.98}$$

By (3.87), it comes

$$\begin{aligned} \mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \boldsymbol{\Pi}_S^2 \boldsymbol{X}\|_2^2 &\leq \|\boldsymbol{E}\|_2^2 \left(1 + \sum_{\substack{S \subset [d], |S|=k \\ \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 > 0}} e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) - k e_k(\Sigma(\boldsymbol{V}_{S,[k]})^2)\right) \\ &\leq \|\boldsymbol{E}\|_2^2 \left(1 + (p-k+1)\, e_{k-1}(\Sigma(\boldsymbol{V}_{:,[k]})^2) - k e_k(\Sigma(\boldsymbol{V}_{:,[k]})^2)\right) \\ &= (1 + (p-k)\, k)\, \|\boldsymbol{E}\|_2^2. \end{aligned}$$

where we again used the double sum trick of (3.92) and Lemma 3.4.

3.7.4 *Proof of Theorem 3.7*

We start with a lemma on evaluations of elementary symmetric polynomials on specific sequences.

Lemma 3.5. *Let $\boldsymbol{\lambda} \in]0,1]^k$ such that*

$$\begin{cases} \lambda_1 \geq \cdots \geq \lambda_k, \\ \Lambda = \sum_{i=1}^{k} \lambda_i \geq k - 1 + \frac{1}{\theta}. \end{cases} \tag{3.99}$$

Then, with the functions ϕ, ψ introduced in Lemma 3.3,

$$\begin{cases} \psi(\boldsymbol{\lambda}) & \geq \frac{1}{\theta}, \\ \phi(\boldsymbol{\lambda}) & \leq k - 1 + \theta. \end{cases} \tag{3.100}$$

Proof. Let $\hat{\boldsymbol{\lambda}} = (1, ..., 1, \Lambda - k + 1) \in \mathbb{R}_+^{*k}$. Then

$$\begin{cases} \lambda_1 \leq \hat{\lambda}_1 \\ \lambda_1 + \lambda_2 \leq \hat{\lambda}_1 + \hat{\lambda}_2 \\ \cdots \\ \sum_{i=1}^{k-1} \lambda_i \leq \sum_{i=1}^{k-1} \hat{\lambda}_i \\ \sum_{i=1}^{k} \lambda_i = \sum_{i=1}^{k} \hat{\lambda}_i \end{cases} \tag{3.101}$$

so that, according to Definition 3.4,

$$\boldsymbol{\lambda} \prec_S \hat{\boldsymbol{\lambda}}. \tag{3.102}$$

Lemma 3.3 ensures the Schur-convexity of ϕ and the Schur-concavity of ψ, so that

$$\phi(\boldsymbol{\lambda}) \leq \phi(\hat{\boldsymbol{\lambda}}) = k - 1 + \frac{1}{\Lambda - k + 1} \leq k - 1 + \theta,$$

and

$$\psi(\boldsymbol{\lambda}) \geq \psi(\hat{\boldsymbol{\lambda}}) = \Lambda - k + 1 \geq \frac{1}{\theta}.$$

□

Frobenius norm bound

Let $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^\intercal$, and π be a permutation of $[d]$ that reorders the leverage scores decreasingly,

$$\ell_{\pi_1}^k \geq \ell_{\pi_2}^k \geq ... \geq \ell_{\pi_d}^k. \tag{3.103}$$

By construction, $T_{p_{\text{eff}}} = [\pi_{p_{\text{eff}}}, ..., \pi_d]$ thus collects the indices of the smallest leverage scores. Finally, denoting by $\boldsymbol{\Pi} = (\delta_{i,\pi_j})_{(i,j)\in[d]\times[d]}$ the matricial representation of permutation π, we let

$$\boldsymbol{K}^\pi = \boldsymbol{\Pi} \boldsymbol{K} \boldsymbol{\Pi}^\intercal = ((\boldsymbol{K}_{\pi_i,\pi_j}))_{1\leq i,j\leq d}.$$

The goal of the proof is to bound

$$\mathbb{E}_{\mathrm{DPP}}\left[\|\boldsymbol{X}-\boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 | S\cap T_{p_{\mathrm{eff}}}=\emptyset\right] = \frac{\sum \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2\|\boldsymbol{X}-\boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2}{\sum \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2}, \tag{3.104}$$

where both sums run over subsets $S\subset[d]$ such that $|S|=k$ and $S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset$. For simplicity, let us write

$$Z_{k,p_{\mathrm{eff}}(\theta)} = \sum_{\substack{S\subset[d],|S|=k\\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset}} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2, \tag{3.105}$$

$$Y_{k,p_{\mathrm{eff}}(\theta)} = \sum_{\substack{S\subset[d],|S|=k\\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset}} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2\,\mathrm{Tr}(\boldsymbol{Z}_S\boldsymbol{Z}_S^\intercal). \tag{3.106}$$

Following steps (3.80) to (3.84) of the previous proof, one obtains

$$\mathbb{E}_{\mathrm{DPP}}\left[\|\boldsymbol{X}-\boldsymbol{\Pi}_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 \mid S\cap T_{p_{\mathrm{eff}}}=\emptyset\right] \le \|\boldsymbol{X}-\boldsymbol{\Pi}_k\boldsymbol{X}\|_{\mathrm{Fr}}^2 + \sigma_{k+1}^2\frac{Y_{k,p_{\mathrm{eff}}(\theta)}}{Z_{k,p_{\mathrm{eff}}(\theta)}}. \tag{3.107}$$

By definition (3.40) of the flatness parameter β,

$$\sigma_{k+1}^2 = \beta\frac{1}{d-k}\sum_{j\ge k+1}\sigma_j^2 = \beta\frac{1}{d-k}\|\boldsymbol{X}-\boldsymbol{\Pi}_k\boldsymbol{X}\|_{\mathrm{Fr}}^2. \tag{3.108}$$

Then, it remains to upper bound the ratio $Y_{k,p_{\mathrm{eff}}(\theta)}/Z_{k,p_{\mathrm{eff}}(\theta)}$ in (3.107), which is the important part of the proof. We first evaluate $Z_{k,p_{\mathrm{eff}}(\theta)}$ and then bound $Y_{k,p_{\mathrm{eff}}(\theta)}$.

The matrix $\boldsymbol{\Pi}\boldsymbol{V}_k\in\mathbb{R}^{d\times k}$ has its rows ordered by decreasing leverage scores. Let $\tilde{\boldsymbol{V}}^{\pi}_{p_{\mathrm{eff}}(\theta)}\in\mathbb{R}^{p_{\mathrm{eff}}(\theta)\times k}$ be the submatrix corresponding to the first $p_{\mathrm{eff}}(\theta)$ rows of $\boldsymbol{\Pi}\boldsymbol{V}_k$. Let also

$$\hat{\boldsymbol{V}}^{\pi}_{p_{\mathrm{eff}}(\theta)} = \begin{pmatrix}\tilde{\boldsymbol{V}}_{\pi,p_{\mathrm{eff}}(\theta)}\\ \boldsymbol{0}_{d-p_{\mathrm{eff}}(\theta),k}\end{pmatrix}$$

be padded with zeros. Then

$$\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)} = \left[\begin{array}{c|c}\tilde{\boldsymbol{V}}_{\pi,p_{\mathrm{eff}}(\theta)}\tilde{\boldsymbol{V}}_{\pi,p_{\mathrm{eff}}(\theta)}^\intercal & \boldsymbol{0}\\ \hline \boldsymbol{0} & \boldsymbol{0}\end{array}\right] = \hat{\boldsymbol{V}}^{\pi}_{p_{\mathrm{eff}}(\theta)}(\hat{\boldsymbol{V}}^{\pi}_{p_{\mathrm{eff}}(\theta)})^\intercal\in\mathbb{R}^{d\times d}. \tag{3.109}$$

The nonzero block of $\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)}$ is a submatrix of $\boldsymbol{K}^{\pi}$, and $\mathrm{rk}\,\boldsymbol{K}^{\pi}=\mathrm{rk}\,\boldsymbol{K}=k$. Hence $\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)}$ has at most k nonzero eigenvalues

$$\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_k\ge 0=\lambda_{k+1}=\cdots=\lambda_d. \tag{3.110}$$

Therefore,

$$e_k(\Lambda(\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)})) = \sum_{\substack{T\subset[d]\\|T|=k}}\prod_{j\in T}\lambda_j = \prod_{i\in[k]}\lambda_i. \tag{3.111}$$

Note moreover that

$$\forall\ell\in[k],\ e_\ell(\Sigma(\hat{\boldsymbol{V}}_{\pi,p_{\mathrm{eff}}(\theta)})^2) = e_\ell(\Lambda(\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)})). \tag{3.112}$$

By construction,

$$Z_{k,p_{\text{eff}}(\theta)} = \sum_{\substack{S\subset[d],|S|=k\\ S\cap T_{p_{\text{eff}}(\theta)}=\emptyset}} \text{Det}(\boldsymbol{V}_{S,[k]})^2 = \sum_{S\subset[d],|S|=k} \text{Det}\left[\left(\hat{\boldsymbol{V}}^{\pi}_{p_{\text{eff}}(\theta)}\right)_{S,:}\right]^2 \tag{3.113}$$

Then, Lemma 3.2 yields

$$Z_{k,p_{\text{eff}}(\theta)} = e_k(\Sigma(\hat{\boldsymbol{V}}_{\pi,p_{\text{eff}}(\theta)})^2) = e_k(\Lambda(\boldsymbol{K}^{\pi}_{p_{\text{eff}}(\theta)})) = \prod_{i\in[k]} \lambda_i. \tag{3.114}$$

Now we bound $Y_{k,p_{\text{eff}}(\theta)}$. We use again principal angles and trigonometric identities. Using (3.85) and (3.90) above, it holds

$$\begin{aligned}
Y_{k,p_{\text{eff}}(\theta)} &= \sum_{\substack{S\subset[d],|S|=k\\ S\cap T_{p_{\text{eff}}(\theta)}=\emptyset}} \text{Det}(\boldsymbol{V}_{S,[k]})^2 \,\text{Tr}(\boldsymbol{Z}_S\boldsymbol{Z}_S^\intercal) \\
&= \sum_{\substack{S\subset[d],|S|=k\\ S\cap T_{p_{\text{eff}}(\theta)}=\emptyset}} \prod_{i\in[k]} \cos^2(\theta_i(S)) \sum_{j\in[k]} \tan^2(\theta_j(S)) \\
&= \sum_{\substack{S\subset[d],|S|=k\\ S\cap T_{p_{\text{eff}}(\theta)}=\emptyset}} e_{k-1}\left(\Sigma(\boldsymbol{V}_{S,[k]})^2\right) - k\, e_k\left(\Sigma(\boldsymbol{V}_{S,[k]}\right)^2 \qquad (3.115) \\
&= \sum_{S\subset[d],|S|=k} e_{k-1}\left(\Sigma\left(\left[\hat{\boldsymbol{V}}^{\pi}_{p_{\text{eff}}(\theta)}\right]_{S,:}\right)^2\right) - k\, e_k\left(\Sigma\left(\left[\hat{\boldsymbol{V}}^{\pi}_{p_{\text{eff}}(\theta)}\right]_{S,:}\right)^2\right) \qquad (3.116)
\end{aligned}$$

By Lemma 3.4 applied to the matrix $\hat{\boldsymbol{V}}_{\pi,p_{\text{eff}}(\theta)}$ combined to (3.113), we get

$$\begin{aligned}
Y_{k,p_{\text{eff}}(\theta)} &\le (p_{\text{eff}}(\theta)-k+1)e_{k-1}(\Sigma(\hat{\boldsymbol{V}}^{\pi}_{p_{\text{eff}}(\theta)})^2) - k\, e_k(\Sigma(\hat{\boldsymbol{V}}^{\pi}_{p_{\text{eff}}(\theta)})^2) \\
&\le (p_{\text{eff}}(\theta)-k+1)e_{k-1}(\Lambda(\boldsymbol{K}^{\pi}_{p_{\text{eff}}(\theta)})) - k\, e_k(\Lambda(\boldsymbol{K}^{\pi}_{p_{\text{eff}}(\theta)})) \\
&\le \Big((p_{\text{eff}}(\theta)-k+1)\phi(\tilde{\boldsymbol{\lambda}}) - k\Big) Z_{k,p_{\text{eff}}(\theta)}. \qquad (3.117)
\end{aligned}$$

where $\tilde{\boldsymbol{\lambda}} = (1,\dots,1,\text{Tr}(\boldsymbol{K}^{\pi}_{p_{\text{eff}}(\theta)})-k+1) \in \mathbb{R}^k$, see Lemma 3.5. Now, as in the proof of Lemma 3.5,

$$\phi(\tilde{\boldsymbol{\lambda}}) = k-1+\frac{1}{\text{Tr}(\boldsymbol{K}^{\pi}_{p_{\text{eff}}(\theta)})-k+1} \le k-1+\theta$$

by (3.44). Thus (3.117) yields

$$\frac{Y_{k,p_{\text{eff}}(\theta)}}{Z_{k,p_{\text{eff}}(\theta)}} \le (p_{\text{eff}}(\theta)-k+1)(k-1+\theta)-k \le (p_{\text{eff}}(\theta)-k+1)(k-1+\theta). \tag{3.118}$$

Finally, plugging (3.118) and (3.108) in (3.107) concludes the proof of (3.47).

Spectral norm bound

We proceed as for the Frobenius norm, using the notation of Section 3.7.3. Lemma 3.1, Equations (3.115) and (3.118) yield

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{DPP}}\left[\|\boldsymbol{X}-\boldsymbol{\Pi}_S^2\boldsymbol{X}\|_2^2 \mid S\cap T_{p_{\mathrm{eff}}}=\emptyset\right] \\
&\qquad = Z_{k,p_{\mathrm{eff}}(\theta)}^{-1} \sum_{\substack{S\subset[d],|S|=k \\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset}} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 \|\boldsymbol{X}-\boldsymbol{\Pi}_S^2\boldsymbol{X}\|_2^2, \\
&\qquad \leq Z_{k,p_{\mathrm{eff}}(\theta)}^{-1} \|\boldsymbol{X}-\boldsymbol{\Pi}_k\boldsymbol{X}\|_2^2 \left(1+\sum_{\substack{S\subset[d],|S|=k \\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset, \\ \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2>0}} \prod_{\ell=1}^{k-1}\sigma_\ell^2(\boldsymbol{V}_{S,[k]}) - k e_k(\Sigma(\boldsymbol{V}_{S,[k]})^2)\right) \\
&\qquad \leq Z_{k,p_{\mathrm{eff}}(\theta)}^{-1} \|\boldsymbol{X}-\boldsymbol{\Pi}_k\boldsymbol{X}\|_2^2 \left(1+\sum_{\substack{S\subset[d],|S|=k \\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset \\ \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2>0}} e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) - k e_k(\Sigma(\boldsymbol{V}_{S,[k]})^2)\right) \\
&\qquad \leq \left(\frac{Y_{k,p_{\mathrm{eff}}(\theta)}}{Z_{k,p_{\mathrm{eff}}(\theta)}}+1\right)\|\boldsymbol{X}-\boldsymbol{\Pi}_k\boldsymbol{X}\|_2^2 \\
&\qquad \leq \left(1+(p_{\mathrm{eff}}(\theta)-k+1)(k-1+\theta)\right)\|\boldsymbol{X}-\boldsymbol{\Pi}_k\boldsymbol{X}\|_2^2,
\end{aligned}
$$

which is the claimed spectral bound.

Bounding the probability of rejection

Recall from Lemma 3.5 that

$$
\hat{\boldsymbol{\lambda}} = \begin{pmatrix} 1 & \dots & 1 & \sum_{i=1}^k \lambda_i - k + 1 \end{pmatrix} \in \mathbb{R}_+^{*k}.
$$

Still with the notation of Section 3.7.3, (3.113) yields

$$
\begin{aligned}
\mathbb{P}(S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset) &= \sum_{\substack{S\subset[d],|S|=k \\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset}} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 \\
&= e_k(\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)}) \qquad (3.119) \\
&= \prod_{i\in[k]}\lambda_i \\
&\geq \psi(\hat{\boldsymbol{\lambda}}), \qquad (3.120)
\end{aligned}
$$

because the normalization constant $\sum_{S\subset[d],|S|=k} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2$ is equal to 1. Lemma 3.5 concludes the proof since

$$
\psi(\hat{\boldsymbol{\lambda}}) \geq \frac{1}{\theta}. \qquad (3.121)
$$

3.7.5 *Proof of Proposition 3.8*

First, Proposition 3.3 gives

$$\mathcal{E}(\boldsymbol{w}_S) \leq \frac{(1 + \max_{i\in[k]} \tan^2 \theta_i(S))\|\boldsymbol{w}^*\|^2\sigma_{k+1}^2}{N} + \frac{k}{N}\nu. \tag{3.122}$$

Now (3.135) further gives

$$\max_{i\in[k]} \tan^2 \theta_i(S) \leq \sum_{i\in[k]} \tan^2 \theta_i(S) = \mathrm{Tr}(\boldsymbol{Z}_S \boldsymbol{Z}_S^\intercal). \tag{3.123}$$

The proof now follows the same lines as for the approximation bounds. First, following the lines of Section 3.7.3, we straightforwardly bound

$$\mathbb{E}_{\mathrm{DPP}} \sum_{i\in[k]} \tan^2(\theta_i(S)) = \sum_{S\subset[d],|S|=k} \prod_{i\in[k]} \cos^2(\theta_i(S)) \sum_{j\in[k]} \tan^2(\theta_j(S)) \tag{3.124}$$

and obtain (3.57). In a similar vein, the same lines as in Section 3.7.4 allow bounding

$$\mathbb{E}_{\mathrm{DPP}} \left[\sum_{i\in[k]} \tan^2(\theta_i(S)) \mid S \cap T_{p_{\mathrm{eff}}} = \emptyset \right] = \sum_{\substack{S\subset[d],|S|=k \\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset}} \prod_{i\in[k]} \cos^2(\theta_i(S)) \sum_{j\in[k]} \tan^2(\theta_j(S). \tag{3.125}$$

and yield (3.58).

APPENDIX

In Section 3.A we recall the definition of principal angles between subspaces along with some results that we use in the proofs of our theoretical results. In Section 3.B, we recall some notions of majorization and Schur-convexity necessary for the proofs. In Section 3.C, we give another possible interpretation of the k-leverage scores. In Section 3.D, we describe an algorithm that generates orthogonal matrices with a prescribed diagonal. This algorithm was used in Section 3.5 to construct matrices $\boldsymbol{X}$ with realistic right eigenvectors.

3.A PRINCIPAL ANGLES AND THE COSINE SINE DECOMPOSITION

3.A.1 *Principal angles*

This section surveys the notion of principal angles between subspaces, see (Golub and Van Loan, 1996, Section 6.4.3) for details.

Definition 3.3. *Let $\mathcal{P}, \mathcal{Q}$ be two subspaces in $\mathbb{R}^d$. Let $p = \dim \mathcal{P}$ and $q = \dim \mathcal{Q}$ and assume that $q \leq p$. To define the vector of principal angles $\boldsymbol{\theta} \in [0, \pi/2]^q$ between $\mathcal{P}$ and $\mathcal{Q}$, let*

$$\cos(\theta_1) = \max \left\{ \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}; \quad \boldsymbol{x} \in \mathcal{P}, \boldsymbol{y} \in \mathcal{Q} \right\} \tag{3.126}$$

be the cosine of the smallest angle between a vector of $\mathcal{P}$ and a vector of $\mathcal{Q}$, and let $(\boldsymbol{x}_1, \boldsymbol{y}_1) \in \mathcal{P} \times \mathcal{Q}$ be a pair of vectors realizing the maximum. For $i \in [2, q]$, define successively

$$\cos(\theta_i) = \max \left\{ \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}; \quad \boldsymbol{x} \in \mathcal{P}, \boldsymbol{y} \in \mathcal{Q}; \boldsymbol{x} \perp \boldsymbol{x}_j, \boldsymbol{y} \perp \boldsymbol{y}_j \,, \forall j \in [1 : i-1] \right\} \tag{3.127}$$

and denote $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{P} \times \mathcal{Q}$ such that $\cos(\theta_i) = \boldsymbol{x}_i^\intercal \boldsymbol{y}_i$.

Note that although the so-called principal vectors $(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i \in [q]}$ are not uniquely defined by (3.126) and (3.127), the principal angles $\boldsymbol{\theta}$ are uniquely defined, see (Björck and Golub, 1973). The following result confirms this, while also providing a way to compute $\boldsymbol{\theta}$.

Proposition 3.8 (Björck and Golub, 1973, Ben-Israel, 1992)**.** *Let $\mathcal{P}$ and $\mathcal{Q}$ and $\boldsymbol{\theta}$ be as in Definition 3.3. Let $\boldsymbol{P} \in \mathbb{R}^{d \times p}$, $\boldsymbol{Q} \in \mathbb{R}^{d \times q}$ be two orthogonal matrices, whose columns are orthonormal bases of $\mathcal{P}$ and $\mathcal{Q}$, respectively. Then*

$$\forall i \in [q], \quad \cos(\theta_i) = \sigma_i(\boldsymbol{Q}^\intercal \boldsymbol{P}). \tag{3.128}$$

In particular

$$\mathrm{Vol}_q^2(\boldsymbol{Q}^\intercal \boldsymbol{P}) = \prod_{i \in [q]} \cos^2(\theta_i). \tag{3.129}$$

An important case for our work arises when $q = k$, $\boldsymbol{Q} = \boldsymbol{V} \in \mathbb{R}^{d \times k}$, and $\boldsymbol{P} = \boldsymbol{S} \in \mathbb{R}^{d \times k}$ is a sampling matrix. The left-hand side of (3.129) then equals $\mathrm{Det}(\boldsymbol{V}_{:,S})^2$.

3.A.2 *The Cosine Sine decomposition*

The Cosine Sine (CS) decomposition is useful for the study of the relative position of two subspaces. It generalizes the notion of cosine, sine and tangent to subspaces. The tangent of principal angles between subspaces were first mentioned in (Zhu and Knyazev, 2013).

Proposition 3.9 (Golub and Van Loan, 1996)**.** *Let $q, d \in \mathbb{N}$ and $\boldsymbol{Q} = \left[\begin{array}{c} \boldsymbol{Q}_1 \\ \hline \boldsymbol{Q}_2 \end{array}\right]$ be a $d \times q$ orthogonal matrix, where $\boldsymbol{Q}_1 \in \mathbb{R}^{q\times q}$ and $\boldsymbol{Q}_2 \in \mathbb{R}^{(d-q)\times q}$. Assume that $\boldsymbol{Q}_1$ is non singular, then there exist orthogonal matrices $\boldsymbol{Y} \in \mathbb{R}^{d\times q}$ and*

$$\boldsymbol{W} = \left[\begin{array}{c|c} \boldsymbol{W}_1 & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{W}_2 \end{array}\right] \in \mathbb{R}^{d\times d}, \tag{3.130}$$

and a matrix

$$\boldsymbol{\Sigma} = \left[\begin{array}{c} \mathcal{C} \\ \hline \mathcal{S} \\ \hline \boldsymbol{0}_{q',q} \end{array}\right] \in \mathbb{R}^{d\times q}, \tag{3.131}$$

where $q' = \max(0, d-2q)$, such that

$$\boldsymbol{Q} = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{Y}^T, \tag{3.132}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{q\times q}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d-q\times d-q}$, and $\mathcal{C}, \mathcal{S} \in \mathbb{R}^{q\times q}$ are diagonal matrices satisfying the identity $\mathcal{C}^2 + \mathcal{S}^2 = \mathbb{I}_q$. In particular, each block $\boldsymbol{Q}_i$ factorizes as

$$\begin{aligned} \boldsymbol{Q}_1 &= \boldsymbol{W}_1 \mathcal{C}\, \boldsymbol{Y}^T \\ \boldsymbol{Q}_2 &= \boldsymbol{W}_2 \left[\begin{array}{c} \mathcal{S} \\ \hline \boldsymbol{0}_{q',q} \end{array}\right] \boldsymbol{Y}^T. \end{aligned} \tag{3.133}$$

The CS decomposition is defined for every orthogonal matrix. An important case is when $\boldsymbol{Q}$ is the product of an orthogonal matrix $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ and a sampling matrix $\boldsymbol{S} \in \mathbb{R}^{d\times k}$, that is $\boldsymbol{Q} = \boldsymbol{V}^\intercal \boldsymbol{S}$.

Corollary 3.1. *Let $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ be an orthogonal matrix and $\boldsymbol{S} \in \mathbb{R}^{d\times k}$ be a sampling matrix. Let*

$$\boldsymbol{Q} = \boldsymbol{V}^\intercal \boldsymbol{S} = \left[\begin{array}{c} \boldsymbol{V}_k^\intercal \boldsymbol{S} \\ \hline \boldsymbol{V}_{k^\perp}^\intercal \boldsymbol{S} \end{array}\right] \tag{3.134}$$

be a $d \times k$ orthogonal matrix, with $\mathrm{Det}(\boldsymbol{V}_k^\intercal \boldsymbol{S})^2 > 0$. Let further $\boldsymbol{Z}_S = \boldsymbol{V}_{k^\perp}^\intercal \boldsymbol{S}(\boldsymbol{V}_k^\intercal \boldsymbol{S})^{-1}$. Then

$$\mathrm{Tr}(\boldsymbol{Z}_S \boldsymbol{Z}_S^\intercal) = \sum_{i\in[k]} \tan^2(\theta_i(S)), \tag{3.135}$$

where the $(\theta_i(S))_{i\in[k]}$ are the principal angles between $\mathrm{Span}(\boldsymbol{V}_k)$ and $\mathrm{Span}(\boldsymbol{S})$.

Proof. Proposition 3.9 applied to the matrix $\boldsymbol{Q} = \boldsymbol{V}^\intercal \boldsymbol{S}$ with $\boldsymbol{Q}_1 = \boldsymbol{V}_k^\intercal \boldsymbol{S}$ and $\boldsymbol{Q}_2 = \boldsymbol{V}_{k^\perp}^\intercal \boldsymbol{S}$ yields

$$\boldsymbol{Q}_1 = \boldsymbol{W}_1 \mathcal{C}\, \boldsymbol{Y}^T \tag{3.136}$$

$$\boldsymbol{Q}_2 = \boldsymbol{W}_2 \left[\begin{array}{c} \mathcal{S} \\ \hline \boldsymbol{0}_{q',q} \end{array}\right] \boldsymbol{Y}^T. \tag{3.137}$$

Thus, the diagonal matrix $\mathcal{C}$ contains the singular values of the matrix $\boldsymbol{V}_k^\intercal \boldsymbol{S}$, which are cosines of the principal angles $(\theta_i(S))_{i\in[k]}$ between $\mathrm{Span}(\boldsymbol{V}_k)$ and $\mathrm{Span}(\boldsymbol{S})$, see Proposition 3.8. The identity $\mathcal{C}^2 + \mathcal{S}^2 = \mathbb{I}_k$ and the fact that $\theta_i(S) \in [0, \frac{\pi}{2}]$ imply that the (diagonal) elements of $\mathcal{S}$ are equal to the sines of the principal angles between $\mathrm{Span}(\boldsymbol{V}_k)$ and $\mathrm{Span}(\boldsymbol{S})$. Let $\mathcal{T} = \mathcal{S}\,\mathcal{C}^{-1} \in \mathbb{R}^{k\times k}$ be the diagonal matrix containing the tangents of the principal angles $(\theta_i(S))_{i\in[k]}$ on its diagonal. Using (3.136) and (3.137), it comes

$$\begin{aligned}\boldsymbol{Z}_S = \boldsymbol{V}_{k^\perp}^\intercal \boldsymbol{S}(\boldsymbol{V}_k^\intercal \boldsymbol{S})^{-1} &= \boldsymbol{W}_2 \begin{bmatrix} \mathcal{S} \\ \hline \boldsymbol{0}_{q',q} \end{bmatrix} \boldsymbol{Y}^\intercal \boldsymbol{Y}\, \mathcal{C}^{-1}\, \boldsymbol{W}_1^\intercal \\ &= \boldsymbol{W}_2 \begin{bmatrix} \mathcal{S} \\ \hline \boldsymbol{0}_{q',q} \end{bmatrix} \mathcal{C}^{-1}\, \boldsymbol{W}_1^\intercal = \boldsymbol{W}_2 \begin{bmatrix} \mathcal{S}\,\mathcal{C}^{-1} \\ \hline \boldsymbol{0}_{q',q} \end{bmatrix} \boldsymbol{W}_1^\intercal. \end{aligned} \tag{3.138}$$

Then,

$$\mathrm{Tr}(\boldsymbol{Z}_S \boldsymbol{Z}_S^\intercal) = \mathrm{Tr}(\boldsymbol{W}_2 \left[\begin{array}{c|c} \mathcal{T}^2 & \boldsymbol{0}_{q,q'} \\ \hline \boldsymbol{0}_{q',q} & \boldsymbol{0}_{q',q'} \end{array}\right] \boldsymbol{W}_2^\intercal) = \sum_{i\in[k]} \tan^2(\theta_i(S)). \tag{3.139}$$

□

3.B MAJORIZATION AND SCHUR CONVEXITY

This section recalls some definitions and results from the theory of majorization and the notions of Schur-convexity and Schur-concavity. We refer to (Marshall et al., 2011) for further details. In this section, a subset $\mathcal{D} \subset \mathbb{R}^d$ is a symmetric domain if $\mathcal{D}$ is stable under coordinate permutations. Furthermore, a function f defined on a symmetric domain $\mathcal{D}$ is called symmetric if it is stable under coordinate permutations.

Definition 3.4. *Let $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}_+^d$. $\boldsymbol{p}$ is said to majorize $\boldsymbol{q}$ according to Schur order and we note $\boldsymbol{q} \prec_S \boldsymbol{p}$ if*

$$\begin{cases} q_{i_1} \leq p_{j_1} \\ q_{i_1} + q_{i_2} \leq p_{j_1} + p_{j_2} \\ \dots \\ \sum_{k=1}^{d-1} q_{i_k} \leq \sum_{k=1}^{d-1} p_{j_k} \\ \sum_{k=1}^{d} q_{i_k} = \sum_{k=1}^{d} p_{j_k} \end{cases} \tag{3.140}$$

where $\boldsymbol{p}, \boldsymbol{q}$ are reordered so that $p_{i_d} \leq \dots \leq p_{i_1}$ and $q_{j_d} \leq \dots \leq q_{j_1}$.

The majorization order has an algebraic characterization using doubly stochastic matrices first proven by Hardy, Littlewood, and Polya in 1929.

Proposition 3.10 (Theorem B.2. Marshall et al., 2011)**.** *The vector $\boldsymbol{p}$ majorizes the vector $\boldsymbol{q}$ if and only if there exists a $d \times d$ doubly stochastic matrix Π such that $\boldsymbol{q} = \boldsymbol{p}\boldsymbol{\Pi}$.*

Example 3.1. *Let $\boldsymbol{p} = (3, 0, 0)$ and $\boldsymbol{q} = (1, 1, 1)$. We check easily that $\boldsymbol{p}$ majorizes $\boldsymbol{q}$. Note that we can 'redistribute' $\boldsymbol{p}$ over $\boldsymbol{q}$ as follows: $\boldsymbol{q} = \frac{1}{3}\boldsymbol{J}\boldsymbol{p}$, where $\boldsymbol{J}$ is a 3×3 matrix of ones. The matrix $\boldsymbol{\Pi} = \frac{1}{3}\boldsymbol{J}$ is a doubly stochastic matrix.*

Schur order compares two vectors using multiple inequalities. To avoid such cumbersome calculations, a scalar metric of inequality in a vector is desired. This is possible using the notion of Schur-convex/concave function.

Definition 3.5. *Let f be a function on a symmetric domain $\mathcal{D} \subset \mathbb{R}_+^d$.*
f is said to be Schur convex if

$$\forall \boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}_+^d, \boldsymbol{q} \prec_S \boldsymbol{p} \implies f(\boldsymbol{q}) \leq f(\boldsymbol{p}). \tag{3.141}$$

f is said to be Schur concave if

$$\forall \boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}_+^d, \boldsymbol{q} \prec_S \boldsymbol{p} \implies f(\boldsymbol{q}) \geq f(\boldsymbol{p}). \tag{3.142}$$

Proposition 3.11 (Theorem A.3, Marshall et al., 2011)**.** *Let f be a symmetric function defined on $\mathbb{R}_+^d$, let $\mathcal{D}$ be a permutation-symmetric domain in $\mathbb{R}_+^d$ and suppose that*

$$\forall x_i, x_j \in \mathbb{R}_+, (x_i - x_j)\left(\frac{\partial f}{\partial x_i} - \frac{\partial f}{\partial x_j}\right) > 0 \tag{3.143}$$

then

$$\forall \boldsymbol{p}, \boldsymbol{q} \in \mathcal{D}, \boldsymbol{q} \prec_S \boldsymbol{p} \implies f(\boldsymbol{q}) \leq f(\boldsymbol{p}), \tag{3.144}$$

and f is Schur convex.

We get a similar result for Schur concavity by switching the orders in the previous proposition.

3.C ANOTHER INTERPRETATION OF THE k-LEVERAGE SCORES

For $i \in [d]$, the SVD of $\boldsymbol{X}$ yields

$$\boldsymbol{X}_{:,i} = \sum_{\ell=1}^{r} V_{i,\ell} \boldsymbol{f}_\ell, \tag{3.145}$$

where $\boldsymbol{f}_\ell = \sigma_\ell \boldsymbol{U}_{:,\ell}$, $\ell \in [r]$, are orthogonal. Thus

$$\boldsymbol{X}_{:,i}^\intercal \boldsymbol{f}_j = V_{i,j} \|\boldsymbol{f}_j\|^2 = V_{i,j} \sigma_j^2. \tag{3.146}$$

Then

$$\frac{V_{i,j}}{\|\boldsymbol{X}_{:,i}\|} = \frac{\boldsymbol{X}_{:,i}^\intercal \boldsymbol{f}_j}{\sigma_j \|\boldsymbol{X}_{:,i}\| \|\boldsymbol{f}_j\|} =: \frac{\cos \eta_{i,j}}{\sigma_j}, \tag{3.147}$$

where $\eta_{i,j} \in [0, \pi/2]$ is the angle formed by $\boldsymbol{X}_{:,i}$ and $\boldsymbol{f}_j$. Finally, (3.146) also yields

$$\ell_i^k = \|\boldsymbol{X}_{:,i}\|^2 \sum_{j=1}^{k} \frac{\cos^2 \eta_{i,j}}{\sigma_j^2}. \tag{3.148}$$

Compared to the length-square distribution in Section 3.2, k-leverage scores thus favour columns that are aligned with the principal features. The weight $1/\sigma_j^2$ corrects the fact that features associated with large singular values are typically aligned with more columns. One could also imagine more arbitrary weights w_j/σ_j^2 in lieu of $1/\sigma_j^2$, or, equivalently, modified k-leverage scores

$$\ell_i^k(\boldsymbol{w}) = \sum_{j=1}^{k} w_j V_{i,j}^2.$$

However, the projection DPP with marginal kernel $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^\intercal$ that we study in this paper is invariant to such reweightings. Indeed, let Y be a random subset of $[d]$ following the distribution of the k-DPP of kernel $\boldsymbol{K}_{\boldsymbol{w}} = \boldsymbol{V}_k \mathrm{Diag}(\boldsymbol{w}_{[k]}) \boldsymbol{V}_k^\intercal$ such that for all $i \in [k]$, $w_i \neq 0$. For any $S \subset [d]$ of cardinality k,

$$\mathbb{P}(Y = S) \propto \mathrm{Det}\left[\boldsymbol{V}_{S,[k]} \,\mathrm{Diag}(\boldsymbol{w}_{[k]})\, \boldsymbol{V}_{[k],S}^\intercal\right] = \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 \prod_{j\in[k]} w_j^2 \propto \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2. \quad (3.149)$$

Such a scaling is thus not a free parameter in $\boldsymbol{K}$.

3.D GENERATING ORTHOGONAL MATRICES WITH PRESCRIBED LEVERAGE SCORES

In this section, we describe an algorithm that samples a random orthonormal matrix with a prescribed profile of k-leverage scores. This algorithm was used to generate the matrices $\boldsymbol{F} = \boldsymbol{V}_k^\intercal \in \mathbb{R}^{k\times d}$ for the toy datasets of Section 3.5. The orthogonality constraint can be expressed as a condition on the spectrum of the matrix $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^\intercal$, namely $\mathrm{Sp}(\boldsymbol{K}) \subset \{0, 1\}$. On the other hand, the constraint on the k-leverage scores can be expressed as a condition on the diagonal of $\boldsymbol{K}$. Thus, the problem of generating an orthogonal matrix with a given profile of k-leverage scores boils down to enforcing conditions on the spectrum and the diagonal of a symmetric matrix $\boldsymbol{K}$.

3.D.1 *Definitions and statement of the problem*

We denote by $(\boldsymbol{f}_i)_{i\in[d]}$ the columns of the matrix $\boldsymbol{F}$. For $n \in \mathbb{N}$, we write $\boldsymbol{0}_n$ the vector containing zeros living in $\mathbb{R}^n$. We say that the vector $\boldsymbol{u} \in \mathbb{R}^n$ interlaces on $\boldsymbol{v} \in \mathbb{R}^n$ and we denote

$$\boldsymbol{u} \sqsubseteq \boldsymbol{v}$$

if $u_n \leq v_n$ and $\forall i \in [1 : n-1]$, $v_{i+1} \leq u_i \leq v_i$.

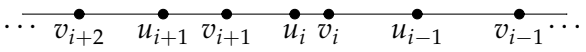


Figure 3.D1 – Illustration of the interlacing of $\boldsymbol{u}$ on $\boldsymbol{v}$.

Definition 3.6. *Let $k, d \in \mathbb{N}$, with $k \leq d$. Let $\boldsymbol{F} \in \mathbb{R}^{k\times d}$ be a full rank matrix*[5]*. Within this section, we denote $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$ the squares of the nonvanishing singular values of the matrix $\boldsymbol{F}$, and $\boldsymbol{\ell} = (\ell_1 = \|\boldsymbol{f}_1\|^2, \ell_2 = \|\boldsymbol{f}_2\|^2, \ldots, \ell_d = \|\boldsymbol{f}_d\|^2)$ are the squared norms of the columns of $\boldsymbol{F}$, which we assume to be ordered decreasingly:*

$$\ell_1 \geq \ell_2 \geq \cdots \geq \ell_d.$$

When the rows of $\boldsymbol{F}$ are orthonormal, we can think of $\boldsymbol{\ell}$ as a vector of leverage scores.

We are interested in the problem of constructing a matrix $\boldsymbol{F}$ with orthonormal rows given its leverage scores.

[5] A *frame*, using the definitions of (Fickus et al., 2011) and (Fickus et al., 2013).

Problem 1. *Let $k, d \in \mathbb{N}$, with $k \leq d$, and let $\boldsymbol{\ell} \in \mathbb{R}_+^d$ such that $\sum_{i=1}^{d} \ell_i = k$. Build a matrix $\boldsymbol{F} \in \mathbb{R}^{k \times d}$ such that*

$$\mathrm{Sp}(\boldsymbol{F}^\intercal \boldsymbol{F}) = [\mathbb{I}_k, \mathbf{0}_{d-k}], \tag{3.150}$$

and

$$\mathrm{Diag}(\boldsymbol{F}^\intercal \boldsymbol{F}) = \boldsymbol{\ell}. \tag{3.151}$$

We actually consider here the generalization of Problem 2 to an arbitrary spectrum.

Problem 2. *Let $k, d \in \mathbb{N}$, with $k \leq d$, and let $\boldsymbol{\ell} \in \mathbb{R}_+^d$ such that $\sum_{i=1}^{d} \ell_i = \sum_{i=1}^{k} \sigma_i^2$. Build a matrix $\boldsymbol{F} \in \mathbb{R}^{k \times d}$ such that*

$$\mathrm{Sp}(\boldsymbol{F}^\intercal \boldsymbol{F}) = [\sigma^2, \mathbf{0}_{d-k}] =: \hat{\boldsymbol{\sigma}}^2 \tag{3.152}$$

and

$$\mathrm{Diag}(\boldsymbol{F}^\intercal \boldsymbol{F}) = \boldsymbol{\ell}. \tag{3.153}$$

Denote by

$$\mathcal{M}_{(\boldsymbol{\ell}, \boldsymbol{\sigma})} = \{\boldsymbol{M} \in \mathbb{R}^{d \times d} \text{ symmetric } / \ \mathrm{Diag}(\boldsymbol{M}) = \boldsymbol{\ell},\ \mathrm{Sp}(\boldsymbol{M}) = \hat{\boldsymbol{\sigma}}^2\}. \tag{3.154}$$

The non-emptiness of $\mathcal{M}_{(\boldsymbol{\ell}, \boldsymbol{\sigma})}$ is determined by a majorization condition between $\boldsymbol{\ell}$ and $\hat{\boldsymbol{\sigma}}$, see Appendix 3.B for definitions. More precisely, we have the following theorem.

Theorem 3.9 (Schur-Horn). *Let $k, d \in \mathbb{N}$, with $k \leq d$, and let $\boldsymbol{\ell} \in \mathbb{R}_+^d$. We have*

$$\mathcal{M}_{(\boldsymbol{\ell}, \boldsymbol{\sigma})} \neq \emptyset \Leftrightarrow \boldsymbol{\ell} \prec_S \hat{\boldsymbol{\sigma}}. \tag{3.155}$$

The proof by Horn, 1954 of the reciprocal in Theorem 3.9 is non constructive. In the next section, we survey algorithms that output an element of $\mathcal{M}_{(\boldsymbol{\ell}, \boldsymbol{\sigma})}$.

3.D.2 *Related work*

Several articles (Raskutti and Mahoney, 2016, Ma et al., 2015) in the randomized linear algebra community propose the use of non Gaussian random matrices to generate matrices with a fast decreasing profile of leverage scores (so-called *heavy hitters*) without controlling the exact profile of the leverage scores.

Dhillon et al., 2005 showed how to generate matrices from $\mathcal{M}_{(\boldsymbol{\ell}, \boldsymbol{\sigma})}$ using Givens rotations; see the algorithm in Figure 3.D2. The idea of the algorithm is to start with a frame with the exact spectrum and repeatedly apply orthogonal matrices (Lines 4 and 6 of Figure 3.D2) that preserve the spectrum while changing the leverage scores of only two columns, setting one of their leverage scores to the desired value. The orthogonal matrices are the so-called *Givens rotations*.

Definition 3.7. *Let $\theta \in [0, 2\pi[$ and $i, j \in [d]$. The Givens rotation $\boldsymbol{G}_{i,j}(\theta) \in \mathbb{R}^{d \times d}$ is defined by*

$$\boldsymbol{G}_{i,j}(\theta) = \begin{bmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & \cos(\theta) & & & & -\sin(\theta) & & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & \sin(\theta) & & & & \cos(\theta) & & \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{bmatrix}. \tag{3.156}$$

GivensAlgorithm(ℓ, σ)

$\boldsymbol{F} \longleftarrow \left[\ \mathrm{Diag}(\sigma) \ \middle| \ \boldsymbol{0} \ \right] \in \mathbb{R}^{k \times d}$
while $\exists i, j, k \in [d], i < k < j : \|\boldsymbol{f}_i\|^2 < \ell_i, \|\boldsymbol{f}_k\|^2 = \ell_k, \|\boldsymbol{f}_j\|^2 > \ell_j$
  if $\ell_i - \|\boldsymbol{f}_i\|^2 \leq \|\boldsymbol{f}_j\|^2 - \ell_j$
    $\boldsymbol{F} \leftarrow \boldsymbol{G}_{i,j}(\theta)\boldsymbol{F}$, where $\|(\boldsymbol{G}_{i,j}(\theta)\boldsymbol{F})_i\|^2 = \ell_i$.
  else
    $\boldsymbol{F} \leftarrow \boldsymbol{G}_{i,j}(\theta)\boldsymbol{F}$, where $\|(\boldsymbol{G}_{i,j}(\theta)\boldsymbol{F})_j\|^2 = \ell_j$,
return $\boldsymbol{F} \in \mathbb{R}^{k \times d}$.

Figure 3.D2 – The pseudocode of the algorithm proposed by Dhillon et al., 2005 for generating a matrix given its leverage scores and spectrum by successively applying Givens rotations.

Figure 3.D3 shows the output of the algorithm in Figure 3.D2, for the input $(\ell, \sigma) = (\ell, \mathbb{1}_k)$ for three different values of ℓ. The main drawbacks of this algorithm are first that it is deterministic, so that it outputs a unique matrix $\boldsymbol{F}$ for a given input (ℓ, σ), and second that the output is a highly structured matrix, as observed on Figure 3.D3.

We propose an algorithm that outputs random, more "generic" matrices belonging to $\mathcal{M}_{(\ell,\sigma)}$. This algorithm is based on a parametrization of $\mathcal{M}_{(\ell,\sigma)}$ using the collection of spectra of all minors of $\boldsymbol{F} \in \mathcal{M}_{(\ell,\sigma)}$. This parametrization was introduced by Fickus et al., 2013, and we recall it in Section 3.D.3. For now, let us simply look at Figure 3.D4, which displays a few outputs of our algorithm for the same input as in Figure 3.D3a. We now obtain different matrices for the same input (ℓ, σ), and these matrices are less structured than the output of Algorithm 3.D2, as required.

3.D.3 *The restricted Gelfand-Tsetlin polytope*

Definition 3.8. *Recall that $(\boldsymbol{f}_i)_{i \in [d]}$ are the columns of the matrix $\boldsymbol{F} \in \mathbb{R}^{k \times d}$. For $r \in [d]$, we further define*

$$\boldsymbol{F}_r = \boldsymbol{F}_{:,[r]} \in \mathbb{R}^{k \times r}, \tag{3.157}$$

$$\boldsymbol{C}_r = \sum_{i\in[r]} \boldsymbol{f}_i \boldsymbol{f}_i^\intercal \in \mathbb{R}^{k\times k}, \tag{3.158}$$

$$\boldsymbol{G}_r = \boldsymbol{F}_r^\intercal \boldsymbol{F}_r \in \mathbb{R}^{r\times r}. \tag{3.159}$$

Furthermore, we note for $r \in [d]$,

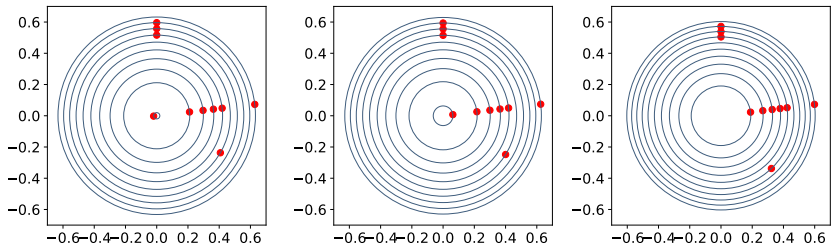
$$(\lambda_{r,i})_{i\in[k]} = \Lambda(\boldsymbol{C}_r), \tag{3.160}$$

$$(\tilde{\lambda}_{r,i})_{i\in[r]} = \Lambda(\boldsymbol{G}_r). \tag{3.161}$$

The $(\lambda_{r,i})_{i\in[k]}$, $r \in [d]$, *are called the outer eigensteps of* $\boldsymbol{F}$, *and we group them in the matrix*

$$\Lambda^{out}(\boldsymbol{F}) = (\lambda_{r,i})_{i\in[k],r\in[d]} \in \mathbb{R}^{k\times d}.$$

Similarly, the $(\tilde{\lambda}_{r,i})_{i\in[r]}$ *are called inner eigensteps of* $\boldsymbol{F}$*: for* $r \in [d]$, $(\lambda_{r,i})_{i\in[k]}$ *and* $(\tilde{\lambda}_{r,i})_{i\in[r]}$ *share the same nonzeros elements.*



(a) (b) (c)

Figure 3.D3 – The output of the algorithm in Figure 3.D2 for $k = 2$, $d = 10$, $\boldsymbol{\sigma} = (1, 1)$, and three different values of $\boldsymbol{\ell}$ that each add to k. Each red dot has coordinates a column of $\boldsymbol{F}$. The blue circles have for radii the prescribed $(\sqrt{\ell_i})$.

Example 3.2. *For* $k = 2$, $d = 4$, *consider the full-rank matrix*

$$\boldsymbol{F} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \tag{3.162}$$

Then

$$\Lambda^{out}(\boldsymbol{F}) = \begin{bmatrix} 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 2 \end{bmatrix}. \tag{3.163}$$

Proposition 3.12. *The outer eigensteps satisfy the following constraints:*

$$\begin{cases} \forall i \in [k],\ \lambda_{0,i} = 0 \\ \forall i \in [k],\ \lambda_{d,i} = \sigma_i^2 \\ \forall r \in [d],\ (\lambda_{r,:}) \sqsubseteq (\lambda_{r+1,:}) \\ \forall r \in [d],\ \sum\limits_{i\in[d]} \lambda_{r,i} = \sum\limits_{i\in[r]} \ell_i \end{cases}. \tag{3.164}$$

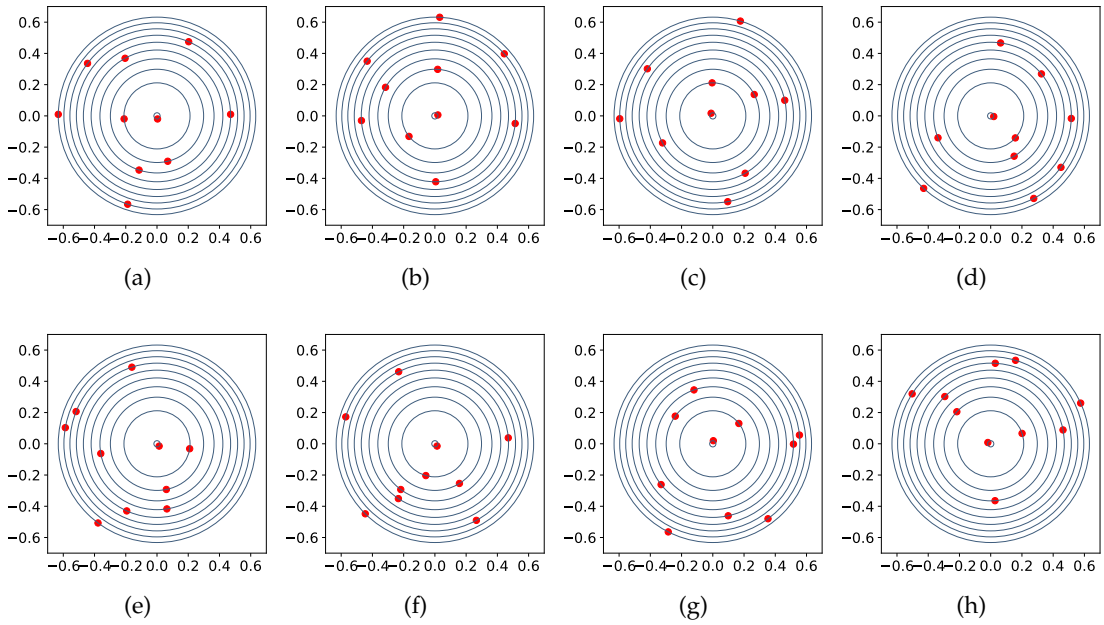


Figure 3.D4 – The output of our algorithm for $k = 2$, $d = 10$, an input $\sigma = (1, 1)$, and ℓ as in Figure 3.D3a. Each red dot has coordinates a column of $\boldsymbol{F}$. The blue circles have for radii the prescribed $(\sqrt{\ell_i})$.

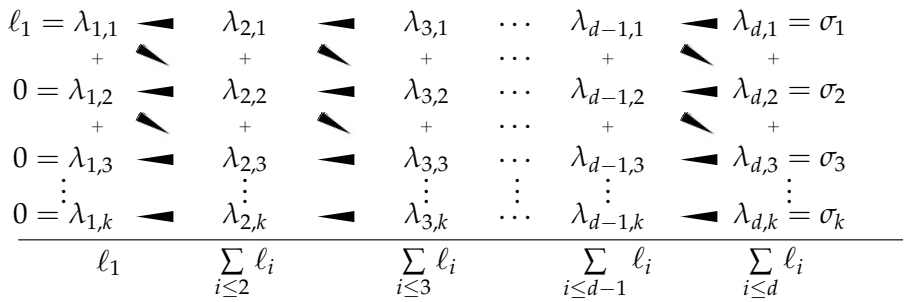


Figure 3.D5 – The interlacing relationships (3.164) satisfied by the outer eigensteps of a frame. Thick triangles are used in place of $\leq$ for improved readability.

In other words, the outer eigensteps are constrained to live in a polytope. We define the restricted Gelfand-Tsetlin polytope $\boldsymbol{GT}_{(k,d)}(\sigma, \ell)$ to be the subset of $\mathbb{R}^{k\times d}$ defined by the equations (3.164). A more graphical summary of the interlacing and sum constraints is given in Figure 3.D5. The restricted GT polytope[6] allows a parametrization of $\mathcal{M}_{(\ell,\sigma)}$ by the following reconstruction result.

Theorem 3.10 (Theorem 3, Fickus et al., 2011)**.** *Every matrix $\boldsymbol{F} \in \mathcal{M}_{(\ell,\sigma)}$ can be constructed as follows:*

- *pick a valid sequence of outer eigensteps noted $\Lambda^{out} \in \boldsymbol{GT}_{(k,d)}(\sigma, \ell)$,*
- *pick $\boldsymbol{f}_1 \in \mathbb{R}^k$ such that*

$$\|\boldsymbol{f}_1\|^2 = \ell_1, \tag{3.165}$$

[6] Note the difference with the Gelfand-Tsetlin polytope in the random matrix literature (Baryshnikov, 2001), where only the spectrum is constrained, not the diagonal.

- *for $r \in [d]$, consider the polynomial $p_r(x) = \prod_{i\in[d]} (x - \lambda_{r,i})$, and for each $r \in [d-1]$, choose $\boldsymbol{f}_{r+1} \in \mathbb{R}^k$ such that*

$$\forall \lambda \in \{\lambda_{r,i}\}_{i\in[d]},\ \|\boldsymbol{P}_{r,\lambda}\boldsymbol{f}_{r+1}\|^2 = -\lim_{x\to\lambda}(x-\lambda)\frac{p_{r+1}(\lambda)}{p_r(\lambda)}, \tag{3.166}$$

where $\boldsymbol{P}_{r,\lambda}$ denotes the orthogonal projection onto the eigenspace $\mathrm{Ker}(\lambda \mathbb{I}_k - \boldsymbol{F}_r\boldsymbol{F}_r^T)$.

Conversely, any matrix $\boldsymbol{F}$ constructed by this process is in $\mathcal{M}_{(\boldsymbol{\ell},\boldsymbol{\sigma})}$.

Fickus et al., 2011 propose an algorithm to construct a vector $\boldsymbol{f}_r$ satisfying Equation (3.166). Finally, an algorithm for the construction of a valid sequence of eigensteps $\Lambda^{\text{out}} \in \boldsymbol{GT}_{(k,d)}(\boldsymbol{\sigma},\boldsymbol{\ell})$ was proposed in (Fickus et al., 2013). This yields the following constructive result.

Theorem 3.11 (Theorem 4.1, Fickus et al., 2013)**.** *Every matrix $\boldsymbol{F} \in \mathcal{M}(\boldsymbol{\sigma},\boldsymbol{\ell})$ can be constructed as follows:*

- *Set $\forall i \in [k]$, $\tilde{\lambda}_{d,i} = \sigma_i^2$,*
- *For $r \in \{d-1,\dots,1\}$, construct $\{\tilde{\lambda}_{r,:}\}$ as follows. For each $i \in \{k,\dots,1\}$, pick*

$$\tilde{\lambda}_{r-1,i} \in [B_{i,r}(\boldsymbol{\ell},\boldsymbol{\sigma}), A_{i,r}(\boldsymbol{\ell},\boldsymbol{\sigma})],$$

where

$$\begin{aligned} A_{i,r}(\boldsymbol{\ell},\boldsymbol{\sigma}) &= \max\left\{\tilde{\lambda}_{r+1,i+1}, \sum_{t=i}^{k}\tilde{\lambda}_{r+1,t} - \sum_{t=i+1}^{k}\tilde{\lambda}_{r,t} - \ell_{r+1}\right\} \\ B_{i,r}(\boldsymbol{\ell},\boldsymbol{\sigma}) &= \min\left\{\tilde{\lambda}_{r+1,i}, \min_{z=1,\dots,i}\left\{\sum_{t=z}^{r}\ell_t - \sum_{t=z+1}^{i}\tilde{\lambda}_{r+1,t} - \sum_{t=i+1}^{k}\tilde{\lambda}_{r,t}\right\}\right\}. \end{aligned} \tag{3.167}$$

Furthermore, any sequence constructed by this algorithm is a valid sequence of inner eigensteps.

Based on these results we propose an algorithm for the generation of orthogonal random matrices with a given profile of leverage scores.

3.D.4 *Our algorithm*

We consider a randomization of the algorithm given in Theorem 3.11. First, we generate a random sequence of valid inner eigensteps Λ^{in} using Algorithm 3.D6. Then we proceed to the reconstruction a frame that admits Λ^{in} as a sequence of eigensteps using the Algorithm proposed in (Fickus et al., 2011).

Note that Equations (3.165) and (3.166) admit several solutions. For example, for $r \in [d]$, and if $\boldsymbol{f}_{r+1}$ satisfies (3.166), $-\boldsymbol{f}_{r+1}$ satisfies this equation too. Fickus et al., 2011 actually prove that the set of solutions of these equations is invariant under a specific action of the orthogonal group $\mathbb{O}(\rho(r,k))$ where $\rho(r,k) \in \mathbb{N}$ nontrivially depends on the eigensteps. In the reconstruction step of our algorithm, we apply a random orthogonal matrix sampled from the Haar measure on $\mathbb{O}(d)$ to the vector $\boldsymbol{f}_1$ and, then, for every $r \in [2:d]$, we apply an independent random orthogonal matrix Ω to a vector $\boldsymbol{f}_{r+1}$, that satisfies (3.166), so that $\Omega\boldsymbol{f}_{r+1}$ still satisfies (3.166).

Figure 3.D4 displays a few samples from our algorithm, which display diversity and no apparent structure, as required for a generator of toy datasets. The question of fully characterizing the distribution of the output of our algorithm is an open question.

RandomEigensteps(ℓ, σ)

$\Lambda^{\text{out}} \longleftarrow \mathbb{O} \in \mathbb{R}^{k \times d}$

$\forall i \in [k]$, $\tilde{\lambda}_{d,i} \longleftarrow \sigma_i$

for $r \in \{d-1, \ldots, 1\}$

  for $i \in \{k, \ldots, 1\}$

    Pick $\tilde{\lambda}_{r-1,i} \sim \mathcal{U}([B_{i,r}(\ell, \sigma), A_{i,r}(\ell, \sigma)])$

return Λ^{out}

Figure 3.D6 – The pseudocode of the generator of random valid eigensteps taking as input (ℓ, σ).

4

KERNEL QUADRATURE USING DPPS

Numerical integration is at the heart of many tasks in applied mathematics, and statistics, like Bayesian inference (Robert and Casella, 2004) or option pricing (Glasserman, 2013). Indeed, all these tasks involve, at some level, the evaluation of an integral

$$\int_{\mathcal{X}} f(x)\mathrm{d}\omega(x). \tag{4.1}$$

Integrals that can be written in closed form are the exception; in general the value of (4.1) is only known through approximations. The problem of approximating an integral find its roots in an older problem: the evaluation of the perimeter or the area of a domain bounded by curves. For example, the problem of squaring a circle occupied minds for centuries. By approximating the volume of a cylindrical granary by a rectangular one, a first estimation of $\pi \approx 256/81 \approx 3.1605$ was already known by Egyptians as it can be witnessed by the Rhind papyrus (Robins and Shute, 1987). Archimedes achieved a better approximation of the value of π by computing the perimeter of the regular 96-gons and obtained (Heath, 2003)

$$3 + \frac{10}{71} < \pi < 3 + \frac{1}{7}.$$

These geometric techniques were further developed after the invention of calculus. For instance, Newton considered the approximation of the integral of a function f on some interval $[a, b]$ using some evaluations of $f(x_1), \ldots, f(x_N)$ of f

$$\int_a^b f(t)\mathrm{d}t \approx \sum_{n\in[N]} w_n f(x_n). \tag{4.2}$$

In modern words, his approach goes as follows: evaluate f on equally distanced nodes $x_1, \ldots, x_N$ and consider the interpolating polynomial p_{N-1}

$$p_{N-1}(t) = \sum_{n\in[N]} f(x_n)\ell_{n,\boldsymbol{x}}(t) \in \mathbb{R}_{N-1}[t], \tag{4.3}$$

where $\ell_{n,\boldsymbol{x}}$ are the Lagrange polynomial [1] polynomials (Burden and Faires, 1997)

$$\ell_{n,\boldsymbol{x}}(t) = \frac{\prod_{m\in[N],m\neq n}(t - x_m)}{\prod_{m\in[N],m\neq n}(x_n - x_m)}. \tag{4.4}$$

Newton then approximates the integral of f by that of p_{N-1}

$$\int_a^b f(t)\mathrm{d}t \approx \sum_{n\in[N]} w_n f(x_n), \tag{4.5}$$

[1] At that time, Newton considered divided differences interpolation polynomial (Burden and Faires, 1997).

where the scalars w_n, also called the *Cotes numbers*, have the expression

$$\forall n \in [N],\ w_n = \int_a^b \ell_{n,x}(t)\mathrm{d}t. \tag{4.6}$$

The approximation (4.5), called the *Newton-Cotes formula*, is an early form of a *quadrature*: the nodes x_n and the weights w_n are independent of the function f. In the case of the Newton-Cotes formula, the explicit values of the weights, for small values of N, were found by Cotes. This formula coincides with some already known quadratures such as Trapezoid rule $(N = 2)$, Simpson's $1/3$ rule $(N = 3)$, Simpson's $3/8$ rule $(N = 4)$... Of course, the use of equidistant nodes is arbitrary, and other configurations can be used in the interpolation step (4.3). This observation gave birth to the Gaussian quadrature and the field of numerical integration using the zeros of orthogonal polynomials. As we shall see, these techniques are not amenable to be generalized to high-dimensional domains.

Monte Carlo methods offer another approach to numerical integration. This class of methods, introduced by Metropolis and Ulam, 1949, relies on randomized averaging. Given a sequence of random variables $x_1, \dots, x_N$ generated from a density g, it comes

$$\mathbb{E}\,\frac{1}{N}\sum_{n\in[N]} f(x_n) = \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x), \tag{4.7}$$

and by the Strong Law of Large Numbers, $\sum_{n\in[N]} f(x_n)/N$ converges almost surely to $\int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x)$. Moreover, if $\int_{\mathcal{X}} f(x)^2\mathrm{d}\omega(x) < +\infty$, the variance of $\sum_{n\in[N]} f(x_n)/N$ scales as $\mathcal{O}(1/N)$, and the estimator satisfies a central limit theorem (Robert and Casella, 2004). Remarkably, this rate of convergence does not depend on the dimension of $\mathcal{X}$ (if $\mathcal{X}$ is a d-dimensional manifold). This property gives an advantage, provided that we know how to sample from the density g, of Monte Carlo methods over deterministic quadrature rules, that scales poorly for high dimension integration problems.

It would be fair to claim that deterministic quadratures are very efficient for smooth low dimensional functions, while this smoothness does not improve the Monte Carlo rate. This observation have fueled a multitude of investigations with a common purpose: improving upon the Monte Carlo rate for smooth functions on high dimensional integration problems. Many approaches were proposed in this rich line of research: Quasi-Monte Carlo methods (Dick and Pillichshammer, 2010), kernel quadrature methods (Hickernell, 1998) (Smola et al., 2007) (Chen et al., 2010) (Bach, 2017) (Briol et al., 2019) or determinantal point processes (Bardenet and Hardy, 2020).

We start with a brief review of existing numerical integration approaches in Section 4.1, with an emphasis on the connections between these methods and existing quadratures using DPPs as illustrated in Figure 4.1. In particular, a special attention will be given to kernel quadrature methods that will be reviewed in Section 4.2.

The contribution of this chapter is the introduction and the theoretical analysis of a new class of quadratures: optimal kernel quadrature based on projection DPPs nodes. The main theoretical results are given in Section 4.3. Section 4.4 is devoted for numerical simulations. The proofs of the theoretical results are detailed in Section 4.6.

The material of this chapter is based on the following article

- A. Belhadji, R. Bardenet, and P. Chainais (2019a). "Kernel quadrature with DPPs". In: *Advances in Neural Information Processing Systems* 32, pp. 12907–12917.

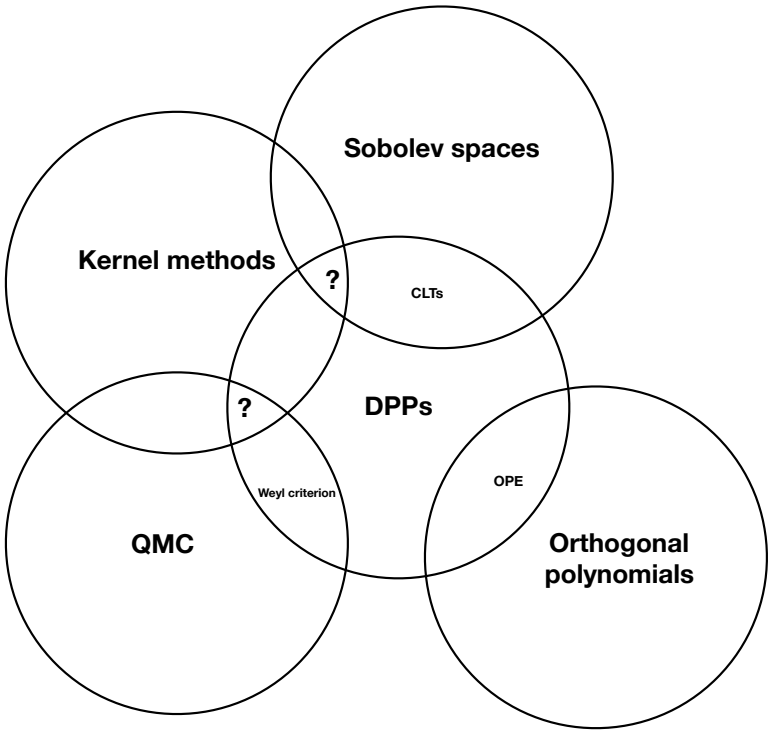


Figure 4.1 – The connections between the different fields of numerical integration with DPPs.

4.1 RELATED WORK

The purpose of this section is to motivate the definition and the theoretical analysis of a class of quadratures based on projection DPPs. This class of quadratures is naturally defined within the RKHS framework. The merit of this class of quadratures is its universality and applicability to a variety of settings. An exhaustive review of existing results on quadratures is outside the scope of this manuscript; nevertheless, it turns out that many landmark quadratures are relevant to the contribution of this chapter. Thus, it is the purpose of this section to highlight these connections.

4.1.1 *Gaussian quadrature*

As we have seen in the introduction, the Newton-Cotes formula is based on equi-distanced nodes on $[a, b]$ and it is exact for polynomials of order smaller than $N-1$ when N nodes are used. Gauss (Gauss, 1815) observed that equi-distance nodes are not optimal, in a sense that will be defined later, and shed the light on the question of designing the quadrature nodes. This observation gave birth to Gaussian quadrature.

The construction

Gauss investigated the maximal order of exactness that could be achieved by a quadrature rule based on a configuration of nodes $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ and the corresponding weights $w_1, \ldots, w_N$. This order is defined by

$$M(\boldsymbol{x}) = \sup \left\{ M \in \mathbb{N}; \exists \boldsymbol{w} \in \mathbb{R}^N, \forall m \in [M], \int_a^b t^m \mathrm{d}t = \sum_{n \in [N]} w_n x_n^m \right\}. \tag{4.8}$$

He proved that for any positive integer N there exists a unique configuration $\boldsymbol{x}_G$, also called the *Gauss nodes*, of cardinality N such that $M(\boldsymbol{x}_G) = 2N-1$. He also proved that $2N-1$ is the maximal order that could be achieved by any configuration of cardinality N. His proof (Gauss, 1815) is based on arguments from continued fractions theory (Khinchin, 1997). Later, Jacobi, 1826 showed, using an alternative proof based on simple manipulations on polynomials, that $M(\boldsymbol{x}) = 2N-1$ if and only if for every polynomial p of order smaller or equal to $N-1$

$$\int_a^b p(t)\pi_{\boldsymbol{x}}(t)\mathrm{d}t = 0, \tag{4.9}$$

where $\pi_{\boldsymbol{x}}$ is the polynomial

$$\pi_{\boldsymbol{x}}(t) = \prod_{n \in [N]} (t - x_n). \tag{4.10}$$

In other words, the Gauss nodes are the roots of a polynomial of degree N that satisfy the condition (4.9) for every polynomial of degree less than $N-1$: the polynomial $\pi_{\boldsymbol{x}}$ is nothing else but the scaled Legendre polynomial of degree N. This characterization (4.9) of the Gauss nodes highlighted the importance of the orthogonality between

polynomials with respect to $\mathrm{d}t$ and motivated the extension of Gaussian quadrature to weighted measures, that is approximating

$$\int_a^b f(t)\mathrm{d}\omega(t) \approx \sum_{n\in[N]} w_n f(x_n). \tag{4.11}$$

In particular, Christoffel, 1877 generalized the Gauss-Jacobi construction to measures $\mathrm{d}\omega(t) = w(t)\mathrm{d}t$ for arbitrary weight functions ω and his approach is generalizable to even a broader class of measures (Szegő, 1939). Typically, $\mathrm{d}\omega$ is assumed to have finite moments of all orders, so that the orthonormal polynomials with respect to $\mathrm{d}\omega$ are well defined by

$$\forall m \in \mathbb{N},\ \deg P_m = m, \tag{4.12}$$

$$\forall m, m' \in \mathbb{N},\ \int_a^b P_m(t)P_{m'}(t)\mathrm{d}\omega(t) = \delta_{m,m'}. \tag{4.13}$$

Again, the roots $x_1, \dots, x_N$ of P_N achieve the maximal order $2N-1$ and the corresponding weights, called the *Christoffel numbers*, satisfy the identity (Shohat, 1929) (see also Theorem 3.4.2 in (Szegő, 1939)):

$$w_n = \frac{1}{\sum_{m=0}^{N-1} P_m(x_n)^2}. \tag{4.14}$$

Observe that in (4.14), every w_n depends only on the corresponding node x_n.

In practice, the evaluation of the nodes and weights of Gaussian quadrature for a given measure $\mathrm{d}\omega$ relies on the three-term recurrence coefficients associated to the polynomials P_m

$$\forall m \in \mathbb{N},\ tP_m(t) = \sqrt{b_{m+1}}P_{m+1}(t) + a_m P_m(t) + \sqrt{b_m}P_{m-1}(t), \tag{4.15}$$

$$tP_0(t) = a_0 P_1(t) + \sqrt{b_0}P_0(t). \tag{4.16}$$

Indeed, let J_N the tridiagonal matrix of order N defined by

$$J_N = \begin{pmatrix} a_1 & \sqrt{b_1} & & 0 \\ \sqrt{b_1} & \ddots & \ddots & \\ & \ddots & \ddots & \sqrt{b_{N-1}} \\ 0 & & \sqrt{b_{N-1}} & a_N \end{pmatrix}. \tag{4.17}$$

The roots of the P_N are the eigenvalues of the tridiagonal matrix J_N, and for every eigenvalue x the corresponding eigenvector is the vector $\big(P_0(x), \dots, P_{N-1}(x)\big)^\intercal$; see Theorem 2.13 in (Golub and Meurant, 2009). In other words, the spectral decomposition of the Jacobi matrix gives the nodes and the weights of the Gaussian quadrature. The matrix J_N is known explicitly for many classic orthogonal polynomials such as Chebyshev, Legendre, Hermite... Building up this matrices amount to the calculation of the moments of the measure $\mathrm{d}\omega$ which is a non trivial task for a general measure $\mathrm{d}\omega$. We refer to (Golub and Meurant, 2009) for more details on the algorithmic construction of Gaussian quadrature.

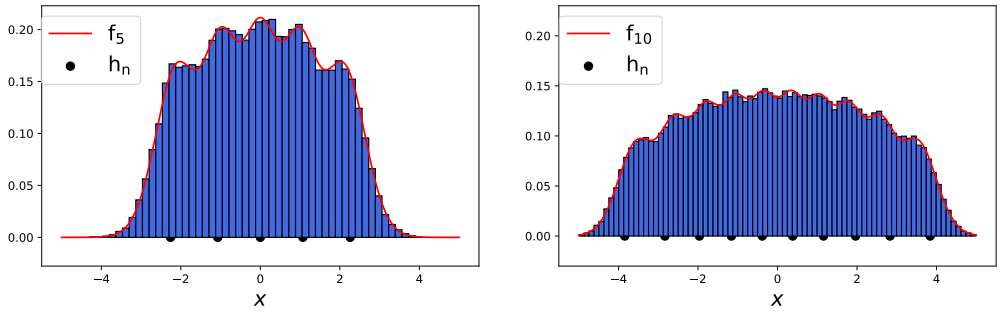


Figure 4.2 – The histogram of the eigenvalues of $\tilde{J}_N$ based on 50000 realizations compared to the first intensity function f_N of the projection DPP and the eigenvalues of J_N, with $N \in \{5, 10\}$.

DPPs as probabilistic relaxations of Gaussian quadrature

We illustrate the construction of Gaussian quadrature based on the spectral decomposition of J_N in the special case of the Gaussian measure on the real line. The aim of this illustration is to highlight a connection between Gaussian quadrature and DPPs.

Now, let $\mathrm{d}\omega(t) = w(t)\mathrm{d}t$ be the Gaussian measure on $\mathbb{R}$ with

$$w(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \tag{4.18}$$

It is well known, that the corresponding orthonormal polynomials are proportional to the probabilist Hermite polynomials, and the corresponding Jacobi matrix writes

$$J_N = \begin{pmatrix} 0 & \sqrt{1} & & 0 \\ \sqrt{1} & \ddots & \ddots & \\ & \ddots & \ddots & \sqrt{N-1} \\ 0 & & \sqrt{N-1} & 0 \end{pmatrix}. \tag{4.19}$$

Now consider the following random matrix

$$\tilde{J}_N = \begin{pmatrix} \mathcal{N}(0,1) & \chi_1 & & 0 \\ \chi_1 & \ddots & \ddots & \\ & \ddots & \ddots & \chi_{(N-1)} \\ 0 & & \chi_{(N-1)} & \mathcal{N}(0,1) \end{pmatrix}, \tag{4.20}$$

where $\mathcal{N}(0,1)$ is a random variable that follows the standard normal distribution, and χ_n is a random variable that follows the chi distribution with n degrees of freedom. Observe that

$$\mathbb{E}\,\tilde{J}_N = J_N. \tag{4.21}$$

The eigenvalues $x_1, \ldots, x_N$ of $\tilde{J}_N$, seen as a vector of $\mathbb{R}^d$, have the following density with respect to the Lebesgue measure of $\mathbb{R}^d$ (Dumitriu and Edelman, 2005)

$$f(x_1, \ldots, x_N) \propto \prod_{n=1}^{N} e^{-x_n^2/2} \prod_{1 \le n < n' \le N} |x_n - x_{n'}|^2. \tag{4.22}$$

In other words, the eigenvalues of $\tilde{J}_N$ form a DPP of marginal kernel

$$\kappa(x,y) = \sum_{n=0}^{N-1} P_n(x)P_n(y), \tag{4.23}$$

and the $\mathrm{d}\omega$ as a reference measure, where the P_n are the normalized probabilist Hermite polynomials [2].

Figure 4.2 shows the histogram of the eigenvalues of the matrix $\tilde{J}_N$ with $N \in \{5, 10\}$ calculated over 50000 realizations of $\tilde{J}_N$ compared to the first intensity function of the projection DPP defined by the kernel κ and the reference measure $\mathrm{d}\omega$. We observe that the histogram matches the first intensity function that have local maxima around the eigenvalues $h_1, \ldots, h_N$ of the matrix J_N (the red dots), that are the roots of P_N. In other words, in this case, the projection DPP can be seen as a probabilistic relaxation of Gaussian quadrature.

This case is far from being anecdotal. Indeed, this probabilistic relaxation is possible for Gaussian quadrature of other measures (Dumitriu and Edelman, 2002) (Killip and Nenciu, 2004), and the resulting point process coincides with a projection DPP; see (Gautier et al., 2020) for details on this topic.

The convergence rates

In general, the integrand f is not a polynomial of low degree so that the quadrature is not exact. One would expect to give a control of the remainder

$$R_{x,w}[f] = \int_a^b f(t)\mathrm{d}t - \sum_{n\in[N]} w_n f(x_n), \tag{4.24}$$

if the function f is well approximated by low degree polynomials. This is the case of analytic functions for which $|R_{x,w}| = \mathcal{O}(\alpha^N)$ for some $\alpha \in]0,1[$. For functions of limited degree of smoothness, Taylor series with integral remainder can be used to give an upper bound on the integration error $|R_{x,w}|$ that scales as $\mathcal{O}(N^{-s})$ for some $s > 1$ that depends on the smoothness of the function f. See Chapter 4 of (Davis and Rabinowitz, 1984) for details on this topic.

Gaussian quadrature in high-dimensional domains

Finally, there were many attempts to extend Gaussian quadrature to high dimensional domains using common zeros of multidimensional orthogonal polynomials; see (Xu, 1994b) for details on this topic. For example, the construction of a high dimensional Gaussian quadrature would require to work with block Jacobi matrices instead of tridiagonal Jacobi matrices (Xu, 1994a). However, this line of research is not developed yet from an algorithmic point of view.

4.1.2 *Quasi-Monte Carlo methods*

As we have seen in Section 4.1.1, a high precision approximation of a unidimensional integral is possible using a deterministic rule such as Gaussian quadrature provided

[2] Using row and column operations on the Vandermonde determinant $\prod |x_n - x_{n'}|^2$, we prove that it is proportional to $\mathrm{Det}\,\boldsymbol{\kappa}(x_1, \ldots, x_N)$.

that the integrand is a polynomial or well approximated by a polynomial. Now, for a smooth function defined on some high-dimensional domain, one can use the tensor product of these rules. However, for a given error, the number of required integrand evaluations grows exponentially with the dimension d. This poor scalability of grids with respect to the dimension d motivated the emergence of a new class of quadratures called quasi-Monte Carlo rules: they are based on "well-spread" low discrepancy sequences and their weights are uniform $w_n = 1/N$.

Uniform distribution and discrepancy functions

A Quasi-Monte Carlo rule is a quadrature based on a sequence of nodes $\boldsymbol{x} = \{x_1, \dots, x_N\} \subset \mathcal{X}$, where $\mathcal{X} = [0,1]^d$ that writes

$$\frac{1}{N} \sum_{n \in [N]} f(x_n), \tag{4.25}$$

for a given function f defined $\mathcal{X}$. The nodes $x_1, \dots, x_N$ are assumed to be "uniformly distributed" in $\mathcal{X}$.

We start by the definition of uniform distribution of an infinite sequence of nodes.

Definition 4.1. *An infinite sequence of elements of* $[0,1]^d$ $(x_n)_{n \in \mathbb{N}^*}$ *is said to be* uniformly distributed modulo one *on* $[0,1]^d$*, if for any subset* $\prod_{\delta \in [d]} [a_\delta, b_\delta] \subset [0,1]^d$ *we have*

$$\lim_{N \to +\infty} \frac{\left| \{x_1, \dots, x_N\} \cap \prod_{\delta \in [d]} [a_\delta, b_\delta] \right|}{N} = \prod_{\delta \in [d]} (b_\delta - a_\delta). \tag{4.26}$$

In other words, every subset $\prod_{\delta \in [d]} [a_\delta, b_\delta]$ gets a "fair share" of a uniformly distributed sequence. In particular, such a sequence satisfy

$$\lim_{N \to +\infty} \frac{1}{N} \sum_{n \in [N]} f(x_n) = \int_{[0,1]^d} f(x) \mathrm{d}x. \tag{4.27}$$

for every Riemann integrable function $f : \mathcal{X} \to \mathbb{R}$; see Theorem 3.3 in (Dick and Pillichshammer, 2010). An equivalent characterization, yet more practical to use, of uniformly distributed modulo one sequences is given by the *Weyl criterion*.

Theorem 4.1. *A sequence* $(x_n)_{n \in \mathbb{N}^*}$ *of elements in* $[0,1]^2$ *is uniformly distributed modulo one if and only if*

$$\forall h \in \mathbb{Z}^d \setminus \{0\}, \quad \lim_{N \to +\infty} \frac{1}{N} \sum_{n=1}^{N} e^{2\pi \mathrm{i} h^\intercal . x_n} = 0. \tag{4.28}$$

Now, for a finite number of nodes, one would measure how far is the sequence from the "uniformity" of (4.26). These measures of uniformity are called *discrepancy functions*.

A discrepancy function of a *finite* configuration $\boldsymbol{x} = \{x_1, \dots, x_N\} \subset \mathcal{X}$ quantifies the uniformity of the nodes of $\boldsymbol{x}$ in $\mathcal{X}$. It is defined with respect to a set of test sets $\mathcal{S}$. A typical choice of $\mathcal{S}$ is the set of the Cartesian products $[\mathbf{0}, u] = \prod_{\delta \in [d]} [0, u_\delta]$ with $u = (u_1, \dots, u_d) \in \mathcal{X}$. In this case, the *local discrepancy* is defined for all $u \in \mathcal{X}$ by

$$\mathcal{D}_{\boldsymbol{x}}(u) = \frac{1}{N} \sum_{n \in [N]} \mathbb{1}_{[\mathbf{0}, u]}(x_n) - \prod_{\delta \in [d]} u_\delta. \tag{4.29}$$

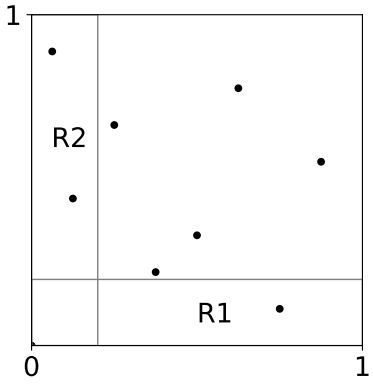


Figure 4.3 – The local discrepancy of a configuration $\boldsymbol{x}$ depends on the set $[\mathbf{0}, u]$.

The term $\frac{1}{N}\sum_{n\in[N]} \mathbb{1}_{[\mathbf{0},u]}(x_n)$ in (4.29) counts the fraction of the nodes x_n that falls into $[\mathbf{0}, u]$; while the second term measures the volume of $[\mathbf{0}, u]$: $\mathcal{D}_{\boldsymbol{x}}$ measures the difference between the relative number of points that belongs to $[\mathbf{0}, u]$ and its volume. A configuration $\boldsymbol{x}$ would be "well-spread" if the discrepancy function $\mathcal{D}_{\boldsymbol{x}}$ takes small values on $\mathcal{X}$. For example, the value of the $\|.\|_p$ norm [3] of $\mathcal{D}_{\boldsymbol{x}}$ is a measure of "well-spreadness" of $\boldsymbol{x}$ called the $\mathbb{L}_p$*-discrepancy* of $\boldsymbol{x}$

$$\Delta_p(\boldsymbol{x}) = \left(\int_{[0,1]^d} |\mathcal{D}_{\boldsymbol{x}}(u)|^p \otimes_{\delta\in[d]} \mathrm{d}u_\delta \right)^{1/p}. \tag{4.30}$$

In particular, the $\|.\|_\infty$ norm of the local discrepancy (4.29) is called the *star-discrepancy* and it is denoted $\Delta_*(\boldsymbol{x})$.

Figure 4.3 illustrates the concept of local discrepancy on the domain $[0,1]^2$: the two rectangles R1 $= [0,1] \times [0, 0.2]$ and R2 $= [0, 0.2] \times [0,1]$ have the same volume, yet the corresponding local discrepancies with respect to the design $\boldsymbol{x}$ are not equal as R2 contains more nodes than R1.

The importance of discrepancy in bounding the integration error shows up in Koksma-Hlawka inequality.

Theorem 4.2 (The Koksma-Hlawka inequality)**.** *Consider $f \in \mathbb{L}_2([0,1]^d)$. Then*

$$\left| \int_{[0,1]^d} f(u)\mathrm{d}u - \frac{1}{N}\sum_{n\in[N]} f(x_n) \right| \leq \Delta_*(\boldsymbol{x}) \|f\|_{\mathrm{HK}}, \tag{4.31}$$

where $\|f\|_{\mathrm{HK}}$ is the variation of f in the sense of Hardy-Krause (HK).

In other words, the upper bound in Koksma-Hlawka inequality is the product of two quantities: the star discrepancy of the configuration $\boldsymbol{x}$, which does not depend on f, and the HK variation of the function f, which does not depend on the configuration $\boldsymbol{x}$.

Koksma, 1942 proved the inequality for $d = 1$. In this case, the inequality holds when f is of bounded variation in $[0,1]$, and the HK variation coincides with the total variation

$$\|f\|_{\mathrm{HK}} = \int_{[0,1]} |f'(x)|\mathrm{d}x. \tag{4.32}$$

[3] This norm is well defined for $p \in [1,\infty]$, $\mathcal{D}_{\boldsymbol{x}} \in \mathbb{L}_p(\mathbb{R}^d)$ because $\mathcal{X} = [0,1]^d$ is a compact set of $\mathbb{R}^d$

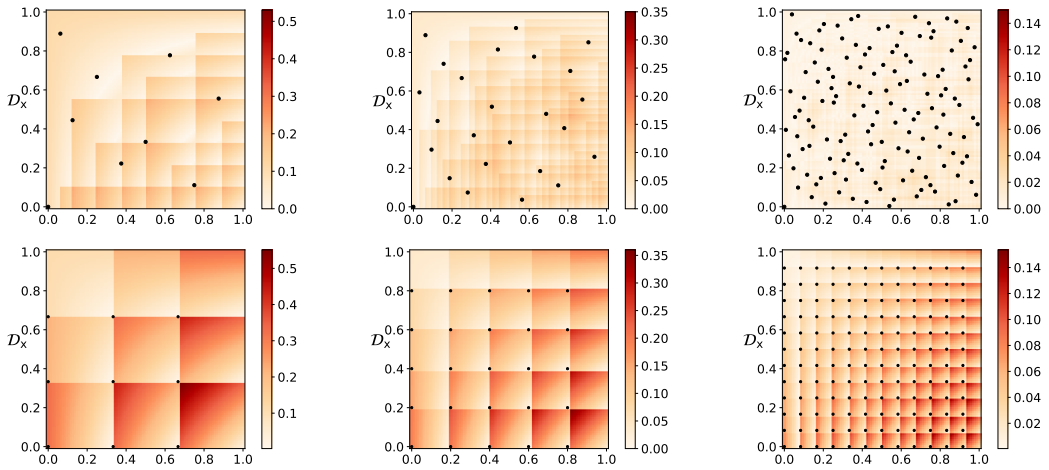


Figure 4.4 – The local discrepancy $\mathcal{D}_x$ in $[0,1]^2$ for two designs: in the top panel, the Halton sequence with $N \in \{9, 25, 144\}$; in the bottom panel, the uniform grid with $N \in \{9, 25, 144\}$.

The multi-dimensional version was proved by Hlawka, 1961. When $d \geq 2$, the HK variation has a more complicated expression through the Vitali variation. For the sake of convenience, we only mention that if f has continuous mixed first partial derivatives, the HK variation reads

$$\|f\|_{\mathrm{HK}} = \sum_{S \subset [d], S \neq \emptyset} \int_{[0,1]^{|S|}} \left| \frac{\partial^{|S|} f(x_S; 1_{\bar{S}})}{\otimes_{s \in S} \partial x_s} \right| \otimes_{s \in S} \mathrm{d}x_s. \tag{4.33}$$

Other variations of the Koksma-Hlawka inequality exist: the star-discrepancy is replaced by the $\mathbb{L}_p$-discrepancies, and the HK variation is replaced by other variation functions; see for example Theorem 3.9 in (Dick and Pillichshammer, 2014). In particular, the $\mathbb{L}_2$-discrepancy has a tractable formula (Warnock, 1972). This is to be compared to the star discrepancy for witch no tractable formula is known: the computation of the star discrepancy is an NP-hard problem (Gnewuch et al., 2009). We shall give, in Section 4.2, an interpretation of the $\mathbb{L}_2$-discrepancy in the context of kernel quadrature.

The decoupling of the upper bound in the Koksma-Hlawka inequality, and its variants, allows to focus only on the discrepancy of the design when the function f is sufficiently regular.

Low discrepancy sequences

From a distance, a uniform grid looks "uniform" in the hypercube $[0,1]^d$. Yet, by considering the star-discrepancy of uniform grids, we reach the counter-intuitive conclusion that the uniform grids are not convenient for numerical integration. Indeed, consider the grid $\boldsymbol{x}_N \subset \mathbb{R}$, where $N = n^d$ for some $n \in \mathbb{N}^*$, defined by

$$\boldsymbol{x}_N = \left\{ \left(\frac{n_1}{n}, \ldots, \frac{n_d}{n} \right), 0 \leq n_\delta < n, \forall \delta \in [d] \right\}. \tag{4.34}$$

Then by Proposition 3.32 in (Dick and Pillichshammer, 2010)

$$\Delta_*(\boldsymbol{x}_N) = 1 - (1 - 1/n)^d \geq \frac{1}{N^{1/d}}. \tag{4.35}$$

In other words, in order to get a star-discrepancy lower than some level $\epsilon \in [0,1[$, $n = \Omega(\epsilon^{1/d})$ nodes are needed: the star discrepancy of the uniform grid scales poorly with the dimension. Looking for designs having better scaling, of discrepancy, in high dimension is a topic of intense research. A very brief review of this line of research is provided in the following.

Van der Corput sequences are the simplest examples of low discrepancy sequences. They are defined on $[0,1]$ and allow the construction of low-discrepancy sequences in $[0,1]^d$. The construction of these sequences goes as follow: consider $b \geq 2$ an integer, and take for every $n \in \mathbb{N}^*$ the b-adic expansion

$$n = \sum_{i \in \mathbb{N}} n_i b^i. \tag{4.36}$$

The b-adic Van der Corput sequence is the sequence $(v_{b,n})_{n \in \mathbb{N}}$, where for every $n \in \mathbb{N}$,

$$v_{b,n} = \sum_{i \in \mathbb{N}} n_i b^{-i}. \tag{4.37}$$

For instance, for $b = 2$, the first elements of the sequence are $0, \frac{1}{2}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \ldots$. While, for $b = 10$, the elements of the sequence goes as follows

$$\frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}, \frac{1}{100}, \frac{11}{100}, \ldots, \frac{91}{100}, \ldots$$

.

Given d Van der Corput sequences associated to integers $b_1, \ldots, b_d$, one can construct a sequence defined on $[0,1]^d$ by concatenation:

$$x_n = (v_{b_1,n}, \ldots, v_{b_d,n}), \ n \in \mathbb{N}^*. \tag{4.38}$$

These are the Halton sequences (Halton, 1964). Another possible construction is the Hammersley configuration defined for a fixed value of N:

$$x_n = ((n-1)/N, v_{b_1,n}, \ldots, v_{b_{d-1},n}), \ n \in \mathbb{N}^*. \tag{4.39}$$

These cryptic configurations have an interesting property: they are low discrepancy sequences on $[0,1]^d$. Indeed, under the assumption that the $b_1, \ldots, b_d$ are pairwise relatively prime, the star discrepancy of the Halton sequence is $\mathcal{O}\big((\log N)^d / N\big)$; while the star discrepancy of the Hammersley sequence is $\mathcal{O}\big((\log N)^{d-1} / N\big)$: for a sufficiently regular function, and for a sufficiently large number of nodes $(N = \Omega(e^d))$, the rate of convergence of the Quasi-Monte Carlo quadrature based on these sequences would be faster compared to the Monte Carlo rate $\mathcal{O}(N^{-1/2})$. Figure 4.4 compares the local discrepancy of the Halton sequence and the uniform grid for $N \in \{9, 25, 144\}$: the Halton sequence have always the smallest star-discrepancy that decreases faster as a function of N.

Now, if the function f have a higher degree of smoothness $s \in \mathbb{N}^*$, i.e. f have square-integrable partial derivatives $\frac{\partial^{u_1+\cdots+u_d} f}{\partial x_1^{u_1} \ldots \partial x_d^{u_d}}$ for every $(u_1, \ldots, u_d) \in \{0, \ldots, s\}^d$, then the QMC rule based on Higher-order digital nets converge faster as $\mathcal{O}(\log(N)^{2sd} N^{-2s})$ (Dick and Pillichshammer, 2014)[Theorem 5]. However, the algorithmic construction of

these sequences is very technical and we refer to (Dick and Pillichshammer, 2010) for details on these constructions.

As it was mentioned in the introduction of this section, this review is limited to the fundamental aspects of Quasi-Monte Carlo methods relevant for later discussions, and many topics have been left out. We refer to (Dick and Pillichshammer, 2014) for more details on the topic.

4.1.3 *DPPs for numerical integration*

Several quadratures are based on nodes that follow the distribution of a Determinantal Point Process. The motivation behind this choice is the strong statistical properties of these quadratures. We review in this section the results and the techniques used in this line of research.

The quadrature based on the Circular Unitary Ensemble

We start by the case of the Circular Unitary Ensemble (CUE) presented in Section 2.2.4. Remember that it corresponds to the set of the eigenvalues of a random matrix chosen from Haar measure on the unitary group $\mathbb{U}_N$. This DPP is defined on the unit complex circle $\mathbb{U} = \{u \in \mathbb{C}, |u| = 1\}$ equipped with Lebesgue measure, and the corresponding kernel

$$\mathfrak{K}(x, y) = \sum_{n \in [N]} e^{2\pi n \mathbf{i}(x-y)}. \tag{4.40}$$

The first theoretical analysis of a quadrature based on the CUE can be tracked to the work of Diaconis and Shahshahani (Diaconis and Shahshahani, 1994). Indeed, the following identity satisfied by any unitary matrix $M \in \mathbb{U}_N$ with eigenvalues $e^{\mathbf{i}\theta_1}, \dots, e^{\mathbf{i}\theta_N}$,

$$\forall \ell \in \mathbb{N}^*, \ \frac{1}{N} \sum_{n \in [N]} (e^{\mathbf{i}\theta_n})^\ell = \frac{1}{N} \operatorname{Tr} M^\ell, \tag{4.41}$$

can be seen as a quadrature formula applied to the functions $z \mapsto z^\ell$: the nodes are the $e^{\mathbf{i}\theta_1}, \dots, e^{\mathbf{i}\theta_N}$ and the weights are uniform. The authors proved that when M is chosen from Haar measure on the unitary group $\mathbb{U}_N$, $\operatorname{Tr} M^\ell$ converges in distribution to $\sqrt{\ell/2}(X + \mathbf{i}Y)$, where X and Y are two independent standard normal random variables. This convergence is reminiscent of Weyl criterion (4.28) of uniformly distributed modulo one. In particular, we conclude that the elements of the CUE are *very regularly spread out* on the unit disc as N goes to infinity; see Figure 2.2 for an illustration. This line of work was pursued by Johansson who extended the analysis to trigonometric series.

Theorem 4.3 (Johansson, 1997)**.** *Let $(e^{\mathbf{i}\theta_n})_{n \in [N]}$ be the CUE and let g be a real-valued function on $\mathbb{U}$ with*

$$g(e^{\mathbf{i}\theta}) = \sum_{m=-\infty}^{+\infty} \hat{g}_m e^{\mathbf{i}m\theta}, \tag{4.42}$$

such that $\hat{g}_0 = 0$ and $\sum_{m \in \mathbb{N}^*} m|\hat{g}_m|^2 < +\infty$. Then

$$\frac{1}{N} \sum_{n \in [N]} g(e^{\mathbf{i}\theta_n}) \xrightarrow[N \to +\infty]{law} \mathcal{N}(0, \frac{2}{N} \sum_{m \in \mathbb{N}^*} m|\hat{g}_m|^2). \tag{4.43}$$

In other words, the quadrature based on CUE nodes converges at the asymptotic rate $\mathcal{O}(1/N^2)$ compared to the rate $\mathcal{O}(1/N)$ of vanilla Monte Carlo, for every function g satisfying the condition $\sum_{m\in\mathbb{N}^*} m|\hat{g}_m|^2 < +\infty$.

Quadratures based on Orthogonal Polynomial Ensembles

Bardenet and Hardy, 2020 proved the first CLT for a DPP based quadrature defined on a domain of arbitrary dimension. This time, the nodes follow an Orthogonal Polynomial Ensemble (OPE): the marginal kernel is defined through orthogonal polynomials with respect to some measure $\mathrm{d}\omega$ that factorizes as

$$\mathrm{d}\omega(x) = \prod_{\delta\in[d]} \mathrm{d}\omega_\delta(x_\delta), \tag{4.44}$$

where the measures $\mathrm{d}\omega_\delta$, defined on $[-1,1]$, belong to what is called the *Nevai class*.

Definition 4.2. *Let $\mathrm{d}\tilde{\omega}$ a measure supported on $[-1,1]$. $\mathrm{d}\tilde{\omega}$ is said to be of Nevai class if the recurrence coefficients, defined in (4.15) and (4.16), for the associated orthonormal polynomials satisfy*

$$\lim_{m\to+\infty} a_m = 1/2, \quad \lim_{m\to+\infty} b_m = 0. \tag{4.45}$$

The simplest example of a Nevai class measure is the measure $\mathrm{d}\tilde{\omega}^*(x) = \tilde{w}^*(x)\mathrm{d}x$ with

$$\tilde{w}^*(x) = \frac{1}{\pi\sqrt{1-x^2}}. \tag{4.46}$$

It corresponds to the particular case when $a_m = a_m^* = \frac{1}{2}$ and $b_m = b_m^* = 0$ for $m \in \mathbb{N}^*$: the corresponding orthonormal polynomials are the Chebyshev polynomials of the first kind, that we denote $(T_m)_{m\in\mathbb{N}}$. In a nutshell, the Nevai class contains measures that have three-term recurrence coefficients converging to the corresponding coefficients of $\mathrm{d}\tilde{\omega}^*$. This class contains, at least, all measures that are absolutely continuous with respect to the Lebesgue measure with strictly positive densities; see Section 1.4 of (Simon, 2010) for details.

Now, the marginal kernel of an OPE is constructed as follows: for $\delta \in [d]$, denote by $(P_m^\delta)_{m\in\mathbb{N}}$ the orthonormal polynomials with respect to $\mathrm{d}\omega_\delta$, and let

$$\mathfrak{K}_N(x,y) = \sum_{n\in[N]} \boldsymbol{P}_{\boldsymbol{u}(n)}(x)\boldsymbol{P}_{\boldsymbol{u}(n)}(y), \tag{4.47}$$

where the $\boldsymbol{u}_n \in \mathbb{N}^d$ are multi-indices ordered with respect to the graded lexicographic order, and $\boldsymbol{P}_{\boldsymbol{u}(n)}$ are multidimensional orthogonal polynomials

$$\boldsymbol{P}_{\boldsymbol{u}(n)}(x) = \prod_{\delta\in[d]} P^\delta_{\boldsymbol{u}(n)_\delta}(x_\delta). \tag{4.48}$$

The authors studied the quadrature based on OPE, and for every node x_n the corresponding weight $w_n = 1/\mathfrak{K}_N(x_n,x_n)$. Among settings considered in the article, we recall the following theorem.

Theorem 4.4 (Theorem 3 in (Bardenet and Hardy, 2020)). *Let $\boldsymbol{x}$ be an OPE with respect to the measure*

$$\mathrm{d}\omega(x) = w(x)\prod_{\delta\in[d]}\mathrm{d}x_\delta := \prod_{\delta\in[d]}(1-x_\delta)^{\alpha_\delta}(1+x_\delta)^{\beta_\delta}\mathbb{1}_{[-1,1]}(x_\delta)\mathrm{d}x_\delta, \tag{4.49}$$

where $\alpha_1,\beta_1,\dots,\alpha_d,\beta_d > -1$; and denote $w^*(x) = \prod_{\delta\in[d]}\tilde{w}^*(x_\delta)$, where $\tilde{w}^*$ is defined in (4.46). Let f be a real valued function defined on $[-1,1]^d$ that vanishes at the boundary of $[-1,1]^d$ and it is of class $\mathcal{C}^1$ in $]-1,1[^d$. Then

$$\lim_{N\to+\infty} N^{1+1/d}\,\mathbb{E}\left(\sum_{n\in[N]}\frac{f(x_n)}{\mathfrak{K}_N(x_n,x_n)} - \int_{[-1,1]^d} f(x)\mathrm{d}\omega(x)\right)^2 = \sigma^2(f), \tag{4.50}$$

and

$$\sqrt{N^{1+1/d}}\left(\sum_{n\in[N]}\frac{f(x_n)}{\mathfrak{K}_N(x_n,x_n)} - \int_{[-1,1]^d} f(x)\mathrm{d}\omega(x)\right) \to \mathcal{N}(0,\sigma^2(f)), \tag{4.51}$$

where

$$\sigma^2(f) = \frac{1}{2}\sum_{m_1,\dots,m_d=0}^{+\infty}(m_1+\cdots+m_d)\widehat{\frac{fw}{w^*}}(m_1,\dots,m_d)^2, \tag{4.52}$$

and

$$\hat{\varphi}(m_1,\dots,m_d) = \int_{[-1,1]^d}\varphi(u_1,\dots,u_d)\prod_{\delta\in[d]}T_{m_\delta}(u_\delta)\frac{1}{\pi\sqrt{1-u_\delta^2}}\mathrm{d}u_\delta. \tag{4.53}$$

In other words, the quadrature based on the OPE with the reference measure $\mathrm{d}\omega$ converges with an asymptotic rate $\mathcal{O}(N^{-1-\frac{1}{d}})$ for functions of class $\mathcal{C}^1$. As we can see, this rate is better than the Monte Carlo rate $\mathcal{O}(N^{-1})$ but the improvement "shrinks" as the dimension increases, because of the term $1/d$. This rate is also valid for tensor products of Nevai measures under some technical conditions as it was proved by the authors. In practice, as we have seen in Section 2.2.5, the numerical sampling of a general OPE requires the computation of the orthonormal polynomials $\boldsymbol{P}_{\boldsymbol{u}(n)}$ with respect to the measure $\mathrm{d}\omega$; which is possible for the measures (4.49), also called *Jacobi measures*.

Moreover, the quadrature, after scaling by the factor $\sqrt{N^{1+1/d}}$, satisfies a CLT with an asymptotic variance $\sigma^2(f)$ that depends on the coefficients of the function fw/w^* on the orthonormal basis defined by $\boldsymbol{T}_{\boldsymbol{u}(n)} = \prod_{\delta\in[d]} T_{\boldsymbol{u}(n)_\delta}$. In particular, when $w = w^*$, the asymptotic variance writes

$$\sigma^2(f) = \frac{1}{2}\sum_{m_1,\dots,m_d=0}^{+\infty}(m_1+\cdots+m_d)\widehat{f}(m_1,\dots,m_d)^2, \tag{4.54}$$

can be interpreted as a semi-norm in a Hilbert space that reflects the regularity of the integrand f. Indeed, the authors proved the inequality [Proposition 1 in (Bardenet and Hardy, 2020)]

$$\sigma^2(f) \le \frac{1}{2}\sum_{\delta\in[d]}\int_{[-1,1]^d}\left(\sqrt{1-x_\delta^2}\partial_\delta f(x_1,\dots,x_d)\right)^2\prod_{\delta\in[d]}\frac{\mathrm{d}x_\delta}{\pi\sqrt{1-x_\delta^2}}, \tag{4.55}$$

the r.h.s of (4.55) depends on the fluctuations of the function f: $\sigma^2(f)$ is a measure of the smoothness of the function f. This connection between the smoothness of the function f and the statistical properties of a quadrature based on DPP nodes was confirmed in another setting.

Another CLT using the Dirichlet kernel

In (Coeurjolly, Mazoyer, and Amblard, 2020), the authors proposed to study a quadrature based on a different projection DPP and using uniform weights. The integrand f was assumed to be square integrable and periodic on the hypercube $[0,1]^d$. The repulsion kernel $\mathfrak{K}$ is constructed as follows. Let $N = \prod_{\delta\in[d]} N_\delta \in \mathbb{N}^*$ for some $N_1, \ldots, N_d \in \mathbb{N}^*$, and the projection DPP be defined by

$$\mathfrak{K}(x,y) = \sum_{m\in\mathcal{M}_N} e^{2\mathbf{i}\pi m^\intercal(x-y)}, \tag{4.56}$$

where

$$\mathcal{M}_N = \prod_{\delta\in[d]} \{0,\ldots,N_\delta - 1\}. \tag{4.57}$$

It can be shown that this is equivalent to define $\mathfrak{K}$ to be the tensor product of Dirichlet kernels; and $\mathfrak{K}$ is called (N,d)-Dirichlet DPP by simplification. The proposed quadrature defines an unbiased estimator of the integral $\int_{[0,1]^d} f(u_1,\ldots,u_d) \otimes_{\delta\in[d]} \mathrm{d}u_\delta$ with a variance that depends on the Fourier coefficients of f.

Proposition 4.1. *[Proposition 3.1, (Coeurjolly et al., 2020)] Let $\mathbf{x}$ be a random set of $[0,1]^d$ that follows an (N,d)-Dirichlet DPP, and $f \in \mathbb{L}_2([0,1]^d)$ and periodic. Define*

$$\hat{I}_N(f) = \frac{1}{N} \sum_{n\in[N]} f(x_n). \tag{4.58}$$

Then, $\hat{I}_N(f)$ is an unbiased estimator of $\int_{[0,1]^d} f(u_1,\ldots,u_d) \otimes_{\delta\in[d]} \mathrm{d}u_\delta$, and

$$\mathbb{V}(\hat{I}_N(f)) = \frac{1}{N} \sum_{m\in\mathbb{Z}^d} |\hat{f}_m|^2 - \frac{1}{N^2} \sum_{m_1,m_2\in\mathcal{M}_N} |\hat{f}_{m_1-m_2}|^2, \tag{4.59}$$

where

$$\hat{f}_m = \int_{[0,1]^d} f(u_1,\ldots,u_d) e^{-2\pi\mathbf{i}m^\intercal u} \otimes_{\delta\in[d]} \mathrm{d}u_\delta. \tag{4.60}$$

This result was the first step for the asymptotic analysis of the variance of $\hat{I}_N(f)$ under the assumption that $N_\delta = N_\delta(N)$, and there exists constants $\alpha_1,\ldots,\alpha_d$ such that

$$\forall \delta \in [d], \lim_{N\to+\infty} N_\delta(N) N^{-1/d} = \alpha_\delta. \tag{4.61}$$

Define $\mathcal{F}_s([0,1]^d)$ to be the subspace of periodic elements of $\mathbb{L}_2([0,1]^d)$ that satisfy the condition

$$\sum_{m\in\mathbb{Z}^d} (1 + \|m\|_\infty)^{2s} |\hat{f}_m|^2 < +\infty. \tag{4.62}$$

Theorem 4.5. *Let $\mathbf{x}$ and f as in Proposition 4.1. Assume that $f \in \mathcal{F}_s([0,1]^d)$, then*

- *if* $s \in [0, 1/2]$, *then*

$$\mathbb{V}(\hat{I}_N(f)) = \mathcal{O}(N^{-1-\frac{2s}{d}}), \tag{4.63}$$

- *If* $s > 1/2$, *then*

$$\lim_{N \to +\infty} N^{1+1/d} \, \mathbb{V}(\hat{I}_N(f)) = \sigma^2(f) := \sum_{m \in \mathbb{Z}^d} \Big(\sum_{\delta=1}^{d} \frac{|m_\delta|}{\alpha_\delta}\Big) |\hat{f}_m|^2. \tag{4.64}$$

- *Moreover, if* $s > 1/2$ *and* $\|f\|_\infty < +\infty$, *then*

$$\sqrt{N^{1+1/d}} \left(\hat{I}_N(f) - \int_{[0,1]^d} f(u_1, \dots, u_d) \otimes_{\delta \in [d]} \mathrm{d}u_\delta \right) \to \mathcal{N}(0, \sigma^2(f)). \tag{4.65}$$

In other words, under the assumption that the integrand f belongs to the functional subspace $\mathcal{F}_s([0,1]^d)$, that can be seen as a Sobolev space, the variance of the estimator $\mathbb{V}(\hat{I}_N(f))$ converges at the rate $\mathcal{O}(N^{-1-\frac{\min(2s,1)}{d}})$; and a CLT holds when $s > 1/2$ and the function is essentially bounded. Again, the improvement upon the Monte Carlo rate worsens with the dimension, and the asymptotic variance involves the constant

$$\sigma^2(f) = \sum_{m \in \mathbb{Z}^d} \left(\sum_{\delta=1}^{d} \frac{|m_\delta|}{\alpha_\delta} \right) |\hat{f}_m|^2, \tag{4.66}$$

that is reminiscent of a Sobolev norm.

To sum up, we reviewed three settings where a DPP-based quadrature leads to a fast asymptotic rate of convergence compared to Monte Carlo. These fast rates are achieved when the integrand belongs to some functional subspace, a Sobolev space, and the asymptotic variance involves a constant that can be interpreted as a norm on the functional subspace.

These observations suggest the possibility to define DPP-based quadratures on other functional spaces. Indeed, we introduce a class of DPP-based quadratures in Section 4.3 for functions living in a generic reproducing kernel Hilbert space, that can be seen as generalizing Sobolev spaces. This new class of quadratures fit within the kernel framework for quadratures that we recall in the following section.

4.2 THE KERNEL QUADRATURE FRAMEWORK

The analysis of quadrature using kernel methods is convenient and elegant and it is expressed using functional analysis tools, similarly to DPPs. Added to that, many results in quasi-Monte Carlo theory have an interpretation within this framework.

The strength of this framework is its flexibility: the integration domain $\mathcal{X}$ is assumed to be a metric space endowed with a Borel measure ω, and we consider approximating an integral $\int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x)$, where $f, g \in \mathbb{L}_2(\mathrm{d}\omega)$ and f is continuous, using a quadrature rule $(\boldsymbol{x}, \boldsymbol{w}) \in \mathcal{X}^N \times \mathbb{R}^N$:

$$\int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x) \approx \sum_{n \in [N]} w_n f(x_n). \tag{4.67}$$

Here $\mathbb{L}_2(\mathrm{d}\omega)$ is the Hilbert space of square integrable, real-valued functions defined on $\mathcal{X}$, with the usual inner product denoted by $\langle \cdot, \cdot \rangle_{\mathrm{d}\omega}$, and the associated norm by $\|.\|_{\mathrm{d}\omega}$.

We recall in the following some classic results on kernel quadrature. In particular, in Section 4.2.1, the definition of an RKHS is recalled along with our assumptions; in Section 4.2.2, Mercer's theorem and its various extensions; in Section 4.2.3, we give some examples of kernels widely used in the literature on kernel quadrature; in Section 4.2.4, the definition of the worst case integration error on an RKHS is recalled; in Section 4.2.5 we recall the definition of optimal kernel quadrature along with some of its properties; and in Section 4.2.6 we review the existing methods of node design for optimal kernel quadrature.

4.2.1 *Reproducing Kernel Hilbert Spaces*

Let $\mathcal{X}$ be a metric space. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is said to be a kernel over $\mathcal{X}$ if k is symmetric and for any finite set of points in $\mathcal{X}$, the matrix of pairwise kernel evaluations is positive semi-definite. A reproducing kernel Hilbert space (RKHS) over $\mathcal{X}$ is a Hilbert space $\mathcal{F}$ endowed with a kernel k that satisfies the following two properties Berlinet and Thomas-Agnan, 2011

- the continuity property: for every $x \in \mathcal{X}$, the evaluation function $f \mapsto f(x)$ is continuous with respect to the norm of $\mathcal{F}$, that is

$$\forall x \in \mathcal{X},\ \exists M_x > 0,\ \ \forall f \in \mathcal{F},\ |f(x)| \leq M_x \|f\|_{\mathcal{F}},$$

- the reproducibility property

$$\forall (x, f) \in \mathcal{X} \times \mathcal{F},\ f(x) = \langle f, k(x, .) \rangle_{\mathcal{F}}.$$

We suppose the following assumptions on the measure $\mathrm{d}\omega$ and the kernel k. We make the following assumption on $\mathrm{d}\omega$.

Assumption 4.1. *ω is a* non-degenerate *measure, i.e. the support of ω is equal to $\mathcal{X}$.*

Assumption 4.2. *k is continuous with respect to the product topology of $\mathcal{X} \times \mathcal{X}$.*

Assumption 4.3. *$x \mapsto k(x, x)$ is integrable with respect to $\mathrm{d}\omega$ so that $\mathcal{F} \subset \mathbb{L}_2(\mathrm{d}\omega)$.*

An example: the periodic Sobolev spaces

Let $\mathcal{X} = [0, 1]$ equipped with the uniform measure ω, and define for $s \in \mathbb{N}^*$ the kernel

$$k_s(x, y) = 1 + 2 \sum_{m \in \mathbb{N}^*} \frac{1}{m^{2s}} \cos(2\pi m(x - y)). \tag{4.68}$$

The kernel k_s can be expressed in closed form using Bernoulli polynomials (Wahba, 1990)

$$k_s(x, y) = 1 + \frac{(-1)^{s-1}(2\pi)^{2s}}{(2s)!} B_{2s}(\{x - y\}). \tag{4.69}$$

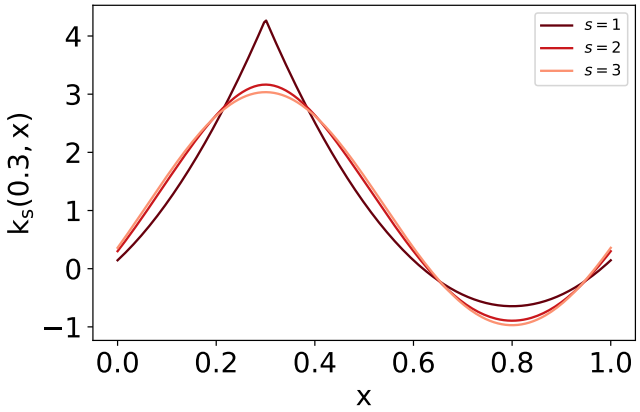


Figure 4.5 – The evaluation of the kernel translate $x \mapsto k_s(0.3, x)$ for $s \in \{1, 2, 3\}$.

The corresponding RKHS $\mathcal{F} = \mathcal{S}_s$ is the periodic Sobolev space of order s: an element of $\mathcal{S}_s$ is a function f defined on $[0, 1]$, that have the derivative of order s defined in the sense of distributions such that $f^{(s)} \in \mathbb{L}_2(\mathrm{d}\omega)$, and

$$\forall i \in \{0, \dots, s-1\}, \ f^{(i)}(0) = f^{(i)}(1). \tag{4.70}$$

Figure 4.5 illustrates the graphs of the kernel translates $x \mapsto k_s(0.3, x)$ for $s \in \{1, 2, 3\}$. The elements of $\mathcal{S}_s$ are characterized by their Fourier series coefficients (see Example 19, Chapter 7 in (Berlinet and Thomas-Agnan, 2011))

$$f \in \mathcal{S}_s \iff \sum_{m \in \mathbb{N}^*} m^{2s} \left[\left(\int_0^{2\pi} f(x) \cos(2\pi m x) \mathrm{d}x \right)^2 + \left(\int_0^{2\pi} f(x) \sin(2\pi m x) \mathrm{d}x \right)^2 \right] < +\infty. \tag{4.71}$$

4.2.2 *The Mercer decomposition*

As we have seen in the example given in the previous section, periodic Sobolev spaces are characterized by the Fourier coefficients of their elements. This property is not specific to periodic Sobolev spaces and it is satisfied by other RKHSs. Indeed, recall that under the Assumption 4.3, the RKHS $\mathcal{F}$ is a subspace of $\mathbb{L}_2(\mathrm{d}\omega)$, and the Mercer's theorem gives a finer description of $\mathcal{F}$ as a subspace of $\mathbb{L}_2(\mathrm{d}\omega)$.

The Mercer's theorem was first proven when $\mathcal{X} = [0, 1]$ and ω is the Lebesgue measure in (Mercer, 1909). A modern proof of this classic result can be found in (Lax, 2002). An extension to a general compact space $\mathcal{X}$ can be found in (Cucker and Zhou, 2007).

Theorem 4.6. *Assume that $\mathcal{X}$ is a compact space and ω is a finite Borel measure on $\mathcal{X}$ with full support (non-degenerate measure), and define the integration operator*

$$\mathbf{\Sigma} f(x) = \int_{\mathcal{X}} k(x, y) f(y) \mathrm{d}\omega(y). \tag{4.72}$$

Then, there exists an orthonormal basis $(e_n)_{n\in\mathbb{N}^*}$ *of* $\mathbb{L}_2(\mathrm{d}\omega)$ *consisting of eigenfunctions of* $\mathbf{\Sigma}$*, and the corresponding eigenvalues are non-negative. The eigenfunctions corresponding to non-vanishing eigenvalues can be taken to be continuous, and the kernel* k *writes*

$$k(x,y) = \sum_{n\in\mathbb{N}^*} \sigma_n e_n(x) e_n(y), \tag{4.73}$$

where the convergence is absolute and uniform.

The existence of the orthonormal basis $(e_m)_{m\in\mathbb{N}^*}$ is a consequence of the spectral theorem applied to the compact operator $\mathbf{\Sigma}$. The existence of a continuous function in the $\mathrm{d}\omega$-class of equivalence of e_m when $\sigma_m > 0$ is a consequence of

$$e_m(.) = \frac{1}{\sigma_m} \int_{\mathcal{X}} k(.,y) e_m(y) \mathrm{d}\omega(.). \tag{4.74}$$

Moreover, by (4.74) we can take $e_m \in \mathcal{F}$ and, as a consequence of Theorem 4.6, the family $(e_m^{\mathcal{F}} = \sqrt{\sigma_m} e_m; \sigma_m > 0)_{m\in\mathbb{N}^*}$ is an o.n.b. of $\mathcal{F}$; see Theorem 4.12 in (Cucker and Zhou, 2007). In particular

$$\|e_m^{\mathcal{F}}\|_{\mathcal{F}}^2 = \sigma_m \|e_m\|_{\mathcal{F}}^2 = 1, \tag{4.75}$$

and for an element $f \in \mathcal{F}$

$$f = \sum_{m\in\mathbb{N}^*} \langle f, e_m\rangle_{\mathrm{d}\omega} e_m = \sum_{\substack{m\in\mathbb{N}^* \\ \sigma_m>0}} \frac{\langle f, e_m\rangle_{\mathrm{d}\omega}}{\sqrt{\sigma_m}} e_m^{\mathcal{F}}, \tag{4.76}$$

so that for $m \in \mathbb{N}^*$ such that $\sigma_m > 0$

$$\langle f, e_m^{\mathcal{F}}\rangle_{\mathcal{F}} = \frac{\langle f, e_m\rangle_{\mathrm{d}\omega}}{\sqrt{\sigma_m}}, \tag{4.77}$$

and

$$\|f\|_{\mathcal{F}}^2 = \sum_{\substack{m\in\mathbb{N}^* \\ \sigma_m>0}} \frac{\langle f, e_m\rangle_{\mathrm{d}\omega}^2}{\sigma_m}. \tag{4.78}$$

In other words, the RKHS $\mathcal{F}$ is equal to the set [4]

$$\left\{ f \in \mathbb{L}_2(\mathrm{d}\omega), \sum_{m\in\mathbb{N}^*} \frac{\langle f, e_m\rangle_{\mathrm{d}\omega}^2}{\sigma_m} < +\infty \right\}. \tag{4.79}$$

Let $g \in \overline{\mathrm{Span}(e_m)_{\{m\in\mathbb{N}^*;\sigma_m>0\}}} \subset \mathbb{L}_2(\mathrm{d}\omega)$. We have

$$\|g\|_{\mathrm{d}\omega}^2 = \sum_{\substack{m\in\mathbb{N}^* \\ \sigma_m>0}} \langle g, e_m\rangle_{\mathrm{d}\omega}^2 = \sum_{\substack{m\in\mathbb{N}^* \\ \sigma_m>0}} \frac{(\sqrt{\sigma_m}\langle g, e_m\rangle_{\mathrm{d}\omega})^2}{\sigma_m} = \sum_{\substack{m\in\mathbb{N}^* \\ \sigma_m>0}} \frac{\langle f, e_m\rangle_{\mathrm{d}\omega}^2}{\sigma_m} = \|f\|_{\mathcal{F}}^2, \tag{4.80}$$

where $f = \mathbf{\Sigma}^{1/2} g$ with $\mathbf{\Sigma}^{1/2}$ is the operator defined by

[4] If there exists $m \in \mathbb{N}^*$ such that $\sigma_m = 0$ and $\langle f, e_m\rangle_{\mathrm{d}\omega} \neq 0$, then we write $\sum_{m\in\mathbb{N}^*} \langle f, e_m\rangle_{\mathrm{d}\omega}^2/\sigma_m = +\infty$ and $f \notin \mathcal{F}$.

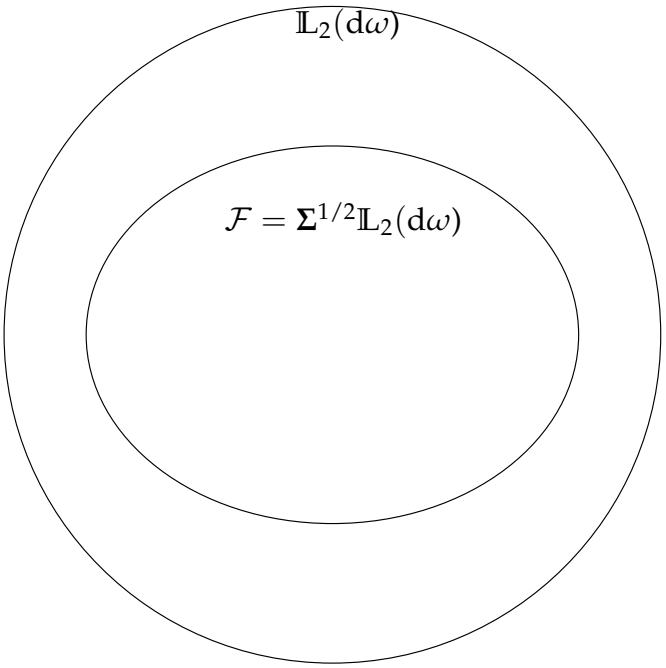


Figure 4.6 – The RKHS $\mathcal{F}$ is an ellipsoid in $\mathbb{L}_2(\mathrm{d}\omega)$.

$$\begin{aligned} \boldsymbol{\Sigma}^{1/2} : \mathbb{L}_2(\mathrm{d}\omega) &\longrightarrow \mathbb{L}_2(\mathrm{d}\omega) \\ \sum_{m\in\mathbb{N}^*} a_m e_m &\longmapsto \sum_{m\in\mathbb{N}^*} a_m \sqrt{\sigma_m} e_m. \end{aligned}$$

In other words, $\boldsymbol{\Sigma}^{1/2}$ is an isomorphism of Hilbert spaces between $\overline{\mathrm{Span}(e_m)_{\{m\in\mathbb{N}^*;\sigma_m>0\}}}$ and $\mathcal{F}$. In particular, for every $f \in \mathcal{F}$, there exists $g \in \mathbb{L}_2(\mathrm{d}\omega)$ such that

$$f = \boldsymbol{\Sigma}^{1/2} g, \tag{4.81}$$

with $\|f\|_{\mathcal{F}} = \|g\|_{\mathrm{d}\omega}$. By Assumption 4.1, ω is non-degenerate, thus $\mathcal{F}$ is dense in $\mathbb{L}_2(\mathrm{d}\omega)$ if and only if all the eigenvalues σ_m are positive. Under this condition, $\boldsymbol{\Sigma}^{1/2}$ is an isomorphism between $\mathbb{L}_2(\mathrm{d}\omega)$ and $\mathcal{F}$. Finally, note that

$$\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}. \tag{4.82}$$

In other words, $\boldsymbol{\Sigma}^{1/2}$ is the positive self-adjoint square root of $\boldsymbol{\Sigma}$.

The RKHS $\mathcal{F}$ can be seen as an ellipsoid in $\mathbb{L}_2(\mathrm{d}\omega)$ as it is illustrated in Figure 4.6: the principal axis correspond to the directions spanned by the eigenfunctions e_m.

The compactness assumption in Theorem 4.6 excludes some classic RKHSs defined on non-compact domains such as the RKHS corresponding to the Gaussian kernel on $\mathbb{R}^d$ equipped with the Gaussian measure. Hence the need to an extension of this result to non-compact spaces.

In Sun, 2005, the author extended Theorem 4.6 to $\mathcal{X} = \cup_{i\in\mathbb{N}}\mathcal{X}_i$, with the $\mathcal{X}_i$s are compacts and $\omega(\mathcal{X}_i) < \infty$. One can also extend Mercer's theorem under a *compact*

embedding assumption (Steinwart and Scovel, 2012): the RKHS $\mathcal{F}$ associated to k is said to be compactly embedded in $\mathbb{L}_2(\mathrm{d}\omega)$ if the operator

$$\begin{aligned} I_{\mathcal{F}} : \mathcal{F} &\longrightarrow \mathbb{L}_2(\mathrm{d}\omega) \\ f &\longmapsto f \end{aligned}$$

is compact. This is guaranteed by Assumption 4.3 (Lemma 2.3, (Steinwart and Scovel, 2012)).

Now, under the compact embedding assumption, the pointwise convergence of the Mercer decomposition to the kernel k is equivalent to the injectivity of the embedding $I_{\mathcal{F}}$ (Theorem 3.1, (Steinwart and Scovel, 2012)). This is guaranteed by Assumption 4.1 (Chapter 4 in (Steinwart and Christmann, 2008)).

These conditions are satisfied by a large class of kernels; particularly the Gaussian kernel on $\mathbb{R}^d$ equipped with the Gaussian measure.

4.2.3 *Examples*

The periodic Sobolev spaces

The Mercer decomposition of k_s with respect to ω is explicitly given by (4.68)

$$\forall x, y \in [0,1],\ k_s(x,y) = 1 + \sum_{m \in \mathbb{N}^*} \left(\frac{1}{m^{2s}} \sqrt{2} \cos(2\pi m x) + \frac{1}{m^{2s}} \sqrt{2} \sin(2\pi m x) \right). \quad (4.83)$$

Indeed, define for $n \in \mathbb{N}^*$, define $e_n^{\mathcal{S}_s} : [0,1] \to \mathbb{R}$ by

$$e_n^{\mathcal{S}_s}(x) = \begin{cases} 1 & \text{if } n = 1 \\ \sqrt{2}\cos(2\pi n/2x) & \text{if } n > 1 \text{ and n is even} \\ \sqrt{2}\sin(2\pi(n-1)/2x) & \text{if } n > 1 \text{ and n is odd.} \end{cases}$$

The family $(e_n^{\mathcal{S}_s})_{n \in \mathbb{N}^*}$ is an o.n.b of $\mathbb{L}_2(\mathrm{d}\omega)$, and $e_n^{\mathcal{S}_s}$ is an eigenfunction of the integration operator $\mathbf{\Sigma}$ associated to the eigenvalue

$$\sigma_n^{\mathcal{S}_s} = \begin{cases} 1 & \text{if } n = 1 \\ \frac{2^{2s}}{n^{2s}} & \text{if } n > 1 \text{ and n is even} \\ \frac{2^{2s}}{(n-1)^{2s}} & \text{if } n > 1 \text{ and n is odd.} \end{cases}$$

The Korobov spaces

For a given $d \in \mathbb{N}^*$, let $\mathcal{X} = [0,1]^d$ and let ω be the uniform measure on $\mathcal{X}$. The Korobov space in dimension d of order s is the tensor product of d copies of the periodic Sobolev space $\mathcal{S}_s$

$$\mathcal{K}_{d,s} = \{ f \in \mathbb{L}_2(\mathrm{d}\omega);\ f = \prod_{\delta \in [d]} f_\delta,\ f_\delta \in \mathcal{S}_s \}, \quad (4.84)$$

and the corresponding kernel $k_{d,s}$ writes

$$\forall x, y \in [0,1]^d,\ k_{d,s}(x,y) = \prod_{\delta \in [d]} k_s(x_\delta, y_\delta). \quad (4.85)$$

An element $f \in \mathcal{K}_{d,s}$ have square-integrable partial derivatives

$$\frac{\partial^{u_1+\cdots+u_d} f}{\partial x_1^{u_1} \dots \partial x_d^{u_d}} \in \mathbb{L}_2(\mathrm{d}\omega), \tag{4.86}$$

for every $(u_1, \dots, u_d) \in \{0, \dots, s\}^d$. Therefore,

$$\forall x, y \in \mathcal{X},\ k_{d,s}(x,y) = \sum_{u \in (\mathbb{N} \setminus \{0\})^d} \prod_{\delta \in [d]} \sigma_{u_\delta}^{\mathcal{S}_s} e_{u_\delta}^{\mathcal{S}_s}(x_\delta) e_{u_\delta}^{\mathcal{S}_s}(y_\delta). \tag{4.87}$$

We choose an order $\prec_{\mathcal{K}_{d,s}}$ on $(\mathbb{N} \setminus \{0\})^d$ that keeps the eigenvalues of $\mathbf{\Sigma}$ decreasing

$$u \prec_{\mathcal{K}_{d,s}} v \iff \prod_{\delta \in [d]} \sigma_{u_\delta}^{\mathcal{S}_s} \leq \prod_{\delta \in [d]} \sigma_{v_\delta}^{\mathcal{S}_s}. \tag{4.88}$$

This order can be implemented as follows. Define for $n \in \mathbb{N}^*$, let

$$\lfloor n \rfloor_1 = \begin{cases} 1 & \text{if } n = 1 \\ \lfloor (n-1)/2 \rfloor & \text{else} \end{cases}, \tag{4.89}$$

and observe that $\sigma_n^{\mathcal{S}_s} = \lfloor n \rfloor_1^{-2s}$. Therefore,

$$u \prec_{\mathcal{K}_{d,s}} v \iff \left(\prod_{i \in [d]} \frac{1}{1 + \lfloor u_i \rfloor_1} \right)^{2s} \leq \left(\prod_{i \in [d]} \frac{1}{1 + \lfloor v_i \rfloor_1} \right)^{2s} \tag{4.90}$$

$$\iff \sum_{i \in [d]} \log(1 + \lfloor v_i \rfloor_1) \leq \sum_{i \in [d]} \log(1 + \lfloor u_i \rfloor_1). \tag{4.91}$$

Let $(u^n)_{n \in \mathbb{N}^*}$ be the family that contains all the elements of $(\mathbb{N} \setminus \{0\})^d$ ordered according to $\prec_{\mathcal{K}_{d,s}}$. The n-th eigenpair of $\mathbf{\Sigma}$ is given by

$$\forall x \in [0,1]^d,\ e_n^{\mathcal{K}_{d,s}}(x) = \prod_{\delta \in [d]} e_{u_\delta^n}^{\mathcal{S}_s}(x_\delta), \tag{4.92}$$

and

$$\sigma_n^{\mathcal{K}_{d,s}} = \prod_{\delta \in [d]} \sigma_{u_\delta^n}^{\mathcal{S}_s}. \tag{4.93}$$

The left of Figure 4.7 illustrates the eigenvalues σ_{N+1} ordered according to the spectral order $\prec_{\mathcal{K}_{d,s}}$ compared to the rate $\mathcal{O}((\log N)^{2s(d-1)} N^{-2s})$ proved in (Bach, 2017): the eigenvalues σ_{N+1} decreases in plateaus and the size of each plateau corresponds to the multiplicity of the respective eigenvalues. Moreover, the rate $\mathcal{O}((\log N)^{2s(d-1)} N^{-2s})$ matches the asymptotic behaviour of σ_N for $N \to +\infty$.

The Gaussian spaces

Let $\mathcal{X} = \mathbb{R}$ equipped with the measure $\mathrm{d}\omega(x) = w(x)\mathrm{d}x$ defined by

$$w(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{4.94}$$

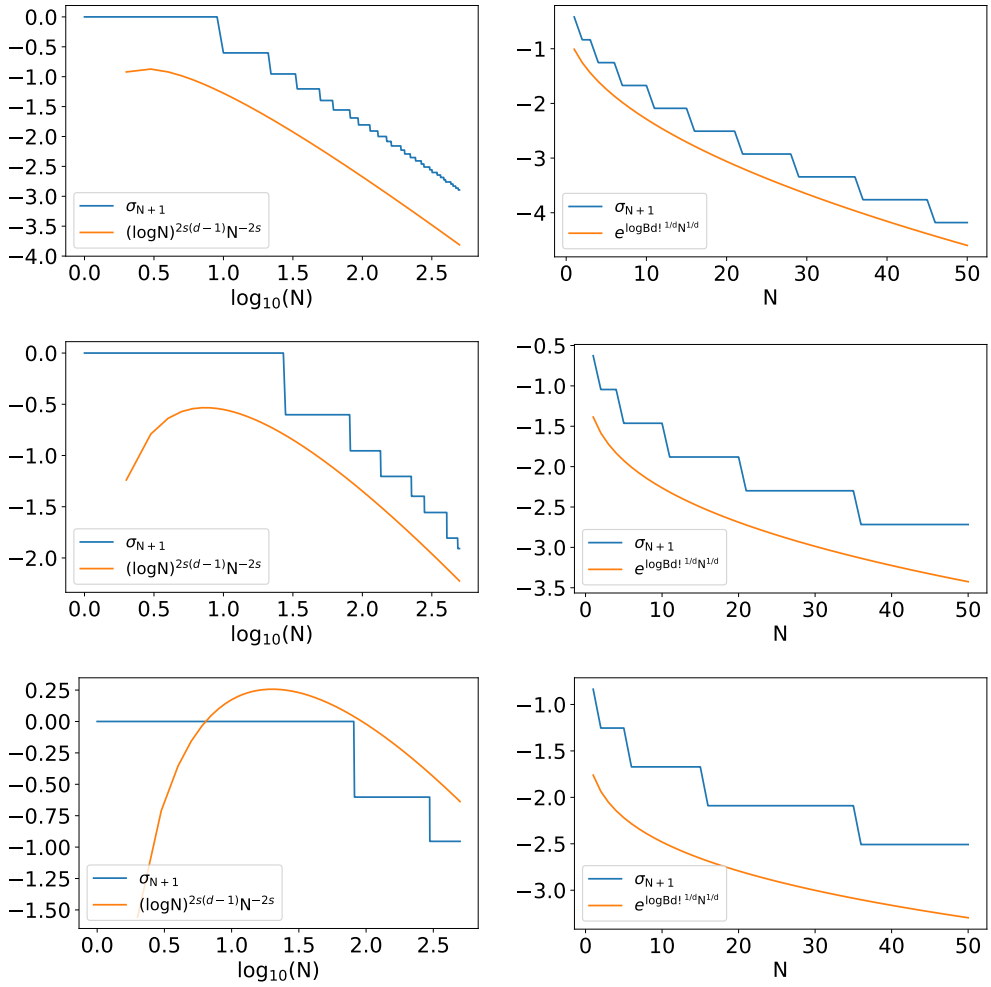


Figure 4.7 – (Left): comparison of σ_{N+1} in the Korobov case according to the spectral order and $(\log N)^{2s(d-1)}N^{-2s}$ for $d \in \{2,3,4\}$ and $s = 1$, (Right): comparison of σ_{N+1} in the Gaussian case according to the spectral order and $\beta^d e^{-\delta d!^{1/d}N^{1/d}}$ for $d \in \{2,3,4\}$ and $\gamma = 1$.

For $\gamma > 0$, consider the kernel

$$k_\gamma(x,y) = e^{-\frac{(x-y)^2}{2\gamma^2}}. \tag{4.95}$$

For notational convenience, we further let

$$a = \frac{1}{4}, \quad b = \frac{1}{2\gamma^2}, \quad c = \sqrt{a^2 + 2ab}, \tag{4.96}$$

and

$$A = a + b + c, \quad B = b/A. \tag{4.97}$$

Now, the Mercer decomposition of k_γ reads Rasmussen, 2003

$$k_\gamma(x,y) = \sum_{m \in \mathbb{N}^*} \sigma_m^{\mathcal{G}_\gamma} e_m^{\mathcal{G}_\gamma}(x) e_m^{\mathcal{G}_\gamma}(y), \tag{4.98}$$

where

$$\sigma_m^{\mathcal{G}_\gamma} = \sqrt{\frac{2a}{A}} B^m, \quad e_m^{\mathcal{G}_\gamma}(x) = \sqrt{\frac{\sqrt{c}}{2^{m-1} m!}} e^{-(c-a)x^2} H_m(\sqrt{2c}x), \tag{4.99}$$

and H_m is the m-th physicists' Hermite polynomial.

The corresponding RKHS is called the Gaussian space and denoted $\mathcal{G}_\gamma$. By extension, we define $\mathcal{G}_{d,\gamma}$, the Gaussian space of dimension d and parameter γ as the tensor product of d copies of the Gaussian space $\mathcal{G}_\gamma$. This RKHS corresponds to the kernel

$$k_{\gamma,d}(x,y) = e^{-\frac{\|x-y\|^2}{2\gamma^2}}, \tag{4.100}$$

and the corresponding Mercer decomposition, with respect to the Gaussian measure on $\mathbb{R}^d$, writes

$$k_{d,\gamma}(x,y) = \sum_{u \in (\mathbb{N} \setminus \{0\})^d} \prod_{\delta \in [d]} \sigma_{u_\delta}^{\mathcal{G}_\gamma} e_{u_\delta}^{\mathcal{G}_\gamma}(x_\delta) e_{u_\delta}^{\mathcal{G}_\gamma}(y_\delta). \tag{4.101}$$

We choose an order $\prec_{\mathcal{G}_{d,\gamma}}$ on $(\mathbb{N} \setminus \{0\})^d$ that keeps the eigenvalues of $\mathbf{\Sigma}$ decreasing

$$u \prec_{\mathcal{G}_{d,\gamma}} v \iff \prod_{\delta \in [d]} \sigma_{u_\delta}^{\mathcal{G}_\gamma} \leq \prod_{\delta \in [d]} \sigma_{v_\delta}^{\mathcal{G}_\gamma}. \tag{4.102}$$

Remember that $\sigma_m^{\mathcal{G}_\gamma} = \sqrt{2a/A} B^m$, therefore, the order $\prec_{\mathcal{G}_{d,\gamma}}$ can be implemented as follows

$$u \prec_{\mathcal{G}_{d,\gamma}} v \iff \prod_{i \in [d]} B^{u_i} \leq \prod_{i \in [d]} B^{v_i} \iff \sum_{i \in [d]} v_i \leq \sum_{i \in [d]} u_i. \tag{4.103}$$

Let $(u^n)_{n \in \mathbb{N}^*}$ be the family that contains all the elements of $(\mathbb{N} \setminus \{0\})^d$ ordered according to $\prec_{\mathcal{G}_{d,\gamma}}$. The n-th eigenpair of $\mathbf{\Sigma}$ is given by

$$e_n^{\mathcal{G}_{d,\gamma}}(x) = \prod_{\delta \in [d]} e_{u_\delta^n}^{\mathcal{G}_\gamma}(x_\delta), \tag{4.104}$$

and

$$\sigma_n^{\mathcal{G}_{d,\gamma}} = \prod_{\delta \in [d]} \sigma_{u_\delta^n}^{\mathcal{G}_\gamma}. \tag{4.105}$$

The right of Figure 4.7 illustrates the eigenvalues σ_{N+1} ordered according to the spectral order $\prec_{\mathcal{G}_{d,\gamma}}$ compared to the rate $\mathcal{O}(e^{\log B d!^{1/d} N^{1/d}})$ proved in (Bach, 2017).

4.2.4 *The integration error*

When the integrand f belongs to $\mathcal{F}$, the integration error reads (Smola et al., 2007)

$$\left| \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x) - \sum_{j\in[N]} w_j f(x_j) \right| = \left| \langle f, \mu_g - \sum_{j\in[N]} w_j k(x_j, .)\rangle_{\mathcal{F}} \right| \leq \|f\|_{\mathcal{F}} \left\| \mu_g - \sum_{j\in[N]} w_j k(x_j, .) \right\|_{\mathcal{F}}, \tag{4.106}$$

where

$$\mu_g = \int_{\mathcal{X}} k(x, .)g(x)\mathrm{d}\omega(x) \tag{4.107}$$

is the so-called *mean element* (Dick and Pillichshammer, 2014; Muandet et al., 2017) of the measure $g\mathrm{d}\omega$ or the *embedding* of g. (4.106) entails that the approximation error of the embedding μ_g by the kernel translates $\sum_{n\in[N]} w_n k(x_n, .)$ in the RKHS norm gives an upper bound on the integration error of the quadrature $(\boldsymbol{x}, \boldsymbol{w})$, independently of the integrand f. It actually is the worst case integration error in the unit ball of $\mathcal{F}$,

$$\sup_{\substack{f\in\mathcal{F} \\ \|f\|_{\mathcal{F}}\leq 1}} \left| \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x) - \sum_{j\in[N]} w_j f(x_j) \right| = \|\mu_g - \sum_{n\in[N]} w_n k(x_n, .)\|_{\mathcal{F}}. \tag{4.108}$$

The characterization (4.108), of the worst case integration error, motivates the use of a kernel in the evaluation of a quadrature. This idea can be tracked to Hickernell (Hickernell, 1996; Hickernell, 1998) who introduced reproducing kernels to the QMC community. Indeed, Hickernell, 1996 derived a tractable formula for the worst-case integration error of a quasi-Monte Carlo rule ($\mathcal{X} = [0,1]^d$, g is the constant function that take the value 1 and $w_n = 1/N$)

$$\sup_{\substack{f\in\mathcal{F} \\ \|f\|_{\mathcal{F}}\leq 1}} \left| \int_{[0,1]^d} f(x)\mathrm{d}x - \frac{1}{N} \sum_{n\in[N]} f(x_n) \right|^2. \tag{4.109}$$

In particular, he proved that (4.109) is equal to

$$\int_{[0,1]^d} \int_{[0,1]^d} k(x, y)\mathrm{d}x\mathrm{d}y - \frac{2}{N} \sum_{n\in[N]} \int_{[0,1]^d} k(x_n, y)\mathrm{d}y + \frac{1}{N^2} \sum_{n,n'=1}^{N} k(x_n, x_{n'}). \tag{4.110}$$

The generalization of (4.110) can be easily derived using properties of reproducing kernels:

$$\|\mu_g - \sum_{n\in[N]} w_n k(x_n, .)\|_{\mathcal{F}}^2 = \|\mu_g\|_{\mathcal{F}}^2 - 2\langle \mu_g, \sum_{n\in[N]} w_n k(x_n, .)\rangle_{\mathcal{F}} + \|\sum_{n\in[N]} w_n k(x_n, .)\|_{\mathcal{F}}^2 \tag{4.111}$$

$$= \|\mu_g\|_{\mathcal{F}}^2 - 2\sum_{n\in[N]} w_n \mu_g(x_n) + \sum_{n,n'=1}^{N} w_n k(x_n, x_{n'}) w_{n'} \tag{4.112}$$

$$= \|\mu_g\|_{\mathcal{F}}^2 - 2\boldsymbol{w}^\intercal \mu_g(\boldsymbol{x}) + \boldsymbol{w}^\intercal \boldsymbol{K}(\boldsymbol{x})\boldsymbol{w}, \tag{4.113}$$

where $\boldsymbol{K}(\boldsymbol{x}) = (k(x_n, x_{n'}))_{(n,n')\in[N]\times[N]} \in \mathbb{R}^{N\times N}$ and $\mu_g(\boldsymbol{x}) = (\mu_g(x_n))_{n\in[N]} \in \mathbb{R}^N$.

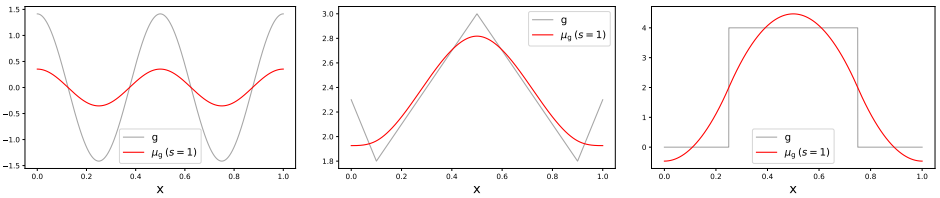


Figure 4.8 – μ_g in the periodic Sobolev space for three different g.

We can check that (4.110) is a consequence of (4.113), by taking $\mathcal{X} = [0,1]^d$, $w_n = 1/N$, $g = 1$ and $\mathrm{d}\omega$ is the uniform measure on $\mathcal{X}$ and observing that

$$\begin{aligned} \|\mu_g\|_{\mathcal{F}}^2 &= \langle \boldsymbol{\Sigma} g, \boldsymbol{\Sigma} g \rangle_{\mathcal{F}} && (4.114)\\ &= \langle g, \boldsymbol{\Sigma} g \rangle_{\mathrm{d}\omega} && (4.115)\\ &= \int_{\mathcal{X}} g(x) \int_{\mathcal{X}} k(x,y) g(y) \mathrm{d}\omega(y) \mathrm{d}\omega(x) && (4.116)\\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} k(x,y) g(x) g(y) \mathrm{d}\omega(y) \mathrm{d}\omega(x), && (4.117) \end{aligned}$$

and that

$$\boldsymbol{w}^\intercal \mu_g(\boldsymbol{x}) = \sum_{n\in[N]} w_n \mu_g(x_n) = \sum_{n\in[N]} w_n \int_{\mathcal{X}} g(y) k(x_n, y) \mathrm{d}\omega(y). \tag{4.118}$$

In other words, quasi-Monte Carlo rules fit within the kernel framework of quadratures.

4.2.5 *Optimal kernel quadrature*

As we have seen in Section 4.2.4, the worst case integration error for a given quadrature $(\boldsymbol{x}, \boldsymbol{w})$ writes:

$$\left\| \mu_g - \sum_{n\in[N]} w_n k(x_n, .) \right\|_{\mathcal{F}}. \tag{4.119}$$

Under the assumption that the configuration $\boldsymbol{x} \subset \mathcal{X}$ is uni-solvent with respect to the kernel k, i.e. the matrix $\boldsymbol{K}(\boldsymbol{x}) = (k(x_n, x_{n'}))_{n,n'\in[N]}$ is non-singular, there exists a unique vector $\boldsymbol{w} \in \mathbb{R}^N$ that minimizes (4.119). Indeed, the square of (4.119) is quadratic on $\boldsymbol{w}$ so that the optimization problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^N} \left\| \mu_g - \sum_{n\in[N]} w_n k(x_n, .) \right\|_{\mathcal{F}}^2 \tag{4.120}$$

have a unique solution $\hat{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x})^{-1} \mu_g(\boldsymbol{x})$. The quadrature $(\boldsymbol{x}, \hat{\boldsymbol{w}}(\boldsymbol{x}))$ is called the *optimal kernel quadrature* in $\boldsymbol{x}$.

Now, under the assumption that $\boldsymbol{K}(\boldsymbol{x})$ is non-singular, the dimension of the subspace $\mathcal{T}(\boldsymbol{x}) = \mathrm{Span}(k(x_n, .))_{n\in[N]}$ is equal to N. Define

$$\boldsymbol{\Phi} : (w_j)_{j\in[N]} \mapsto \sum_{j\in[N]} w_j k(x_j, .) \tag{4.121}$$

the reconstruction operator[5]. The optimal approximation of μ_g by an element of $\mathcal{T}(\boldsymbol{x})$, with respect to the RKHS norm, is $\boldsymbol{\Phi}\hat{\boldsymbol{w}}$, and the optimal approximation error writes

$$\|\mu_g - \boldsymbol{\Phi}\hat{\boldsymbol{w}}\|_{\mathcal{F}}^2 = \|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2, \tag{4.122}$$

where $\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^*\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^*$ is the orthogonal projection onto $\mathcal{T}(\boldsymbol{x})$ with $\boldsymbol{\Phi}^*$ the dual[6] of $\boldsymbol{\Phi}$. $\boldsymbol{\Phi}\hat{\boldsymbol{w}}$ is the orthogonal projection of μ_g on the subspace $\mathcal{T}(\boldsymbol{x})$. Moreover, we can prove easily that $\boldsymbol{\Phi}\hat{\boldsymbol{w}}$ interpolates μ_g on the nodes $x_1, \dots, x_N$:

$$\forall n \in [N],\ \boldsymbol{\Phi}\hat{\boldsymbol{w}}(x_n) = \mu_g(x_n). \tag{4.123}$$

We define the *interpolation error* of μ_g

$$\|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}. \tag{4.124}$$

As we have seen, under the assumption that the matrix $\boldsymbol{K}(\boldsymbol{x})$ is non-singular, the computation of the weights of the optimal kernel quadrature is a relatively simple task. Yet the task of designing a good configuration $\boldsymbol{x}$ is not trivial as we shall see in the next section where we review the main results on the literature about design problems for optimal kernel quadrature.

The Bayesian interpretation of optimal kernel quadrature

The optimal kernel quadrature $(\boldsymbol{x}, \hat{\boldsymbol{w}})$ has an interpretation within *Bayesian Quadrature* introduced by (Larkin, 1972). This method aims to quantifies the uncertainty of numerical integration; see (Diaconis, 1988) for a historical review of this method.

This method goes as follows. Consider a fixed set of nodes $\boldsymbol{x} = \{x_1, \dots, x_N\}$ and put a Gaussian process prior on the integrand f with mean 0 and kernel k. The posterior distribution over f after conditioning on $(f(x_1), \dots, f(x_N))$ is a Gaussian process of mean $m_{\boldsymbol{x}}$ and kernel $k_{\boldsymbol{x}}$, where

$$m_{\boldsymbol{x}}(x) = k(x, \boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} f(\boldsymbol{x}), \tag{4.125}$$

and

$$k_{\boldsymbol{x}}(x, x') = k(x, x') - k(x, \boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} k(x', \boldsymbol{x}), \tag{4.126}$$

with $k(x, \boldsymbol{x}) = (k(x, x_n))_{n\in[N]} \in \mathbb{R}^N$. Therefore, the random variable

$$\int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x), \tag{4.127}$$

is Gaussian with

$$\mathbb{E}_{f\sim \mathrm{GP}(m_{\boldsymbol{x}}, k_{\boldsymbol{x}})} \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega = \mu_g(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} f(\boldsymbol{x}) = \sum_{n\in[N]} \hat{w}_n f(x_n), \tag{4.128}$$

and

$$\mathbb{V}_{f\sim \mathrm{GP}(m_{\boldsymbol{x}}, k_{\boldsymbol{x}})} \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega = \|\mu_g - \sum_{n\in[N]} \hat{w}_n k(x_n, .)\|_{\mathcal{F}}^2. \tag{4.129}$$

In other words, the optimal kernel quadrature $(\boldsymbol{x}, \hat{\boldsymbol{w}})$ controls both the mean and the variance of $\int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x)$. This link was leveraged in (Briol et al., 2019) to establish rates of posterior contraction of Bayesian quadrature based on rates of convergence of optimal kernel quadrature.

[5] The reconstruction operator $\boldsymbol{\Phi}$ depends on the nodes x_j, although our notation doesn't reflect it for simplicity.

[6] For $\mu \in \mathcal{F}$,$\boldsymbol{\Phi}^*\mu = (\mu(x_j))_{j\in[N]}$. $\boldsymbol{\Phi}^*\boldsymbol{\Phi}$ is an operator from $\mathbb{R}^N$ to $\mathbb{R}^N$ that can be identified with $\boldsymbol{K}(\boldsymbol{x})$.

4.2.6 *Designs for optimal kernel quadrature*

We review in this section the literature on the design of configurations for optimal kernel quadrature.

Deterministic configurations

Bojanov, 1981 proved that, for $g = 1$, the interpolation of μ_g using the uniform grid over $\mathcal{X} = [0,1]$ has an error in $\mathcal{O}(N^{-2s})$ if $\mathcal{F}$ is the periodic Sobolev space of order s, and that any set of nodes leads to that rate at least. A similar rate was proved for g not constant (Novak et al., 2015) even though it is only asymptotically optimal in that case.

As for QMC sequences, the rates are naturally inherited if the uniform weights, of a QMC rule, are replaced by the respective optimal weights $\hat{\boldsymbol{w}}$, as observed by Briol et al., 2019. In particular, Briol et al., 2019 emphasize that the bound for higher-order digital nets attains the almost optimal rate in this RKHS. However, this inheritance argument does not explain the fast $\mathcal{O}(\log(N)^{2sd}N^{-2s})$ rates observed empirically for optimal kernel quadrature based on Halton sequences; see (Oettershagen, 2017) and Section 4.4.2.

Scattered data approximation (Wendland, 2004) is another field where quantitative error bounds for optimal kernel quadrature on $\mathcal{X} \subset \mathbb{R}^d$ are investigated; see (Schaback and Wendland, 2006) for a modern review. In a few words, these bounds typically depend on quantities such as the *fill-in distance* $\varphi(\boldsymbol{x}) = \sup_{y\in\mathcal{X}} \min_{i\in[N]} \|y - x_i\|_2$, so that the interpolation error converges to zero as $N \to \infty$ if $\varphi(\boldsymbol{x})$ goes to zero. Any node set can be considered, as long as $\varphi(\boldsymbol{x})$ is small. Using these techniques, Briol et al., 2019 proved optimal rates for optimal kernel quadrature in RKHS based on multidimensional Sobolev spaces using some sequences. Note that the application of these techniques is restricted to compact domains: the fill-in distance is infinite if $\mathcal{X}$ is not compact.

Beside the hypercube, optimal kernel quadrature has been considered on the hypersphere equipped with the uniform measure (Ehler et al., 2019), or on $\mathbb{R}^d$ equipped with the Gaussian measure (Karvonen and Särkkä, 2019). In these works, the design construction is adhoc for the space $\mathcal{X}$ and g is usually assumed to be constant.

Sequential algorithms

Sequential Bayesian quadrature (Huszár and Duvenaud, 2012) offers another approach to designing a configuration of nodes $\boldsymbol{x} = \{x_1, \dots, x_N\}$ for optimal kernel quadrature: based on a configuration $\boldsymbol{x}_t$ look for $x^* \in \mathcal{X}$ such that

$$\|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x}_t\cup\{x^*\})}\mu_g\|_{\mathcal{F}} = \min_{x\in\mathcal{X}} \|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x}_t\cup\{x\})}\mu_g\|_{\mathcal{F}}^2, \tag{4.130}$$

and put $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t \cup \{x^*\}$.

Implementing this algorithm corresponds to solving sequentially the non-convex problems (4.130), with many local minima. In practice, costly approximations must be employed to run the algorithm (Hinrichs and Oettershagen, 2016; Oettershagen, 2017). We will discuss these implementation issues for similar sequential algorithms in Chapter 5.

Ridge leverage score quadrature

The regularization of the problem (4.120) offers an alternative approach for the design of the configuration $\boldsymbol{x}$ as it was shown in (Bach, 2017). This work highlights the fundamental role played by the spectral decomposition of the operator $\boldsymbol{\Sigma}$ in designing and analyzing kernel quadrature rules. In fact, the author proposed to sample the nodes (x_j) i.i.d. from some proposal distribution q, and then take the vector of weights $\boldsymbol{w}$ to be the vector $\boldsymbol{w}_\lambda$ that solve the optimization problem

$$\min_{w\in\mathbb{R}^N} \left\| \mu_g - \sum_{j\in[N]} \frac{w_j}{q(x_j)^{1/2}} k(x_j,.) \right\|_{\mathcal{F}}^2 + \lambda N \|w\|_2^2, \tag{4.131}$$

for some regularization parameter $\lambda > 0$.

Within this analysis, any proposal distribution q, with $q > 0$, could be used to sample N i.i.d. nodes that constitutes the design $\boldsymbol{x}$: this follows from the following result.

Proposition 4.2 (Proposition 1 in Bach, 2017)**.** *Let $\delta \in]0,1]$, and denote*

$$d_{\max}(q,\lambda) = \sup_{x\in\mathcal{X}} \frac{1}{q(x)} \langle k(x,.), (\boldsymbol{\Sigma} + \lambda \mathbb{I}_{\mathbb{L}_2(\mathrm{d}\omega)})^{-1} k(x,.)\rangle_{\mathbb{L}_2(\mathrm{d}\omega)}. \tag{4.132}$$

Assume that $N \geq 5 d_{\max}(q,\lambda)\log(16 d_{\max}(q,\lambda)/\delta)$, then

$$\mathbb{P}\left(\sup_{\|g\|_{\mathrm{d}\omega}\leq 1} \inf_{\|\boldsymbol{w}\|^2 \leq \frac{4}{N}} \left\| \mu_g - \sum_{j\in[N]} \frac{w_j}{q_\lambda(x_j)^{1/2}} k(x_j,.) \right\|_{\mathcal{F}}^2 \leq 4\lambda \right) \geq 1-\delta. \tag{4.133}$$

In other words, Proposition 4.2 gives a uniform control on the approximation error μ_g by the subspace spanned by the $k(x_j,.)$ for g belonging to the unit ball of $\mathbb{L}_2(\mathrm{d}\omega)$, where the (x_j) are sampled i.i.d. from q. The required number of nodes is equal to $\mathcal{O}(d_{\max}(q,\lambda)\log d_{\max}(q,\lambda))$ for a given approximation error λ. This bound is relevant if an upper bound of the *maximal ridge leverage score* $d_{\max}(q,\lambda)$ is known. This is possible in some cases; yet in general this quantity is hard to control. Nevertheless, the author showed that one may use a specific choice of q, namely the ridge leverage score distribution q_λ^* defined as

$$q_\lambda^*(x) \propto \langle k(x,.), \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\Sigma} + \lambda \mathbb{I}_{\mathbb{L}_2(\mathrm{d}\omega)})^{-1}\boldsymbol{\Sigma}^{-1/2} k(x,.)\rangle_{\mathbb{L}_2(\mathrm{d}\omega)} = \sum_{m\in\mathbb{N}^*} \frac{\sigma_m}{\sigma_m+\lambda} e_m(x)^2. \tag{4.134}$$

Indeed, for this specific choice of q

$$\begin{aligned} d_{\max}(q_\lambda^*,\lambda) &= \int_{\mathcal{X}} \langle k(x,.), \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\Sigma} + \lambda \mathbb{I}_{\mathbb{L}_2(\mathrm{d}\omega)})^{-1}\boldsymbol{\Sigma}^{-1/2} k(x,.)\rangle_{\mathbb{L}_2(\mathrm{d}\omega)} \mathrm{d}\omega(x) \\ &= \sum_{m\in\mathbb{N}^*} \frac{\sigma_m}{\sigma_m+\lambda} \\ &= \operatorname{Tr} \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1}, \end{aligned} \tag{4.135}$$

and the quantity

$$d_\lambda = \operatorname{Tr} \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1}, \tag{4.136}$$

is called the *effective degree of freedom*. This quantity is more amenable to analysis as it depends only on λ and on the eigenvalues of $\boldsymbol{\Sigma}$. An instantiation of Proposition 4.2 using the ridge leverage score distribution q_λ^* yields the following result.

Proposition 4.3 (Proposition 2 in Bach, 2017)**.** *Let $\delta \in]0,1]$. Assume that $N \geq 5d_\lambda \log(16 d_\lambda/\delta)$, then*

$$\mathbb{P}\left(\sup_{\|g\|_{\mathrm{d}\omega} \leq 1} \inf_{\|\boldsymbol{w}\|^2 \leq \frac{4}{N}} \left\| \mu_g - \sum_{j\in[N]} \frac{w_j}{q_\lambda(x_j)^{1/2}} k(x_j, .) \right\|_{\mathcal{F}}^2 \leq 4\lambda \right) \geq 1 - \delta. \tag{4.137}$$

Now, for fixed λ, the approximation error in Proposition 4.3 does not go to zero when N increases. One theoretical workaround is to make $\lambda = \lambda(N)$ decrease with N. However, the coupling of N and λ through d_λ makes it very intricate to derive a convergence rate from Proposition 4.3. Moreover, the optimal density q_λ^* is in general only available as the limit (4.134), which makes sampling and evaluation difficult.

4.3 PROJECTION DPP FOR OPTIMAL KERNEL QUADRATURE

In this section, we give our first contribution of this chapter (Belhadji et al., 2019a). We propose to study optimal kernel quadrature based on nodes that follows the distribution of a projection DPP. We follow in the footsteps of Bach, 2017, as reviewed in Section 4.2.6, but with two main differences:

- we study the un-regularized optimization problem (4.120),
- we use a projection DPP rather than independent sampling to obtain the nodes.

In a nutshell, we consider nodes $(x_j)_{j\in[N]}$ that are drawn from the projection DPP with reference measure $\mathrm{d}\omega$ and marginal kernel

$$\mathfrak{K}(x,y) = \sum_{n\in[N]} e_n(x) e_n(y), \tag{4.138}$$

where we recall that $(e_n)_{n\in\mathbb{N}^*}$ is an o.n.b of $\mathbb{L}_2(\mathrm{d}\omega)$ that diagonalizes $\boldsymbol{\Sigma}$; see Section 4.2.2. As seen in Chapter 2, this boils down to sample a random configuration $\boldsymbol{x}$ in $\mathcal{X}^N$ from the probability distribution

$$\frac{1}{N!} \mathrm{Det}(\mathfrak{K}(x_i, x_j)_{i,j\in[N]}) \prod_{i\in[N]} \mathrm{d}\omega(x_i). \tag{4.139}$$

As we shall see, among many possible choices of kernels, the one defined in (4.138) stands out for several reasons: i) it is intuitive and comes with a theoretical analysis, ii) it defines a random configuration such that the un-regularized optimization problem (4.120) admits a unique solution with probability one, iii) it is amenable to exact sampling in many settings.

We explain more in detail these three reasons in the following and we start by the first one. This projection DPP defines a uni-solvent configuration $\boldsymbol{x}$ with respect to the kernel k: the subspace $\mathcal{T}(\boldsymbol{x}) = \mathrm{Span}(k(x_n, .))_{n\in[N]}$ is N-dimensional a.s., or equivalently, the matrix $\boldsymbol{K}(\boldsymbol{x})$ is invertible a.s. This is a consequence of the following result.

Proposition 4.4. *Assume that the matrix $\boldsymbol{E}(\boldsymbol{x}) = (e_i(x_j))_{i,j\in[N]}$ is invertible, then $\boldsymbol{K}(\boldsymbol{x})$ is invertible.*

The proof of Proposition 4.4 is given in Section 4.6.3. Since the pdf (4.139) of the projection DPP with kernel (4.138) is proportional to $\mathrm{Det}^2 \boldsymbol{E}(\boldsymbol{x})$, the following corollary immediately follows.

Corollary 4.1. *Let $\boldsymbol{x} = \{x_1, \dots, x_N\}$ be a projection DPP with reference measure $\mathrm{d}\omega$ and kernel $\mathfrak{K}$ defined in (4.138). Then $\boldsymbol{K}(\boldsymbol{x})$ is a.s. invertible, so that (4.120) has unique solution $\hat{\boldsymbol{w}} = \boldsymbol{K}(\boldsymbol{x})^{-1}\mu_g(x_j)_{j\in[N]}$ a.s.*

As we have seen in Section 4.2.5, under the assumption that $\boldsymbol{K}(\boldsymbol{x})$ is non singular the optimal approximation of μ_g by an element of $\mathcal{T}(\boldsymbol{x})$, with respect to the RKHS norm, is $\boldsymbol{\Phi}\hat{\boldsymbol{w}}$. The optimal approximation error then writes

$$\|\mu_g - \boldsymbol{\Phi}\hat{\boldsymbol{w}}\|_{\mathcal{F}}^2 = \|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2, \tag{4.140}$$

where $\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^*\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^*$ is the orthogonal projection onto $\mathcal{T}(\boldsymbol{x})$ with $\boldsymbol{\Phi}^*$ the dual[7] of $\boldsymbol{\Phi}$: $\boldsymbol{\Phi}\hat{\boldsymbol{w}}$ is the orthogonal projection of μ_g on the subspace $\mathcal{T}(\boldsymbol{x})$.

Now, the second reason to introduce the projection DPP associated to the the kernel (4.138) is the possibility of the theoretical analysis of the expected value of

$$\|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2. \tag{4.141}$$

We explain this point more in detail after presenting our theoretical guarantees in the following section. The details of the proofs are given in Section 4.6.

4.3.1 *The main results*

In the section, we present the theoretical analysis of $\|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2$ under the projection DPP $(\mathfrak{K}, \mathrm{d}\omega)$, where $\mathfrak{K}$ is defined in (4.138). We start with a result that we obtained in (Belhadji, Bardenet, and Chainais, 2019a), which quantifies the expected interpolation error in terms of the spectrum of the integration operator $\boldsymbol{\Sigma}$.

Theorem 4.7. *Let $\boldsymbol{x} = \{x_1, \dots, x_N\}$ be a random set that follows the distribution of the projection DPP $(\mathfrak{K}, \mathrm{d}\omega)$. Let $g \in \mathbb{L}_2(\mathrm{d}\omega)$ such that $\|g\|_{\mathrm{d}\omega} \leq 1$. Then,*

$$\mathbb{E}_{\mathrm{DPP}}\|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 \leq 2\sigma_{N+1} + 2\|g\|_{\mathrm{d}\omega,1}^2\left(Nr_N + \sum_{\ell=2}^{N}\frac{\sigma_1}{\ell!^2}\left(\frac{Nr_N}{\sigma_1}\right)^{\ell}\right), \tag{4.142}$$

where $\|g\|_{\mathrm{d}\omega,1} = \sum_{n\in[N]}|\langle e_n, g\rangle_{\mathrm{d}\omega}|$ and $r_N = \sum_{m\geq N+1}\sigma_m$. Moreover,

$$\mathbb{E}_{\mathrm{DPP}}\sup_{\|g\|_{\mathrm{d}\omega}\leq 1}\|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 \leq 2\sigma_{N+1} + 2N\left(Nr_N + \sum_{\ell=2}^{N}\frac{\sigma_1}{\ell!^2}\left(\frac{Nr_N}{\sigma_1}\right)^{\ell}\right). \tag{4.143}$$

Since the publication of (Belhadji et al., 2019a), we obtained a refinement of Theorem 4.7 that we give in the following.

Theorem 4.8. *Let $\boldsymbol{x} = \{x_1, \dots, x_N\}$ be a random set that follows the distribution of the projection DPP $(\mathfrak{K}, \mathrm{d}\omega)$. Let $g \in \mathbb{L}_2(\mathrm{d}\omega)$ such that $\|g\|_{\mathrm{d}\omega} \leq 1$. Then,*

$$\mathbb{E}_{\mathrm{DPP}}\|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 \leq 2\sigma_{N+1} + 2\|g\|_{\mathrm{d}\omega,1}^2 Nr_N, \tag{4.144}$$

where $\|g\|_{\mathrm{d}\omega,1} = \sum_{n\in[N]}|\langle e_n, g\rangle_{\mathrm{d}\omega}|$ and $r_N = \sum_{m\geq N+1}\sigma_m$. Moreover,

$$\mathbb{E}_{\mathrm{DPP}}\sup_{\|g\|_{\mathrm{d}\omega}\leq 1}\|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 \leq 2\sigma_{N+1} + 2N^2 r_N. \tag{4.145}$$

[7] For $\mu \in \mathcal{F}$, $\boldsymbol{\Phi}^*\mu = (\mu(x_j))_{j\in[N]}$. $\boldsymbol{\Phi}^*\boldsymbol{\Phi}$ is an operator from $\mathbb{R}^N$ to $\mathbb{R}^N$ that can be identified with $\boldsymbol{K}(\boldsymbol{x})$.

Let us comment on the upper bounds of Theorem 4.8. The first term in the upper bounds (4.144) and (4.145) scales as σ_{N+1}. As we shall in Section 5.2.3 of Chapter 5, σ_{N+1} is the lower bound of interpolation in a sense that will be defined. The second terms in the upper bounds (4.144) and (4.145) stem from our proof technique. Indeed, the constant $\|g\|_{\mathrm{d}\omega,1}$ in (4.144) is the ℓ_1 norm of the coefficients of the projection of g onto $\mathrm{Span}(e_n)_{n\in[N]}$ in $\mathbb{L}_2(\mathrm{d}\omega)$. For example, for $g = e_n$, $\|g\|_{\mathrm{d}\omega,1} = 1$ if $n \in [N]$ and $\|g\|_{\mathrm{d}\omega,1} = 0$ if $n \geq N+1$. In the worst case, $\|g\|_{\mathrm{d}\omega,1} \leq \sqrt{N}\|g\|_{\mathrm{d}\omega} \leq \sqrt{N}$. Thus, we can obtain a uniform bound for $\|g\|_{\mathrm{d}\omega} \leq 1$ but with a supplementary factor N. Now, if $\sum_{m\in\mathbb{N}^*} |\langle g, e_m\rangle_{\mathrm{d}\omega}| < +\infty$ for a particular $g \in \mathbb{L}_2(\mathrm{d}\omega)$, the r.h.s. of (4.144) is $\mathcal{O}(Nr_N)$ because $\sigma_{N+1} \leq Nr_N$. Similarly, the upper bound in (4.145) is dominated by $\mathcal{O}(N^2 r_N)$.

Both these bounds converge to 0 if the convergence of (σ_n) to 0 is fast enough. For example, if $\sigma_m = m^{-2s}$, $\mathcal{O}(r_N) = N^{1-2s}$ thus $\mathcal{O}(Nr_N) = N^{2-2s}$ and $\mathcal{O}(N^2 r_N) = N^{3-2s}$; and convergence is guaranteed if $s > 1$ or if $s > 3/2$ respectively. Another example is given by $\sigma_m = \beta\alpha^m$ with $0 < \alpha < 1$ and $\beta > 0$ then $Nr_N = N\frac{\beta}{1-\alpha}\alpha^{N+1} = o(1)$ and $N^2 r_N = N^2\frac{\beta}{1-\alpha}\alpha^{N+1} = o(1)$.

We have assumed that $\mathcal{F}$ is dense in $\mathbb{L}_2(\mathrm{d}\omega)$ but Theorem 4.8 is valid also when $\mathcal{F}$ is finite-dimensional. In this case, denote $N_0 = \dim \mathcal{F}$. Then, for $n > N_0$, $\sigma_n = 0$ and $r_{N_0} = 0$, so that (4.144) implies

$$\|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}} = 0 \text{ a.s.} \tag{4.146}$$

when $N = N_0$.

In comparison with Bach, 2017, we emphasize that the dependence of our bound on the eigenvalues of the kernel k, via r_N, is explicit. This is in contrast with Proposition 4.3 that depends on the eigenvalues of $\mathbf{\Sigma}$ through the degree of freedom d_λ so that the necessary number of samples N diverges when $\lambda \to 0$. On the contrary, our quadrature requires a finite number of points for $\lambda = 0$.

Another advantage of DPPs is that they can be sampled exactly. Because of the orthonormality of (ψ_n), one can use the HKPV algorithm; see Section 2.2.5. The main limitation is the availability of good proposals for the successive rejection sampling routines.

Theorem 4.8 gives a slightly sharper bound than Theorem 4.7, since if $Nr_N = o(1)$, then the right-hand side of (4.142) is $Nr_N + o(Nr_N)$. The proof of Theorem 4.8 follows the same steps as the proof of Theorem 4.7. We give the proofs of the two results simultaneously in the following section and we highlight the difference between the two in the end of the section.

4.3.2 *Bounding the interpolation error under the projection DPP*

In this section, we give the skeleton of the proofs of Theorem 4.7 and Theorem 4.8. The details of the proofs are deferred to Section 4.6.

The starting point of the analysis is the observation that the interpolation error is the RKHS norm of the residual $\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g$. Since the subspace $\mathcal{T}(\boldsymbol{x})$ can be arbitrary, as $\boldsymbol{x}$ spans the set of uni-solvent of $\mathcal{X}$ of cardinality N, it is hard to reckon how large the corresponding residual can be. For this reason, we consider the principal subspace

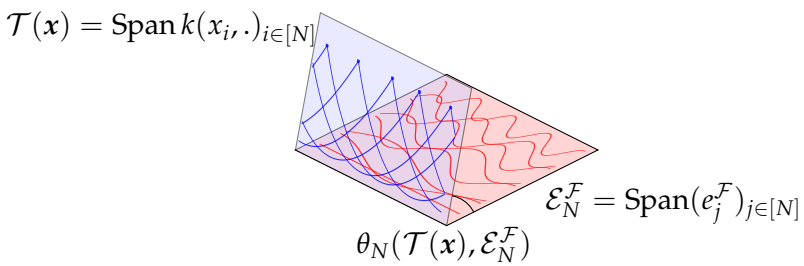


Figure 4.9 – Illustration of the largest principal angle between the subspaces $\mathcal{T}(\boldsymbol{x})$ and $\mathcal{E}_N^{\mathcal{F}}$ in the case of the periodic Sobolev space of order 1.

$\mathcal{E}_N^{\mathcal{F}} = \operatorname{Span}(e_n^{\mathcal{F}})_{n\in[N]}$ and the corresponding *filtering error* $\|\mu_g - \mathbf{\Pi}_{\mathcal{E}_N^{\mathcal{F}}}\mu_g\|_{\mathcal{F}}$. The filtering error is easier to quantify because we can prove easily that

$$\|\mu_g - \mathbf{\Pi}_{\mathcal{E}_N^{\mathcal{F}}}\mu_g\|_{\mathcal{F}}^2 = \sum_{m\geq N+1} \langle \mu_g, e_m^{\mathcal{F}}\rangle_{\mathcal{F}}^2 = \sum_{m\geq N+1} \sigma_m \langle g, e_m\rangle_{\mathrm{d}\omega}^2. \tag{4.147}$$

In particular, when $\|g\|_{\mathrm{d}\omega} \leq 1$, $\|\mu_g - \mathbf{\Pi}_{\mathcal{E}_N^{\mathcal{F}}}\mu_g\|_{\mathcal{F}}^2 \leq \sigma_{N+1}$, with equality when $g = e_{n+1}$. Moreover, if the two subspaces $\mathcal{E}_N^{\mathcal{F}}$ and $\mathcal{T}(\boldsymbol{x})$ are close to each other in some sense, we expect to have $\|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2$ close to σ_{N+1} when $\|g\|_{\mathrm{d}\omega} \leq 1$. As a first step, we have the following lemma that gives an upper bound of $\|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2$ using the $\|\mu_{e_n^{\mathcal{F}}} - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_{e_n^{\mathcal{F}}}\|_{\mathcal{F}}^2$.

Lemma 4.1. *Assume that* $\|g\|_{\mathrm{d}\omega} \leq 1$ *then*

$$\|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 \leq 2\left(\sigma_{N+1} + \|g\|_{\mathrm{d}\omega,1}^2 \max_{n\in[N]} \|\mu_{e_n^{\mathcal{F}}} - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_{e_n^{\mathcal{F}}}\|_{\mathcal{F}}^2\right). \tag{4.148}$$

The term $2\sigma_{N+1}$ converges to 0 when N goes to $+\infty$ and we shall see later that it scales as a specific lower bound. Now it remains to upper bound the $\|\mu_{e_n^{\mathcal{F}}} - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_{e_n^{\mathcal{F}}}\|_{\mathcal{F}}^2$ for $n \in [N]$ that also have the following expression

$$\|\mu_{e_n^{\mathcal{F}}} - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_{e_n^{\mathcal{F}}}\|_{\mathcal{F}}^2 = \sigma_n \|e_n^{\mathcal{F}} - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}e_n^{\mathcal{F}}\|_{\mathcal{F}}^2. \tag{4.149}$$

To upper bound the right-hand side of (4.149), we note that $\sigma_n\|e_n^{\mathcal{F}} - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}e_n^{\mathcal{F}}\|_{\mathcal{F}}^2$ is the product of two terms: σ_n is a decreasing function of n while $\|e_n^{\mathcal{F}} - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}e_n^{\mathcal{F}}\|_{\mathcal{F}}^2$ is the interpolation error of the eigenfunction $e_n^{\mathcal{F}}$, measured in the $\|.\|_{\mathcal{F}}$ norm. We can bound the latter interpolation error uniformly in $n \in [N]$ using the largest principal angle between $\mathcal{T}(\boldsymbol{x})$ and $\mathcal{E}_N^{\mathcal{F}} = \operatorname{Span}(e_n^{\mathcal{F}})_{n\in[N]}$; see Section 4.6.2 for details on principal angles between subspaces in Hilbert spaces. In particular, remember that the maximal principal angle is defined through its cosine

$$\cos^2 \theta_N(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}}) = \inf_{\substack{\boldsymbol{u}\in\mathcal{T}(\boldsymbol{x}), \boldsymbol{v}\in\mathcal{E}_N^{\mathcal{F}} \\ \|u\|_{\mathcal{F}}=1, \|v\|_{\mathcal{F}}=1}} \langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\mathcal{F}}. \tag{4.150}$$

We can then define successively the $N-1$ principal angles $\theta_n(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}}) \in \left[0, \frac{\pi}{2}\right]$ for $n \in [N-1]$ between the subspaces $\mathcal{E}_N^{\mathcal{F}}$ and $\mathcal{T}(\boldsymbol{x})$. These angles quantify the relative position of these two subspaces; see Figure 4.9 for a visual illustration. Now, we have the following lemma.

Lemma 4.2. *Let* $\boldsymbol{x} = \{x_1, \dots, x_N\} \subset \mathcal{X}$ *such that* $\operatorname{Det} \boldsymbol{E}(\boldsymbol{x}) \neq 0$. *Then*

$$\max_{n \in [N]} \| e_n^{\mathcal{F}} - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}} \|_{\mathcal{F}}^2 \leq \frac{1}{\cos^2 \theta_N(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}})} - 1. \tag{4.151}$$

In particular

$$\max_{n \in [N]} \| e_n^{\mathcal{F}} - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}} \|_{\mathcal{F}}^2 \leq \prod_{n \in [N]} \frac{1}{\cos^2 \theta_n(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}})} - 1, \tag{4.152}$$

and

$$\max_{n \in [N]} \| e_n^{\mathcal{F}} - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}} \|_{\mathcal{F}}^2 \leq \sum_{n \in [N]} \frac{1}{\cos^2 \theta_n(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}})} - N. \tag{4.153}$$

To sum up, we have so far bounded the approximation error by the geometric quantities in the right-hand side of (4.152) and (4.153). The proof of Theorem 4.7 uses the symmetrization (4.152), while the proof of Theorem 4.8 uses the symmetrization (4.153).

As we have seen in Chapter 3, projection DPPs shine in taking expectations of such symmetric geometric quantities in finite dimensional vector spaces. We will show in the following that this is also true in RKHSs.

Proposition 4.5. *Let* $\boldsymbol{x} = \{x_1, \dots, x_N\} \subset \mathcal{X}$ *such that* $\operatorname{Det}^2 \boldsymbol{E}(\boldsymbol{x}) \neq 0$. *Then,*

$$\prod_{\ell \in [N]} \frac{1}{\cos^2 \theta_\ell \left(\mathcal{E}_N^{\mathcal{F}}, \mathcal{T}(\boldsymbol{x}) \right)} = \frac{\operatorname{Det} \boldsymbol{K}(\boldsymbol{x})}{\operatorname{Det}^2 \boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})}, \tag{4.154}$$

and

$$\sum_{\ell \in [N]} \frac{1}{\cos^2 \theta_\ell \left(\mathcal{E}_N^{\mathcal{F}}, \mathcal{T}(\boldsymbol{x}) \right)} = \operatorname{Tr} \left(\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal^{-1}} \boldsymbol{K}(\boldsymbol{x}) \boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{-1} \right). \tag{4.155}$$

Proposition 4.5 allows to obtain tractable formulas, in terms of the eigenvalues of the kernel k, of the expectation of the right-hand side of (4.152) and (4.153). This is achieved in the following two results.

The first one concerns the multiplicative symmetrization.

Proposition 4.6. *Let* $\boldsymbol{x} = \{x_1, \dots, x_N\}$ *be a random set that follows the projection DPP* $(\mathrm{d}\omega, \mathfrak{K})$. *Then,*

$$\mathbb{E}_{\mathrm{DPP}} \prod_{n \in [N]} \frac{1}{\cos^2 \theta_n \left(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}} \right)} = \sum_{\substack{T \subset \mathbb{N}^* \\ |T| = N}} \frac{\prod_{t \in T} \sigma_t}{\prod_{n \in [N]} \sigma_n}. \tag{4.156}$$

The second result concerns the additive symmetrization.

Proposition 4.7. *Let* $\boldsymbol{x} = \{x_1, \dots, x_N\}$ *be a random set that follows the projection DPP* $(\mathrm{d}\omega, \mathfrak{K})$. *Then,*

$$\mathbb{E}_{\mathrm{DPP}} \sum_{\ell \in [N]} \frac{1}{\cos^2 \theta_\ell \left(\mathcal{E}_N^{\mathcal{F}}, \mathcal{T}(\boldsymbol{x}) \right)} - N = \sum_{v \in [N]} \frac{1}{\sigma_v} \sum_{w \in \mathbb{N}^* \setminus [N]} \sigma_w. \tag{4.157}$$

The bound of Proposition 4.6, once reported in Lemma 4.2 and Lemma 4.1, already yields Theorem 4.7 in the special case where the kernel k is "saturated": $\sigma_1 = \cdots = \sigma_N$. This seems a very restrictive condition, but Proposition 4.8 below shows that we can always reduce the analysis to that case. In fact, let the kernel $\tilde{k}$ be defined by

$$\tilde{k}(x,y) = \sum_{n\in[N]} \sigma_1 e_n(x)e_n(y) + \sum_{n\geq N+1} \sigma_n e_n(x)e_n(y) = \sum_{n\in\mathbb{N}^*} \tilde{\sigma}_n e_n(x)e_n(y), \tag{4.158}$$

and let $\tilde{\mathcal{F}}$ be the corresponding RKHS. Then one has the following inequality.

Proposition 4.8. *Let* $\tilde{\mathcal{T}}(\boldsymbol{x}) = \operatorname{Span}\left(\tilde{k}(x_j,.)\right)_{j\in[N]}$ *and* $\boldsymbol{\Pi}_{\tilde{\mathcal{T}}(\boldsymbol{x})}$ *the orthogonal projection onto* $\tilde{\mathcal{T}}(\boldsymbol{x})$ *in* $(\tilde{\mathcal{F}}, \langle .,. \rangle_{\tilde{\mathcal{F}}})$. *Then,*

$$\forall n \in [N],\ \sigma_n \|e_n^{\mathcal{F}} - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 \leq \sigma_1 \|e_n^{\tilde{\mathcal{F}}} - \boldsymbol{\Pi}_{\tilde{\mathcal{T}}(\boldsymbol{x})} e_n^{\tilde{\mathcal{F}}}\|_{\tilde{\mathcal{F}}}^2. \tag{4.159}$$

Simply put, capping the first eigenvalues of k yields a new kernel $\tilde{k}$ that captures the interaction between the terms σ_n and $\|e_n^{\mathcal{F}} - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2$, so that we only have to deal with the term $\|e_n^{\tilde{\mathcal{F}}} - \boldsymbol{\Pi}_{\tilde{\mathcal{T}}(\boldsymbol{x})} e_n^{\tilde{\mathcal{F}}}\|_{\tilde{\mathcal{F}}}^2$. Combining Proposition 4.6 with Proposition 4.8 applied to the kernel $\tilde{k}$ yields Theorem 4.7.

The same steps are followed to obtain the Theorem 4.8 using Proposition 4.7. The additive symmetrization gives a neater upper bound compared to the multiplicative symmetrization. Nevertheless the proof of Proposition 4.7 is more technical.

4.4 NUMERICAL SIMULATIONS

This section is devoted to numerical simulations that illustrate the main results of Section 4.3.1.

4.4.1 *The periodic Sobolev space*

Let $\mathcal{S}_s$ the periodic Sobolev space of order $s \in \mathbb{N}^*$ defined in Section 4.2.3. We consider a set of three numerical experiments. In the first one, we consider the approximation of μ_g when $g \equiv 1$, that is $g \equiv e_1$ [8], using several quadratures; in the second one, we consider the approximation of μ_g when g is an arbitrary eigenfunction of $\boldsymbol{\Sigma}$ using the optimal kernel quadrature based on the projection DPP; in the third one, we consider the worst-case interpolation error of the μ_g when g spans the unit ball of $\mathbb{L}_2(\mathrm{d}\omega)$.

The reconstruction of the embedding of the first eigenfunction

We take $g \equiv e_1$ so that the embedding $\mu_g = e_1$. We compare the following algorithms: (i) the quadrature rule DPPKQ we propose in Theorem 4.7 (optimal kernel quadrature based on the projection DPP $(\mathfrak{K}, \mathrm{d}\omega)$), (ii) the quadrature rule DPPUQ based on the same projection DPP but with uniform weights $(w_n = 1/N)$ (Johansson, 1997), (iii) the kernel quadrature rule (4.131) of Bach, 2017, which we denote LVSQ for *leverage score quadrature*, without regularization $(\lambda = 0)$[9], (iv) sequential Bayesian quadrature

[8] We write e_m rather than $e_m^{\mathcal{S}_s}$ to make the notation lighter.

[9] The optimal proposal q_λ^* is not defined, in a strict sense, when $\lambda = 0$. However, it is constant for $\lambda > 0$; so we implement LVSQ $(\lambda = 0)$ using i.i.d. nodes from the uniform distribution of $[0,1]$.

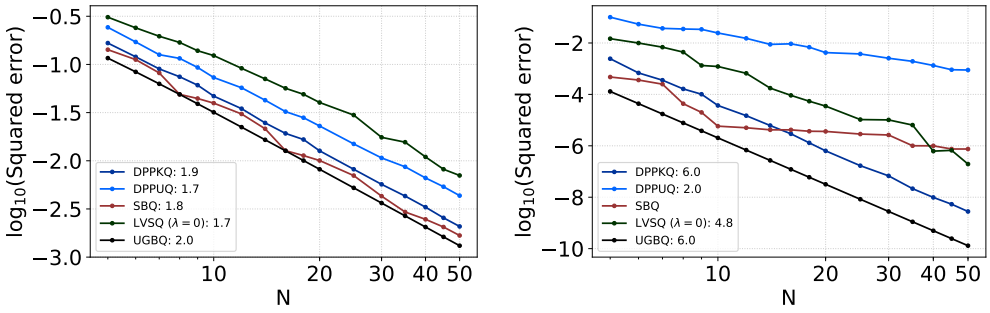


Figure 4.10 – Squared approximation error vs. number of nodes N in the case of the periodic Sobolev space for $s = 1$ (left) and $s = 3$ (right).

(SBQ) (Huszár and Duvenaud, 2012) with regularization to avoid numerical instability, and (v) optimal kernel quadrature on the uniform grid (UGKQ). We take $N \in [5, 50]$. Figure 4.10 shows log-log plots of the square of the worst case quadrature error (4.108) w.r.t. N, averaged over 50 samples for each point, for $s \in \{1, 3\}$.

We observe that the approximation errors of all first four quadratures converge to 0 with different rates. Both UGKQ and DPPKQ converge to 0 with a rate of $\mathcal{O}(N^{-2s})$, which indicates that our $\mathcal{O}(N^{2-2s})$ bound in Theorem 4.7 is not tight in the Sobolev case. Meanwhile, the rate of DPPUQ is $\mathcal{O}(N^{-2})$ across the three values of s: it does not adapt to the regularity of the integrands. This corresponds to the CLT proven by Johansson, 1997. LVSQ without regularization converges to 0 slightly slower than $\mathcal{O}(N^{-2s})$. SBQ is the only one that seems to plateau for $s = 3$, although it consistently has the best performance for low N.

Overall, in the Sobolev case, DPPKQ and UGKQ have the best convergence rate. UGKQ – known to be optimal in this case (Bojanov, 1981) – has a better constant.

The reconstruction of the embeddings of the other eigenfunctions

We consider g to be an eigenfunction $\mathbf{\Sigma}$: $g \in \{e_5, e_{10}, e_{14}, e_{15}, e_{16}, e_{17}\}$, so that $\mu_g = \mathbf{\Sigma} g$ is tractable. We observe the interpolation error under the projection DPP (DPPKQ) for $N \in [5, 50]$. Figure 4.11 shows log-log plots of the square of the interpolation error (4.124) of the embedding μ_{e_n} w.r.t. N, denoted $\epsilon_n(N)$, averaged over 50 samples for each point, for $s \in \{1, 3\}$.

We observe that for every e_{N_0}, the interpolation error goes through two phases: in the first phase ($N < N_0$) the interpolation error is practically equal to *the initial error* that is $\|\mu_{e_{N_0}}\|^2_{\mathcal{F}} = \sigma_{N_0}$; in the second phase ($N \geq N_0$) the interpolation error decreases to 0 at the rate $\mathcal{O}(N^{-2s})$. These observations indicate again that the $\mathcal{O}(N^{2-2s})$ bound in Theorem 4.7 is not tight. Moreover, we observe that given two eigenfunctions e_n and $e_{n'}$ corresponding to the same eigenvalue, the expected interpolation error is practically the same. This is the case, for instance, of e_{14} and e_{15} or e_{16} and e_{17}.

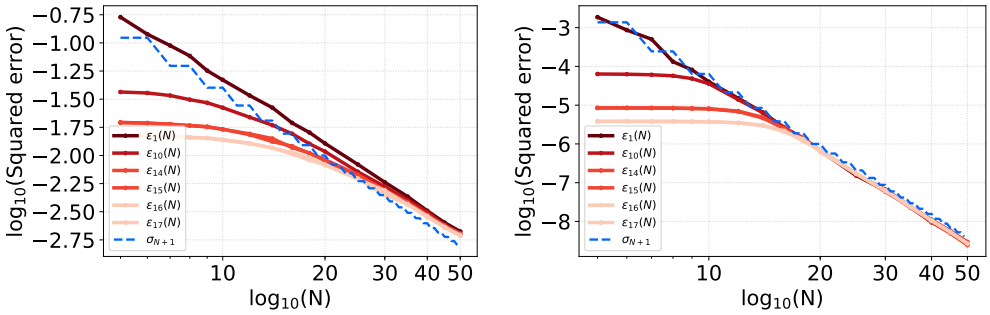


Figure 4.11 – Squared interpolation error for the embeddings of the eigenfunctions vs. number of nodes N in the periodic Sobolev space for $s = 1$ (left) and $s = 3$ (right).

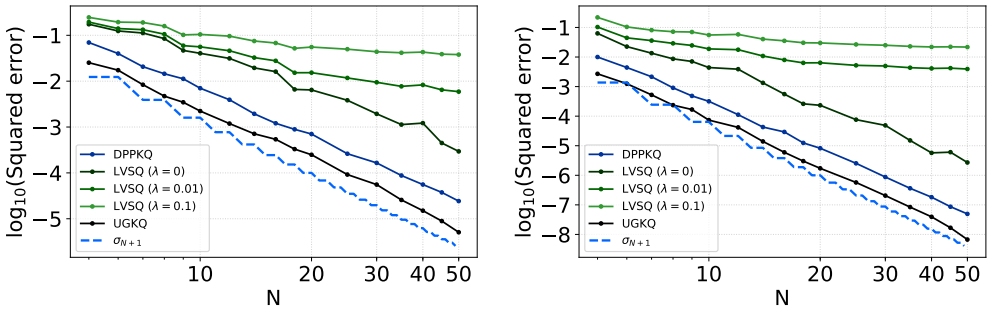


Figure 4.12 – Squared worst-case interpolation error on $\mathcal{U}$ error vs. number of nodes N for $s = 1$ (left) and $s = 3$ (right).

The uniform reconstruction of the embeddings of the unit ball

In the previous experiments, we investigated empirically $\|\mu_g - \mathbf{\Pi}_{\mathcal{T}(x)}\, \mu_g\|_{\mathcal{F}}^2$ for some functions g in the unit ball of $\mathbb{L}_2(\mathrm{d}\omega)$. Remember, that the bound (4.145) in Theorem 4.8 deals with the worst case interpolation error defined by

$$\sup_{\|g\|_{\mathrm{d}\omega} \leq 1} \|\mu_g - \mathbf{\Pi}_{\mathcal{T}(x)} \mu_g\|_{\mathcal{F}}^2. \tag{4.160}$$

We investigate this quantity by considering the following surrogate

$$\sup_{g \in \mathcal{U}} \|\mu_g - \mathbf{\Pi}_{\mathcal{T}(x)} \mu_g\|_{\mathcal{F}}^2, \tag{4.161}$$

where $\mathcal{U}$ is a finite subset of the intersection of the unit ball of $\mathbb{L}_2(\mathrm{d}\omega)$ with the subspace $\mathrm{Span}(e_n)_{n \in [50]}$[10]. We take $|\mathcal{U}| = 2000$.

Again, we compare the following quadratures defined in the first experiment: (i) the quadrature rule DPPKQ, (ii) LVSQ with regularization parameter $\lambda \in \{0, 0.01, 0.1\}$, and (iii) UGKQ. We take $N \in [5, 50]$. Figure 4.12 show log-log plots of the square of the worst-case interpolation error on $\mathcal{U}$ w.r.t. N, averaged over 50 samples for each point, for $s \in \{1, 3\}$.

We observe that the square of the worst case interpolation errors, for the three quadratures DPPKQ, UGKQ and LVSQ ($\lambda = 0$), converges to 0. DPPKQ converges

[10] The square of the interpolation error when $g \in \mathrm{Span}(e_n^{\mathcal{S}_s})_{n \in [50]}^{\perp}$ starts from an initial error lower than σ_{51}.

at the rate $\mathcal{O}(N^{-2s})$ faster than the theoretical rate $\mathcal{O}(N^{3-2s})$ of Theorem 4.8. UGKQ seems to converge at the rate $\mathcal{O}(N^{-2s})$ with a better constant than DPPKQ. Finally, LVSQ converges at a rate slower than $\mathcal{O}(N^{-2s})$ when $\lambda = 0$; while it plateaus, below the level λ, when $\lambda \in \{0.01, 0.1\}$ in accordance with Proposition 4.3.

Among the three "ridgeless" ($\lambda = 0$) algorithms, DPPKQ is the only one to have guarantees for the worst case interpolation error (4.160) on the unit ball of $\mathbb{L}_2(\mathrm{d}\omega)$although the empirical investigation shows that the rates are pessimistic.

4.4.2 *The Korobov spaces*

Now, we consider the Korobov space $\mathcal{K}_{d,s}$ defined in Section 4.2.3. We sampled from the corresponding DPP using the generic sampling algorithm in Hough et al., 2006. The implementation requires to sample from density $x \mapsto \frac{1}{N} \mathfrak{K}(x, x)$, which is possible using the uniform density as a proposal in the successive rejection sampling steps. Indeed, the analytical expression of $\mathfrak{K}(x, x)$ involves the squares of cosine and sine functions that can be upper bounded by constants.

We consider two numerical experiments. In the first one, we consider the approximation of μ_g when $g \equiv 1$ using several quadratures; in the second one, we consider the approximation of μ_g when g is an eigenfunction of $\mathbf{\Sigma}$ using DPPKQ, LVSQ ($\lambda = 0$) and UGKQ.

The reconstruction of the embedding of the first eigenfunction

We still take $g \equiv 1$ so that $\mu_g \equiv 1$. We compare (i) our DPPKQ, (ii) LVSQ without regularization ($\lambda = 0$)[11], (iii) the optimal kernel quadrature based on the uniform grid UGKQ, (iv) the optimal kernel quadrature SGKQ based on the sparse grid from (Smolyak, 1963), (v) the optimal kernel quadrature based on the Halton sequence HaltonKQ (Halton, 1964). We take $N \in [5, 1000]$ and $s = 1$. The results are shown in Figure 4.13. This time, UGKQ suffers from the dimension with a rate in $\mathcal{O}(N^{-2s/d})$, while DPPKQ, HaltonKQ and LVSQ ($\lambda = 0$) all perform similarly well. They scale as $\mathcal{O}((\log N)^{2s(d-1)} N^{-2s})$, which is a tight upper bound on σ_{N+1}, see (Bach, 2017). SGKQ seems to lag slightly behind with a rate $\mathcal{O}((\log N)^{2(s+1)(d-1)} N^{-2s})$ (Holtz, 2008; Smolyak, 1963).

The reconstruction of the embeddings of the other eigenfunctions

We consider now g to be an eigenfunction $\mathbf{\Sigma}$, $g \in \{e_1, \ldots, e_8, e_{11}, e_{12}, e_{21}, e_{22}\}$. We observe the interpolation errors for the following algorithms (i) our DPPKQ, (ii) LVSQ without regularization ($\lambda = 0$), (iii) the optimal kernel quadrature based on the uniform grid UGKQ for $N \in [5, 1000]$. Figure 4.14 shows log-log plots of the square of the interpolation errors of the corresponding embedding w.r.t. N, averaged over 50 samples (in the case of DPPKQ and LVSQ) for each point, for $s \in \{1, 3\}$; respectively for the DPPKQ, LVSQ and UGKQ.

Once again, we observe that under DPPKQ, the empirical expected interpolation error curves manifest two phases: the plateau of the initial error ($N < N_0$) followed by

[11] The optimal proposal q_λ^* is not defined, in a strict sense, when $\lambda = 0$. However, it is constant for $\lambda > 0$; so we implement LVSQ ($\lambda = 0$) using i.i.d. nodes from the uniform distribution of $[0, 1]^d$.

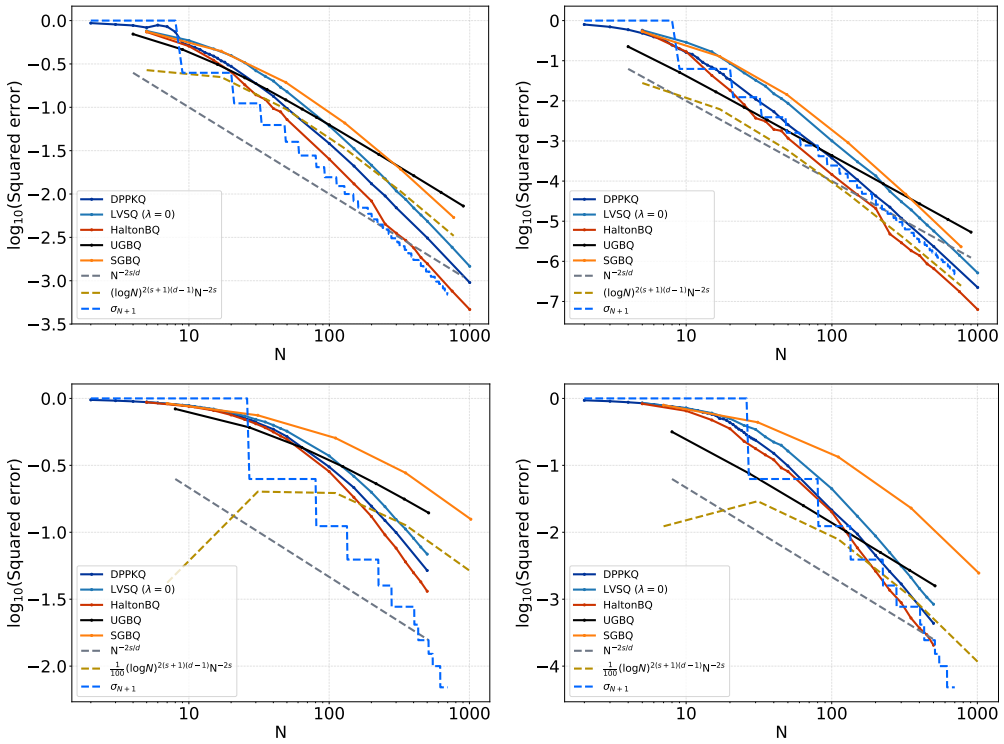


Figure 4.13 – Squared approximation error vs. number of nodes N in the case of the Korobov space for $s = 1$ (left) and $s = 3$ (right); and for $d = 2$ (top) and $d = 3$ (bottom).

the convergence at the rate $\mathcal{O}(\sigma_{N+1})$. Moreover, we observe again that eigenfunctions with the same eigenvalue have similar curves; and for sufficiently large values of N all the curves seem to coincide. As for LVSQ, the empirical curves seem to have the same rates as those of DPPKQ yet with worst constants. Finally, the empirical curves of UGKQ converges to 0 at the rate $\mathcal{O}(N^{-2s/d})$ and the curves do not seem to coincide for large values of N.

Once again, the rates predicted by the bounds in Theorem 4.7 are pessimistic compared to the empirical rates of DPPKQ.

4.4.3 *The unidimensional Gaussian kernel*

Now, we consider the Gaussian space $\mathcal{G}_\gamma$ defined in Section 4.2.3.

Figure 4.15 compiles the empirical performance of DPPKQ for $g \in \{e_1, e_5, e_{10}, e_{15}\}$ compared to the theoretical bound of Theorem 4.7 , crude Monte Carlo with i.i.d. sampling from the measure ω (MC), optimal kernel quadrature based on i.i.d. samples from ω (MCKQ), sequential Bayesian Quadrature (SBQ). From random quadrature, we take the average over 50 samples. We take $N \in [5, 50]$ and $\gamma = \frac{1}{2}$. Note that, this time, only the y-axis is on the log scale for better display, and that LVSQ is not plotted since we don't know how to sample from the continuous ridge leverage score distribution q^*_λ.

We observe that, for $g = e_{N_0}$, the quadratures converge to 0 at different rates after the plateau of the initial error $N \geq N_0$. DPPKQ converges as $\mathcal{O}(\sigma_{N+1})$ while the discussion below Theorem 4.7 let us expect a slightly slower rate $\mathcal{O}(N\sigma_{N+1})$ denoted

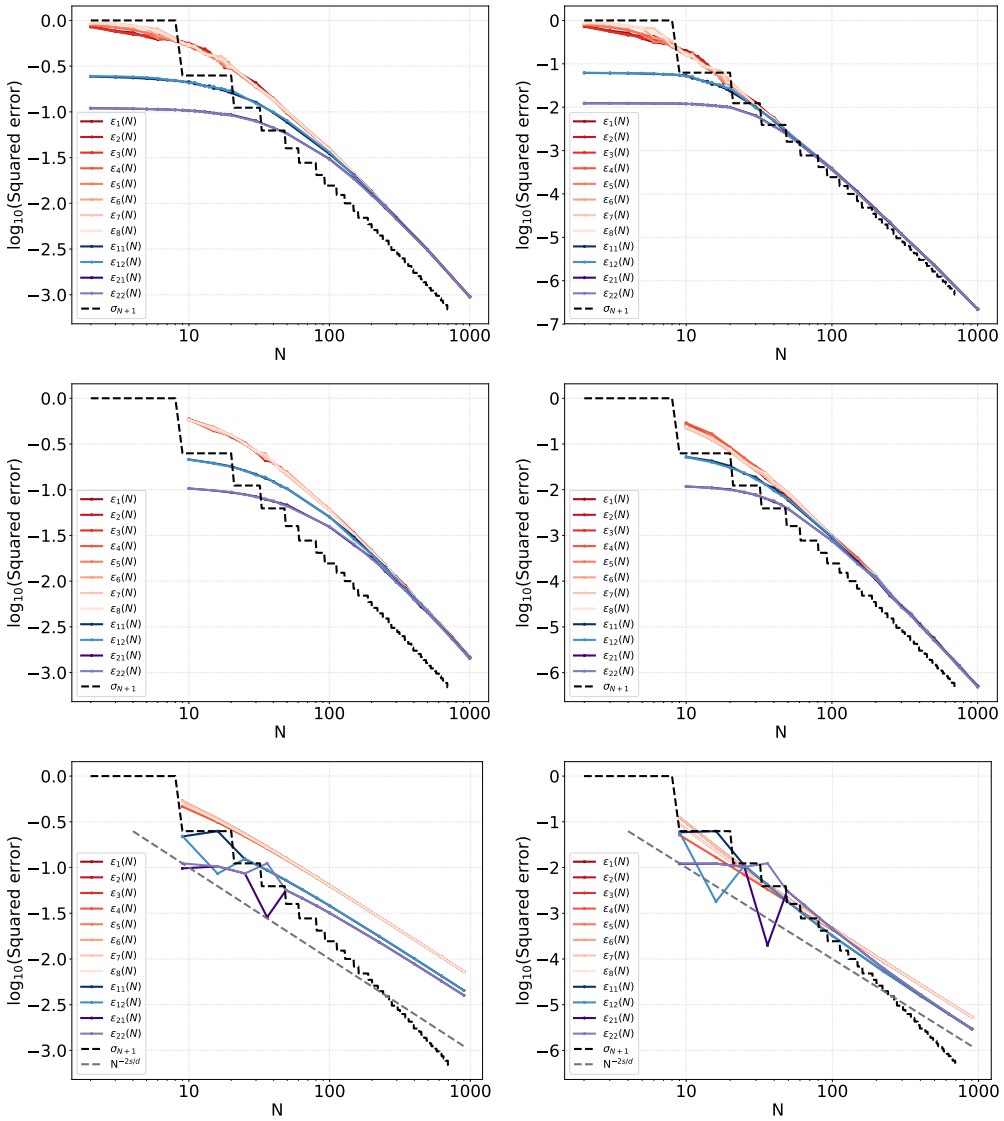
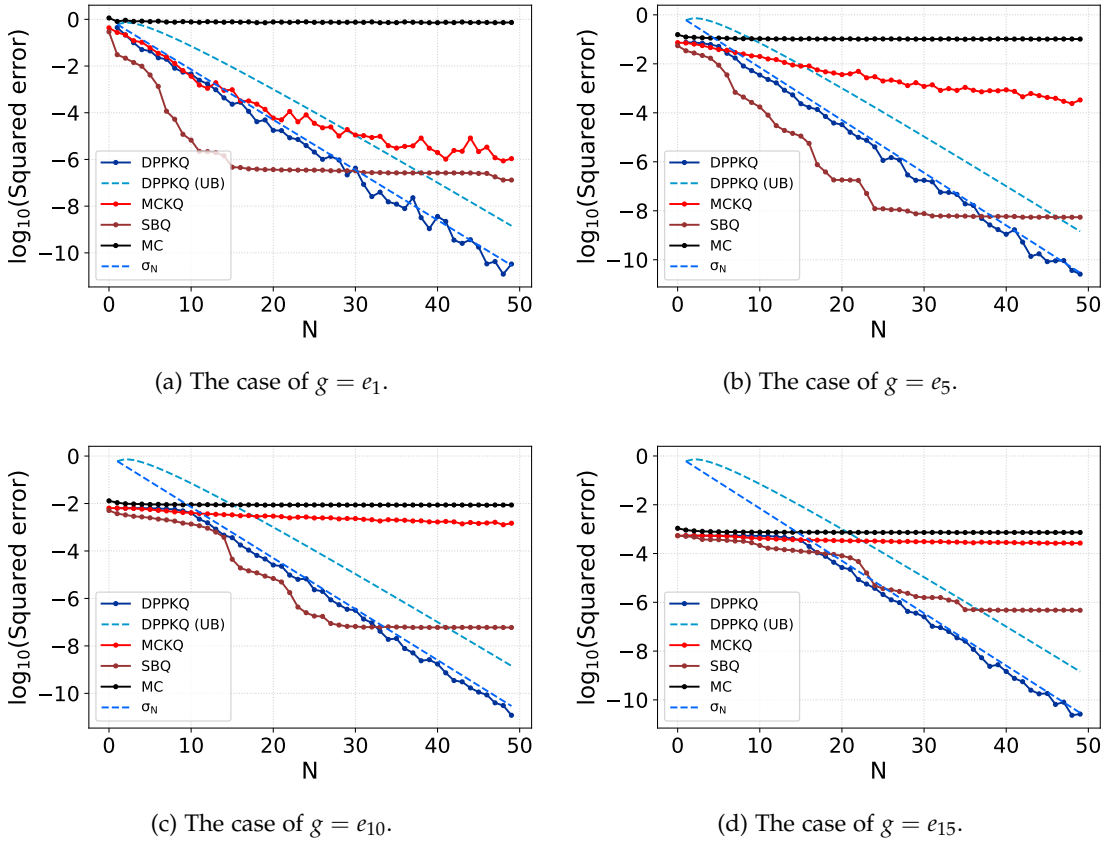


Figure 4.14 – Squared interpolation error for the embeddings of the eigenfunctions vs. number of nodes N in the Korobov space for $s = 1$ (left) and $s = 3$ (right) under DPPKQ (top), LVSQ (middle) and UGKQ (bottom).



(a) The case of $g = e_1$.

(b) The case of $g = e_5$.

(c) The case of $g = e_{10}$.

(d) The case of $g = e_{15}$.

Figure 4.15 – The squared error $\|\mu_g - \sum_{n\in[N]} w_n k(x_n, .)\|_{\mathcal{F}}^2$ vs. the number of nodes N in the Gaussian space $\mathcal{G}_\gamma$ with $\gamma = 1/2$.

DPPKQ (UB) in the figure. MCKQ improves upon Monte Carlo that converges as $\mathcal{O}(N^{-1})$. Yet, it seems that the larger is the value of N_0, the slower is the convergence of MCKQ to 0 compared to the convergence of DPPKQ that does not depend on the value of N_0. The same observations apply for SBQ, that has the smallest error for small values of N, yet its convergence slowdowns for large values of N. Moreover, its performance deteriorates for large values of N_0.

We conclude that DPPKQ has the most consistent behaviour among the four quadratures that we have compared. Moreover, the squared worst integration error scales as $\mathcal{O}(\sigma_{N+1})$, which is slightly better than the rate predicted by Theorem 4.8.

4.4.4 *The multidimensional Gaussian kernel*

We consider the case of the multidimensional Gaussian space $\mathcal{G}_{d,\gamma}$ with $d \in \{2, 3\}$ and $\gamma = 1$. We take $g = e_1$. The results are compiled in Figure 4.16. Once again, the numerical simulations show that the empirical rate of DPPKQ scales as $\mathcal{O}(\sigma_{N+1})$.

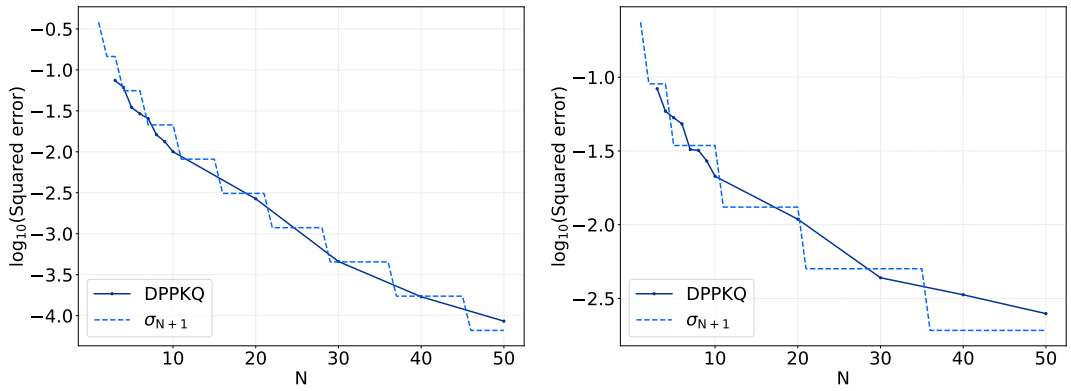


Figure 4.16 – The squared interpolation error in the Gaussian space $\mathcal{G}_{d,\gamma}$ with $d \in \{2,3\}$ and $\gamma = 1$.

4.5 DISCUSSION

We have introduced a new class of quadratures: the optimal kernel quadrature based on nodes that follows the distribution of the projection DPP$(\mathfrak{K}, \mathrm{d}\omega)$. Moreover, we gave theoretical guarantees of this class of quadratures. In particular, we proved that for $g \in \mathbb{L}_2(\mathrm{d}\omega)$ such that $\|g\|_{\mathrm{d}\omega} \leq 1$, the rate of convergence $\mathbb{E}_{\mathrm{DPP}} \|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2$ scales as $\mathcal{O}(Nr_{N+1})$, and $\mathbb{E}_{\mathrm{DPP}} \sup_{\|g\|_{\mathrm{d}\omega}\leq 1} \|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2$ scales as $\mathcal{O}(N^2 r_{N+1})$. The numerical simulations that we conducted suggest that these rates scale as $\mathcal{O}(\sigma_{N+1})$, which means that the theoretical guarantees we gave are pessimistic, especially for RKHSs with finite degree of smoothness like the periodic Sobolev spaces and the Korobov spaces.

Similarly to the quadratures reviewed in Section 4.1.3, this new class of quadratures is suitable for functions living in a Hilbert space; however, the two settings are different. Indeed, the previous results deal with functions living in Sobolev spaces that does not correspond to some RKHS ($s \leq d/2$). This point will be discussed more in details in Section 6.2.3.

Unlike the CLTs reviewed in Section 4.1.3, our theoretical rates are non-asymptotic. Moreover, the technique of the proof is new and relies on controlling the largest principal angle between the subspaces $\tilde{\mathcal{T}}(\boldsymbol{x})$ and $\mathcal{E}_N^{\tilde{\mathcal{F}}}$ under the projection DPP. To achieve this control, we have introduced the symmetrizations (4.152) and (4.153) of (4.151), that are rather loose majorizations. Our motivation is that the expected value of each of these symmetric quantities is tractable under the DPP. Compared to the multiplicative symmetrization, the additive symmetrization gives a neater upper bound of $\mathbb{E}_{\mathrm{DPP}} 1/\cos^2\theta_N(\mathcal{E}_N^{\tilde{\mathcal{F}}}, \tilde{\mathcal{T}}(\boldsymbol{x}))$ without improving the rate of convergence. Getting rid of these symmetrizations could make the bound much tighter. We discuss this perspective in Section 6.2.4.

Now, compared to other algorithms and quadratures, DPPKQ seem to give consistent empirical performance: the squared interpolation error scales as $\mathcal{O}(\sigma_N)$ after a stagnation at the plateau of the initial error. DPPKQ offers an implementable solution with guarantees in the ridgeless regime ($\lambda = 0$), in comparison LVSQ have guarantees when $\lambda > 0$, yet the continuous ridge leverage score distribution is not be

tractable in general. Moreover, DPPKQ seems to be numerically stable compared to other algorithms such as SBQ that require some regularization.

Now all that remains to be done is to sharpen the theoretical bounds of DPPKQ. We can already guess that the successive majorizations are to blame for these non optimal bounds. We will see in Chapter 5, an alternative distribution under which we can express $\mathbb{E}\,\|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|^2_{\mathcal{F}}$ in a tractable way without loose majorization. In addition, this distribution can be approximated using an MCMC algorithm without the need to the spectral decomposition of the integration operator.

4.6 PROOFS

This section contains the detailed proofs of the results appearing in this chapter.

We start by Section 4.6.1 that contains the proof of an adapted result, borrowed from the literature, on leverage score changes under rank 1 updates. This result will be used later.

Section 4.6.2 introduces the principal angles between subspaces in Hilbert spaces. In particular, we prove the existence of these angles along with some of their properties that will be used to prove Lemma 4.2 and Proposition 4.5.

Section 4.6.3 contains the proof of Proposition 4.4 that we use to ensure that $\boldsymbol{K}(\boldsymbol{x})$ is almost surely invertible when $\boldsymbol{x} = \{x_1, \dots, x_N\}$ is a projection DPP with reference measure $\mathrm{d}\omega$ and kernel $\mathfrak{K}$.

The rest of Section 4.6 deals with Theorem 4.7, our upper bound on the approximation error of DPP-based kernel quadrature. The proof is rather long, but can be decomposed in four steps, which we now introduce for ease of reading.

First, in Section 4.6.4, we prove Lemma 4.1 that highlights the importance of the term $\max_{n\in[N]} \sigma_n \|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})^\perp} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2$ that relates to the approximation error of the space spanned by $(e_n^{\mathcal{F}})_{n\in[N]}$ by the (random) subspace $\mathcal{T}(\boldsymbol{x})$.

Second, we prove Proposition 4.6 which is proven thanks (4.154) in Proposition 4.5 and Proposition 4.11. This is Section 4.6.6.

Third, in Section 4.6.5, we bound the geometric term $\max_{n\in[N]} \sigma_n \|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})^\perp} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2$ for a fixed configuration $\boldsymbol{x}$ by $\max_{n\in[N]} \sigma_1 \|\boldsymbol{\Pi}_{\tilde{\mathcal{T}}(\boldsymbol{x})^\perp} e_n^{\tilde{\mathcal{F}}}\|_{\tilde{\mathcal{F}}}^2$. This is done in Proposition 4.8, which in turn requires two intermediate results, Lemma 4.4 and Proposition 4.10.

Fourth, Theorem 4.7 is obtained in Section 4.6.7, using the results of the previous steps.

Finally, we prove the refined result of Theorem 4.8 using Proposition 4.7 in Section 4.6.8.

4.6.1 *A borrowed result: leverage scores changes under rank 1 updates*

In this section we prove a lemma inspired from Lemma 5 in Cohen et al., 2015. This lemma concerns the changes of leverage scores under rank 1 updates.

Let $N, M \in \mathbb{N}^*$, $M \geq N$. Let $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ be a matrix of full rank. For $i \in [M]$, denote $\boldsymbol{a}_i$ the i-th column of the matrix $\boldsymbol{A}$. The i-th leverage score of the matrix $\boldsymbol{A}$ is defined by

$$\tau_i(\boldsymbol{A}) = \boldsymbol{a}_i^\intercal (\boldsymbol{A}\boldsymbol{A}^\intercal)^{-1} \boldsymbol{a}_i, \tag{4.162}$$

while the cross-leverage score between the i-th column and the j-th column is defined by

$$\tau_{i,j}(\boldsymbol{A}) = \boldsymbol{a}_i^\intercal (\boldsymbol{A}\boldsymbol{A}^\intercal)^{-1} \boldsymbol{a}_j. \tag{4.163}$$

It holds (Drineas et al., 2006)

$$\forall i \in [M],\ \tau_i(\boldsymbol{A}) \in [0,1], \tag{4.164}$$

and we have the following result.

Lemma 4.3. *Let $N, M \in \mathbb{N}^*$, $M \geq N$. Let $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ of full rank and $\rho \in \mathbb{R}_+^*$ and $i \in [M]$. Let $\boldsymbol{W} \in \mathbb{R}^{M\times M}$ a diagonal matrix such that $\boldsymbol{W}_{i,i} = \sqrt{1+\rho}$ and $\boldsymbol{W}_{j,j} = 1$ for $j \neq i$. Then*

$$\tau_i(\boldsymbol{AW}) = \frac{(1+\rho)\tau_i(\boldsymbol{A})}{1+\rho\tau_i(\boldsymbol{A})} \geq \tau_i(\boldsymbol{A}), \tag{4.165}$$

and

$$\forall j \in [M] - \{i\},\ \tau_j(\boldsymbol{AW}) = \tau_j(\boldsymbol{A}) - \frac{\rho\tau_{i,j}(\boldsymbol{A})^2}{1+\rho\tau_i(\boldsymbol{A})} \leq \tau_j(\boldsymbol{A}). \tag{4.166}$$

The proof of this lemma is similar to Lemma 5 in Cohen et al., 2015. We recall the proof for completeness.

Proof. (Adapted from Cohen et al., 2015) The Sherman-Morrison formula applied to $\boldsymbol{AWW}^\intercal\boldsymbol{A}^\intercal$ and the vector $\sqrt{\rho}\boldsymbol{a}_i$ yields

$$\begin{aligned}(\boldsymbol{AWW}^\intercal\boldsymbol{A}^\intercal)^{-1} &= (\boldsymbol{AA}^\intercal + \rho\boldsymbol{a}_i\boldsymbol{a}_i^\intercal)^{-1} &(4.167)\\ &= (\boldsymbol{AA}^\intercal)^{-1} - \frac{(\boldsymbol{AA}^\intercal)^{-1}\rho\boldsymbol{a}_i\boldsymbol{a}_i^\intercal(\boldsymbol{AA}^\intercal)^{-1}}{1+\rho\boldsymbol{a}_i^\intercal(\boldsymbol{AA}^\intercal)^{-1}\boldsymbol{a}_i}. &(4.168)\end{aligned}$$

By definition of $\tau_i(\boldsymbol{AW})$

$$\begin{aligned}\tau_i(\boldsymbol{AW}) &= \sqrt{1+\rho}\boldsymbol{a}_i^\intercal(\boldsymbol{AWW}^\intercal\boldsymbol{A}^\intercal)^{-1}\boldsymbol{a}_i\sqrt{1+\rho} &(4.169)\\ &= (1+\rho)\boldsymbol{a}_i^\intercal\left((\boldsymbol{AA}^\intercal)^{-1} - \frac{(\boldsymbol{AA}^\intercal)^{-1}\rho\boldsymbol{a}_i\boldsymbol{a}_i^\intercal(\boldsymbol{AA}^\intercal)^{-1}}{1+\rho\boldsymbol{a}_i^\intercal(\boldsymbol{AA}^\intercal)^{-1}\boldsymbol{a}_i}\right)\boldsymbol{a}_i\\ &= (1+\rho)\left(\tau_i(\boldsymbol{A}) - \frac{\rho\tau_i(\boldsymbol{A})^2}{1+\rho\tau_i(\boldsymbol{A})}\right)\\ &= (1+\rho)\frac{\tau_i(\boldsymbol{A})}{1+\rho\tau_i(\boldsymbol{A})}.\end{aligned}$$

Now let $j \in [M] - \{i\}$. By definition of $\tau_j(\boldsymbol{AW})$

$$\begin{aligned}\tau_j(\boldsymbol{AW}) &= \boldsymbol{a}_j^\intercal(\boldsymbol{AWW}^\intercal\boldsymbol{A}^\intercal)^{-1}\boldsymbol{a}_j &(4.170)\\ &= \boldsymbol{a}_j^\intercal\left((\boldsymbol{AA}^\intercal)^{-1} - \frac{(\boldsymbol{AA}^\intercal)^{-1}\rho\boldsymbol{a}_i\boldsymbol{a}_i^\intercal(\boldsymbol{AA}^\intercal)^{-1}}{1+\rho\boldsymbol{a}_i^\intercal(\boldsymbol{AA}^\intercal)^{-1}\boldsymbol{a}_i}\right)\boldsymbol{a}_j\\ &= \tau_j(\boldsymbol{A}) - \frac{\rho\tau_{i,j}(\boldsymbol{A})^2}{1+\rho\tau_i(\boldsymbol{A})}\\ &\leq \tau_j(\boldsymbol{A}).\end{aligned}$$

□

4.6.2 *Principal angles in Hilbert spaces*

The notion of principal angles between subspaces of a Hilbert spaces can be extended in many ways. See for example (Deutsch, 1995). We give in the following theorem one extension of this notion. This extension is restricted to subspaces of finite dimension, which is more than enough for our proofs.

Theorem 4.9. *Let $\mathcal{H}$ be a Hilbert space. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two finite-dimensional subspaces of $\mathcal{H}$ with $N = \dim \mathcal{P}_1 = \dim \mathcal{P}_2$. Denote $\mathbf{\Pi}_{\mathcal{P}_1}$ and $\mathbf{\Pi}_{\mathcal{P}_2}$ the orthogonal projections of $\mathcal{H}$ onto these two subspaces. There exist two orthonormal bases for $\mathcal{P}_1$ and $\mathcal{P}_2$ denoted $(\boldsymbol{v}_i^1)_{i\in[N]}$ and $(\boldsymbol{v}_i^2)_{i\in[N]}$, and a set of angles $\theta_i(\mathcal{P}_1, \mathcal{P}_2) \in [0, \frac{\pi}{2}]$ such that*

$$\cos\theta_N(\mathcal{P}_1, \mathcal{P}_2) \leq \cdots \leq \cos\theta_1(\mathcal{P}_1, \mathcal{P}_2), \tag{4.171}$$

and for $i \in [1, ..., N]$

$$\langle \boldsymbol{v}_i^1, \boldsymbol{v}_i^2 \rangle_{\mathcal{H}} = \cos\theta_i(\mathcal{P}_1, \mathcal{P}_2), \tag{4.172}$$

and

$$\mathbf{\Pi}_{\mathcal{P}_1} \boldsymbol{v}_i^2 = \cos\theta_i(\mathcal{P}_1, \mathcal{P}_2) \boldsymbol{v}_i^1, \tag{4.173}$$

and

$$\mathbf{\Pi}_{\mathcal{P}_2} \boldsymbol{v}_i^1 = \cos\theta_i(\mathcal{P}_1, \mathcal{P}_2) \boldsymbol{v}_i^2. \tag{4.174}$$

In particular

$$\cos\theta_N(\mathcal{P}_1, \mathcal{P}_2) = \inf_{\boldsymbol{v}\in\mathcal{P}_1, \|\boldsymbol{v}\|_{\mathcal{H}}=1} \|\mathbf{\Pi}_{\mathcal{P}_2}\boldsymbol{v}\|_{\mathcal{H}} = \inf_{\boldsymbol{v}\in\mathcal{P}_2, \|\boldsymbol{v}\|_{\mathcal{H}}=1} \|\mathbf{\Pi}_{\mathcal{P}_1}\boldsymbol{v}\|_{\mathcal{H}}. \tag{4.175}$$

Proof. Consider the operator $\boldsymbol{\pi} = \mathbf{\Pi}_{\mathcal{P}_2}\mathbf{\Pi}_{\mathcal{P}_1}$. Since $\mathcal{P}_1$ and $\mathcal{P}_2$ are finite-dimensional subspaces of $\mathcal{H}$, $\boldsymbol{\pi}$ is a finite-rank operator. Therefore, using the singular value decomposition, there exists two orthonormal families $(\boldsymbol{v}_n^1)_{n\in[R]}$ and $(\boldsymbol{v}_n^2)_{n\in[R]}$ such that

$$\forall x \in \mathcal{H},\ \boldsymbol{\pi}x = \sum_{n\in[R]} \sigma_n(\boldsymbol{\pi}) \langle x, \boldsymbol{v}_n^1 \rangle_{\mathcal{H}} \boldsymbol{v}_n^2, \tag{4.176}$$

where R is the rank of $\boldsymbol{\pi}$, and the $(\sigma_n(\boldsymbol{\pi}))_{n\in[R]}$ is the non-increasing sequence of positive singular values of $\boldsymbol{\pi}$. Let $m \in [R]$ then

$$\boldsymbol{v}_m^2 = \frac{1}{\sigma_m(\boldsymbol{\pi})} \boldsymbol{\pi}\boldsymbol{v}_m^1 = \frac{1}{\sigma_m(\boldsymbol{\pi})} \mathbf{\Pi}_{\mathcal{P}_2}\mathbf{\Pi}_{\mathcal{P}_1}\boldsymbol{v}_m^1 \in \mathcal{P}_2. \tag{4.177}$$

Moreover,

$$\forall x \in \mathcal{P}_1^{\perp},\ \sum_{n\in[N]} \sigma_n(\boldsymbol{\pi}) \langle x, \boldsymbol{v}_n^1 \rangle_{\mathcal{H}} \boldsymbol{v}_n^2 = \mathbf{\Pi}_{\mathcal{P}_2}\mathbf{\Pi}_{\mathcal{P}_1} x = 0, \tag{4.178}$$

therefore and for every $m \in [R]$, $\boldsymbol{v}_m^1 \in (\mathcal{P}_1^{\perp})^{\perp} = \mathcal{P}_1$.

We can complete the family $(\boldsymbol{v}_n^1)_{n\in[R]}$ in $\mathcal{P}_1$ to an o.n.b. $(\boldsymbol{v}_n^1)_{n\in[N]}$, and similarly we complete the family $(\boldsymbol{v}_n^2)_{n\in[R]}$ in $\mathcal{P}_2$ to an o.n.b. $(\boldsymbol{v}_n^2)_{n\in[N]}$.

We have for $m > R$,

$$\mathbf{\Pi}_2 \boldsymbol{v}_m^1 = \boldsymbol{\pi}\boldsymbol{v}_m^1 = \sum_{n\in[R]} \sigma_n(\boldsymbol{\pi}) \langle \boldsymbol{v}_m^1, \boldsymbol{v}_n^1 \rangle_{\mathcal{H}} \boldsymbol{v}_n^2 = 0, \tag{4.179}$$

therefore, $\boldsymbol{v}_m^1 \in \mathcal{P}_2^\perp$. In particular,

$$\forall m \in \{R+1,\dots,N\},\ \langle \boldsymbol{v}_m^1, \boldsymbol{v}_m^2\rangle_{\mathcal{H}} = 0 = \cos\theta_m(\mathcal{P}_1,\mathcal{P}_2), \tag{4.180}$$

where we take $\theta_m(\mathcal{P}_1,\mathcal{P}_2) = \pi/2$ for $m \in \{R+1,\dots,N\}$.

Now, $\boldsymbol{\pi}$ is the product of two orthogonal projections, then the $\sigma_n(\boldsymbol{\pi})$ are included in $]0,1]$. We denote for $n \in [R]$, $\theta_n(\mathcal{P}_1,\mathcal{P}_2) \in [0,\pi/2[$ such that

$$\cos\theta_n(\mathcal{P}_1,\mathcal{P}_2) = \sigma_n(\boldsymbol{\pi}). \tag{4.181}$$

By definition, $(\sigma_n(\boldsymbol{\pi}))_{n\in[R]}$ is a non-increasing sequence, then $(\theta_n(\mathcal{P}_1,\mathcal{P}_2))_{n\in[R]}$ is a non-decreasing sequence. Therefore, $(\theta_n(\mathcal{P}_1,\mathcal{P}_2))_{n\in[N]}$ is a non-decreasing sequence.

Now, by (4.176)

$$\cos\theta_n(\mathcal{P}_1,\mathcal{P}_2)\boldsymbol{v}_n^2 = \sigma_n(\boldsymbol{\pi})\boldsymbol{v}_n^2 = \boldsymbol{\pi}\boldsymbol{v}_n^2 = \boldsymbol{\Pi}_{\mathcal{P}_2}\boldsymbol{\Pi}_{\mathcal{P}_1}\boldsymbol{v}_n^1 = \boldsymbol{\Pi}_{\mathcal{P}_2}\boldsymbol{v}_n^1, \tag{4.182}$$

and

$$\begin{aligned}\cos\theta_n(\mathcal{P}_1,\mathcal{P}_2) &= \langle \boldsymbol{v}_n^2, \cos\theta_n(\mathcal{P}_1,\mathcal{P}_2)\boldsymbol{v}_n^2\rangle_{\mathcal{H}}\\ &= \langle \boldsymbol{v}_n^2, \boldsymbol{\Pi}_{\mathcal{P}_2}\boldsymbol{v}_n^1\rangle_{\mathcal{H}}\\ &= \langle \boldsymbol{\Pi}_{\mathcal{P}_2}\boldsymbol{v}_n^2, \boldsymbol{v}_n^1\rangle_{\mathcal{H}}\\ &= \langle \boldsymbol{v}_n^2, \boldsymbol{v}_n^1\rangle_{\mathcal{H}}.\end{aligned} \tag{4.183}$$

By taking the adjoint in (4.176) we get

$$\boldsymbol{\pi}^* x = \boldsymbol{\Pi}_{\mathcal{P}_1}\boldsymbol{\Pi}_{\mathcal{P}_2} = \sum_{n\in[R]} \sigma_n(\boldsymbol{\pi})\langle x, \boldsymbol{v}_n^2\rangle_{\mathcal{H}}\boldsymbol{v}_n^1, \tag{4.184}$$

and we can prove similarly that

$$\forall n \in [R],\ \cos\theta_n(\mathcal{P}_1,\mathcal{P}_2)\boldsymbol{v}_n^1 = \boldsymbol{\pi}^*\boldsymbol{v}_n^1 = \boldsymbol{\Pi}_{\mathcal{P}_1}\boldsymbol{\Pi}_{\mathcal{P}_2}\boldsymbol{v}_n^2 = \boldsymbol{\Pi}_{\mathcal{P}_1}\boldsymbol{v}_n^2, \tag{4.185}$$

and

$$\forall n > R,\ \boldsymbol{\Pi}_{\mathcal{P}_1}\boldsymbol{v}_n^2 = 0 = \cos\theta_n(\mathcal{P}_1,\mathcal{P}_2)\boldsymbol{v}_n^1. \tag{4.186}$$

□

We give in the following, the consequences of Theorem 4.9. In particular, we prove Lemma 4.2 and Proposition 4.5 in the sequel .

We start by the following result that shows that the principal angles are somewhat independent of the choice of orthonormal bases. It can be found in Björck and Golub, 1973; Miao and Ben-Israel, 1992 for the finite dimensional case. We give here the proof for the general case.

Proposition 4.9. *Let $(\boldsymbol{w}_i^1)_{i\in[N]}$ be any orthonormal basis of $\mathcal{P}_1$ and $(\boldsymbol{w}_i^2)_{i\in[N]}$ be any orthonormal basis of $\mathcal{P}_2$, and let $\boldsymbol{W} = (\langle \boldsymbol{w}_i^1, \boldsymbol{w}_j^2\rangle_{\mathcal{H}})_{1\le i,j\le N}$ and $\boldsymbol{G} = \boldsymbol{W}\boldsymbol{W}^\intercal$. Then the eigenvalues of $\boldsymbol{G}$ are the $\cos^2\theta_i(\mathcal{P}_1,\mathcal{P}_2)$.*

Proof of Proposition 4.9

Let $(\boldsymbol{v}_i^i)_{i\in[N]}$, $i \in \{1,2\}$, be the bases of Theorem 4.9. Let $\boldsymbol{U}^1 \in \mathbb{O}_N(\mathbb{R})$ be such that

$$\forall i \in [N],\ \boldsymbol{w}_i^1 = \sum_{j\in[N]} u_{i,j}^1 \boldsymbol{v}_j^1. \tag{4.187}$$

Similarly, there exists a matrix $\boldsymbol{U}^2 \in \mathbb{O}_N(\mathbb{R})$ such that

$$\forall i \in [N],\ \boldsymbol{w}_i^2 = \sum_{j\in[N]} u_{i,j}^2 \boldsymbol{v}_j^2. \tag{4.188}$$

This implies that

$$\boldsymbol{W} = \boldsymbol{U}^1 \boldsymbol{V} \boldsymbol{U}^{2\intercal}, \tag{4.189}$$

where $\boldsymbol{V} = (\langle \boldsymbol{v}_i^1, \boldsymbol{v}_j^2 \rangle_{\mathcal{H}})_{1\leq i,j\leq N}$. Then

$$\boldsymbol{G} = \boldsymbol{W}\boldsymbol{W}^\intercal = \boldsymbol{U}^1 \boldsymbol{V}\boldsymbol{V}^\intercal \boldsymbol{U}^{1\intercal}. \tag{4.190}$$

Thus the eigenvalues of $\boldsymbol{G}$ are the eigenvalues of $\boldsymbol{V}\boldsymbol{V}^\intercal$. By Theorem 4.9, the diagonal elements of $\boldsymbol{V}$ are

$$v_{i,i} = \langle \boldsymbol{v}_i^1, \boldsymbol{v}_i^2 \rangle_{\mathcal{H}} = \cos\theta_i(\mathcal{P}_1, \mathcal{P}_2). \tag{4.191}$$

We finish the proof by showing that the anti-diagonal elements satisfy

$$v_{i,j} = \langle \boldsymbol{v}_i^1, \boldsymbol{v}_j^2 \rangle_{\mathcal{H}} = 0. \tag{4.192}$$

By (4.173),

$$\forall i \in [N],\ \sum_{j\in[N]} \langle \boldsymbol{v}_i^2, \boldsymbol{v}_j^1 \rangle_{\mathcal{H}}^2 = \|\boldsymbol{\Pi}_{\mathcal{P}_1} \boldsymbol{v}_i^2\|_{\mathcal{H}}^2 = \cos^2\theta_i(\mathcal{P}_1, \mathcal{P}_2). \tag{4.193}$$

Then

$$\sum_{i\in[N]} \sum_{j\in[N]} \langle \boldsymbol{v}_i^2, \boldsymbol{v}_j^1 \rangle_{\mathcal{H}}^2 = \sum_{i\in[N]} \cos^2\theta_i(\mathcal{P}_1, \mathcal{P}_2) = \sum_{i\in[N]} \langle \boldsymbol{v}_i^2, \boldsymbol{v}_i^1 \rangle_{\mathcal{H}}^2. \tag{4.194}$$

Thus

$$\sum_{\substack{i,j\in[N]\\ i\neq j}} \langle \boldsymbol{v}_i^2, \boldsymbol{v}_i^1 \rangle_{\mathcal{H}}^2 = 0. \tag{4.195}$$

Finally, $\boldsymbol{V}$ is a diagonal matrix and the eigenvalues of $\boldsymbol{G}$ are the $\cos^2\theta_i(\mathcal{P}_1, \mathcal{P}_2)$.

Proof of Lemma 4.2

Let $\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathcal{X}^N$ such that $\operatorname{Det} \boldsymbol{E}(\boldsymbol{x}) \neq 0$. By Proposition 4.4, $\boldsymbol{K}(\boldsymbol{x})$ is non singular. Thus $\dim \mathcal{T}(\boldsymbol{x}) = N$.

We have

$$\forall n \in [N],\ \|e_n^{\mathcal{F}} - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 = 1 - \|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2, \tag{4.196}$$

so that

$$\max_{n\in[N]} \|e_n^{\mathcal{F}} - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 = \max_{n\in[N]} \left(1 - \|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2\right) = 1 - \min_{n\in[N]} \|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2. \tag{4.197}$$

Since $\{e_1^{\mathcal{F}}, \dots, e_N^{\mathcal{F}}\} \subset \{\mu \in \mathcal{F}, \|\mu\|_{\mathcal{F}} = 1\}$, then

$$1 - \min_{n\in[N]} \|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 \leq 1 - \inf_{\substack{\mu\in\mathcal{E}_N^{\mathcal{F}} \\ \|\mu\|_{\mathcal{F}}=1}} \|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu\|_{\mathcal{F}}^2. \tag{4.198}$$

Therefore, by Theorem 4.9, we have

$$\begin{aligned}
\max_{n\in[N]} \|e_n^{\mathcal{F}} - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 &\leq 1 - \inf_{\substack{\mu\in\mathcal{E}_N^{\mathcal{F}} \\ \|\mu\|_{\mathcal{F}}=1}} \|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu\|_{\mathcal{F}}^2 && (4.199)\\
&\leq 1 - \cos^2\theta_N(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}}) && (4.200)\\
&\leq \frac{1}{\cos^2\theta_N(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}})} - 1, && (4.201)
\end{aligned}$$

because $\cos^2\theta_N(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}}) \in]0, 1]$ and

$$\forall x \in]0, 1], \ 1 - x \leq \frac{1}{x} - 1. \tag{4.202}$$

Moreover,

$$\forall n \in [N], \ \frac{1}{\cos^2\theta_n(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}})} \geq 1, \tag{4.203}$$

and

$$\forall n \in [N], \ \frac{1}{\cos^2\theta_n(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}})} - 1 \geq 0. \tag{4.204}$$

We conclude that

$$\max_{n\in[N]} \|e_n^{\mathcal{F}} - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 \leq \prod_{n\in[N]} \frac{1}{\cos^2\theta_n(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}})} - 1, \tag{4.205}$$

and

$$\begin{aligned}
\max_{n\in[N]} \|e_n^{\mathcal{F}} - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 &\leq \sum_{n\in[N]} \left(\frac{1}{\cos^2\theta_n(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}})} - 1\right) && (4.206)\\
&\leq \sum_{n\in[N]} \frac{1}{\cos^2\theta_n(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}})} - N. && (4.207)
\end{aligned}$$

Proof of Proposition 4.5

The condition $\mathrm{Det}^2 \boldsymbol{E}(\boldsymbol{x}) \neq 0$ yields by Proposition 4.4 that $\boldsymbol{K}(\boldsymbol{x})$ is non singular. Thus $\dim \mathcal{T}(\boldsymbol{x}) = N$. Let $(t_i)_{i\in[N]}$ be an orthonormal basis of $\mathcal{T}(\boldsymbol{x})$ with respect to $\langle ., . \rangle_{\mathcal{F}}$. Define

$$\boldsymbol{W} = (\langle e_n^{\mathcal{F}}, t_i \rangle_{\mathcal{F}})_{(n,i)\in[N]\times[N]}. \tag{4.208}$$

Using Proposition 4.9, and the fact that $(e_n^{\mathcal{F}})_{n\in[N]}$ is an orthonormal basis of $\mathcal{E}_N^{\mathcal{F}}$, the eigenvalues of the matrix $\boldsymbol{W}\boldsymbol{W}^\intercal$ are the $\cos^2\theta_\ell\left(\mathcal{E}_N^{\mathcal{F}}, \mathcal{T}(\boldsymbol{x})\right)$. Therefore

$$\prod_{\ell\in[N]} \cos^2\theta_\ell\left(\mathcal{E}_N^{\mathcal{F}}, \mathcal{T}(\boldsymbol{x})\right) = \mathrm{Det}(\boldsymbol{W}\boldsymbol{W}^\intercal), \tag{4.209}$$

and if $\boldsymbol{W}\boldsymbol{W}^\intercal$ is non singular then

$$\sum_{\ell\in[N]} \frac{1}{\cos^2\theta_\ell\left(\mathcal{E}_N^{\mathcal{F}}, \mathcal{T}(\boldsymbol{x})\right)} = \mathrm{Tr}(\boldsymbol{W}\boldsymbol{W}^\intercal)^{-1}. \tag{4.210}$$

Now, write for $i \in [N]$,

$$t_i = \sum_{j\in[N]} c_{i,j} k(x_j, .). \tag{4.211}$$

Thus

$$\langle e_n^{\mathcal{F}}, t_i\rangle_{\mathcal{F}} = \sum_{j\in[N]} c_{i,j}\langle e_n^{\mathcal{F}}, k(x_j, .)\rangle_{\mathcal{F}} \tag{4.212}$$

$$= \sum_{j\in[N]} c_{i,j} e_n^{\mathcal{F}}(x_j). \tag{4.213}$$

Then

$$\boldsymbol{W} = \boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})\boldsymbol{C}(\boldsymbol{x})^\intercal, \tag{4.214}$$

where

$$\boldsymbol{C}(\boldsymbol{x}) = (c_{i,j})_{1\leq i,j\leq N}. \tag{4.215}$$

In particular

$$\boldsymbol{W}\boldsymbol{W}^\intercal = \boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})\boldsymbol{C}(\boldsymbol{x})^\intercal\boldsymbol{C}(\boldsymbol{x})\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^\intercal, \tag{4.216}$$

Now, $(t_i)_{i\in[N]}$ is an orthonormal basis of $\mathcal{T}(\boldsymbol{x})$, then by (4.211)

$$\delta_{i,i'} = \langle t_i, t_{i'}\rangle_{\mathcal{F}} = \sum_{j\in[N]}\sum_{j'\in[N]} c_{i,j}c_{i',j'}k(x_j, x_{j'}). \tag{4.217}$$

Therefore

$$\boldsymbol{C}(\boldsymbol{x})\boldsymbol{K}(\boldsymbol{x})\boldsymbol{C}(\boldsymbol{x})^\intercal = \mathbb{I}_N. \tag{4.218}$$

Thus

$$\boldsymbol{W}\boldsymbol{W}^\intercal = \boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})\boldsymbol{K}(\boldsymbol{x})^{-1}\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^\intercal. \tag{4.219}$$

In particular $\boldsymbol{W}\boldsymbol{W}^T$ is non singular. Moreover, we have

$$\prod_{\ell\in[N]} \frac{1}{\cos^2\theta_\ell\left(\mathcal{E}_N^{\mathcal{F}}, \mathcal{T}(\boldsymbol{x})\right)} = \frac{1}{\mathrm{Det}(\boldsymbol{W}\boldsymbol{W}^T)} = \frac{\mathrm{Det}\,\boldsymbol{K}(\boldsymbol{x})}{\mathrm{Det}^2\,\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})}, \tag{4.220}$$

and

$$\sum_{\ell\in[N]} \frac{1}{\cos^2\theta_\ell\left(\mathcal{E}_N^{\mathcal{F}}, \mathcal{T}(\boldsymbol{x})\right)} = \mathrm{Tr}\left(\boldsymbol{W}\boldsymbol{W}^\intercal\right)^{-1} = \mathrm{Tr}\left(\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal^{-1}}\boldsymbol{K}(\boldsymbol{x})\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{-1}\right). \tag{4.221}$$

4.6.3 *Proof of Proposition 4.4*

Proof. Recall the Mercer decomposition of k:

$$k(x, y) = \sum_{m\in\mathbb{N}^*} \sigma_m e_m(x)e_m(y), \tag{4.222}$$

where the convergence is point-wise on $\mathcal{X} \times \mathcal{X}$. Define for $M \in \mathbb{N}^*$, $M \geq N$ the M-th truncated kernel

$$k_M(x,y) = \sum_{m \in [M]} \sigma_m e_m(x) e_m(y). \tag{4.223}$$

By (4.222)

$$\forall x, y \in \mathcal{X}, \lim_{M \to \infty} k_M(x,y) = k(x,y). \tag{4.224}$$

Let $\boldsymbol{x} = (x_1, \dots, x_N) \in \mathcal{X}^N$ such that $\operatorname{Det} \boldsymbol{E}(\boldsymbol{x}) \neq 0$, and define

$$\boldsymbol{K}_M(\boldsymbol{x}) = (k_M(x_i, x_j))_{i,j \in [N]}. \tag{4.225}$$

By the continuity of the function $\boldsymbol{M} \in \mathbb{R}^{N \times N} \mapsto \operatorname{Det} \boldsymbol{M}$ and by (4.224)

$$\lim_{M \to \infty} \operatorname{Det} \boldsymbol{K}_M(\boldsymbol{x}) = \operatorname{Det} \boldsymbol{K}(\boldsymbol{x}). \tag{4.226}$$

Thus to prove that $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) > 0$, it is enough to prove that the $\operatorname{Det} \boldsymbol{K}_M(\boldsymbol{x})$ is larger than a positive real number for M large enough. We write

$$\boldsymbol{K}_M(\boldsymbol{x}) = \boldsymbol{F}_M(\boldsymbol{x})^\intercal \boldsymbol{\Sigma}_M \boldsymbol{F}_M(\boldsymbol{x}), \tag{4.227}$$

with $\boldsymbol{F}_M(\boldsymbol{x}) = (e_i(x_j))_{(i,j) \in [M] \times [N]}$ and $\boldsymbol{\Sigma}_M$ is a diagonal matrix containing the first M eigenvalues (σ_m). The Cauchy-Binet identity yields

$$\begin{aligned} \operatorname{Det} \boldsymbol{K}_M(\boldsymbol{x}) &= \sum_{T \subset [M], |T| = N} \prod_{i \in T} \sigma_i \operatorname{Det}^2 (e_i(x_j))_{(i,j) \in T \times [N]} && (4.228) \\ &\geq \prod_{i \in [N]} \sigma_i \operatorname{Det}^2 \boldsymbol{E}(\boldsymbol{x}) > 0. && (4.229) \end{aligned}$$

Therefore,

$$\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) = \lim_{M \to \infty} \operatorname{Det} \boldsymbol{K}_M(\boldsymbol{x}) \geq \prod_{i \in [N]} \sigma_i \operatorname{Det}^2 \boldsymbol{E}(\boldsymbol{x}) > 0. \tag{4.230}$$

so that $\boldsymbol{K}(\boldsymbol{x})$ is invertible. □

4.6.4 *Proof of Lemma 4.1*

Proof. First, we prove that

$$\|\boldsymbol{\Sigma}^{-1/2} \mu_g\|_{\mathcal{F}}^2 = \|g\|_{\mathrm{d}\omega}^2 \leq 1. \tag{4.231}$$

Recall that

$$\mu_g = \int_{\mathcal{X}} g(y) k(., y) \mathrm{d}\omega(y) = \boldsymbol{\Sigma} g. \tag{4.232}$$

Then $\boldsymbol{\Sigma}^{-1/2} \mu_g = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma} g = \boldsymbol{\Sigma}^{1/2} g \in \mathcal{F}$, so that $\|\boldsymbol{\Sigma}^{-1/2} \mu_g\|_{\mathcal{F}}^2 = \|g\|_{\mathrm{d}\omega}^2 \leq 1$. Therefore, there exists $\tilde{\mu}_g \in \mathcal{F}$ such that $\|\tilde{\mu}_g\|_{\mathcal{F}} \leq 1$ and $\mu_g = \boldsymbol{\Sigma}^{1/2} \tilde{\mu}_g$. Moreover, since $(e_n^{\mathcal{F}})_{n \in \mathbb{N}^*}$ is an o.n.b. of $\mathcal{F}$, we have

$$\mathcal{F} = \mathcal{E}_N^{\mathcal{F}} \oplus \operatorname{Span}(e_n^{\mathcal{F}})_{n \geq N+1}, \tag{4.233}$$

and we write

$$\tilde{\mu}_g = \tilde{\mu}_{g,N} + \tilde{\mu}_{g,N^\perp}, \tag{4.234}$$

with $(\tilde{\mu}_{g,N}, \tilde{\mu}_{g,N^\perp}) \in \mathcal{E}_N^{\mathcal{F}} \times \mathrm{Span}(e_n^{\mathcal{F}})_{n \geq N+1}$. Observe that

$$\|\boldsymbol{\Sigma}^{1/2}\tilde{\mu}_{g,N^\perp}\|_{\mathcal{F}}^2 = \sum_{n \geq N+1} \sigma_n \langle \tilde{\mu}_g, e_n^{\mathcal{F}} \rangle_{\mathcal{F}}^2 \leq \sigma_{N+1} \|\tilde{\mu}_g\|_{\mathcal{F}}^2 \leq \sigma_{N+1}. \tag{4.235}$$

Now, the approximation error writes

$$\begin{aligned}
\|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp}\mu_g\|_{\mathcal{F}}^2 &= \|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp}\boldsymbol{\Sigma}^{1/2}\tilde{\mu}_g\|_{\mathcal{F}}^2 && (4.236)\\
&= \|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp}\boldsymbol{\Sigma}^{1/2}(\tilde{\mu}_{g,N} + \tilde{\mu}_{g,N^\perp})\|_{\mathcal{F}}^2 \\
&= \|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp}\boldsymbol{\Sigma}^{1/2}\tilde{\mu}_{g,N}\|_{\mathcal{F}}^2 + \|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp}\boldsymbol{\Sigma}^{1/2}\tilde{\mu}_{g,N^\perp}\|_{\mathcal{F}}^2 && (4.237)\\
&\qquad + 2\langle \boldsymbol{\Pi}_{\mathcal{T}(x)^\perp}\boldsymbol{\Sigma}^{1/2}\tilde{\mu}_{g,N}, \boldsymbol{\Pi}_{\mathcal{T}(x)^\perp}\boldsymbol{\Sigma}^{1/2}\tilde{\mu}_{g,N^\perp}\rangle_{\mathcal{F}} \\
&\leq 2\left(\|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp}\boldsymbol{\Sigma}^{1/2}\tilde{\mu}_{g,N}\|_{\mathcal{F}}^2 + \|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp}\boldsymbol{\Sigma}^{1/2}\tilde{\mu}_{g,N^\perp}\|_{\mathcal{F}}^2\right).
\end{aligned}$$

The operator $\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp}$ is an orthogonal projection. Then by (4.235)

$$\|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp}\boldsymbol{\Sigma}^{1/2}\tilde{\mu}_{g,N^\perp}\|_{\mathcal{F}}^2 \leq \|\boldsymbol{\Sigma}^{1/2}\tilde{\mu}_{g,N^\perp}\|_{\mathcal{F}}^2 \leq \sigma_{N+1}. \tag{4.238}$$

Moreover

$$\forall n \in [N],\ \boldsymbol{\Sigma}^{1/2} e_n^{\mathcal{F}} = \sqrt{\sigma}_n e_n^{\mathcal{F}}. \tag{4.239}$$

Thus

$$\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp}\boldsymbol{\Sigma}^{1/2}\tilde{\mu}_{g,N} = \boldsymbol{\Pi}_{\mathcal{T}(x)^\perp} \sum_{n \in [N]} \sqrt{\sigma_n}\langle \tilde{\mu}_g, e_n^{\mathcal{F}} \rangle_{\mathcal{F}} e_n^{\mathcal{F}} = \sum_{n \in [N]} \langle \tilde{\mu}_g, e_n^{\mathcal{F}} \rangle_{\mathcal{F}} \sqrt{\sigma_n} \boldsymbol{\Pi}_{\mathcal{T}(x)^\perp} e_n^{\mathcal{F}}. \tag{4.240}$$

Then

$$\begin{aligned}
\|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp}\boldsymbol{\Sigma}^{1/2}\tilde{\mu}_{g,N}\|_{\mathcal{F}}^2 &= \| \sum_{n \in [N]} \langle \tilde{\mu}_g, e_n^{\mathcal{F}} \rangle_{\mathcal{F}} \sqrt{\sigma_n} \boldsymbol{\Pi}_{\mathcal{T}(x)^\perp} e_n^{\mathcal{F}} \|_{\mathcal{F}}^2 && (4.241)\\
&= \sum_{n \in [N]} \sum_{m \in [N]} \langle \tilde{\mu}_g, e_n^{\mathcal{F}} \rangle_{\mathcal{F}} \langle \tilde{\mu}_g, e_m^{\mathcal{F}} \rangle_{\mathcal{F}} \sqrt{\sigma_n}\sqrt{\sigma_m} \langle \boldsymbol{\Pi}_{\mathcal{T}(x)^\perp} e_n^{\mathcal{F}}, \boldsymbol{\Pi}_{\mathcal{T}(x)^\perp} e_m^{\mathcal{F}} \rangle_{\mathcal{F}} \\
&\leq \sum_{n \in [N]} \sum_{m \in [N]} \langle \tilde{\mu}_g, e_n^{\mathcal{F}} \rangle_{\mathcal{F}} \langle \tilde{\mu}_g, e_m^{\mathcal{F}} \rangle_{\mathcal{F}} \sqrt{\sigma_n}\sqrt{\sigma_m} \|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp} e_n^{\mathcal{F}}\|_{\mathcal{F}} \|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp} e_m^{\mathcal{F}}\|_{\mathcal{F}} \\
&\leq \left(\sum_{n \in [N]} \sum_{m \in [N]} |\langle \tilde{\mu}_g, e_n^{\mathcal{F}} \rangle_{\mathcal{F}}| \cdot |\langle \tilde{\mu}_g, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}| \right) \max_{n \in [N]} \sigma_n \|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 \\
&\leq \left(\sum_{n \in [N]} |\langle \tilde{\mu}_g, e_n^{\mathcal{F}} \rangle_{\mathcal{F}}| \right)^2 \max_{n \in [N]} \sigma_n \|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2. && (4.242)
\end{aligned}$$

Remarking that $\|g\|_{d\omega,1} = \sum_{n \in [N]} |\langle \tilde{\mu}_g, e_n^{\mathcal{F}} \rangle_{\mathcal{F}}|$ concludes the proof of (4.148) and therefore Lemma 4.1. □

4.6.5 *Proof of Proposition 4.8*

Proposition 4.8 gives an upper bound to the term $\max_{n \in [N]} \sigma_n \|\boldsymbol{\Pi}_{\mathcal{T}(x)^\perp} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2$ that appears in Lemma 4.1. We first prove a technical result, Lemma 4.4, and then combine it with Proposition 4.10 to finish the proof. We conclude with the proof of Proposition 4.10.

A preliminary lemma

Let $\boldsymbol{x} = (x_1, \dots, x_N) \in \mathcal{X}^N$. Recall that $\boldsymbol{K}(\boldsymbol{x}) = (k(x_i, x_j))_{1 \leq i,j \leq N}$ and denote similarly $\tilde{\boldsymbol{K}}(\boldsymbol{x}) = (\tilde{k}(x_i, x_j))_{1 \leq i,j \leq N}$. In the following, we define

$$\Delta_n^{\mathcal{F}}(\boldsymbol{x}) = e_n^{\mathcal{F}}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} e_n^{\mathcal{F}}(\boldsymbol{x}) \tag{4.243}$$

$$\Delta_n^{\tilde{\mathcal{F}}}(\boldsymbol{x}) = e_n^{\tilde{\mathcal{F}}}(\boldsymbol{x})^\intercal \tilde{\boldsymbol{K}}(\boldsymbol{x})^{-1} e_n^{\tilde{\mathcal{F}}}(\boldsymbol{x}) \tag{4.244}$$

Lemma 4.4 below shows that each term of the form $\Delta_n^{\mathcal{F}}(\boldsymbol{x})$ measures the squared norm of the projection of $e_n^{\mathcal{F}}$ on $\mathcal{T}(\boldsymbol{x})$. The same holds for $\Delta_n^{\tilde{\mathcal{F}}}(\boldsymbol{x})$ and the projection of $e_n^{\tilde{\mathcal{F}}}$ onto $\tilde{\mathcal{T}}(\boldsymbol{x})$.

Indeed, $\|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})^\perp} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 = 1 - \|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2$ since $\|e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 = 1$. Thus it is sufficient to prove that $\|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 = \Delta_n^{\mathcal{F}}(\boldsymbol{x})$.

Lemma 4.4. *For $n \in \mathbb{N}^*$, let $e_n^{\mathcal{F}}(\boldsymbol{x}), e_n^{\tilde{\mathcal{F}}}(\boldsymbol{x}) \in \mathbb{R}^N$ the vectors of the evaluations of $e_n^{\mathcal{F}}$ and $e_n^{\tilde{\mathcal{F}}}$ on the elements of $\boldsymbol{x}$ respectively. Then*

$$\|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})^\perp} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 = 1 - \Delta_n^{\mathcal{F}}(\boldsymbol{x}), \tag{4.245}$$

$$\|\boldsymbol{\Pi}_{\tilde{\mathcal{T}}_N(\boldsymbol{x})^\perp} e_n^{\tilde{\mathcal{F}}}\|_{\tilde{\mathcal{F}}}^2 = 1 - \Delta_n^{\tilde{\mathcal{F}}}(\boldsymbol{x}). \tag{4.246}$$

We give the proof of (4.245); the proof of (4.246) follows the same lines.

Proof. Let us write

$$\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}} = \sum_{i \in [N]} c_i k(x_i, .), \tag{4.247}$$

where the c_i are the elements of the vector $\boldsymbol{c} = \boldsymbol{K}(\boldsymbol{x})^{-1} e_n^{\mathcal{F}}(\boldsymbol{x})$. Then

$$\begin{aligned} \|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 &= \| \sum_{i \in [N]} c_i k(x_i, .)\|_{\mathcal{F}}^2 \\ &= \boldsymbol{c}^\intercal \boldsymbol{K}(\boldsymbol{x}) \boldsymbol{c} \\ &= e_n^{\mathcal{F}}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} \boldsymbol{K}(\boldsymbol{x}) \boldsymbol{K}(\boldsymbol{x})^{-1} e_n^{\mathcal{F}}(\boldsymbol{x}) \\ &= e_n^{\mathcal{F}}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} e_n^{\mathcal{F}}(\boldsymbol{x}) \\ &= \Delta_n^{\mathcal{F}}(\boldsymbol{x}). \end{aligned} \tag{4.248}$$

□

End of the proof of Proposition 4.8

Proof. By Lemma 4.4, the inequality (4.159) in Proposition 4.8 is equivalent to

$$\forall n \in [N],\ \sigma_n \left(1 - \Delta_n^{\mathcal{F}}(\boldsymbol{x})\right) \leq \sigma_1 \left(1 - \Delta_n^{\tilde{\mathcal{F}}}(\boldsymbol{x})\right). \tag{4.249}$$

As an intermediate remark, note that in the special case $n = 1$, by construction $\tilde{k} - k$ is positive definite kernel on $\mathcal{X}$, therefore

$$\boldsymbol{K}(\boldsymbol{x}) \prec \tilde{\boldsymbol{K}}(\boldsymbol{x}), \tag{4.250}$$

where $\prec$ is the Loewner order, the partial order defined by the convex cone of positive semi-definite matrices. Thus

$$\tilde{\boldsymbol{K}}(\boldsymbol{x})^{-1} \prec \boldsymbol{K}(\boldsymbol{x})^{-1}. \tag{4.251}$$

Noting that $\tilde{\sigma}_1 = \sigma_1$ and that

$$e_1^{\mathcal{F}} = \sqrt{\sigma_1} e_1 = \sqrt{\tilde{\sigma}_1} e_1 = e_1^{\tilde{\mathcal{F}}}. \tag{4.252}$$

yields (4.249) for $n = 1$:

$$1 - e_1^{\mathcal{F}}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} e_1^{\mathcal{F}}(\boldsymbol{x}) \leq 1 - e_1^{\tilde{\mathcal{F}}}(\boldsymbol{x})^\intercal \tilde{\boldsymbol{K}}(\boldsymbol{x})^{-1} e_1^{\tilde{\mathcal{F}}}(\boldsymbol{x}). \tag{4.253}$$

For $n \neq 1$, the proof is much more subtle. Indeed, a naive application of the inequality (4.251) would lead to the following inequality

$$1 - e_n^{\tilde{\mathcal{F}}}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} e_n^{\tilde{\mathcal{F}}}(\boldsymbol{x}) \leq 1 - e_n^{\tilde{\mathcal{F}}}(\boldsymbol{x})^\intercal \tilde{\boldsymbol{K}}(\boldsymbol{x})^{-1} e_n^{\tilde{\mathcal{F}}}(\boldsymbol{x}). \tag{4.254}$$

Since $\forall n \in \mathbb{N}$, $e_n^{\tilde{\mathcal{F}}} = \sqrt{\sigma_1 / \sigma_n} e_n^{\mathcal{F}}$, we get

$$1 - \sigma_1 e_n^{\mathcal{F}}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} e_n^{\mathcal{F}}(\boldsymbol{x}) \leq 1 - \sigma_n e_n^{\tilde{\mathcal{F}}}(\boldsymbol{x})^\intercal \tilde{\boldsymbol{K}}(\boldsymbol{x})^{-1} e_n^{\tilde{\mathcal{F}}}(\boldsymbol{x}), \tag{4.255}$$

and hence the unsatisfactory inequality

$$1 - \sigma_1 \Delta_n^{\mathcal{F}}(\boldsymbol{x}) \leq 1 - \sigma_n \Delta_n^{\tilde{\mathcal{F}}}(\boldsymbol{x}) \tag{4.256}$$

We can prove a better inequality by applying a sequence of rank-one updates to the kernel k to build N intermediate kernels $k^{(\ell)}$ that lead to N inequalities sharp enough to prove (4.249) for $n \neq 1$. Then inequality (4.249) will result as a corollary of Proposition 4.10 below. To this aim, we define N RKHS $\tilde{\mathcal{F}}_\ell$, $1 \leq \ell \leq N$, that interpolate between $\mathcal{F}$ and $\tilde{\mathcal{F}}$. For $\ell \in [N]$, define the kernel $\tilde{k}^{(\ell)}$ by

$$\tilde{k}^{(\ell)}(x, y) = \sum_{m \in [\ell]} \sigma_1 e_m(x) e_m(y) + \sum_{m \geq \ell + 1} \sigma_m e_m(x) e_m(y), \tag{4.257}$$

and let $\tilde{\mathcal{F}}_\ell$ the RKHS corresponding to the kernel $\tilde{k}^{(\ell)}$. For $\boldsymbol{x} \in \mathcal{X}^N$, define $\tilde{\boldsymbol{K}}^{(\ell)}(\boldsymbol{x}) = (\tilde{k}^{(\ell)}(x_i, x_j))_{1 \leq i,j \leq N}$. Similar to previous notations, we define as well

$$\Delta_n^{\tilde{\mathcal{F}}_\ell}(\boldsymbol{x}) = e_n^{\tilde{\mathcal{F}}_\ell}(\boldsymbol{x})^\intercal \tilde{\boldsymbol{K}}^{(\ell)}(\boldsymbol{x})^{-1} e_n^{\tilde{\mathcal{F}}_\ell}(\boldsymbol{x}). \tag{4.258}$$

Now we have the following useful proposition.

Proposition 4.10. *For $n \in [N] \setminus \{1\}$, we have*

$$\sigma_n \left(1 - \Delta_n^{\tilde{\mathcal{F}}_{n-1}}(\boldsymbol{x})\right) \leq \sigma_1 \left(1 - \Delta_n^{\tilde{\mathcal{F}}_n}(\boldsymbol{x})\right), \tag{4.259}$$

and

$$\forall\, \ell \in [N] \setminus \{1, n\},\ 1 - \Delta_n^{\tilde{\mathcal{F}}_{\ell-1}}(\boldsymbol{x}) \leq 1 - \Delta_n^{\tilde{\mathcal{F}}_\ell}(\boldsymbol{x}). \tag{4.260}$$

For ease of reading, we first show that inequality (4.249) and therefore Proposition 4.8 is easily deduced from this Proposition 4.10 and then give its proof.

Let $n \in [N]$ such that $n \neq 1$. We first remark that $\mathcal{F} = \tilde{\mathcal{F}}_1$ and use $(n-2)$ times inequality (4.260) of Proposition 4.10:

$$\begin{aligned} \sigma_n \left(1 - \Delta_n^{\mathcal{F}}(\boldsymbol{x})\right) &= \sigma_n \left(1 - \Delta_n^{\tilde{\mathcal{F}}_1}(\boldsymbol{x})\right) \\ &\leq \sigma_n \left(1 - \Delta_n^{\tilde{\mathcal{F}}_{n-1}}(\boldsymbol{x})\right). \end{aligned} \tag{4.261}$$

Then we use (4.259) that is connected to the rank-one update from the kernel $k^{(n-1)}$ to $k^{(n)}$ so that

$$\sigma_n\left(1-\Delta_n^{\tilde{\mathcal{F}}_{n-1}}(\boldsymbol{x})\right) \leq \sigma_1\left(1-\Delta_n^{\tilde{\mathcal{F}}_n}(\boldsymbol{x})\right). \tag{4.262}$$

Then we apply (4.260) to the r.h.s. again $N-n-1$ times to finally get:

$$\begin{aligned}\sigma_n\left(1-\Delta_n^{\mathcal{F}}(\boldsymbol{x})\right) &\leq \sigma_1\left(1-\Delta_n^{\tilde{\mathcal{F}}_N}(\boldsymbol{x})\right) \\ &\leq \sigma_1\left(1-\Delta_n^{\tilde{\mathcal{F}}}(\boldsymbol{x})\right),\end{aligned} \tag{4.263}$$

since $\tilde{k}^{(N)}=\tilde{k}$ and $\tilde{\mathcal{F}}_N=\tilde{\mathcal{F}}$. This concludes the proof of the desired inequality (4.249) and therefore of Proposition 4.8. □

Proof of Proposition 4.10

Proof. (Proposition 4.10) Let $n \in [N] \setminus \{1\}$, and $M \in \mathbb{N}$ such that $M \geq N$. Let $\boldsymbol{A}_\ell \in \mathbb{R}^{N\times M}$ defined by

$$\forall (i,m) \in [N]\times[M],\ (\boldsymbol{A}_\ell)_{i,m} = e_m^{\tilde{\mathcal{F}}_\ell}(x_i).^{12} \tag{4.264}$$

For $\ell \in [N]$ define

$$\tilde{\boldsymbol{K}}_M^{(\ell)}(\boldsymbol{x}) = \boldsymbol{A}_\ell^\intercal \boldsymbol{A}_\ell. \tag{4.265}$$

Let $\boldsymbol{W}_\ell \in \mathbb{R}^{M\times M}$ the diagonal matrix defined by

$$\boldsymbol{W}_\ell = \mathrm{diag}(\underbrace{1,\ldots,1}_{\ell-1}, \sqrt{\frac{\sigma_1}{\sigma_\ell}}, 1\ldots,1) \tag{4.266}$$

Then one has the simple relation

$$\boldsymbol{A}_{\ell+1} = \boldsymbol{A}_\ell \boldsymbol{W}_\ell, \tag{4.267}$$

which prepares the use of Lemma 4.3 in Section 4.6.1. By definition of the n-th leverage score of the matrix $\boldsymbol{A}$, see (4.162) in Section 4.6.1,

$$e_n^{\tilde{\mathcal{F}}_\ell}(\boldsymbol{x})^\intercal \tilde{\boldsymbol{K}}_M^{(\ell)}(\boldsymbol{x})^{-1} e_n^{\tilde{\mathcal{F}}_\ell}(\boldsymbol{x}) = e_n^{\tilde{\mathcal{F}}_\ell}(\boldsymbol{x})^\intercal \left(\boldsymbol{A}_\ell^\intercal \boldsymbol{A}_\ell\right)^{-1} e_n^{\tilde{\mathcal{F}}_\ell}(\boldsymbol{x}) = \tau_n\left(\boldsymbol{A}_\ell\right). \tag{4.268}$$

Define similarly $\Delta_{n,M}^{\tilde{\mathcal{F}}_\ell}(\boldsymbol{x}) = e_n^{\tilde{\mathcal{F}}_\ell}(\boldsymbol{x})^\intercal \tilde{\boldsymbol{K}}_M^{(\ell)}(\boldsymbol{x})^{-1} e_n^{\tilde{\mathcal{F}}_\ell}(\boldsymbol{x})$. Thanks to (4.165) of Lemma 4.3 and (4.267) and for $\ell = n$

$$\tau_n\Big(\boldsymbol{A}_n\Big) = \tau_n\left(\boldsymbol{A}_{n-1}\boldsymbol{W}_n\right) = \frac{(1+\rho_n)\tau_n\Big(\boldsymbol{A}_{n-1}\Big)}{1+\rho_n\tau_n\Big(\boldsymbol{A}_{n-1}\Big)}, \tag{4.269}$$

where $\rho_n = \dfrac{\sigma_1}{\sigma_n} - 1$. Thus

$$1-\tau_n\Big(\boldsymbol{A}_n\Big) = 1 - \frac{(1+\rho_n)\tau_n\Big(\boldsymbol{A}_{n-1}\Big)}{1+\rho_n\tau_n\Big(\boldsymbol{A}_{n-1}\Big)} = \frac{1-\tau_n\Big(\boldsymbol{A}_{n-1}\Big)}{1+\rho_n\tau_n\Big(\boldsymbol{A}_{n-1}\Big)}. \tag{4.270}$$

[12] The matrix $\boldsymbol{A}_\ell$ depends on $\boldsymbol{x}$.

Then

$$\begin{aligned}\sigma_1\left(1-\tau_n\left(\boldsymbol{A}_n\right)\right) &= \sigma_1\frac{1-\tau_n\left(\boldsymbol{A}_{n-1}\right)}{1+\rho_n\tau_n\left(\boldsymbol{A}_{n-1}\right)} \\ &= \sigma_n(1+\rho_n)\frac{1-\tau_n\left(\boldsymbol{A}_{n-1}\right)}{1+\rho_n\tau_n\left(\boldsymbol{A}_{n-1}\right)} \\ &= \frac{1+\rho_n}{1+\rho_n\tau_n\left(\boldsymbol{A}_{n-1}\right)}\sigma_n\left(1-\tau_n\left(\boldsymbol{A}_{n-1}\right)\right) \\ &\geq \sigma_n\left(1-\tau_n\left(\boldsymbol{A}_{n-1}\right)\right),\end{aligned} \tag{4.271}$$

since $\rho_n \geq 0$ and $\tau_n\left(\boldsymbol{A}_{n-1}\right) \in [0,1]$ thanks to (4.164). This proves that for $M \in \mathbb{N}^*$ such that $M \geq N$,

$$\sigma_n\left(1-\Delta_{n,M}^{\tilde{\mathcal{F}}_{n-1}}(\boldsymbol{x})\right) \leq \sigma_1\left(1-\Delta_{n,M}^{\tilde{\mathcal{F}}_n}(\boldsymbol{x})\right). \tag{4.272}$$

Now,

$$\lim_{M\to\infty}\tilde{\boldsymbol{K}}_M^{(n+1)}(\boldsymbol{x}) = \tilde{\boldsymbol{K}}^{(n+1)}(\boldsymbol{x}), \tag{4.273}$$

$$\lim_{M\to\infty}\tilde{\boldsymbol{K}}_M^{(n)}(\boldsymbol{x}) = \tilde{\boldsymbol{K}}^{(n)}(\boldsymbol{x}). \tag{4.274}$$

Moreover the application $\boldsymbol{X} \mapsto \boldsymbol{X}^{-1}$ is continuous in $GL_N(\mathbb{R})$. This proves the inequality (4.259) of Proposition 4.10. To prove the inequality (4.260), we start by using (4.166):

$$\forall \ell \in [N] \setminus \{1,n\},\ \tau_n\left(\boldsymbol{A}(\ell)\right) = \tau_n\left(\boldsymbol{A}_{\ell-1}\boldsymbol{W}_\ell\right) \leq \tau_n\left(\boldsymbol{A}_{\ell-1}\right). \tag{4.275}$$

which implies that

$$\forall \ell \in [N] \setminus \{1,n\}, 1-\tau_n\left(\boldsymbol{A}_{\ell-1}\right) \leq 1-\tau_n\left(\boldsymbol{A}_\ell\right). \tag{4.276}$$

Then for $M \geq N$,

$$\forall\, \ell \in\ [N] \setminus \{1,n\},\ 1-\Delta_{n,M}^{\tilde{\mathcal{F}}_{\ell-1}}(\boldsymbol{x}) \leq 1-\Delta_{n,M}^{\tilde{\mathcal{F}}_\ell}(\boldsymbol{x}). \tag{4.277}$$

As above, we conclude the proof by considering the limit $M \to \infty$

$$\forall\, \ell \in\ [N] \setminus \{1,n\},\ \lim_{M\to\infty}\tilde{\boldsymbol{K}}_M^{(\ell)}(\boldsymbol{x}) = \tilde{\boldsymbol{K}}^{(\ell)}(\boldsymbol{x}). \tag{4.278}$$

This proves inequality (4.260) and concludes the proof of Proposition 4.10. □

4.6.6 *Proof of Proposition 4.6*

In this section, $\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathcal{X}^N$ is the realization of the DPP of Theorem 4.7. Let $\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x}) = (e_i^{\mathcal{F}}(x_j))_{1\leq i,j\leq N}$ and $\boldsymbol{E}(\boldsymbol{x}) = (e_i(x_j))_{1\leq i,j\leq N}$, and $\boldsymbol{K}(\boldsymbol{x}) = (k(x_i,x_j))_{1\leq i,j\leq N}$. Moreover, let $\mathcal{E}_N^{\mathcal{F}} = \mathrm{Span}(e_m^{\mathcal{F}})_{m\in[N]}$ and $\mathcal{T}(\boldsymbol{x}) = \mathrm{Span}\,(k(x_i,.))_{i\in[N]}$.

We first prove the following proposition.

Proposition 4.11.

$$\frac{1}{N!}\int_{\mathcal{X}^N} \operatorname{Det} \boldsymbol{K}(x_1,\dots,x_N) \otimes_{j\in[N]} \mathrm{d}\omega(x_j) = \sum_{\substack{T\subset\mathbb{N}^* \\ |T|=N}} \prod_{t\in T} \sigma_t. \tag{4.279}$$

Proof. Let $\boldsymbol{x} = (x_1,\dots,x_N) \in \mathcal{X}^N$. From (4.224)

$$\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) = \lim_{M\to\infty} \operatorname{Det} \boldsymbol{K}_M(\boldsymbol{x}). \tag{4.280}$$

Moreover,

$$\operatorname{Det} \boldsymbol{K}_M(\boldsymbol{x}) = \sum_{T\subset[M],|T|=N} \prod_{i\in T} \sigma_i \operatorname{Det}^2(e_i(x_j))_{(i,j)\in T\times[N]}. \tag{4.281}$$

Now, for $T \subset [M]$ such that $|T| = N$, $(e_t)_{t\in T}$ is an orthonormal family of $\mathbb{L}_2(\mathrm{d}\omega)$, then by Hough et al., 2006 Lemma 21:

$$\int_{\mathcal{X}^N} \operatorname{Det}^2(e_t(x_j)) \otimes_{j\in[N]} \mathrm{d}\omega(x_j) = N!. \tag{4.282}$$

Thus

$$\begin{aligned}\frac{1}{N!}\int_{\mathcal{X}^N} \operatorname{Det} \boldsymbol{K}_M(\boldsymbol{x}) \otimes_{j\in[N]} \mathrm{d}\omega(x_j) &= \frac{1}{N!} \sum_{T\subset[M],|T|=N} \prod_{t\in T} \sigma_t \int_{\mathcal{X}^N} \operatorname{Det}^2(e_t(x_j)) \otimes_{j\in[N]} \mathrm{d}\omega(x_j) \\ &= \sum_{T\subset[M],|T|=N} \prod_{t\in T} \sigma_t.\end{aligned} \tag{4.283}$$

Now, $\sum_{n\in\mathbb{N}^*} \sigma_n < \infty$ implies that $\sum_{T\subset\mathbb{N}^*,|T|=N} \prod_{t\in T} \sigma_t < \infty$. In fact, for $\ell \in [N]$ let p_ℓ the ℓ-th symmetric polynomial. By Maclaurin's inequality (Steele, 2004), and for any vector $\boldsymbol{\nu} \in \mathbb{R}_+^M$

$$\left(\frac{p_\ell(\boldsymbol{\nu})}{\binom{M}{\ell}}\right)^{\frac{1}{\ell}} \leq \frac{p_1(\boldsymbol{\nu})}{M}. \tag{4.284}$$

Thus

$$\begin{aligned} p_\ell(\boldsymbol{\nu}) &\leq \frac{\binom{M}{\ell}}{M^\ell} p_1(\boldsymbol{\nu})^\ell \\ &\leq \frac{M!}{\ell!(M-\ell)!M^\ell} p_1(\boldsymbol{\nu})^\ell \\ &\leq \frac{M(M-1)\dots(M-\ell+1)}{\ell!M^\ell} p_1(\boldsymbol{\nu})^\ell \\ &\leq \frac{1}{\ell!} p_1(\boldsymbol{\nu})^\ell. \end{aligned} \tag{4.285}$$

This inequality is independent of the dimension M thus it can be extended for $\boldsymbol{\nu} \in \mathbb{R}_+^{\mathbb{N}^*}$ with $\sum_{n\in\mathbb{N}^*} \nu_n < \infty$. Therefore

$$\sum_{T\subset\mathbb{N}^*,|T|=N} \prod_{t\in T} \sigma_t \leq \frac{1}{N!}\Big(\sum_{n\in\mathbb{N}^*} \sigma_n\Big)^N < \infty. \tag{4.286}$$

Furthermore,

$$\forall M \in \mathbb{N}^*,\ \forall \boldsymbol{x} \in \mathcal{X}^N,\ 0 \leq \operatorname{Det} \boldsymbol{K}_M(\boldsymbol{x}) \leq \operatorname{Det} \boldsymbol{K}_{M+1,N}(\boldsymbol{x}). \tag{4.287}$$

Then by monotone convergence theorem, $\boldsymbol{x} \mapsto \frac{1}{N!} \operatorname{Det} \boldsymbol{K}(\boldsymbol{x})$ is mesurable and

$$\begin{aligned} \int_{\mathcal{X}^N} \frac{1}{N!} \operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) \otimes_{j\in[N]} \mathrm{d}\omega(x_j) &= \lim_{M\to\infty} \int_{\mathcal{X}^N} \frac{1}{N!} \operatorname{Det} \boldsymbol{K}_M(\boldsymbol{x}) \otimes_{j\in[N]} \mathrm{d}\omega(x_j) \\ &= \lim_{M\to\infty} \sum_{T\subset[M],|T|=N} \prod_{t\in T} \sigma_t \\ &= \sum_{T\subset\mathbb{N}^*,|T|=N} \prod_{t\in T} \sigma_t. \end{aligned} \tag{4.288}$$

□

End of the proof of Proposition 4.6

Proof. Remember that under the distribution of the projection DPP of kernel $\mathfrak{K}$

$$\mathbb{P}\left(\operatorname{Det} \boldsymbol{E}(\boldsymbol{x}) \neq 0\right) = 1. \tag{4.289}$$

Then by Proposition 4.5 and the fact that $\operatorname{Det}^2 \boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x}) = \prod_{n\in[N]} \sigma_n \operatorname{Det}^2 \boldsymbol{E}(\boldsymbol{x})$

$$\prod_{\ell\in[N]} \frac{1}{\cos^2 \theta_\ell\left(\mathcal{E}_N^{\mathcal{F}}, \mathcal{T}(\boldsymbol{x})\right)} = \frac{\operatorname{Det} \boldsymbol{K}(\boldsymbol{x})}{\operatorname{Det}^2 \boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})} = \frac{1}{\prod_{n\in[N]} \sigma_n} \frac{\operatorname{Det} \boldsymbol{K}(\boldsymbol{x})}{\operatorname{Det}^2 \boldsymbol{E}(\boldsymbol{x})}. \tag{4.290}$$

Then, taking the expectation with respect to $\boldsymbol{x}$ resulting from a DPP of kernel $\mathfrak{K}$

$$\begin{aligned} \mathbb{E}_{\mathrm{DPP}} \prod_{\ell\in[N]} \frac{1}{\cos^2 \theta_\ell\left(\mathcal{E}_N^{\mathcal{F}}, \mathcal{T}(\boldsymbol{x})\right)} &= \frac{1}{N!} \int_{\mathcal{X}^N} \operatorname{Det}^2 \boldsymbol{E}(\boldsymbol{x}) \prod_{\ell\in[N]} \frac{1}{\cos^2 \theta_\ell\left(\mathcal{E}_N^{\mathcal{F}}, \mathcal{T}(\boldsymbol{x})\right)} \otimes_{i=1}^N \mathrm{d}\omega(x_i) \\ &= \frac{1}{N!} \int_{\mathcal{X}^N} \operatorname{Det}^2 \boldsymbol{E}(\boldsymbol{x}) \frac{1}{\prod_{n\in[N]} \sigma_n} \frac{\operatorname{Det} \boldsymbol{K}(\boldsymbol{x})}{\operatorname{Det}^2 \boldsymbol{E}(\boldsymbol{x})} \otimes_{i=1}^N \mathrm{d}\omega(x_i) \\ &= \frac{1}{\prod_{n\in[N]} \sigma_n} \frac{1}{N!} \int_{\mathcal{X}^N} \operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) \otimes_{i=1}^N \mathrm{d}\omega(x_i). \end{aligned} \tag{4.291}$$

Now, by Lemma 4.11

$$\frac{1}{N!} \int_{\mathcal{X}^N} \operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) \otimes_{i=1}^N \mathrm{d}\omega(x_i) = \sum_{\substack{T\subset\mathbb{N}^* \\ |T|=N}} \prod_{t\in T} \sigma_t. \tag{4.292}$$

Therefore,

$$\mathbb{E}_{\mathrm{DPP}} \prod_{\ell\in[N]} \frac{1}{\cos^2 \theta_\ell\left(\mathcal{E}_N^{\mathcal{F}}, \mathcal{T}(\boldsymbol{x})\right)} = \sum_{\substack{T\subset\mathbb{N}^* \\ |T|=N}} \frac{\prod_{t\in T} \sigma_t}{\prod_{n\in[N]} \sigma_n}. \tag{4.293}$$

□

4.6.7 *Proof of Theorem 4.7*

Proof. Thanks to Proposition 4.8 and Lemma 4.2 (for $\tilde{\mathcal{F}}$ and $\tilde{k}$)

$$\max_{n\in[N]} \sigma_n \|\mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})^\perp} e_n^{\mathcal{F}}\|_{\mathcal{F}}^2 \leq \sigma_1 \cdot \max_{n\in[N]} \|\mathbf{\Pi}_{\tilde{\mathcal{T}}(\boldsymbol{x})^\perp} e_n^{\tilde{\mathcal{F}}}\|_{\tilde{\mathcal{F}}}^2 \tag{4.294}$$

$$\leq \sigma_1 \cdot \left(\prod_{n\in[N]} \frac{1}{\cos^2 \theta_n(\tilde{\mathcal{T}}(\boldsymbol{x}), \mathcal{E}_N^{\tilde{\mathcal{F}}})} - 1 \right). \tag{4.295}$$

Then Proposition 4.6 applied to $\tilde{\mathcal{F}}$ with kernel $\tilde{k}$ yields

$$\mathbb{E}_{\mathrm{DPP}} \prod_{n\in[N]} \frac{1}{\cos^2 \theta_n \left(\mathcal{E}_N^{\tilde{\mathcal{F}}}, \tilde{\mathcal{T}}_N(\boldsymbol{x}) \right)} = \sum_{\substack{T\subset\mathbb{N}^* \\ |T|=N}} \frac{\prod_{t\in T} \tilde{\sigma}_t}{\prod_{n\in[N]} \tilde{\sigma}_n}. \tag{4.296}$$

Every subset $T \subset \mathbb{N}^*$ such that $|T| = N$ can be written as $T = V \cup W$ with $V \subset [N]$ and $W \subset \mathbb{N}^* \setminus [N]$, and this decomposition is unique. Then

$$\frac{\prod_{t\in T} \tilde{\sigma}_t}{\prod_{n\in[N]} \tilde{\sigma}_n} = \frac{\prod_{v\in V} \tilde{\sigma}_v \prod_{w\in W} \tilde{\sigma}_w}{\prod_{n\in[N]} \tilde{\sigma}_n} = \frac{\prod_{w\in W} \tilde{\sigma}_w}{\prod_{n\in[N]\setminus V} \tilde{\sigma}_n}. \tag{4.297}$$

Therefore

$$\begin{aligned}
\sum_{\substack{T\subset\mathbb{N}^* \\ |T|=N}} \frac{\prod_{t\in T} \tilde{\sigma}_t}{\prod_{n\in[N]} \tilde{\sigma}_n} &= \sum_{\substack{T\subset\mathbb{N}^* \\ |T|=N \\ T=V\cup W}} \frac{\prod_{w\in W} \tilde{\sigma}_w}{\prod_{n\in[N]\setminus V} \tilde{\sigma}_n} \\
&= \sum_{V\subset[N]} \sum_{\substack{W\subset\mathbb{N}^*\setminus[N] \\ |W|=N-|V|}} \frac{\prod_{w\in W} \tilde{\sigma}_w}{\prod_{n\in[N]\setminus V} \tilde{\sigma}_n} \\
&= \sum_{0\leq\ell\leq N} \left[\sum_{\substack{V\subset[N] \\ |V|=\ell}} \prod_{n\in[N]\setminus V} \frac{1}{\tilde{\sigma}_n} \right] \left[\sum_{\substack{W\subset\mathbb{N}^*\setminus[N] \\ |W|=N-\ell}} \prod_{w\in W} \tilde{\sigma}_w \right] \\
&= \sum_{0\leq\ell\leq N} \left[\sum_{\substack{V\subset[N] \\ |V|=N-\ell}} \prod_{n\in V} \frac{1}{\tilde{\sigma}_n} \right] \left[\sum_{\substack{W\subset\mathbb{N}^*\setminus[N] \\ |W|=N-\ell}} \prod_{w\in W} \tilde{\sigma}_w \right] \\
&= \sum_{0\leq\ell\leq N} p_{N-\ell} \left(\left(\frac{1}{\tilde{\sigma}_m} \right)_{m\in[N]} \right) p_{N-\ell} \left((\tilde{\sigma}_m)_{m\geq N+1} \right) \\
&= \sum_{0\leq\ell\leq N} p_\ell \left(\left(\frac{1}{\tilde{\sigma}_m} \right)_{m\in[N]} \right) p_\ell \left((\tilde{\sigma}_m)_{m\geq N+1} \right),
\end{aligned} \tag{4.298}$$

where for $\ell \in [N]$, p_ℓ is the ℓ-th symmetric polynomial with the convention that $p_0 = 1$.

Finally, thanks to (4.285) above

$$\sum_{\substack{T\subset\mathbb{N}^*\\|T|=N}}\frac{\prod_{t\in T}\tilde{\sigma}_t}{\prod_{n\in[N]}\tilde{\sigma}_n}\leq 1+\sum_{\ell\in[N]}\frac{1}{\ell!^2}\left(\sum_{m\in[N]}\frac{1}{\tilde{\sigma}_m}\sum_{m\geq N+1}\tilde{\sigma}_m\right)^\ell \tag{4.299}$$

$$\leq 1+\sum_{\ell\in[N]}\frac{1}{\ell!^2}\left(\frac{N}{\sigma_1}\sum_{m\geq N+1}\sigma_m\right)^\ell.$$

As a consequence, by writing $r_N=\sum_{m\geq N+1}\sigma_m$,

$$\mathbb{E}_{\mathrm{DPP}}\left[\max_{n\in[N]}\sigma_n\|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})^\perp}e_n^{\mathcal{F}}\|_{\mathcal{F}}^2\right]\leq\sigma_1\cdot\sum_{\ell=1}^{N}\frac{1}{\ell!^2}\left(\frac{Nr_N}{\sigma_1}\right)^\ell \tag{4.300}$$

which can be plugged in Lemma 4.1 to conclude the proof. □

4.6.8 *The proof of Theorem 4.8 and Proposition 4.7*

The proof of Theorem 4.8 follows the same steps as the proof of Theorem 4.7 with the exception that we use Proposition 4.7 instead of Proposition 4.6 (for $\tilde{\mathcal{F}}$ and $\tilde{k}$).

Proof of Proposition 4.7

Our objective in this section is to give a tractable expression of

$$\mathbb{E}_{\mathrm{DPP}}\sum_{\ell\in[N]}\frac{1}{\cos^2\theta_\ell\left(\mathcal{E}_N^{\mathcal{F}},\mathcal{T}(\boldsymbol{x})\right)}. \tag{4.301}$$

Let $\boldsymbol{x}=(x_1,\ldots,x_N)\in\mathcal{X}^N$ such that $\mathrm{Det}^2\boldsymbol{E}(\boldsymbol{x})>0$, and consider the characteristic polynomial of the matrix $\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal^{-1}}\boldsymbol{K}(\boldsymbol{x})\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{-1}$

$$\chi_{\boldsymbol{x}}(t)=\mathrm{Det}\left(\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal^{-1}}\boldsymbol{K}(\boldsymbol{x})\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{-1}-t\mathbb{I}_N\right)\in\mathbb{R}_N[t]. \tag{4.302}$$

By Proposition 4.5, the roots of $\chi_{\boldsymbol{x}}(t)$ are the $1/\cos^2\theta_n(\mathcal{T}(\boldsymbol{x}),\mathcal{E}_N^{\mathcal{F}})$ for $n\in[N]$.

Now, for any polynomial $P\in\mathbb{R}_N[t]$, denote by $a_n(P)$ its coefficients

$$P(t)=\sum_{n=0}^{N}a_n(P)t^n. \tag{4.303}$$

Then by Vieta's formula,

$$a_{N-1}(\chi_{\boldsymbol{x}}(t))=(-1)^{N-1}\,\mathrm{Tr}\left(\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal^{-1}}\boldsymbol{K}(\boldsymbol{x})\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{-1}\right). \tag{4.304}$$

Therefore,

$$\begin{aligned}I_{k,N}&=\mathbb{E}_{\mathrm{DPP}}\,\mathrm{Tr}\left(\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal^{-1}}\boldsymbol{K}(\boldsymbol{x})\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{-1}\right)\\&=(-1)^{N-1}\,\mathbb{E}_{\mathrm{DPP}}\,a_{N-1}(\chi_{\boldsymbol{x}}(t))\\&=(-1)^{N-1}a_{N-1}(\mathbb{E}_{\mathrm{DPP}}\,\chi_{\boldsymbol{x}}(t)),\end{aligned} \tag{4.305}$$

where $\mathbb{E}_{\mathrm{DPP}}\,\chi_{\boldsymbol{x}}(t)$ is the expected characteristic polynomial. Therefore, in order to calculate $I_{k,N}$, it is sufficient to calculate $\mathbb{E}_{\mathrm{DPP}}\,\chi_{\boldsymbol{x}}(t)$. Since $\mathbb{E}_{\mathrm{DPP}}\,\chi_{\boldsymbol{x}}(t)$ is a polynomial, it is completely determined in the interval $[-1/2, 1/2]$ [13].

Now, let $t \in [-1/2, 1/2]$

$$\begin{aligned}
\mathbb{E}_{\mathrm{DPP}}\,\chi_{\boldsymbol{x}}(t) &= \frac{1}{N!}\int_{\mathcal{X}^N} \mathrm{Det}^2\,\boldsymbol{E}(\boldsymbol{x})\,\mathrm{Det}\left(\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal^{-1}}\boldsymbol{K}(\boldsymbol{x})\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{-1} - t\mathbb{I}_N\right)\otimes_{j\in[N]}\mathrm{d}\omega(x_j)\\
&= \frac{1}{N!\prod\limits_{n\in[N]}\sigma_n}\int_{\mathcal{X}^N} \mathrm{Det}^2\,\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})\,\mathrm{Det}\left(\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal^{-1}}\boldsymbol{K}(\boldsymbol{x})\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{-1} - t\mathbb{I}_N\right)\otimes_{j\in[N]}\mathrm{d}\omega(x_j)\\
&= \frac{1}{N!\prod\limits_{n\in[N]}\sigma_n}\int_{\mathcal{X}^N} \mathrm{Det}\left(\boldsymbol{K}(\boldsymbol{x}) - t\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal}\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})\right)\otimes_{j\in[N]}\mathrm{d}\omega(x_j). \qquad (4.306)
\end{aligned}$$

Now, define the kernel k_t

$$k_t(x,y) = k(x,y) - t\sum_{n\in[N]}\sigma_n e_n(x)e_n(y). \qquad (4.307)$$

Then

$$\mathbb{E}_{\mathrm{DPP}}\,\chi_{\boldsymbol{x}}(t) = \frac{1}{N!\prod\limits_{n\in[N]}\sigma_n}\int_{\mathcal{X}^N}\mathrm{Det}\left(\boldsymbol{K}_t(\boldsymbol{x})\right)\otimes_{j\in[N]}\mathrm{d}\omega(x_j). \qquad (4.308)$$

Since, $t \in [-1/2, 1/2]$, k_t defines a positive-definite kernel and we can apply Lemma 4.11

$$\phi(t) = N!\sum_{\substack{U\subset\mathbb{N}^*\\|U|=N}}\prod_{u\in U}\sigma_u(t), \qquad (4.309)$$

where the $\sigma_u(t)$ are the eigenvalues of the integration operator corresponding to the kernel k_t with respect to the measure $\mathrm{d}\omega$:

$$\sigma_u(t) = \sigma_u - \mathbb{1}_{[N]}(u)\sigma_u t. \qquad (4.310)$$

Now, every subset $U \subset \mathbb{N}^*$ such that $|U| = N$ can be written as $U = V \cup W$ with $V \subset [N]$ and $W \subset \mathbb{N}^* \setminus [N]$, and this decomposition is unique. Then

$$\prod_{u\in U}\sigma_u(t) = \prod_{v\in V}\sigma_v(t)\prod_{w\in W}\sigma_w(t). \qquad (4.311)$$

Therefore

$$\begin{aligned}
\sum_{\substack{U\subset\mathbb{N}^*\\|U|=N}}\prod_{u\in U}\sigma_u(t) &= \sum_{\substack{U\subset\mathbb{N}^*\\|U|=N\\U=V\cup W}}\prod_{v\in V}\sigma_v(t)\prod_{w\in W}\sigma_w(t) \qquad (4.312)\\
&= \sum_{V\subset[N]}\sum_{\substack{W\subset\mathbb{N}^*\setminus[N]\\|W|=N-|V|}}\prod_{v\in V}\sigma_v(t)\prod_{w\in W}\sigma_w(t)\\
&= \sum_{V\subset[N]}\sum_{\substack{W\subset\mathbb{N}^*\setminus[N]\\|W|=N-|V|}}(\prod_{v\in V}\sigma_v)(1-t)^{|V|}(\prod_{w\in W}\sigma_w)\\
&= \sum_{\ell=0}^{N}(1-t)^{\ell}\sum_{\substack{V\subset[N]\\|V|=\ell}}\prod_{v\in V}\sigma_v\sum_{\substack{W\subset\mathbb{N}^*\setminus[N]\\|W|=N-|V|}}\prod_{w\in W}\sigma_w. \qquad (4.313)
\end{aligned}$$

[13] The reason behind this choice will appear later

Finally,

$$\mathbb{E}_{\mathrm{DPP}}\,\gamma_{\boldsymbol{x}}(t) = \frac{1}{\prod\limits_{n\in[N]}\sigma_n}\sum_{\ell=0}^{N}(1-t)^{\ell}\sum_{\substack{V\subset[N]\\|V|=\ell}}\prod_{v\in V}\sigma_v\sum_{\substack{W\subset\mathbb{N}^*\setminus[N]\\|W|=N-|V|}}\prod_{w\in W}\sigma_w. \tag{4.314}$$

Remember that for any polynomial $P\in\mathbb{R}_N[t]$,

$$a_{n-1}(P) = \frac{1}{(N-1)!}P^{(n-1)}(0). \tag{4.315}$$

Now,

$$\mathbb{E}_{\mathrm{DPP}}\,\gamma_{\boldsymbol{x}}(t) = \frac{1}{\prod\limits_{n\in[N]}\sigma_n}\Big[(1-t)^N\sum_{\substack{V\subset[N]\\|V|=N}}\prod_{v\in V}\sigma_v + (1-t)^{N-1}\sum_{\substack{V\subset[N]\\|V|=N-1}}\prod_{v\in V}\sigma_v\sum_{w\in\mathbb{N}^*\setminus[N]}\sigma_w\Big] + p(t), \tag{4.316}$$

where p is a polynomial of degree smaller than $N-2$. Therefore

$$a_{N-1}(\mathbb{E}_{\mathrm{DPP}}\,\gamma_{\boldsymbol{x}}(t)) = \frac{(-1)^{N-1}}{\prod\limits_{v\in[N]}\sigma_v}\left(N\prod_{v\in[N]}\sigma_v + \sum_{\substack{V\subset[N]\\|V|=N-1}}\prod_{v\in V}\sigma_v\sum_{w\in\mathbb{N}^*\setminus[N]}\sigma_w\right). \tag{4.317}$$

Finally, we get

$$\begin{aligned}\mathbb{E}_{\mathrm{DPP}}\,\mathrm{Tr}\left(\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal^{-1}}\boldsymbol{K}(\boldsymbol{x})\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{-1}\right) &= (-1)^{N-1}a_{N-1}(\mathbb{E}_{\mathrm{DPP}}\,\gamma_{\boldsymbol{x}}(t)) &(4.318)\\ &= \frac{1}{\prod\limits_{v\in[N]}\sigma_v}\left(N\prod_{v\in[N]}\sigma_v + \sum_{\substack{V\subset[N]\\|V|=N-1}}\prod_{v\in V}\sigma_v\sum_{w\in\mathbb{N}^*\setminus[N]}\sigma_w\right)\\ &= N + \sum_{\substack{V\subset[N]\\|V|=N-1}}\frac{1}{\sigma_v}\sum_{w\in\mathbb{N}^*\setminus[N]}\sigma_w. &(4.319)\end{aligned}$$

5

KERNEL INTERPOLATION USING VOLUME SAMPLING

5.1 INTRODUCTION

In this chapter, we propose and analyze the interpolation based on a random design drawn from a distribution called *continuous volume sampling*, which favors designs $\boldsymbol{x}$ with a large value of $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x})$. After introducing this new distribution, we prove non-asymptotic guarantees on the interpolation error which depend on the spectrum of the kernel k. We show here that continuous volume sampling enjoys error bounds that scale as lower bounds, as well as some additional interpretable geometric properties, while having a joint density that can be evaluated as soon as one can evaluate the RKHS kernel k. In particular, this opens the possibility of Markov chain Monte Carlo samplers (Rezaei and Gharan, 2019). This is to be compared with the projection DPP introduced in Chapter 4 that requires access to the Mercer decomposition of k.

As we have seen in Chapter 3, volume sampling has been used in matrix subsampling for linear regression and low-rank approximations. Like Chapter 4, this chapter connects the discrete problem of sub-sampling from a matrix and the continuous problem of interpolating functions in an RKHS.

The rest of the chapter is organized as follows. Section 5.2 reviews kernel-based interpolation. In Section 5.3, we define continuous volume sampling and relate it to projection determinantal point processes. Section 5.4 contains our main results while Section 5.5 contains sketches of all proofs with pointers to the Section 5.8 for the details of the proofs. Section 5.6 numerically illustrates our main result. We conclude this chapter in Section 5.7.

The material of this chapter is based on the following article

- A. Belhadji, R. Bardenet, and P. Chainais (2020b). "Kernel interpolation with continuous volume sampling". In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 725–735.

NOTATION AND ASSUMPTIONS. We keep the same notation as Chapter 4.

For $N \in \mathbb{N}^*$, we will often sum over the sets

$$\mathcal{U}_N^m = \{U \subset \mathbb{N}^*, |U| = N,\ m \notin U\}, \tag{5.1}$$

$$\mathcal{U}_N = \{U \subset \mathbb{N}^*, |U| = N\}. \tag{5.2}$$

Finally, define the approximation error

$$\mathcal{E}(\mu; \boldsymbol{x}, \boldsymbol{w}) = \|\mu - \sum_{i \in [N]} w_i k(x_i, .)\|_{\mathcal{F}}, \tag{5.3}$$

where $[N] = \{1, \dots, N\}$. If $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) > 0$, let $\hat{w} = \boldsymbol{K}(\boldsymbol{x})^{-1}\mu(\boldsymbol{x})$ and define the interpolation error

$$\mathcal{E}(\mu; \boldsymbol{x}) = \|\mu - \sum_{i \in [N]} \hat{w}_i k(x_i, .)\|_{\mathcal{F}} \tag{5.4}$$

$$= \|\mu - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu\|_{\mathcal{F}}. \tag{5.5}$$

Finally, we keep the assumptions

Assumption 5.1. $\mathrm{d}\omega$ *is a* non-degenerate *measure, i.e. the support of* $\mathrm{d}\omega$ *is equal to* $\mathcal{X}$.

Assumption 5.2. k *is continuous with respect to the product topology of* $\mathcal{X} \times \mathcal{X}$.

Assumption 5.3. $x \mapsto k(x, x)$ *is integrable with respect to* $\mathrm{d}\omega$ *so that* $\mathcal{F} \subset \mathbb{L}_2(\mathrm{d}\omega)$.

5.2 THREE TOPICS ON KERNEL INTERPOLATION

The literature on kernel interpolation is prolific and cannot be covered in details in this chapter. This section reviews three topics on this field to better situate our contributions.

5.2.1 *Kernel interpolation beyond embeddings*

Besides the approximation of the embeddings μ_g discussed in Section 4.2.5, theoretical guarantees for the kernel interpolation of a general $\mu \in \mathcal{F}$ are sought *per se*. The Shannon reconstruction formula for bandlimited signals (Shannon, 1948) is implicitly an interpolation by the sinc kernel. The RKHS approach for sampling in signal processing was introduced in (Yao, 1967) for the Hilbert space of bandlimited signals; see also (Nashed and Walter, 1991) for generalizations. Remarkably, in those RKHSs, every $\mu \in \mathcal{F}$ is an embedding μ_g for some $g \in \mathbb{L}_2(\mathrm{d}\omega)$: k is a projection kernel of infinite rank. In general, for a trace-class kernel, the subspace spanned by the embeddings μ_g is strictly included in $\mathcal{F}$. More precisely, every μ_g satisfies

$$\|\boldsymbol{\Sigma}^{-1/2}\mu_g\|_{\mathcal{F}} = \|\boldsymbol{\Sigma}^{1/2}g\|_{\mathcal{F}} = \|g\|_{\mathbb{L}_2(\mathrm{d}\omega)} < +\infty.$$

This condition is more restrictive than what is required for a generic μ to belong to $\mathcal{F}$, i.e., $\|\mu\|_{\mathcal{F}} < +\infty$, so that kernel interpolation is more general than optimal kernel quadrature. Compared to Chapter 4 where we focused on kernel interpolation of the embeddings μ_g, we study in this chapter kernel interpolation for any $\mu \in \mathcal{F}$.

5.2.2 *Optimization algorithms*

Optimization approaches offer a variety of algorithms for the design of the interpolation nodes. De Marchi, 2003 and De Marchi et al., 2005 proposed greedily maximizing the so-called *power function*

$$p(x; \boldsymbol{x}) = \left[k(x, x) - k_{\boldsymbol{x}}(x)^{\intercal}\boldsymbol{K}(\boldsymbol{x})^{-1}k_{\boldsymbol{x}}(x)\right]^{1/2}, \tag{5.6}$$

where $k_{\boldsymbol{x}}(x) = (k(x, x_i))_{i \in [N]}$. This algorithm leads to an interpolation error that goes to zero with N for a kernel of class $\mathcal{C}^2$ (De Marchi et al., 2005). Later, Santin

and Haasdonk, 2017 proved better convergence rates for smoother kernels. Again, these results assume that the domain $\mathcal{X}$ is compact. Other greedy algorithms were proposed in the context of Bayesian quadrature (BQ) such as Sequential BQ (Huszár and Duvenaud, 2012), or Frank-Wolfe BQ (Briol et al., 2015). These algorithms sequentially minimize $\mathcal{E}(\mu_g; \boldsymbol{x})$, for a fixed $g \in \mathbb{L}_2(\mathrm{d}\omega)$. The nodes are thus adapted to one particular μ_g by construction. In general, each step of these greedy algorithms requires to solve a non-convex problem with many local minima (Oettershagen, 2017)[Chapter 5]. In practice, costly approximations must be employed such as local search in a random grid (Lacoste-Julien et al., 2015).

An alternative approach, that is very related to our contribution and has raised a lot of recent interest, is to observe that the squared power function (5.6) can be upper bounded by the inverse of $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x})$ (Schaback, 2005; Tanaka, 2019). Designs that maximize $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x})$ are called *Fekete points*; see e.g. (Bos and Maier, 2002; Bos and De Marchi, 2011). Tanaka, 2019 proposed to approximate $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x})$ using the Mercer decomposition of k, followed by a rounding of the solution of a D-experimental design problem, yet without a theoretical analysis of the interpolation error. Karvonen et al., 2019 proved that for the uni-dimensional Gaussian kernel, the approximate objective function of (Tanaka, 2019) is actually convex. Moreover, Karvonen et al., 2019 analyze their interpolation error; see also Section 5.4.2. Finally, we emphasize that these algorithms require the knowledge of a Mercer-type decomposition of k so that they cannot be implemented for any kernel; moreover, the approximate objective function may be non-convex in general.

Table 5.1 puts in perspective our work compared to the optimization approaches: the projection DPP of Chapter 4 is the stochastic version of approximate Fekete points, continuous volume sampling of this chapter is the stochastic version of Fekete points.

Objective function / Approach	Optimization	Sampling
$\operatorname{Det}^2 \boldsymbol{E}_N(\boldsymbol{x})$	Approximate Fekete points (Karvonen et al., 2019)	Projection DPP (Belhadji et al., 2019a)
$\operatorname{Det} \boldsymbol{K}(\boldsymbol{x})$	Fekete points (Tanaka, 2019)	CVS (Belhadji et al., 2020b)

Table 5.1 – The determinantal approaches for the construction of kernel interpolation nodes.

5.2.3 *Lower bounds*

When investigating upper bounds for kernel interpolation errors, it is useful to remember existing lower bounds, so as to evaluate the tightness of one's results. In particular, N-widths theory (Pinkus, 2012) implies lower bounds for kernel interpolation errors, which once again show the importance of the spectrum of $\boldsymbol{\Sigma}$.

The N-width of $\mathcal{S} = \{\mu_g = \boldsymbol{\Sigma} g,\ \|g\|_{\mathbb{L}_2(\mathrm{d}\omega)} \leq 1\}$ with respect to the couple $(\mathbb{L}_2(\mathrm{d}\omega), \mathcal{F})$ (Pinkus, 2012, Chapter 1.7) is defined as the square root of

$$d_N(\mathcal{S})^2 = \inf_{\substack{Y \subset \mathcal{F} \\ \dim Y = N}} \sup_{\|g\|_{\mathrm{d}\omega} \leq 1} \inf_{y \in Y} \|\boldsymbol{\Sigma} g - y\|_{\mathcal{F}}^2.$$

In interpolation, we do use a subspace $Y \subset \mathcal{F}$ spanned by N independent functions $k(x_i, .)$, so that

$$\sup_{\|g\|_{\mathrm{d}\omega} \leq 1} \mathcal{E}(\boldsymbol{\Sigma} g; \boldsymbol{x})^2 \geq d_N(\mathcal{S})^2. \tag{5.7}$$

Applying (Pinkus, 2012, Theorem 2.2, Chapter 4) to the adjoint of the embedding operator $I_{\mathcal{F}} : \mathcal{F} \to \mathbb{L}_2(\mathrm{d}\omega)$, it comes $d_N(\mathcal{S})^2 = \sigma_{N+1}$. One may object that some QMC sequences seem to breach this lower bound. For example, in the Korobov space $(d = 2, s \geq 1)$, $\sigma_{N+1} = \mathcal{O}(\log(N)^{2s} N^{-2s})$ (Bach, 2017), while the interpolation of μ_g with $g = 1$ using a Fibonacci lattice leads to an error in $\mathcal{O}(\log(N) N^{-2s}) = o(\sigma_{N+1})$ (Bilyk et al., 2012)[Theorem 4]. But this is the rate for one particular μ_g, and it cannot be achieved uniformly in g.

5.3 VOLUME SAMPLING AND DPPS

In this section, we introduce the *continuous volume sampling* (VS) and compare it to projection DPPs.

5.3.1 *Continuous volume sampling*

Definition 5.1 (Continuous volume sampling). *Let $N \in \mathbb{N}^*$ and $\boldsymbol{x} = \{x_1, \dots, x_N\} \subset \mathcal{X}$. We say that $\boldsymbol{x}$ follows the volume sampling distribution, if $(x_1, \dots, x_N)$ is a random variable of $\mathcal{X}^N$, the law of which is absolutely continuous with respect to $\otimes_{i \in [N]} \mathrm{d}\omega$, and the density writes*

$$f_{\mathrm{VS}}(x_1, \dots, x_N) \propto \operatorname{Det} \boldsymbol{K}(\boldsymbol{x}). \tag{5.8}$$

Two remarks are in order. First, under Assumption 5.3, the density f_{VS} in (5.8) indeed integrates to 1. Indeed, Hadamard's inequality yields

$$\begin{aligned} \int_{\mathcal{X}^N} \operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) \otimes_{i \in [N]} \mathrm{d}\omega(x_i) &\leq \int_{\mathcal{X}^N} \prod_{i \in [N]} k(x_i, x_i) \otimes \mathrm{d}\omega(x_i) \\ &= \left(\int_{\mathcal{X}} k(x, x) \mathrm{d}\omega(x) \right)^N < +\infty. \end{aligned}$$

Second, the determinant in (5.8) is invariant to permutations, so that continuous volume sampling can indeed be seen as defining a random set $\boldsymbol{x} = \{x_1, \dots, x_N\}$.

In the following, we denote, for any symmetric and continuous kernel $\tilde{k}$ satisfying Assumption 5.3,

$$Z_N(\tilde{k}) := \int_{\mathcal{X}^N} \operatorname{Det} \tilde{\boldsymbol{K}}(\boldsymbol{x}) \otimes \mathrm{d}\omega(x_i). \tag{5.9}$$

5.3.2 *Continuous volume sampling as a mixture of DPPs*

In the following result, we show that continuous volume sampling is a mixture of projection DPPs. This is an extension of Theorem 2.2 to continuous domain.

Proposition 5.1. *For $U \subset \mathbb{N}^*$ define the projection kernel*

$$\mathfrak{K}_U(x,y) = \sum_{u\in U} e_u(x)e_u(y). \tag{5.10}$$

For $N \in \mathbb{N}^*$, we have

$$f_{\mathrm{VS}}(x_1,\dots,x_N) \propto \sum_{U\in\mathcal{U}_N}\prod_{u\in U}\sigma_u \operatorname{Det}(\mathfrak{K}_U(x_i,x_j))_{(i,j)}, \tag{5.11}$$

and the normalization constant is equal to

$$Z_N(k) = \mathrm{N}!\sum_{U\in\mathcal{U}_N}\prod_{u\in U}\sigma_u. \tag{5.12}$$

In other words, f_{VS} is a mixture of the projection DPPs associated to the kernels $\mathfrak{K}_U$ and the reference measure $\mathrm{d}\omega$.

The proof of this proposition is given in Section 5.8.2.

The largest weight in the mixture (5.11) corresponds to the projection DPP of kernel $\mathfrak{K}_{[N]}$ proposed in Chapter 4. The following lemma gives an upper bound on this maximal weight using the eigenvalues of $\mathbf{\Sigma}$.

Proposition 5.2. *For $N \in \mathbb{N}^*$, define*

$$\delta_N = \prod_{n\in[N]}\sigma_n \Big/ \sum_{U\in\mathcal{U}_N}\prod_{u\in U}\sigma_u. \tag{5.13}$$

Then for all $N \in \mathbb{N}^*$, $\delta_N \le \sigma_N/r_N$.

In particular, if the eigenvalues of $\mathbf{\Sigma}$ decreases polynomially, then $\delta_N = \mathcal{O}(1/N)$, so that as N grows, continuous volume sampling is becoming more and more different from the projection DPP of Chapter 4. In contrast, if the eigenvalues decays exponentially, then $\delta_N = \mathcal{O}(1)$.

5.3.3 *Numerical simulation*

A projection DPP can be sampled exactly using the HKPV algorithm as long as one can evaluate the corresponding projection kernel; see Section 2.2.5. This suggests using the mixture in Proposition 5.1 to implement continuous volume sampling. Yet, such an algorithm requires explicit knowledge of the Mercer decomposition of the kernel or at least a decomposition onto an orthonormal basis of $\mathcal{F}$ as in (Karvonen et al., 2019). This is a strong requirement that is undesirable in practice.

The fact that the value of $\operatorname{Det}\boldsymbol{K}(\boldsymbol{x})$ only depends on the pointwise evaluation of k suggests that continuous volume sampling is *fully kernelized*, in the sense that a sampling algorithm should be able to bypass the need for a decomposition of the kernel. One could proceed by rejection sampling. Yet the acceptance ratio would likely scale poorly with N.

An alternative solution would be to use a variant of the HKPV algorithm to approximate the continuous volume sampling distribution. Indeed we have for $\boldsymbol{x} \in \mathcal{X}^N$ such that $\operatorname{Det}\boldsymbol{K}(\boldsymbol{x}) > 0$ and $k(x_1,x_1) > 0$

$$\begin{aligned}\operatorname{Det}\boldsymbol{K}(\boldsymbol{x}) &= k(x_1,x_1) \times \frac{\operatorname{Det}\boldsymbol{K}(\{x_1,x_2\})}{k(x_1,x_1)} \times \cdots \times \frac{\operatorname{Det}\boldsymbol{K}(\boldsymbol{x})}{\operatorname{Det}\boldsymbol{K}(\boldsymbol{x}\setminus\{x_N\})}\\ &= p(x_1;\emptyset)\times p(x_2;\{x_1\})\times\cdots\times p(x_N;\{x_1,\dots,x_{N-1}\}),\end{aligned} \tag{5.14}$$

A-HKPV$(k, \mathrm{d}\omega)$

$n \longleftarrow 1$
$\mu(x) \propto p(x;\emptyset)\mathrm{d}\omega(x)$ ▷ *Initialize the distribution*
$\boldsymbol{x} \longleftarrow \emptyset$
while $n \leq N$
    Pick x_n in $\mathcal{X}$ from μ ▷ *Sample from the distribution*
    $\boldsymbol{x} \longleftarrow \boldsymbol{x} \cup \{x_n\}$
    $n \longleftarrow n+1$
    $\mu(x) \propto p(x;\boldsymbol{x})\mathrm{d}\omega(x)$ ▷ *Update the distribution*
return $\boldsymbol{x}$

Figure 5.1 – Pseudocode of approximate continuous volume sampling based on a variant of the HKPV.

where p is the power function defined in (5.6). Algorithm 5.1 proposed in (Rezaei and Gharan, 2019) uses this decomposition to approximate continuous volume sampling. The output of the algorithm follows a density $\tilde{f}_{\mathrm{VS}}$ with respect to $\prod_{n\in[N]} \mathrm{d}\omega(x_n)$ that satisfies (Rezaei and Gharan, 2019)[Lemma 4.4]

$$\forall (x_1, \dots, x_N) \in \mathcal{X}^N, \ \tilde{f}_{\mathrm{VS}}(x_1, \dots, x_N) \leq N!^2 f_{\mathrm{VS}}(x_1, \dots, x_N). \tag{5.15}$$

This bound is irrelevant for large values of N. In other words, the approximation of continuous volume sampling by the Algorithm 5.1 is very loose. Note however that $\tilde{f}_{\mathrm{VS}} = f_{\mathrm{VS}}$ if k is a projection kernel that defines an integration operator of rank N: Algorithm 5.1 coincides with Algorithm 2.5 in this case.

A workaround would be to use $\tilde{f}_{\mathrm{VS}}$ as the density of the initial state of an MCMC sampler proposed in (Rezaei and Gharan, 2019). This MCMC algorithm is based on a Gibbs sampler chain: given a state $\boldsymbol{x} = \{x_1, \dots, x_N\}$, remove a node x_n chosen uniformly at random and add a node y with a probability proportional to $\det \boldsymbol{K}(\boldsymbol{x}')$ where $\boldsymbol{x}' = \{x_1, \dots, x_{n-1}, y, x_{n+1}, \dots, x_N\}$.

For this Markov chain, the authors were able to derive bounds for the mixing time:

$$\tau_{P_0}(\eta) = \min\{t | \ \|P_t - P_{\mathrm{VS}}\|_{\mathrm{TV}} \leq \eta\},$$

where $\|.\|_{\mathrm{TV}}$ is the total variation distance, P_t is the distribution of the Markov chain after t steps and P_{VS} is the distribution of the continuous volume sampling. In particular, they proved that the mixing time scales as $\mathcal{O}(N^5 \log(N))$. They also proved bounds on the expected number of rejections, which shows the feasibility of the implementation of the Gibbs steps. This sequential algorithm can be implemented in fully kernelized way without the need for the Mercer decomposition of k.

The power function: between optimization and sampling.

As we have seen in Section 5.2.2, the sequential maximization of the power function was studied in the literature. The numerical implementation of this algorithm relies on solving the optimization problem

$$\max_{x\in\mathcal{X}} p(x;\boldsymbol{x}), \tag{5.16}$$

for an arbitrary configuration $\boldsymbol{x} \subset \mathcal{X}$. Unfortunately, (5.16) is a non-convex problem with many local maxima; see Figure 5.2 for an illustration on the periodic Sobolev spaces of orders $s \in \{1, 3, 5\}$. In practice, (5.16) may be replaced by the surrogate problem

$$\max_{x \in \tilde{\mathcal{X}}} p(x; \boldsymbol{x}), \tag{5.17}$$

where $\tilde{\mathcal{X}}$ is a large finite subset of $\mathcal{X}$. However, these heuristics do not scale to high-dimensional domains. An alternative strategy would be to replace the optimization step (5.16) by the sampling step

$$x \sim p(x; \boldsymbol{x}). \tag{5.18}$$

We can observe that selecting the nodes sequentially using (5.18) is equivalent to running Algorithm in Figure 5.1. Now, the implementation of (5.18) is possible via rejection sampling.

Indeed, by observing that

$$\forall x \in \mathcal{X}, \;\; p(x; \boldsymbol{x}) \leq k(x, x), \tag{5.19}$$

we may use

$$x \mapsto \frac{1}{\int\limits_{\mathcal{X}} k(x, x) \mathrm{d}\omega(x)} k(x, x) \tag{5.20}$$

as a proposal.

Interestingly, if $\mathcal{X} \subset \mathbb{R}^d$ is bounded and k is translation invariant, then the expected number of rejections scales as $\mathcal{O}(1/\sum_{n=N+1}^{+\infty} \sigma_n)$, where N is the cardinality of the configuration $\boldsymbol{x}$ (Rezaei and Gharan, 2019)[Lemma 5.3]. In other words, the complexity [1] of the step (5.18) using the proposal (5.20) increases as N increases. Moreover, for a fixed value of N, the smoother the kernel, the higher the complexity of the step (5.18). Figure 5.2 illustrates the sharpness of the bound (5.19) on the periodic Sobolev spaces of orders $s \in \{1, 2, 3\}$ for $N \in \{0, 1, 2, 3, 4, 5\}$. We observe that for a fixed value of N, the smoother the kernel, the larger is the gap between the proposal (the diagonal of the kernel) and the power function.

These observations suggest that the complexity of the step (5.18) decreases as the dimension of the domain increases for a given family of kernels. To illustrate this phenomenon, we consider the case of Korobov spaces $\mathcal{K}_{d,s}$ defined for a fixed value of $s \in \mathbb{N}^*$. As we have seen in Section 4.2.3, the larger d, the "richer" the spectrum of the integration operator $\boldsymbol{\Sigma}$. Therefore, for a fixed value of N, the rejection rate

$$R_{d,s}(N) = 1/\sum_{n=N+1}^{+\infty} \sigma_n^{\mathcal{K}_{d,s}} \tag{5.21}$$

is decreasing with respect to d. Figure 5.3 illustrates this phenomenon for $s \in \{1, 2\}$ and $d \in \{2, 3, 4\}$. We observe that $R_{4,s}(N) < R_{3,s}(N) < R_{2,s}(N)$ for the two values of s: the larger the dimension d, the smaller the complexity of the step (5.18) using the proposal (5.20). In other words, the algorithm in Figure 5.1, that is an approximation of continuous volume sampling, is a viable alternative to the greedy maximization of the power function in high dimensional domains.

[1] The complexity expressed in term of the expected value of the number of rejections.

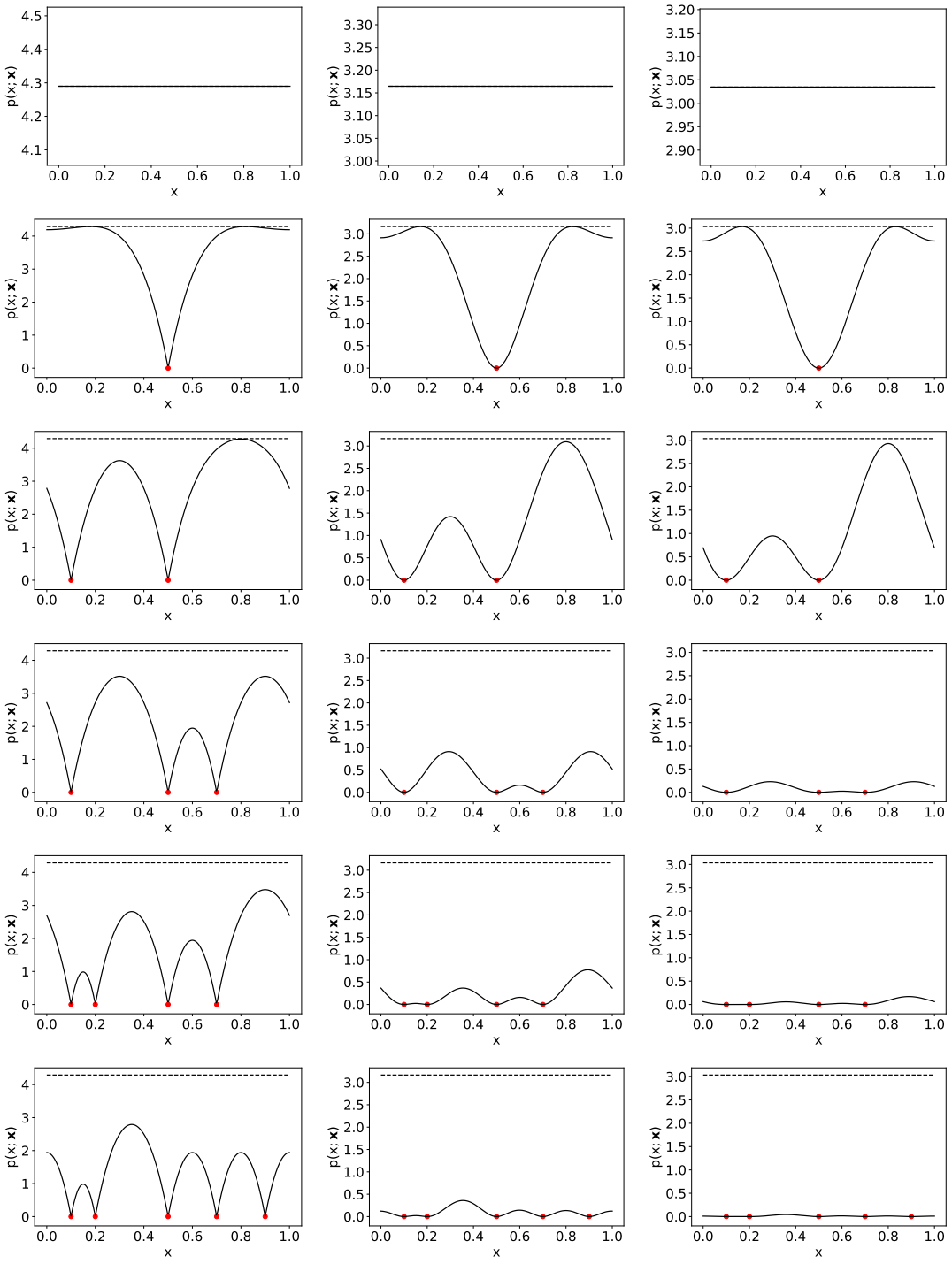


Figure 5.2 – The power function $p(x;\boldsymbol{x})$ in the case of the periodic Sobolev space of orders $s = 1$ (left), $s = 2$ (middle) and $s = 3$ (right), compared to the diagonal of the kernel $x \mapsto k_s(x,x) = k_s(0,0)$ (the dashed line).

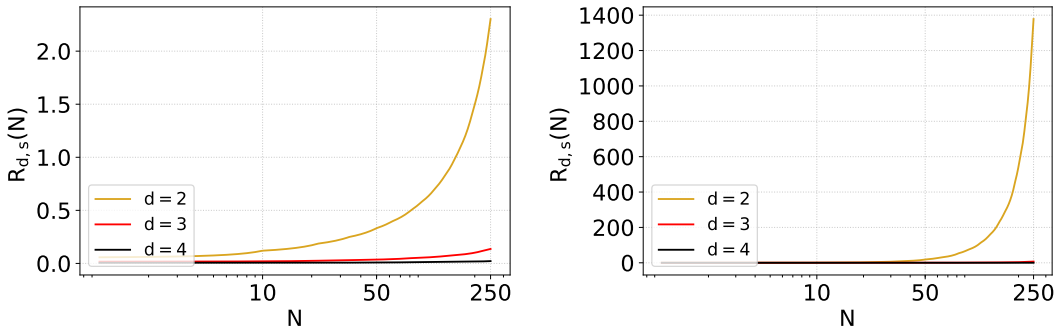


Figure 5.3 – $R_{d,s}(N)$ as a function of N for $s = 1$ (left) and $s = 2$ (right).

As we shall see in Section 5.4, we prove that the squared interpolation error, under continuous volume sampling, scales as $\mathcal{O}(\sigma_{N+1})$ for the embeddings μ_g. In other words, there is a trade-off between the approximation problem (reducing the interpolation error) and the sampling problem (reducing the complexity of sampling). We leave investigating alternative samplers that circumvent this trade-off for future work.

5.4 MAIN RESULTS

In this section, we give a theoretical analysis of kernel interpolation on nodes that follow the continuous volume sampling distribution. We state our main result in Section 5.4.1, an uniform-in-g upper bound of $\mathbb{E}_{\mathrm{VS}} \|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2$. We give an upper bound for a general $\mu \in \mathcal{F}$ in Section 5.4.2.

5.4.1 *The interpolation error for embeddings μ_g*

The main theorem of this chapter decomposes the expected error for an embedding μ_g in terms of the expected errors $\epsilon_m(N)$ for eigenfunctions of the kernel.

Theorem 5.1. *Let $g = \sum_{m\in\mathbb{N}^*} g_m e_m \in \mathbb{L}_2(\mathrm{d}\omega)$ that satisfies $\|g\|_{\mathrm{d}\omega} \leq 1$. Then under Assumption 5.3,*

$$\mathbb{E}_{\mathrm{VS}} \|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 = \sum_{m\in\mathbb{N}^*} g_m^2 \epsilon_m(N), \tag{5.22}$$

where

$$\epsilon_m(N) = \sigma_m \frac{\sum_{U\in\, \mathcal{U}_N^m} \prod_{u\in U} \sigma_u}{\sum_{U\in\, \mathcal{U}_N} \prod_{u\in U} \sigma_u}. \tag{5.23}$$

In particular, the sequence $(\epsilon_m(N))_{m\in\mathbb{N}^*}$ is non-increasing and

$$\sup_{\|g\|_{\mathrm{d}\omega}\leq 1} \mathbb{E}_{\mathrm{VS}} \|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 \leq \sup_{m\in\mathbb{N}^*} \epsilon_m(N) = \epsilon_1(N). \tag{5.24}$$

Moreover,

$$\epsilon_1(N) \leq \sigma_N \left(1 + \beta_N\right), \tag{5.25}$$

where $\beta_N = \min_{M\in[2:N]} \left[(N - M + 1)\sigma_N\right]^{-1} \sum_{m\geq M} \sigma_m$.

In other words, under continuous volume sampling, $\epsilon_1(N)$ is a uniform upper bound on the expected squared interpolation error of *any* embedding μ_g such that $\|g\|_{\mathrm{d}\omega} \leq 1$. We shall see in Section 5.5.1 that $\epsilon_m(N) = \mathbb{E}_{\mathrm{VS}} \|\mu_{e_m} - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\, \mu_{e_m}\|_{\mathcal{F}}^2$.

Now, for $N_0 \in \mathbb{N}^*$, a simple counting argument yields, for $m \geq N_0$, $\epsilon_m(N) \leq \sigma_{N_0}$. Actually, for $m \geq N_0$, $\|\mu_{e_m}\|_{\mathcal{F}}^2 \leq \sigma_{N_0}$, independently of the nodes.

Inequality (5.25) is less trivial and makes continuous volume sampling distribution worth of interest: the upper bound goes to 0 as $N \to +\infty$, below the initial error σ_{N_0}. Moreover, the convergence rate is $\mathcal{O}(\sigma_N)$, matching the lower bound of Section 5.2.3 if the sequence $(\beta_N)_{N\in\mathbb{N}^*}$ is bounded. In the following proposition, we prove that it is the case as soon as the spectrum decreases polynomially (e.g., Sobolev spaces of finite smoothness) or exponentially (e.g., the Gaussian kernel).

Proposition 5.3. *If $\sigma_m = m^{-2s}$ with $s > 1/2$ then*

$$\forall N \in \mathbb{N}^*,\ \beta_N \leq \left(1 + \frac{1}{2s-1}\right)\left(1 + \frac{1}{2s-1}\right)^{2s-1}. \tag{5.26}$$

If $\sigma_m = \alpha^m$, with $\alpha \in [0, 1[$, then

$$\forall N \in \mathbb{N}^*,\ \beta_N \leq \frac{\alpha}{1-\alpha}. \tag{5.27}$$

In both cases, the proof uses the fact that

$$\beta_N \leq [(N - M_N + 1)\sigma_N]^{-1} \sum_{m \geq M_N} \sigma_m, \tag{5.28}$$

for a well designed sequence M_N. For example, if $\sigma_m = m^{-2s}$, we take $M_N = \lceil N/c \rceil$ with $c > 1$; if $\sigma_m = \alpha^m$ we take $M_N = N$. We give a detailed proof in Section 5.8.

For a general kernel, if an asymptotic equivalent of σ_N is known (Widom, 1963; Widom, 1964; Bach, 2017), it should be possible to give an explicit construction of M_N. Indeed,

$$\beta_N \leq \frac{\sigma_{M_N}}{\sigma_N} + [(N - M_N + 1)\sigma_N]^{-1} \sum_{m \geq N+1} \sigma_m, \tag{5.29}$$

and M_N should be chosen to control both terms in the RHS. Figure 5.4 illustrates the upper bound of Theorem 5.1 and the constant of Proposition 5.3 in case of the periodic Sobolev space of order $s = 3$. We observe that $\mathbb{E}_{\mathrm{VS}}\, \mathcal{E}(\mu_{e_m}; \boldsymbol{x})^2$ respects the upper bound: it starts from the initial error level σ_m and decreases according to the upper bound for $N \geq m$.

5.4.2 *The interpolation error of any element of $\mathcal{F}$*

Theorem 5.1 dealt with the interpolation of an embedding μ_g of some function $g \in \mathbb{L}_2(\mathrm{d}\omega)$. We now give a bound on the interpolation error for any $\mu \in \mathcal{F}$. We need the following assumption, which is relatively weak regarding Proposition 5.3 and the discussion that follows.

Assumption 5.4. *There exists $B > 0$ such that $\beta_N \leq B$.*

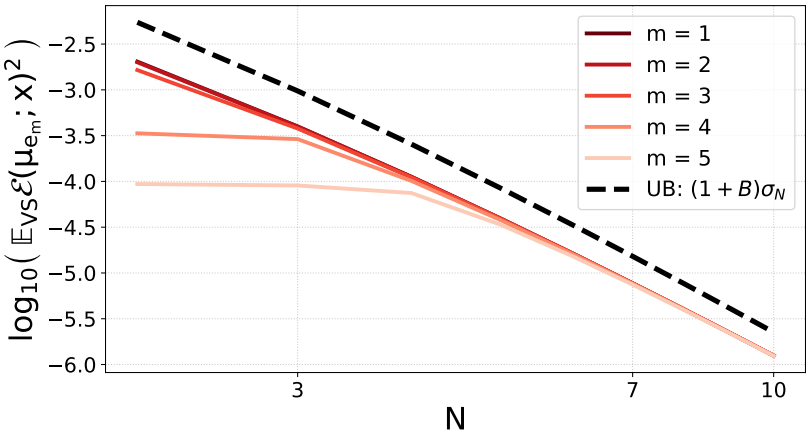


Figure 5.4 – The value of $\mathbb{E}_{\mathrm{VS}}\,\mathcal{E}(\mu_{e_m};\boldsymbol{x})^2$ for $m \in \{1,2,3,4,5\}$ for the periodic Sobolev space of order $s = 3$, compared to the theoretical upper bound (UB) of Theorem 5.1.

Theorem 5.2. *Let $\mu \in \mathcal{F}$. Assume that $\|\boldsymbol{\Sigma}^{-r}\mu\|_{\mathcal{F}}^2 \triangleq \sum_m \langle \mu, e_m^{\mathcal{F}}\rangle_{\mathcal{F}}^2/\sigma_m^{2r} < +\infty$ for some $r \in [0,1/2]$. Then, under Assumption 5.4,*

$$\mathbb{E}_{\mathrm{VS}}\,\mathcal{E}(\mu;\boldsymbol{x})^2 \leq (2+B)\sigma_N^{2r}\|\boldsymbol{\Sigma}^{-r}\mu\|_{\mathcal{F}}^2 = \mathcal{O}(\sigma_N^{2r}).$$

In other words, the expected interpolation error depends on the smoothness parameter r. For $r = 1/2$, we exactly recover the rate of Theorem 5.1. In contrast, for $r < 1/2$, the rate $\mathcal{O}(\sigma_N^{2r})$ is slower. For $r = 0$, our bound is constant with N. The condition $\|\boldsymbol{\Sigma}^{-r}\mu\|_{\mathcal{F}} < +\infty$ is satisfied by the elements of the fractional space

$$\boldsymbol{\Sigma}^{r+1/2}\mathbb{L}_2(\mathrm{d}\omega) = \left\{\boldsymbol{\Sigma}^{r+1/2}g;\ g \in \mathbb{L}(\mathrm{d}\omega)\right\}. \tag{5.30}$$

These subspaces interpolate between the RKHS $\mathcal{F}$ ($r = 0$) and the subspace of the embeddings $\boldsymbol{\Sigma}\mathbb{L}_2(\mathrm{d}\omega)$ ($r = 1/2$); see Figure 5.5.

Let us comment on this bound in two classical cases. First, consider the unidimensional Sobolev space of order s. Assumption 5.4 is satisfied by Proposition 5.3 and the squared error scales as $\mathcal{O}(N^{-4sr})$. Moreover, for this family of RKHSs, $\|\boldsymbol{\Sigma}^{-r}.\|_{\mathcal{F}}$ can be seen as the norm in the Sobolev space of order $(2r+1)s$, and we recover a result in (Schaback and Wendland, 2006)[Theorem 7.8] for quasi-uniform designs. By using the norm in the RKHS $\mathcal{F}$ of rougher functions, we upper bound the interpolation error of μ belonging to the smoother RKHS $\boldsymbol{\Sigma}^r\mathcal{F}$. Second, we emphasize again that our result is agnostic to the choice of the kernel, as long as Assumption 5.4 holds. In particular, Theorem 5.2 applies to the Gaussian kernel: the rate is slower $\mathcal{O}(\sigma_N^{2r})$ yet still exponential. Finally, recall that for $f \in \mathcal{F}$

$$|f(x)|^2 = |\langle f, k(x,.)\rangle_{\mathcal{F}}|^2 \leq \|f\|_{\mathcal{F}}^2 k(x,x), \tag{5.31}$$

so that, bounds on the RKHS norm imply bounds on the uniform norm if the kernel k is bounded. Therefore, for $r \in [0,1/2]$, our result improves on the rate $\mathcal{O}(N^2\sigma_N^{2r})$ of approximate Fekete points (Karvonen et al., 2019).

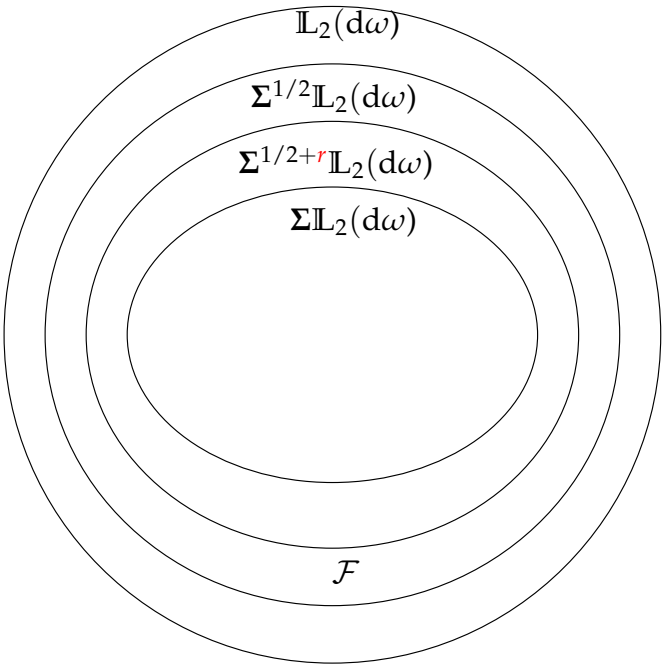


Figure 5.5 – The subspace $\boldsymbol{\Sigma}\mathbb{L}_2(\mathrm{d}\omega)$ is a particular case of the fractional spaces $\boldsymbol{\Sigma}^{1/2+r}\mathbb{L}_2(\mathrm{d}\omega)$, with $r \in [0, 1/2]$, that are subspaces of the RKHS $\mathcal{F} = \boldsymbol{\Sigma}^{1/2}\mathbb{L}_2(\mathrm{d}\omega)$.

5.4.3 *Asymptotic unbiasedness of kernel quadrature*

As explained in Chapter 4, kernel interpolation is widely used for the design of quadratures. In that setting, one more advantage of continuous volume sampling is the consistency of its estimator. This is the purpose of the following result.

Theorem 5.3. *Let $f \in \mathcal{F}$, and $g \in \mathbb{L}_2(\mathrm{d}\omega)$. Define the bias of the optimal kernel quadrature based on nodes that follow the continuous volume sampling distribution*

$$\mathcal{B}_N(f,g) = \mathbb{E}_{\mathrm{VS}}\left(\int_{\mathcal{X}} f(x)g(x)\,\mathrm{d}\omega(x) - \sum_{i\in N} \hat{w}_i f(x_i)\right), \tag{5.32}$$

then

$$\mathcal{B}_N(f,g) = \sum_{n\in\mathbb{N}^*} \langle f, e_n\rangle_{\mathrm{d}\omega}\langle g, e_n\rangle_{\mathrm{d}\omega}\left(1 - \mathbb{E}_{\mathrm{VS}}\,\tau_n^{\mathcal{F}}(\boldsymbol{x})\right). \tag{5.33}$$

Moreover, $\mathcal{B}_N(f,g) \to 0$ as $N \to +\infty$.

Compared to the upper bound on the integration error given by

$$\left|\int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x) - \sum_{n\in[N]} \hat{w}_n f(x_n)\right| \leq \|f\|_{\mathcal{F}}\|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(x)}\mu_g\|_{\mathcal{F}}, \tag{5.34}$$

the bias term in Theorem 5.3 takes into account the interaction between f and g. For example, if

$$\forall n \in \mathbb{N}^*, \langle f, e_n\rangle_{\mathrm{d}\omega}\langle g, e_n\rangle_{\mathrm{d}\omega} = 0, \tag{5.35}$$

the quadrature is unbiased for every N. In particular, when $g = e_{n_0}$, the estimator is unbiased for every function $f \in e_{n_0}^{\perp}$.

Using Theorem 5.3 we can draw the parallel with the determinantal representation (1.26) of the pseudo-inverse of a rectangular matrix. This parallelism is summarized in Table 5.2.

Discrete	Continuous
$\boldsymbol{y}$	μ_g
$\boldsymbol{X}$	$\boldsymbol{\Sigma}$
$[N]$	$\mathcal{X}$
S	$\boldsymbol{x}$
$\boldsymbol{w}_S^*$	$\sum_{n\in[N]} \hat{w}_n \delta_{x_n}$
$\boldsymbol{w}^*$	g
$\mathbb{E}_{S\sim \mathrm{VS}}\, \boldsymbol{w}_S^* = \boldsymbol{w}^*$	$\lim_{N\to+\infty} \mathbb{E}_{\boldsymbol{x}\sim \mathrm{CVS}} \sum_{n\in[N]} \hat{w}_n \delta_{x_n}(f) = \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x)$

Table 5.2 – A parallelism between the determinantal representation (1.26) of (Ben-Tal and Teboulle, 1990) and Theorem 5.3.

It will be interesting to investigate the implications of this (continuous) determinantal representation in the field of numerical methods of linear integral equations (Kress, 1999).

5.5 SKETCH OF THE PROOFS

The proof of Theorem 5.1 decomposes into three steps. First, in Section 5.5.1, we write $\mathcal{E}(\mu_g;\boldsymbol{x})^2$ as a function of the square of the interpolation errors $\mathcal{E}(\mu_{e_m};\boldsymbol{x})^2$ of the embeddings μ_{e_m}. Then, in Section 5.5.2, we give closed formulas for $\mathbb{E}_{\mathrm{VS}}\, \mathcal{E}(\mu_{e_m};\boldsymbol{x})^2$ in terms of the eigenvalues of $\boldsymbol{\Sigma}$. Finally, the inequality (5.25) is proved using an upper bound on the ratio of symmetric polynomials (Guruswami and Sinop, 2012). The details are given in Section 5.8.5. Finally, The proofs of Theorem 5.2 and Theorem 5.3 are straightforward consequences of Theorem 5.1. The details are given in Section 5.8.10 and Section 5.8.11.

5.5.1 *Decomposing the interpolation error*

Let $\boldsymbol{x} \in \mathcal{X}^N$ such that $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) > 0$. For $m_1, m_2 \in \mathbb{N}^*$, let the *cross-leverage score* between m_1 and m_2 associated to $\boldsymbol{x}$ be

$$\tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x}) = e^{\mathcal{F}}_{m_1}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} e^{\mathcal{F}}_{m_2}(\boldsymbol{x}). \tag{5.36}$$

When $m_1 = m_2 = m$, we speak of the m-th leverage score[2] associated to $\boldsymbol{x}$, and simply write $\tau^{\mathcal{F}}_m(\boldsymbol{x})$. By Lemma 5.2, the m-th leverage score is related to the interpolation error of the m-th eigenfunction $e^{\mathcal{F}}_m$. Indeed,

$$\|e^{\mathcal{F}}_m - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e^{\mathcal{F}}_m\|^2_{\mathcal{F}} = 1 - \tau^{\mathcal{F}}_m(\boldsymbol{x}) \in [0,1]. \tag{5.37}$$

[2] Our definition is consistent with the leverage scores used in matrix subsampling (Drineas et al., 2006). Loosely speaking, $\tau^{\mathcal{F}}_m(\boldsymbol{x})$ is the leverage score of the m-th column of the semi-infinite matrix $(e^{\mathcal{F}}_n(x_i))_{(i,n)\in[N]\times\mathbb{N}^*}$.

Similarly, for the cross-leverage score,

$$\langle \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})} e^{\mathcal{F}}_{m_1}, \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})} e^{\mathcal{F}}_{m_2} \rangle_{\mathcal{F}} = \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x}) \in [-1,1]. \tag{5.38}$$

For $g \in \mathbb{L}_2(\mathrm{d}\omega)$, the interpolation error of the embedding μ_g can be expressed using the (cross-)leverage scores.

Lemma 5.1. *If* $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) > 0$, *then,*

$$\mathcal{E}(\mu_g; \boldsymbol{x})^2 = \sum_{m \in \mathbb{N}^*} g_m^2 \sigma_n \left(1 - \tau^{\mathcal{F}}_m(\boldsymbol{x})\right) - \sum_{m_1 \neq m_2 \in \mathbb{N}^*} g_{m_1} g_{m_2} \sqrt{\sigma_{m_1}} \sqrt{\sigma_{m_2}} \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x}). \tag{5.39}$$

In particular, with probability one, a configuration $\boldsymbol{x}$ sampled from the continuous volume sampling distribution in Definition 5.1 satisfies (5.39). Furthermore, we shall see that the expected value of the (cross-) leverage scores has a simple expression.

5.5.2 *Explicit formulas for expected leverage scores*

Proposition 5.4 expresses expected leverage scores in terms of the spectrum of the integration operator.

Proposition 5.4. *For* $m \in \mathbb{N}^*$,

$$\mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_m(\boldsymbol{x}) = \frac{1}{\sum\limits_{U \in \mathcal{U}_N} \prod\limits_{u \in U} \sigma_u} \sum_{\substack{U \in \mathcal{U}_N \\ m \in U}} \prod_{u \in U} \sigma_u. \tag{5.40}$$

Moreover, for $m_1, m_2 \in \mathbb{N}^*$ *such that* $m_1 \neq m_2$, *we have*

$$\mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x}) = 0. \tag{5.41}$$

In Section 5.8.5, we combine Lemma 5.1 with Proposition 5.4. This concludes the proof of Theorem 5.1 by Beppo Levi's monotone convergence theorem.

It remains to prove Proposition 5.4. Again, we proceed in two steps. First, our Proposition 5.5 yields a characterization of $\mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_m(\boldsymbol{x})$ and $\mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x})$ in terms of the spectrum of three perturbed versions of the integration operator $\mathbf{\Sigma}$. Second, we give explicit forms of these spectra in Proposition 5.6 below. The idea is to express $\mathbb{E}_{\mathrm{VS}}\, \tau_m(\boldsymbol{x})^{\mathcal{F}}$ as the normalization constant (5.12) of a perturbation of the kernel k. The same goes for $\mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x})$.

Let $t \in \mathbb{R}_+$ and $\mathbf{\Sigma}_t$, $\mathbf{\Sigma}_t^+$ and $\mathbf{\Sigma}_t^-$ be the integration operators[3] on $\mathbb{L}_2(\mathrm{d}\omega)$, respectively associated to the kernels

$$k_t(x,y) = k(x,y) + t e^{\mathcal{F}}_m(x) e^{\mathcal{F}}_m(y), \tag{5.42}$$

$$k_t^+(x,y) = k(x,y) + t \left(e^{\mathcal{F}}_{m_1}(x) + e^{\mathcal{F}}_{m_2}(x)\right)\left(e^{\mathcal{F}}_{m_1}(y) + e^{\mathcal{F}}_{m_2}(y)\right), \tag{5.43}$$

$$k_t^-(x,y) = k(x,y) + t \left(e^{\mathcal{F}}_{m_1}(x) - e^{\mathcal{F}}_{m_2}(x)\right)\left(e^{\mathcal{F}}_{m_1}(y) - e^{\mathcal{F}}_{m_2}(y)\right). \tag{5.44}$$

[3] We drop from the notation the dependencies on m, m_1 and m_2 for simplicity.

By Assumption 5.3, and by the fact that $(e_m)_{m\in\mathbb{N}^*}$ is an orthonormal basis of $\mathbb{L}_2(\mathrm{d}\omega)$, all three kernels also have integrable diagonals (see Assumption 5.3). In particular, they define RKHSs that can be embedded in $\mathbb{L}_2(\mathrm{d}\omega)$. Moreover, recalling the definition (5.12) of the normalization constant Z_N of continuous volume sampling, the following quantities are finite

$$\phi_m(t) = \frac{1}{N!} Z_N(k_t), \quad \phi^+_{m_1,m_2}(t) = \frac{1}{N!} Z_N(k^+_t), \text{ and } \quad \phi^-_{m_1,m_2}(t) = \frac{1}{N!} Z_N(k^-_t). \tag{5.45}$$

Remember that by Proposition 5.1,

$$\phi_m(t) = N! \sum_{U\in\, \mathcal{U}_N} \prod_{u\, \in U} \tilde{\sigma}_u(t), \tag{5.46}$$

where $\{\tilde{\sigma}_u(t),\, u \in \mathbb{N}^*\}$ is the set of eigenvalues[4] of $\mathbf{\Sigma}_t$. Similar identities are valid for $\phi^+_{m_1,m_2}(t)$ and $\phi^-_{m_1,m_2}(t)$ with the eigenvalues of $\mathbf{\Sigma}^+_t$ and $\mathbf{\Sigma}^-_t$ respectively.

Proposition 5.5. *The functions ϕ_m, $\phi^+_{m_1,m_2}$ and $\phi^-_{m_1,m_2}$ are right differentiable in zero. Furthermore,*

$$\mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_m(\boldsymbol{x}) = \frac{1}{Z_N(k)} \frac{\partial \phi_m}{\partial t}\bigg|_{t=0^+},$$

and

$$\mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x}) = \frac{1}{4Z_N(k)} \left(\frac{\partial \phi^+_{m_1,m_2}}{\partial t} - \frac{\partial \phi^-_{m_1,m_2}}{\partial t} \right)\bigg|_{t=0^+}.$$

The details of the proof are postponed to Section 5.8.8. We complete this proposition with a description of the spectrum of the operators $\mathbf{\Sigma}_t$, $\mathbf{\Sigma}^+_t$ and $\mathbf{\Sigma}^-_t$ using the spectrum of $\mathbf{\Sigma}$.

Proposition 5.6. *The eigenvalues of $\mathbf{\Sigma}_t$ write*

$$\tilde{\sigma}_u(t) = \begin{cases} \sigma_u & \text{if } u \neq m, \\ (1+t)\sigma_u & \text{if } u = m. \end{cases} \tag{5.47}$$

Moreover, the eigenvalues of $\mathbf{\Sigma}^+_t$ and $\mathbf{\Sigma}^-_t$ satisfy

$$\{\tilde{\sigma}^+_u(t),\, u \in \mathbb{N}^*\} = \{\tilde{\sigma}^-_u(t),\, u \in \mathbb{N}^*\}. \tag{5.48}$$

The proof is based on the observation that the perturbations in (5.42), (5.43), and (5.44) only affect a principal subspace of dimension 1 or 2; see Section 5.8.7.

Combining the characterization of $\mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_m(\boldsymbol{x})$ and $\mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x})$ given in Proposition 5.5, and Proposition 5.6, we prove Proposition 5.4; see details in Section 5.8.9.

[4] For a given value of t, the eigenvalues $\tilde{\sigma}_u(t)$ are not necessarily decreasing in u. We give explicit formulas for these eigenvalues in Proposition 5.6, and the order satisfied for $t = 0$ is not necessarily preserved for $t > 0$. This does not change anything to the argument since these eigenvalues only appear in quantities such as $\phi_m(t)$ which are invariant under permutation of the eigenvalues.

5.6 NUMERICAL SIMULATIONS

5.6.1 *Comparing the $\epsilon_m(N)$ to the σ_N.*

In this section, we illustrate the bound of Theorem 5.1 and the constants of Proposition 5.3 on more examples. Remember that by Theorem 5.1:

$$\mathbb{E}_{\mathrm{VS}}\, \mathcal{E}(\mu_{e_m}; \boldsymbol{x})^2 = \epsilon_m(N), \tag{5.49}$$

and

$$\mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x}) = 1 - \frac{\epsilon_m(N)}{\sigma_m}, \tag{5.50}$$

where

$$\epsilon_m(N) = \sigma_m \frac{\sum\limits_{U \in\, \mathcal{U}_N^m} \prod\limits_{u \in U} \sigma_u}{\sum\limits_{U \in\, \mathcal{U}_N} \prod\limits_{u \in U} \sigma_u}. \tag{5.51}$$

We compare in the following $\epsilon_m(N)$ to σ_N in the two cases treated in Proposition 5.3: i) the case $\sigma_m = m^{-2s}$ for some $s > 1/2$, ii) the case $\sigma_m = \alpha^m$ for some $\alpha \in [0, 1[$.

For numerical simulations, we make the following approximation

$$\epsilon_m(N) \approx \sigma_m \left(\sum_{\substack{U \subset [M] \\ |U| = N}} \prod_{u \in U} \sigma_u \right)^{-1} \sum_{\substack{U \subset [M] \\ |U| = N,\, m \notin U}} \prod_{u \in U} \sigma_u, \tag{5.52}$$

for an $M \geq N$ sufficiently large. The numerator and denominator of the right hand side of (5.52) can be calculated efficiently using Vieta's formulas.

Eigenvalues with polynomial decay

Consider

$$\forall m \in \mathbb{N}^*,\ \sigma_m = m^{-2s}, \tag{5.53}$$

with $s \in \{1, 2, 3, 4, 5\}$. Figure 5.6 illustrates the expected value of the m-th leverage score $\mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x})$ (left panels) and the expected interpolation error $\mathbb{E}_{\mathrm{VS}}\, \mathcal{E}(\mu_{e_m}; \boldsymbol{x})^2$ (right panels), both as functions of N for different values of m.

We observe that for low values of s, $\mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x})$ depends smoothly on N. On the other hand, $\mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x})$ undergoes a sharp transition at $N = m$ for high values of s: the reconstruction of the m-th eigenfunction is almost perfect for N slightly larger than m. Moreover, $\mathbb{E}_{\mathrm{VS}}\, \mathcal{E}(\mu_{e_m}; \boldsymbol{x})^2$ respects the upper bound of Theorem 5.1; the constant B of Proposition 5.3 is small for high values of s and converges to e when $s \to +\infty$.

Eigenvalues with exponential decay

Consider now

$$\forall m \in \mathbb{N}^*,\ \sigma_m = \alpha^N, \tag{5.54}$$

with $\alpha \in \{0.7, 0.5, 0.2\}$.

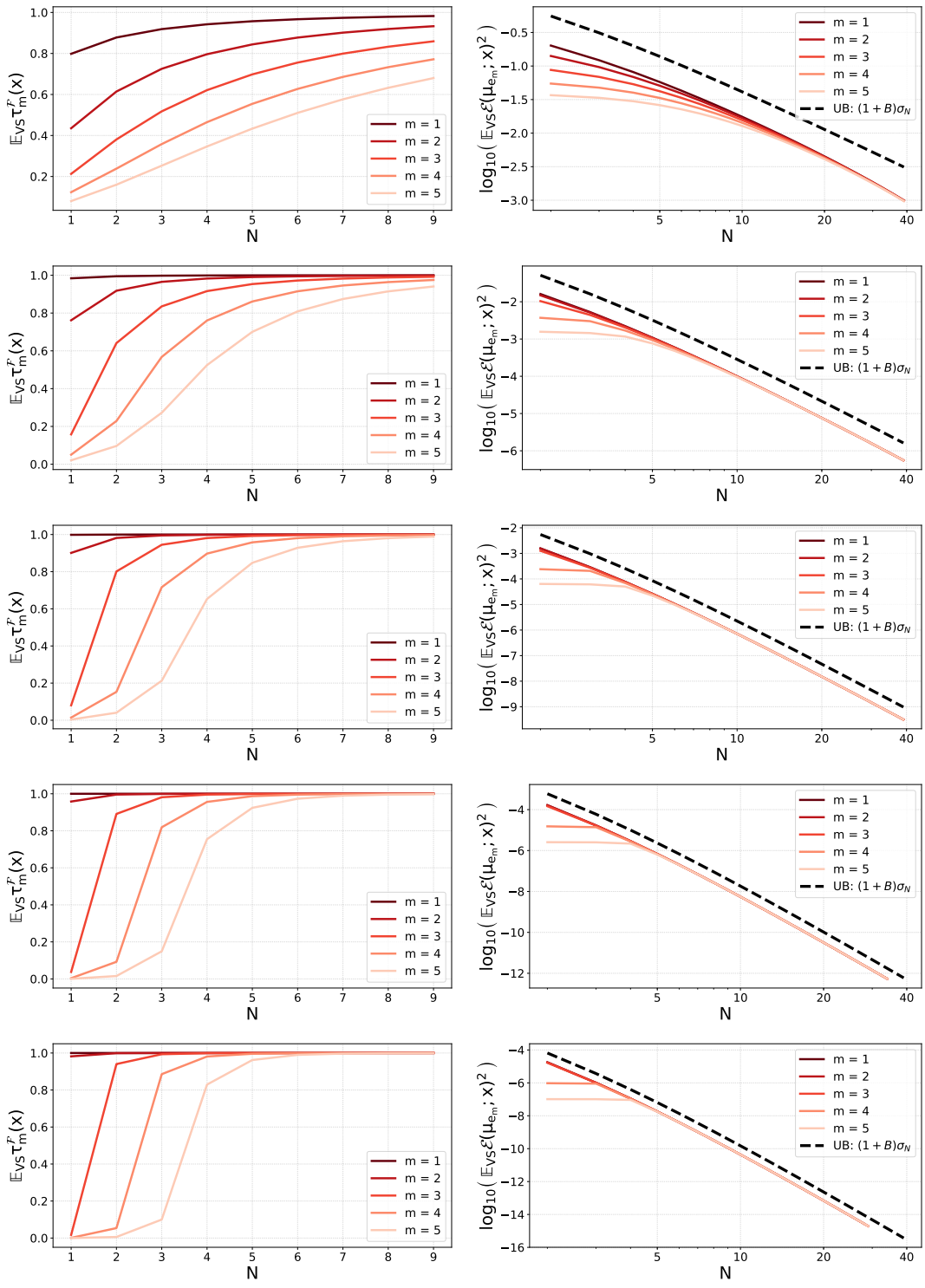


Figure 5.6 – The expected value of the m-th leverage score $\mathbb{E}_{\mathrm{VS}}\,\tau_m^{\mathcal{F}}(\boldsymbol{x})$ (left panels) and the expected interpolation error $\mathbb{E}_{\mathrm{VS}}\,\mathcal{E}(\mu_{e_m};\boldsymbol{x})^2$ (right panels), under the distribution of continuous volume sampling, for $m \in \{1,2,3,4,5\}$ and the uni-variate periodic Sobolev kernel. Rows correspond to increasing values of the smoothness parameter $s = 1,2,3,4,5$.

Figure 5.7 illustrates the expected value of the m-th leverage score $\mathbb{E}_{\mathrm{VS}}\,\tau_m^{\mathcal{F}}(\boldsymbol{x})$ (left panels) and the expected interpolation error $\mathbb{E}_{\mathrm{VS}}\,\mathcal{E}(\mu_{e_m};\boldsymbol{x})^2$ (right panels), both as functions of N, for different values of m.

We make the same observations on the dependency of $\mathbb{E}_{\mathrm{VS}}\,\tau_m^{\mathcal{F}}(\boldsymbol{x})$ on N as in the case of polynomially decaying spectrum. The rougher the kernel (i.e., the lower the value of α), the smoother the transition of $\mathbb{E}_{\mathrm{VS}}\,\tau_m^{\mathcal{F}}(\boldsymbol{x})$ as a function of N. Moreover, $\mathbb{E}_{\mathrm{VS}}\,\mathcal{E}(\mu_{e_m};\boldsymbol{x})^2$ respects the upper bound of Theorem 5.1; the constant B of Proposition 5.3 is small for low values of α and converges to 0 when $\alpha \to 0$.

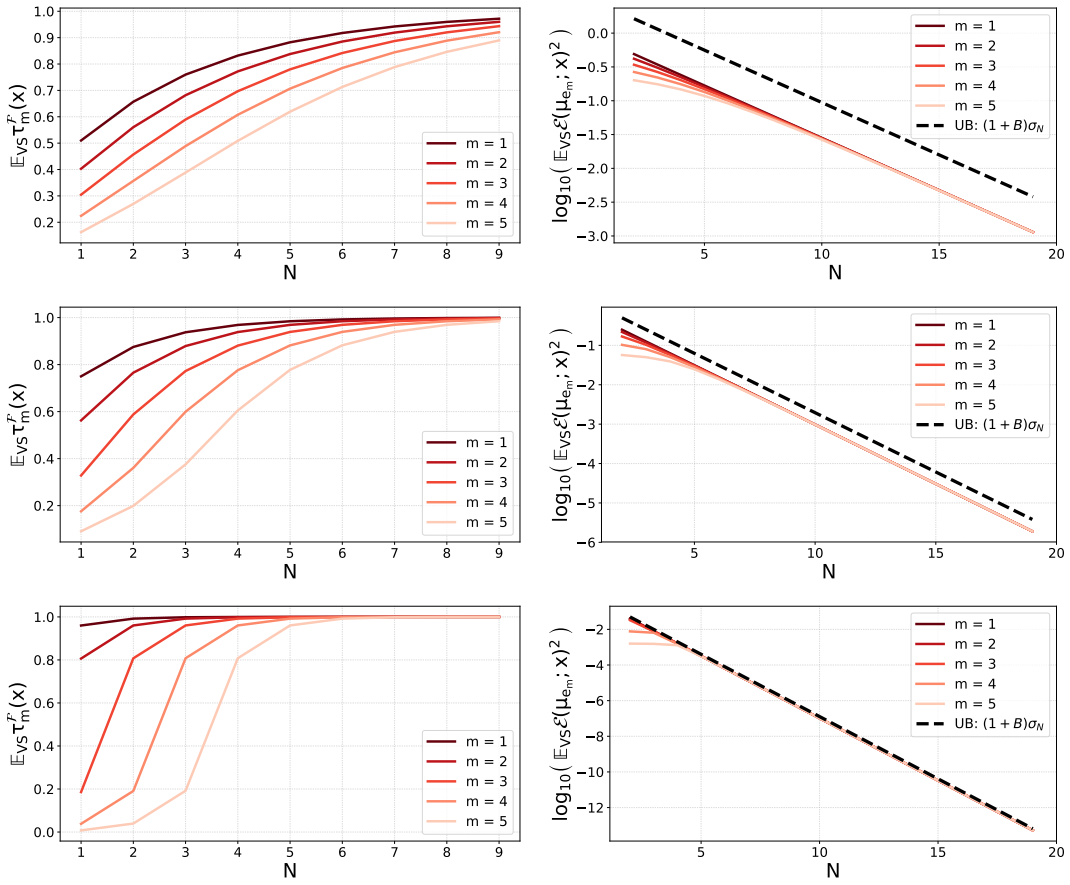


Figure 5.7 – The expected value of the m-th leverage score $\mathbb{E}_{\mathrm{VS}}\,\tau_m^{\mathcal{F}}(\boldsymbol{x})$ and the expected interpolation error $\mathbb{E}_{\mathrm{VS}}\,\mathcal{E}(\mu_{e_m};\boldsymbol{x})^2$ under the distribution of continuous volume sampling for $m \in \{1,2,3,4,5\}$. Every row corresponds to a uni-dimensional Gaussian space ($\sigma_m = \alpha^m$) with a parameter $\alpha \in \{0.7, 0.5, 0.2\}$.

5.6.2 *An MCMC simulation*

To illustrate Theorem 5.1, we In Theorem 5.1, (5.22) decomposes the expected interpolation error of any μ_g in terms of the interpolation error $\epsilon_m(N)$ of the μ_{e_m}. Therefore, it is sufficient to numerically check the values of the $\epsilon_m(N)$. As an illustration we consider $g \in \{e_1, e_5, e_7\}$ in (5.22), so that $\mu_{e_m} = \boldsymbol{\Sigma} e_m = \sigma_m e_m$, with $m \in \{1,5,7\}$. We use the Gibbs sampler proposed by (Rezaei and Gharan, 2019) to approximate continuous volume sampling.

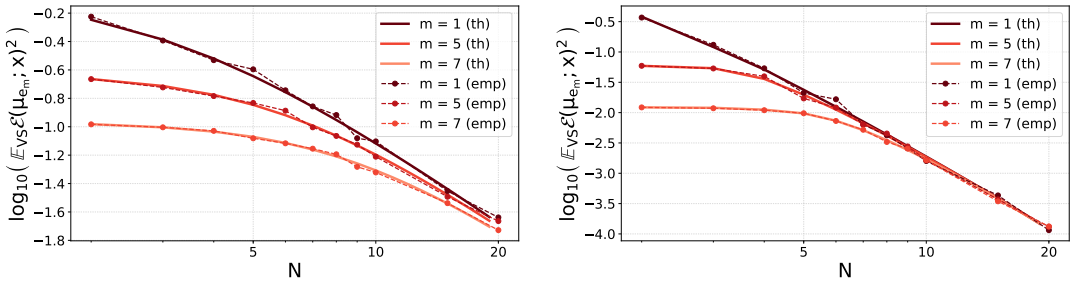


Figure 5.8 – The empirical estimate of $\mathbb{E}_{\mathrm{VS}}\,\mathcal{E}(\mu_{e_m};\boldsymbol{x})^2$ for $m \in \{1,5,7\}$ compared to its expression (5.22) in the case of the periodic Sobolev space. The smoothness is $s=1$ (left), $s=2$ (right).

We consider various numbers of points $N \in [2,20]$. Figure 5.8 shows log-log plots of the theoretical (th) value of $\mathbb{E}_{\mathrm{VS}}\,\|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2$ compared to its empirical (emp) counterpart, vs. N, for $s \in \{1,2\}$. For each N, the estimate is an average over 50 independent samples, each sample resulting from 500 individual Gibbs iterations.

For both values of the smoothness parameter s, we observe a close fit of the estimate with the actual expected error: the mismatch between the two is due to approximate sampling.

5.7 DISCUSSION

We deal with interpolation in RKHSs using random nodes and optimal weights. This problem is intimately related to kernel quadrature, though interpolation is more general. We introduced continuous volume sampling (CVS), a repulsive point process that is a mixture of DPPs, although not a DPP itself. CVS comes with a set of advantages. First, interpretable bounds on the interpolation error can be derived under minimalistic assumptions. Our bounds are close to optimal since they share the same decay rate as known lower bounds. Moreover, we provide explicit evaluations of the constants appearing in our bounds for some particular RKHSs (e.g., Sobolev, Gaussian).

Second, while the eigen-decomposition of the integration operator plays an important role in the analysis, the definition of the density function of volume sampling only involves kernel evaluations. In that sense, CVS is a fully kernelized approach. Unlike previous work on random design, this may permit sampling without knowing the Mercer decomposition of the kernel (Rezaei and Gharan, 2019), as demonstrated in Section 5.6.

We highlighted the trade-off between the interpolation problem and the sampling problem. In particular, approximate continuous volume sampling given by Algorithm 5.1 may replace greedy maximization in high-dimensional domains.

Investigating other efficient samplers and their impact on bounds is deferred to future work; the current chapter is a theoretical motivation for further methodological research.

5.8 PROOFS

5.8.1 *Technical results borrowed from other papers*

Thoughout our proof, we use a few technical results from the literature, which we gather here for ease of reference.

The Jacobi identity

The following proposition is a direct consequence of the rank one-update for determinants.

Proposition 5.7 (Jacobi identity). *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{N\times N}$. If $\operatorname{Det} \boldsymbol{A} \neq 0$, then*

$$\partial_t \operatorname{Det}(\boldsymbol{A} + t\boldsymbol{B})|_{t=0} = \operatorname{Det}(\boldsymbol{A}) \operatorname{Tr}(\boldsymbol{A}^{-1}\boldsymbol{B}). \tag{5.55}$$

In particular, we have

$$\partial_t \operatorname{Det}(\boldsymbol{A} + t\boldsymbol{B})|_{t=0^+} = \operatorname{Det}(\boldsymbol{A}) \operatorname{Tr}(\boldsymbol{A}^{-1}\boldsymbol{B}). \tag{5.56}$$

The Markov brothers' inequality

The following proposition is known as the Markov brother's inequality, see e.g. (Shadrin, 2004).

Proposition 5.8 (Markov brothers). *Let P be a polynomial of degree smaller than N. Then*

$$\max_{\tau\in[-1,1]} |P'(\tau)| \leq N^2 \max_{\tau\in[-1,1]} |P(\tau)|. \tag{5.57}$$

We shall actually use a straightforward corollary.

Corollary 5.1. *Let P be a polynomial of degree smaller than N. Then*

$$\max_{\tau\in[0,1]} |P'(\tau)| \leq 2N^2 \max_{\tau\in[0,1]} |P(\tau)|. \tag{5.58}$$

Proof. Define the polynomial $Q(x) = P((x+1)/2)$, so that

$$Q'(x) = \frac{1}{2} P'((x+1)/2), \quad x \in [-1,1]. \tag{5.59}$$

In particular,

$$\max_{\tau\in[0,1]} |P(\tau)| = \max_{\tau\in[-1,1]} |Q(\tau)|,$$

so that

$$\max_{\tau\in[0,1]} |P'(\tau)| = \max_{\tau\in[-1,1]} 2|Q'(\tau)| \leq 2N^2 \max_{\tau\in[-1,1]} |Q(\tau)| \leq 2N^2 \max_{\tau\in[0,1]} |P(\tau)|. \tag{5.60}$$

□

An inequality on the ratio of symmetric polynomials

Recall that, for $d \in \mathbb{N}^*$, $\mathbb{R}^d$ is naturally embedded in the set of sequences $\mathbb{R}^{\mathbb{N}^*}$.

Now, let $M \in \mathbb{N}^*$, and let $\boldsymbol{\lambda} \in \mathbb{R}_+^{\mathbb{N}^*}$ such that $\sum_{m \in \mathbb{N}^*} \lambda_m < +\infty$. By MacLaurin's inequality [5], see e.g. (Steele, 2004, Chapter 12),

$$\forall M \in \mathbb{N}^*, \sum_{U \in \mathcal{U}_M} \prod_{u \in U} \lambda_u \leq \frac{1}{M!} \left(\sum_{m \in \mathbb{N}^*} \lambda_m \right)^M < +\infty. \tag{5.61}$$

In the following, we denote by $p_M(\boldsymbol{\lambda})$ the elementary symmetric polynomial of order M on the sequence $\boldsymbol{\lambda}$,

$$p_M(\boldsymbol{\lambda}) = \sum_{U \in \mathcal{U}_M} \prod_{u \in U} \lambda_u. \tag{5.62}$$

In particular, the following identity relates p_M and p_{M+1}.

$$\forall M \geq 2, \forall m \in \mathbb{N}^*, p_M(\boldsymbol{\lambda}) = \lambda_m p_{M-1}(\boldsymbol{\lambda}^{\overline{\{m\}}}) + p_M(\boldsymbol{\lambda}^{\overline{\{m\}}}), \tag{5.63}$$

where we denote, for $S \subset \mathbb{N}^*$, $\boldsymbol{\lambda}^{\overline{S}} = (\lambda_m^{\overline{S}})_{m \in \mathbb{N}^*} = (\lambda_m \mathbb{1}_{\{m \notin S\}})_{m \in \mathbb{N}^*}$. Proposition 5.9 further relates two consecutive elementary polynomials.

Proposition 5.9 (Theorem 3.1 of Guruswami and Sinop, 2012)**.** *Let $M \in \mathbb{N}^*$ and $L \geq M+1$. Let $\boldsymbol{\lambda} \in \mathbb{R}_+^L$ be a nonincreasing sequence*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_L. \tag{5.64}$$

Assume that $\lambda_L > 0$, then

$$\forall M' \leq M, \quad \frac{p_{M+1}(\boldsymbol{\lambda})}{p_M(\boldsymbol{\lambda})} \leq \frac{\sum_{m \geq M'+1} \lambda_m}{M+1-M'}. \tag{5.65}$$

We will actually use an immediate consequence of Proposition 5.9.

Corollary 5.2. *Let $M \in \mathbb{N}^*$ and $\boldsymbol{\lambda} \in \mathbb{R}_+^{\mathbb{N}^*}$ be a nonincreasing sequence such that $\sum \lambda_m < +\infty$ and $\lambda_m > 0$ for all $m \in \mathbb{N}^*$. Then (5.65) still holds.*

Proof. Define, for $L \in \mathbb{N}^*$,

$$\boldsymbol{\lambda}_L = (\lambda_\ell)_{\ell \in [L]} \in \mathbb{R}_+^L. \tag{5.66}$$

By Proposition 5.9,

$$\forall M' \leq M, \ \forall L \geq M+1, \ \frac{p_{M+1}(\boldsymbol{\lambda}_L)}{p_M(\boldsymbol{\lambda}_L)} \leq \frac{1}{M+1-M'} \sum_{m=M'+1}^{L} \lambda_m \tag{5.67}$$

$$\leq \frac{1}{M+1-M'} \sum_{m=M'+1}^{+\infty} \lambda_m. \tag{5.68}$$

Letting $L \to \infty$ allows us to conclude. □

For the last result, recall the definition of the (cross-)leverage scores τ_{m_1,m_2} in (5.36). We slightly adapt a result by Belhadji et al., 2019a .

[5] The inequality is usually stated for $\boldsymbol{\lambda} \in \mathbb{R}_+^d$ for some $d \in \mathbb{N}^*$. Taking limits immediatedly yields (5.61).

Lemma 5.2. *Let $\boldsymbol{x} \in \mathcal{X}^N$ satisfy $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) > 0$. For $m, m_1, m_2 \in \mathbb{N}^*$ such that $m_1 \neq m_2$,*

$$\tau_m^{\mathcal{F}}(\boldsymbol{x}) = \|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_m^{\mathcal{F}}\|_{\mathcal{F}}^2 = e_m^{\mathcal{F}}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} e_m^{\mathcal{F}}(\boldsymbol{x}), \tag{5.69}$$

and

$$\tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x}) = \langle \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_{m_1}^{\mathcal{F}}, \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_{m_2}^{\mathcal{F}} \rangle_{\mathcal{F}} = e_{m_1}^{\mathcal{F}}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} e_{m_2}^{\mathcal{F}}(\boldsymbol{x}). \tag{5.70}$$

In particular,

$$\tau_m^{\mathcal{F}}(\boldsymbol{x}) \text{ and } |\tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x})| \text{ are in } [0,1]. \tag{5.71}$$

Proof. The proof is straightforward following the same lines as the proof of Lemma 4.4.

The operator $\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}$ is an orthogonal projection with respect to $\langle ., . \rangle_{\mathcal{F}}$ and

$$\|e_m^{\mathcal{F}}\|_{\mathcal{F}} = \|e_{m_1}^{\mathcal{F}}\|_{\mathcal{F}} = \|e_{m_2}^{\mathcal{F}}\|_{\mathcal{F}} = 1, \tag{5.72}$$

so that (5.71) follows from the Cauchy-Schwarz inequality. □

5.8.2 *Proof of Proposition 5.1*

Proposition 5.1 states that continuous volume sampling is a mixture of projection determinantal point processes. We adapt a result in (Kulesza and Taskar, 2012, Chapter 5) for finite volume sampling to the infinite-dimensional case. The idea of the proof is to apply the Cauchy-Binet identity to a sequence of kernels of finite rank that approximate k.

First, recall from Section 5.1 the Mercer decomposition of k,

$$k(x,y) = \lim_{M\to\infty} \sum_{m\in[M]} \sigma_m e_m(x) e_m(y) = \lim_{M\to\infty} k_M(x,y), \quad \forall x, y \in \mathcal{X}. \tag{5.73}$$

where kernel k_M has rank M.

Now, let $\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathcal{X}^N$, and define $\boldsymbol{K}_M(\boldsymbol{x}) = (k_M(x_i, x_j))_{i,j\in[N]}$. By continuity of the determinant and by (5.73), it comes

$$\lim_{M\to\infty} \operatorname{Det} \boldsymbol{K}_M(\boldsymbol{x}) = \operatorname{Det} \boldsymbol{K}(\boldsymbol{x}). \tag{5.74}$$

By construction,

$$\boldsymbol{K}_M(\boldsymbol{x}) = \boldsymbol{F}_M(\boldsymbol{x})^\intercal \boldsymbol{\Sigma}_M \boldsymbol{F}_M(\boldsymbol{x}), \tag{5.75}$$

where $\boldsymbol{F}_M(\boldsymbol{x}) = (e_m(x_i))_{(m,i)\in[M]\times[N]}$ and Σ_M is a diagonal matrix containing the first M eigenvalues $(\sigma_m)_{m\in[M]}$ on its diagonal. The Cauchy-Binet identity yields

$$\operatorname{Det} \boldsymbol{K}_M(\boldsymbol{x}) = \sum_{\substack{U\subset[M] \\ |U|=N}} \operatorname{Det}^2 (e_u(x_i))_{(u,i)\in U\times[N]} \prod_{u\in U} \sigma_u. \tag{5.76}$$

Let now $\lambda_u = \prod_{u\in U} \sigma_u$ and $\boldsymbol{E}_U(\boldsymbol{x}) = (e_u(x_i))_{(u,i)\in U\times[N]}$, we combine (5.74) and (5.76) to obtain

$$\begin{aligned}
\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) &= \lim_{M\to\infty} \sum_{\substack{U\subset[M]\\|U|=N}} \lambda_u \operatorname{Det}^2(e_u(x_i))_{(u,i)\in U\times[N]} && (5.77)\\
&= \sum_{U\in\mathcal{U}_N} \lambda_u \operatorname{Det}^2(e_u(x_i))_{(u,i)\in U\times[N]} && (5.78)\\
&= \sum_{U\in\mathcal{U}_N} \lambda_u \operatorname{Det}\left(\boldsymbol{E}_U(\boldsymbol{x})^\intercal \boldsymbol{E}_U(\boldsymbol{x})\right) && (5.79)\\
&= \sum_{U\in\mathcal{U}_N} \lambda_u \operatorname{Det}(\mathfrak{K}_U(x_i,x_j))_{i,j\in[N]}, && (5.80)
\end{aligned}$$

where $\mathfrak{K}_U(x,y) \triangleq \sum_{u\in U} e_u(x)e_u(y)$. Since $\mathfrak{K}_U$ is a projection kernel, writing the determinant as a sum over permutations easily yields, for all $U \in \mathcal{U}_N$,

$$\int_{\mathcal{X}^N} \operatorname{Det}(\mathfrak{K}_U(x_i,x_j))_{i,j\in[N]} \otimes_{i\in[N]} \mathrm{d}\omega(x_i) = N!, \tag{5.81}$$

see e.g. Lemma 21 in (Hough et al., 2006). Finally, the monotone convergence theorem allows us to conclude

$$\int_{\mathcal{X}^N} \operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) \otimes_{i\in[N]} \mathrm{d}\omega(x_i) = N! \sum_{\substack{U\subset\mathbb{N}^*\\|U|=N}} \prod_{u\in U} \sigma_u. \tag{5.82}$$

5.8.3 *Proof of Proposition 5.2*

Proposition 5.2 gives an upper bound on the biggest weight δ_N in the mixture of Proposition 5.1. The proof is straightforward, as

$$\begin{aligned}
r_N \prod_{\ell\in[N]} \sigma_\ell &= \sigma_N \sum_{m\geq N+1} \sigma_m \prod_{\ell\in[N-1]} \sigma_\ell\\
&\leq \sigma_N \sum_{\substack{U\subset\mathbb{N}^*\\|U|=N}} \prod_{u\in U} \sigma_u. && (5.83)
\end{aligned}$$

This immediately yields $\delta_N \leq \sigma_N / r_N$.

5.8.4 *Proof of Lemma 5.1*

Lemma 5.1 decomposes the interpolation error in terms of (cross-)leverage scores. Let $g \in \mathbb{L}_2(\mathrm{d}\omega)$ satisfy $\|g\|_{\mathrm{d}\omega} \leq 1$. Since $\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}$ is an orthogonal projection with respect to $\langle .,. \rangle_{\mathcal{F}}$, we have

$$\|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 = \|\mu_g\|_{\mathcal{F}}^2 - \|\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 \tag{5.84}$$

Now, $\mu_g = \boldsymbol{\Sigma} g = \sum_{m\in\mathbb{N}^*} \sqrt{\sigma_m} g_m e_m^{\mathcal{F}}$, so that (5.84) becomes

$$\begin{aligned}
\|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 &= \sum_{m\in\mathbb{N}^*} \sigma_m g_m^2 - \left\| \sum_{m\in\mathbb{N}^*} \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} \sqrt{\sigma_m} g_m e_m^{\mathcal{F}} \right\|_{\mathcal{F}}^2\\
&= \sum_{m\in\mathbb{N}^*} \sigma_m g_m^2 - \sum_{m_1,m_2} g_{m_1} g_{m_2} \sqrt{\sigma_{m_1}} \sqrt{\sigma_{m_2}} \langle \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_{m_1}^{\mathcal{F}}, \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_{m_2}^{\mathcal{F}} \rangle_{\mathcal{F}}. && (5.85)
\end{aligned}$$

Lemma 5.2 allows us to recognize leverage scores in (5.85). Taking out of the second sum in (5.85) the terms for which $m_1 = m_2$ to put them in the first sum concludes the proof of Lemma 5.1.

5.8.5 *Proof of Theorem 5.1*

The proof of (5.22) relies on the identity

$$\mathbb{E}_{\mathrm{VS}} \|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 = \sum_{m\in\mathbb{N}^*} g_m^2 \epsilon_m(N), \tag{5.86}$$

and the fact that $(\epsilon_m(N))$ is a non-increasing sequence. We prove these two results in turn, after what we prove (5.25).

Proof of (5.86)

Let $\boldsymbol{x} \in \mathcal{X}^N$ such that $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) > 0$. Lemma 5.1 yields

$$\|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 = \sum_{m\in\mathbb{N}^*} g_m^2\sigma_m\Big(1 - \tau_m^{\mathcal{F}}(\boldsymbol{x})\Big) - \sum_{\substack{m_1,m_2\in\mathbb{N}^*\\ m_1\neq m_2}} g_{m_1}g_{m_2}\sqrt{\sigma_{m_1}}\sqrt{\sigma_{m_2}}\tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x}). \tag{5.87}$$

First, we prove that

$$\mathbb{E}_{\mathrm{VS}} \sum_{m\in\mathbb{N}^*} g_m^2\sigma_m\Big(1 - \tau_m^{\mathcal{F}}(\boldsymbol{x})\Big) = \sum_{m\in\mathbb{N}^*} g_m^2\sigma_m\Big(1 - \mathbb{E}_{\mathrm{VS}}\,\tau_m^{\mathcal{F}}(\boldsymbol{x})\Big). \tag{5.88}$$

By Lemma 5.2,

$$\forall m \in \mathbb{N}^*,\ g_m^2\sigma_m\Big(1 - \tau_m^{\mathcal{F}}(\boldsymbol{x})\Big) \geq 0, \tag{5.89}$$

so that (5.88) follows from the Beppo Levi's monotone convergence theorem.

Second, it remains to prove that

$$\mathbb{E}_{\mathrm{VS}} \sum_{\substack{m_1,m_2\in\mathbb{N}^*\\ m_1\neq m_2}} g_{m_1}g_{m_2}\sqrt{\sigma_{m_1}}\sqrt{\sigma_{m_2}}\tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x}) = 0. \tag{5.90}$$

Again, Lemma 5.2 guarantees that, for $m_1, m_2 \in \mathbb{N}^*$ such that $m_1 \neq m_2$,

$$|g_{m_1}g_{m_2}\sqrt{\sigma_{m_1}}\sqrt{\sigma_{m_2}}\tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x})| \leq |g_{m_1}g_{m_2}|\sqrt{\sigma_{m_1}}\sqrt{\sigma_{m_2}}. \tag{5.91}$$

Since

$$\begin{aligned}\sum_{m_1\neq m_2\in\mathbb{N}^*} |g_{m_1}g_{m_2}|\sqrt{\sigma_{m_1}}\sqrt{\sigma_{m_2}} &\leq \left(\sum_{m\in\mathbb{N}^*} |g_m|\sqrt{\sigma_m}\right)^2\\ &\leq \sum_{m\in\mathbb{N}^*} g_m^2 \sum_{m\in\mathbb{N}^*} \sigma_m\\ &< +\infty,\end{aligned} \tag{5.92}$$

the dominated convergence theorem yields

$$\mathbb{E}_{\mathrm{VS}} \sum_{\substack{m_1,m_2\in\mathbb{N}^*\\ m_1\neq m_2}} g_{m_1}g_{m_2}\sqrt{\sigma_{m_1}}\sqrt{\sigma_{m_2}}\tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x}) = \sum_{\substack{m_1,m_2\in\mathbb{N}^*\\ m_1\neq m_2}} g_{m_1}g_{m_2}\sqrt{\sigma_{m_1}}\sqrt{\sigma_{m_2}}\,\mathbb{E}_{\mathrm{VS}}\,\tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x}),$$

but this is equal to zero by Proposition 5.4.

Proof that $(\epsilon_m(N))$ is nonincreasing

Let $m \in \mathbb{N}^*$. By definition,

$$\epsilon_m(N) = \sigma_m \frac{\sum_{U \in \mathcal{U}_N^m} \prod_{u \in U} \sigma_u}{\sum_{U \in \mathcal{U}_N} \prod_{u \in U} \sigma_u} = \sigma_m \frac{p_N(\sigma^{\overline{\{m\}}})}{p_N(\sigma)}, \tag{5.93}$$

where we use a notation introduced in Section 5.8.1. This leads to

$$\epsilon_m(N) = \sigma_m \frac{\sigma_{m+1} p_{N-1}(\sigma^{\overline{\{m,m+1\}}}) + p_N(\sigma^{\overline{\{m,m+1\}}})}{p_N(\sigma)}, \tag{5.94}$$

and, similarly,

$$\epsilon_{m+1}(N) = \sigma_{m+1} \frac{\sigma_m p_{N-1}(\sigma^{\overline{\{m,m+1\}}}) + p_N(\sigma^{\overline{\{m,m+1\}}})}{p_N(\sigma)}. \tag{5.95}$$

Taking the ratio, it comes

$$\frac{\epsilon_m(N)}{\epsilon_{m+1}(N)} = \frac{\sigma_m \left(\sigma_{m+1} p_{N-1}\left(\sigma^{\overline{\{m,m+1\}}}\right) + p_N\left(\sigma^{\overline{\{m,m+1\}}}\right) \right)}{\sigma_{m+1} \left(\sigma_m p_{N-1}\left(\sigma^{\overline{\{m,m+1\}}}\right) + p_N\left(\sigma^{\overline{\{m,m+1\}}}\right) \right)} \tag{5.96}$$

$$= \frac{1 + \frac{1}{\sigma_{m+1}} \frac{p_N\left(\sigma^{\overline{\{m,m+1\}}}\right)}{p_{N-1}\left(\sigma^{\overline{\{m,m+1\}}}\right)}}{1 + \frac{1}{\sigma_m} \frac{p_N\left(\sigma^{\overline{\{m,m+1\}}}\right)}{p_{N-1}\left(\sigma^{\overline{\{m,m+1\}}}\right)}} \geq 1, \tag{5.97}$$

because $1/\sigma_{m+1} \geq 1/\sigma_m$.

Proof of (5.25)

We have $\epsilon_1(N) = \epsilon_N(N)\epsilon_1(N)/\epsilon_N(N) \leq \sigma_N \epsilon_1(N)/\epsilon_N(N)$ since a simple counting argument yields $\epsilon_N(N) \leq \sigma_N$. Along the lines of Section 5.8.5,

$$\frac{\epsilon_1(N)}{\epsilon_N(N)} = \frac{1 + \frac{1}{\sigma_N} \frac{p_N\left(\sigma^{\overline{\{1,N\}}}\right)}{p_{N-1}\left(\sigma^{\overline{\{1,N\}}}\right)}}{1 + \frac{1}{\sigma_1} \frac{p_N\left(\sigma^{\overline{\{1,N\}}}\right)}{p_{N-1}\left(\sigma^{\overline{\{1,N\}}}\right)}} \leq 1 + \frac{1}{\sigma_N} \frac{p_N\left(\sigma^{\overline{\{1,N\}}}\right)}{p_{N-1}\left(\sigma^{\overline{\{1,N\}}}\right)}. \tag{5.98}$$

Now, $\sigma^{\overline{\{1,N\}}}$ is a sequence of positive real numbers and the $\mathbf{\Sigma}$ is trace-class. Then, by Corollary 5.2, for $M \in [N-1]$,

$$\frac{p_N\left(\sigma^{\overline{\{1,N\}}}\right)}{p_{N-1}\left(\sigma^{\overline{\{1,N\}}}\right)} \leq \frac{1}{N-M} \sum_{m \geq M} \sigma_{m+2} = \frac{1}{N+1-(M+1)} \sum_{m+1 \geq M+1} \sigma_{m+2}. \tag{5.99}$$

Taking $M' = M+1$ concludes the proof of (5.25).

5.8.6 *Proof of Proposition 5.3*

The case of a polynomially-decreasing spectrum

Assume that $\sigma_m = m^{-2s}$ with $s > 1/2$. Let $N \in \mathbb{N}^*$ and $M_N = \lceil N/c \rceil \in \{2, \dots, N\}$, with $c \in [1, N[$. We have

$$\min_{M \in [2:N]} \frac{\sum_{m \geq M} \sigma_{m+1}}{(N - M + 1)\sigma_N} \leq \frac{\sum_{m \geq M_N} \sigma_{m+1}}{(N - M_N + 1)\sigma_N} \tag{5.100}$$

$$\leq \frac{\sum_{m \geq \lceil N/c \rceil} \sigma_{m+1}}{(N - \lceil N/c \rceil + 1)\sigma_N} \tag{5.101}$$

$$\leq \frac{\sum_{m \geq \lceil N/c \rceil} \sigma_{m+1}}{(N - N/c + 1)\sigma_N} \tag{5.102}$$

$$\leq \frac{\sum_{m \geq \lceil N/c \rceil} (m+1)^{-2s}}{(N - N/c + 1)N^{-2s}}. \tag{5.103}$$

$$\tag{5.104}$$

Now,

$$\forall m \in \mathbb{N}^*, \ (m+1)^{-2s} \leq \int_m^{m+1} t^{-2s} \mathrm{d}t = \frac{1}{2s-1}(m^{1-2s} - (m+1)^{1-2s}), \tag{5.105}$$

so that

$$\sum_{m \geq \lceil N/c \rceil} (m+1)^{-2s} \leq \frac{1}{2s-1} \lceil N/c \rceil^{1-2s}. \tag{5.106}$$

Recall that $2s > 1$, so that

$$\frac{1}{2s-1} \lceil N/c \rceil^{1-2s} \leq \frac{1}{2s-1} (N/c)^{1-2s}, \tag{5.107}$$

and

$$\min_{M \in [2:N]} \frac{\sum_{m \geq M} \sigma_{m+1}}{(N - M + 1)\sigma_N} \leq \frac{1}{2s-1} \frac{(N/c)^{1-2s}}{(N - N/c + 1)N^{-2s}} \tag{5.108}$$

$$\leq \frac{c^{2s}}{2s-1} \frac{N}{(cN - N + c)}. \tag{5.109}$$

Note that c is a free parameter that belongs to $[1, N]$ [6] that we can optimize in the upper bound: $\frac{c^{2s}}{2s-1} \frac{N}{(cN-N+c)}$. For this purpose, denote

$$\phi_N(c) = \frac{c^{2s}}{2s-1} \frac{N}{(cN - N + c)}. \tag{5.110}$$

For every $N \in \mathbb{N}^*$, ϕ_N is differentiable in $]0, +\infty[$ and

$$\phi_N'(c) = \frac{N}{2s-1} \frac{c^{2s-1}}{(cN - N + c)^2} \left((2s-1)(N+1)c - 2sN \right), \tag{5.111}$$

so that ϕ_N' vanishes in $c_N^* = \frac{2s}{2s-1} \frac{N}{N+1}$; it is negative in $]0, c_N^*[$ and positive in $]c_N^*, +\infty[$. We distinguish three cases:

[6] The inequality in (5.109) is valid for $c = N$ by continuity.

If $c_N^* < 1$, $N < 2s-1$ and ϕ_N' is positive on $[1,N]$ so that ϕ_N increases in $[1,N]$ and we take $c=1$ in (5.109):

$$\phi_N(1) = \frac{N}{2s-1} < 1. \tag{5.112}$$

If $c_N^* \in [1,N]$, c_N^* is the unique minimizer of ϕ_N in $[1,N]$ and we take $c = c_N^*$ in (5.109) so that:

$$\phi_N(c_N^*) = \left(\frac{2s}{2s-1}\right)^{2s} \left(\frac{N}{N+1}\right)^{2s} \tag{5.113}$$

$$\leq \left(\frac{2s}{2s-1}\right)^{2s} \tag{5.114}$$

$$\leq \left(1+\frac{1}{2s-1}\right)\left(1+\frac{1}{2s-1}\right)^{2s-1}. \tag{5.115}$$

Finally, if $c_N^* > N$, $N < \frac{1}{2s-1}$, ϕ_N is decreasing in $[1,N]$ and we take $c=N$ in (5.109) so that:

$$\phi_N(N) = \frac{N^{2s-1}}{2s-1} \leq \frac{1}{2s-1}\left(\frac{1}{2s-1}\right)^{2s-1} \tag{5.116}$$

$$\leq \left(1+\frac{1}{2s-1}\right)\left(1+\frac{1}{2s-1}\right)^{2s-1}. \tag{5.117}$$

In the three cases, β_N is upper bounded by $\left(1+\frac{1}{2s-1}\right)\left(1+\frac{1}{2s-1}\right)^{2s-1}$. The artificial two-factor form of (5.115) and (5.117) is there to make limits clearer. In particular, the RHS goes to e as $s \to \infty$.

The case of an exponentially decreasing spectrum

Assume that $\sigma_m = \alpha^m$ with $\alpha \in [0,1[$. Let $N \in \mathbb{N}^*$, and $M_N = N \in \{2,\dots,N\}$. We have

$$\min_{M\in[2:N]} \frac{\sum_{m\geq M}\sigma_{m+1}}{(N-M+1)\sigma_N} \leq \frac{\sum_{m\geq M_N}\sigma_{m+1}}{(N-M_N+1)\sigma_N} \tag{5.118}$$

$$\leq \frac{\sum_{m\geq N}\sigma_{m+1}}{\sigma_N} \tag{5.119}$$

$$\leq \frac{\sum_{m\geq N}\alpha^{m+1}}{\alpha^N} \tag{5.120}$$

$$\leq \alpha^{N+1}\frac{\sum_{m\geq 0}\alpha^m}{\alpha^N} \tag{5.121}$$

$$\leq \frac{\alpha}{1-\alpha}. \tag{5.122}$$

5.8.7 *Proof of Proposition 5.6*

We start with deriving the spectrum of the trace-class, self-adjoint operator

$$\mathbf{\Sigma}_t = \mathbf{\Sigma} + t e_m^{\mathcal{F}} \otimes e_m^{\mathcal{F}}, \tag{5.123}$$

where $e_m^{\mathcal{F}} \otimes e_m^{\mathcal{F}}$ is defined by

$$\forall g \in \mathbb{L}_2(\mathrm{d}\omega),\ \ e_m^{\mathcal{F}} \otimes e_m^{\mathcal{F}}\, g(\cdot) = e_m^{\mathcal{F}}(\cdot)\int_{\mathcal{X}} g(y) e_m^{\mathcal{F}}(y)\mathrm{d}\omega(y). \tag{5.124}$$

The two operators $\mathbf{\Sigma}$ and $e_m^{\mathcal{F}} \otimes e_m^{\mathcal{F}}$ are co-diagonalizable in the basis $(e_m)_{m\in\mathbb{N}^*}$, thus their linear combination $\mathbf{\Sigma}_t$ diagonalizes in this basis too. In other words, for $u \in \mathbb{N}^*$, e_u is an eigenfunction of $\mathbf{\Sigma}_t$ and

$$\mathbf{\Sigma}_t e_u = \mathbf{\Sigma} e_u + t e_m^{\mathcal{F}} \otimes e_m^{\mathcal{F}}(e_u) = (\sigma_u + t\delta_{u,m}\sigma_u)e_u. \tag{5.125}$$

Therefore, the set $\{\sigma_u(1 + t\delta_{u,m}), u \in \mathbb{N}^*\}$ is included in the spectrum of $\mathbf{\Sigma}_t$. Since $(e_m)_{m\in\mathbb{N}^*}$ is an orthonormal basis of $\mathbb{L}_2(\mathrm{d}\omega)$ and correspond to the eigenfunctions of $\mathbf{\Sigma}_t$ associated to the elements of $\{\sigma_u(1 + t\delta_{u,m}), u \in \mathbb{N}^*\}$, then the spectrum of $\mathbf{\Sigma}_t$ is exactly the set $\{\sigma_u(1 + t\delta_{u,m}), u \in \mathbb{N}^*\}$.[7] We now turn to deriving the spectrum of the trace-class, self-adjoint operator $\mathbf{\Sigma}_t^+$; the case of $\mathbf{\Sigma}_t^-$ follows the same lines and will be omitted for brevity. We will prove that there exists an orthonormal basis $(f_m)_{m\in\mathbb{N}^*}$ of $\mathbb{L}_2(\mathrm{d}\omega)$ such that every f_m is an eigenfunction of $\mathbf{\Sigma}_t^+$. If $t = 0$, $\mathbf{\Sigma}_t^+ = \mathbf{\Sigma}$ and $(e_m)_{m\in\mathbb{N}^*}$ is already an orthonormal basis of $\mathbb{L}_2(\mathrm{d}\omega)$. We assume in the following that $t > 0$.

Consider the operator Δ_t^+ defined on $\mathbb{L}_2(\mathrm{d}\omega)$ by

$$\Delta_t^+ g(\cdot) = t\left(e_{m_1}^{\mathcal{F}}(\cdot) + e_{m_2}^{\mathcal{F}}(\cdot)\right)\int_{\mathcal{X}} g(y)\left(e_{m_1}^{\mathcal{F}}(y) + e_{m_2}^{\mathcal{F}}(y)\right)\mathrm{d}\omega(y). \tag{5.126}$$

We can write $\mathbf{\Sigma}_t^+ = \mathbf{\Sigma} + \Delta_t^+$, but this time, if $t > 0$, $\mathbf{\Sigma}$ and Δ_t^+ do not commute. In particular, they are not co-diagonalizable, and a more detailed analysis is necessary. First, by construction of Δ_t^+,

$$\Delta_t^+ e_m = 0, \quad m \notin \{m_1, m_2\},$$

so that for any $m \notin \{m_1, m_2\}$, $\mathbf{\Sigma}_t^+$ and $\mathbf{\Sigma}$ have e_m for eigenfunction, with the same eigenvalue σ_m. Observe that

$$\mathbb{L}_2(\mathrm{d}\omega) = \mathrm{Span}(e_{m_1}, e_{m_2}) \oplus \mathrm{Span}(e_m)_{m\notin\{m_1,m_2\}}. \tag{5.127}$$

Therefore, the rest of the proof consists in completing $(e_m)_{m\notin\{m_1,m_2\}}$ into an orthonormal basis of $\mathbb{L}_2(\mathrm{d}\omega)$, by finding two orthonormal eigenfunctions of $\mathbf{\Sigma}_t^+$ in $\mathrm{Span}(e_{m_1}, e_{m_2})$. Since we assumed in Section 5.1 that the eigenvalues of $\mathbf{\Sigma}$ are nonzero, we note that $\mathrm{Span}(e_{m_1}, e_{m_2}) = \mathrm{Span}(e_{m_1}^{\mathcal{F}}, e_{m_2}^{\mathcal{F}})$. Expressing the new eigenfunctions in terms of $e_{m_1}^{\mathcal{F}}$ and $e_{m_2}^{\mathcal{F}}$ will turn out to be more convenient.

First, note that

$$\begin{aligned}
\mathbf{\Sigma}_t^+ e_{m_1}^{\mathcal{F}}(\cdot) &= \mathbf{\Sigma} e_{m_1}^{\mathcal{F}}(\cdot) + t\int_{\mathcal{X}}\left(e_{m_1}^{\mathcal{F}}(\cdot) + e_{m_2}^{\mathcal{F}}(\cdot)\right)\left(e_{m_1}^{\mathcal{F}}(y) + e_{m_2}^{\mathcal{F}}(y)\right)e_{m_1}^{\mathcal{F}}(y)\mathrm{d}\omega(y) && (5.128)\\
&= \sigma_{m_1} e_{m_1}^{\mathcal{F}}(\cdot) + t\sigma_{m_1}\left(e_{m_1}^{\mathcal{F}}(\cdot) + e_{m_2}^{\mathcal{F}}(\cdot)\right) && (5.129)\\
&= (1+t)\sigma_{m_1} e_{m_1}^{\mathcal{F}} + t\sigma_{m_1} e_{m_2}^{\mathcal{F}}. && (5.130)
\end{aligned}$$

Similarly,

$$\mathbf{\Sigma}_t^+ e_{m_2}^{\mathcal{F}}(.) = t\sigma_{m_2} e_{m_1}^{\mathcal{F}} + (1+t)\sigma_{m_2} e_{m_2}^{\mathcal{F}}. \tag{5.131}$$

[7] $\mathbf{\Sigma}_t$ is self-adjoint, and has no zero eigenvalue by assumption. Thus, any new eigenfunction that is not in our basis needs to be orthogonal to all basis elements, and is thus zero.

Now, let $v = \lambda_1 e^{\mathcal{F}}_{m_1} + \lambda_2 e^{\mathcal{F}}_{m_2}$, so that, by (5.130) and (5.131),

$$\begin{aligned}\boldsymbol{\Sigma}^+_t v &= \lambda_1 \Big((1+t)\sigma_{m_1} e^{\mathcal{F}}_{m_1} + t\sigma_{m_1} e^{\mathcal{F}}_{m_2} \Big) + \lambda_2 \Big((1+t)\sigma_{m_2} e^{\mathcal{F}}_{m_2} + t\sigma_{m_2} e^{\mathcal{F}}_{m_1} \Big) \\ &= \Big(\lambda_1 (1+t)\sigma_{m_1} + \lambda_2 t\sigma_{m_2} \Big) e^{\mathcal{F}}_{m_1} + \Big(\lambda_2 (1+t)\sigma_{m_2} + \lambda_1 t\sigma_{m_1} \Big) e^{\mathcal{F}}_{m_2}, \end{aligned} \tag{5.132}$$

Solving for eigenvalues, we look for $\mu \in \mathbb{R}$ such that $\boldsymbol{\Sigma}^+_t v = \mu v$, or equivalently

$$\left\{ \begin{array}{rcl} (1+t)\sigma_{m_1}\lambda_1 + t\sigma_{m_2}\lambda_2 &=& \mu\lambda_1, \\ t\sigma_{m_1}\lambda_1 + (1+t)\sigma_{m_2}\lambda_2 &=& \mu\lambda_2. \end{array} \right.$$

This is just saying that μ should be an eigenvalue of the matrix

$$\begin{pmatrix} (1+t)\sigma_{m_1} & t\sigma_{m_2} \\ t\sigma_{m_1} & (1+t)\sigma_{m_2} \end{pmatrix}, \tag{5.133}$$

which yields two solutions,

$$\mu^+_1 = (1+t)\frac{\sigma_{m_1} + \sigma_{m_2}}{2} + \frac{1}{2}\sqrt{(1+t)^2(\sigma_{m_1} - \sigma_{m_2})^2 + 4\sigma_{m_1}\sigma_{m_2}t^2}, \tag{5.134}$$

and

$$\mu^+_2 = (1+t)\frac{\sigma_{m_1} + \sigma_{m_2}}{2} - \frac{1}{2}\sqrt{(1+t)^2(\sigma_{m_1} - \sigma_{m_2})^2 + 4\sigma_{m_1}\sigma_{m_2}t^2}. \tag{5.135}$$

These solutions are distinct since $t > 0$, and the corresponding normalized eigenfunctions v^+_1 and v^+_2 are orthogonal with respect to $\langle ., . \rangle_{\mathrm{d}\omega}$ since $\boldsymbol{\Sigma}^+_t$ is self-adjoint. Finally, we define the set of eigenfunctions of $\boldsymbol{\Sigma}^+_t$ by the system $(e_m)_{m \notin \{m_1, m_2\}} \cup (v^+_1, v^+_2)$ that is an orthonormal basis of $\mathbb{L}_2(\mathrm{d}\omega)$. Therefore, the spectrum of the compact operator $\boldsymbol{\Sigma}^+_t$ is exactly the set

$$\{\sigma_m,\ m \notin \{m_1, m_2\}\} \cup \{\mu^+_1, \mu^+_2\}. \tag{5.136}$$

Along the same lines, one can show that the eigenvalues of $\boldsymbol{\Sigma}^-_t$ restricted to $\mathrm{Span}(e^{\mathcal{F}}_{m_1}, e^{\mathcal{F}}_{m_2})$ satisfy

$$\lambda^2 - (1+t)(\sigma_{m_1} + \sigma_{m_2})\lambda - \sigma_{m_1}\sigma_{m_2}t^2 = 0. \tag{5.137}$$

For $t > 0$, this equation again admits two distinct solutions

$$\hat{\mu}^-_1 = (1+t)\frac{\sigma_{m_1} + \sigma_{m_2}}{2} + \frac{1}{2}\sqrt{(1+t)^2(\sigma_{m_1} - \sigma_{m_2})^2 + 4\sigma_{m_1}\sigma_{m_2}t^2}, \tag{5.138}$$

and

$$\hat{\mu}^-_2 = (1+t)\frac{\sigma_{m_1} + \sigma_{m_2}}{2} - \frac{1}{2}\sqrt{(1+t)^2(\sigma_{m_1} - \sigma_{m_2})^2 + 4\sigma_{m_1}\sigma_{m_2}t^2}. \tag{5.139}$$

so that the spectrum of $\boldsymbol{\Sigma}^-_t$ is exactly the set

$$\{\sigma_m,\ m \notin \{m_1, m_2\}\} \cup \{\mu^-_1, \mu^-_2\} = \{\sigma_m,\ m \notin \{m_1, m_2\}\} \cup \{\mu^+_1, \mu^+_2\}. \tag{5.140}$$

In other words, the two operators $\boldsymbol{\Sigma}^+_t$ and $\boldsymbol{\Sigma}^-_t$ share the same eigenvalues.

5.8.8 *Proof of Proposition 5.5*

The expected value of the m-th leverage score

Let $m \in \mathbb{N}^*$. On the one hand, recall that $\tau_m^{\mathcal{F}}(\boldsymbol{x}) = e_m^{\mathcal{F}}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} e_m^{\mathcal{F}}(\boldsymbol{x})$, so that, by Definition 5.1,

$$\mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x}) = \left(N! \sum_{U \in \mathcal{U}_N} \prod_{u \in U} \sigma_u \right)^{-1} \int_{\mathcal{X}^N} e_m^{\mathcal{F}}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} e_m^{\mathcal{F}}(\boldsymbol{x}) \operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) \otimes_{i \in [N]} \mathrm{d}\omega(x_i). \tag{5.141}$$

We have

$$\begin{aligned} \operatorname{Det} \boldsymbol{K}(\boldsymbol{x})\, e_m^{\mathcal{F}}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} e_m^{\mathcal{F}}(\boldsymbol{x}) &= \operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) \operatorname{Tr}\left(e_m^{\mathcal{F}}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} e_m^{\mathcal{F}}(\boldsymbol{x}) \right) \\ &= \operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) \operatorname{Tr}\left(\boldsymbol{K}(\boldsymbol{x})^{-1} e_m^{\mathcal{F}}(\boldsymbol{x}) e_m^{\mathcal{F}}(\boldsymbol{x})^\intercal \right) \\ &= \partial_t \operatorname{Det}(\boldsymbol{K}(\boldsymbol{x}) + t e_m^{\mathcal{F}}(\boldsymbol{x}) e_m^{\mathcal{F}}(\boldsymbol{x})^\intercal)|_{t=0^+}, \end{aligned} \tag{5.142}$$

where the last line follows from the Jacobi identity of Theorem 5.7.

On the other hand, for $t > 0$ and with the notation of Section 5.5.2, let

$$\boldsymbol{K}_t(\boldsymbol{x}) := (k_t(x_i, x_j))_{i,j \in [N]} = \boldsymbol{K}(\boldsymbol{x}) + t e_m^{\mathcal{F}}(\boldsymbol{x}) e_m^{\mathcal{F}}(\boldsymbol{x})^\intercal. \tag{5.143}$$

Since

$$\begin{aligned} \int_{\mathcal{X}} k_t(x,x) \mathrm{d}\omega(x) &= \int_{\mathcal{X}} k(x,x) \mathrm{d}\omega(x) + t \int_{\mathcal{X}} e_m^{\mathcal{F}}(x)^2 \mathrm{d}\omega(x) \\ &= \sum_{n \in \mathbb{N}^*} \sigma_n + t\sigma_m < \infty, \end{aligned} \tag{5.144}$$

Hadamard's inequality yields the integrability of $\psi(.,t) : \boldsymbol{x} \mapsto \operatorname{Det} \boldsymbol{K}_t(\boldsymbol{x})$. Finally, observe that

$$\phi_m(t) := Z_N(k_t) = \int_{\mathcal{X}^N} \psi(\boldsymbol{x}, t) \otimes_{i \in [N]} \mathrm{d}\omega(x_i). \tag{5.145}$$

If we prove that ϕ_m is right differentiable in zero, and that we can justify the interchange of the derivation and the integration operations, we will have equated the right derivative of ϕ_m in zero and (5.141) using (5.142); this will achieve proving the first equation in Proposition 5.5. To this purpose, we need to prove that $t \mapsto \psi(\boldsymbol{x}, t)$ is right differentiable at zero, it is locally dominated by an integrable function and its derivative is locally dominated by an integrable function. Now, observe that $t \mapsto \psi(\boldsymbol{x}, t)$ is a polynomial of degree smaller than N, so that it is differentiable, and Corollary 5.1 yields

$$\max_{\tau \in [0,1]} |\partial_t \psi(\boldsymbol{x}, \tau)| \leq 2N^2 \max_{\tau \in [0,1]} |\psi(\boldsymbol{x}, \tau)|. \tag{5.146}$$

In other words, to dominate $\tau \mapsto |\partial_t \psi(\boldsymbol{x}, \tau)|$ uniformly on $[0,1]$, it is sufficient to dominate $\tau \mapsto |\psi(\boldsymbol{x}, \tau)|$ uniformly there. Now, let $\tau \in [0,1]$, we have

$$\boldsymbol{K}_1(\boldsymbol{x}) - \boldsymbol{K}_\tau(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x}) + e_m^{\mathcal{F}}(\boldsymbol{x}) e_m^{\mathcal{F}}(\boldsymbol{x})^\intercal - \boldsymbol{K}(\boldsymbol{x}) - \tau e_m^{\mathcal{F}}(\boldsymbol{x}) e_m^{\mathcal{F}}(\boldsymbol{x})^\intercal \tag{5.147}$$

$$= (1 - \tau) e_m^{\mathcal{F}}(\boldsymbol{x}) e_m^{\mathcal{F}}(\boldsymbol{x})^\intercal \in \mathcal{S}_N^+. \tag{5.148}$$

Thus

$$0 \preceq \boldsymbol{K}_\tau(\boldsymbol{x}) \preceq \boldsymbol{K}_1(\boldsymbol{x}) \tag{5.149}$$

in the Loewner order, so that for any $\tau \in [0,1]$,

$$|\psi(\boldsymbol{x},\tau)| = \psi(\boldsymbol{x},\tau) = \operatorname{Det} \boldsymbol{K}_\tau(\boldsymbol{x}) \leq \operatorname{Det} \boldsymbol{K}_1(\boldsymbol{x}) = \psi(\boldsymbol{x},1). \tag{5.150}$$

We conclude by observing that $\boldsymbol{x} \mapsto \psi(\boldsymbol{x},1)$ is integrable on $\mathcal{X}^N$ by Proposition 5.1, and the fact that

$$\int_{\mathcal{X}} k_1(x,x)\mathrm{d}\omega(x) < +\infty. \tag{5.151}$$

The expected value of cross-leverage scores

Let $m_1, m_2 \in \mathbb{N}^*$ such that $m_1 \neq m_2$. We have

$$\begin{aligned}
\tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x}) &= e^{\mathcal{F}}_{m_1}(\boldsymbol{x})^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1} e^{\mathcal{F}}_{m_2}(\boldsymbol{x}) \\
&= \frac{1}{4}\left(e^{\mathcal{F}}_{m_1}(\boldsymbol{x}) + e^{\mathcal{F}}_{m_2}(\boldsymbol{x})\right)^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1}\left(e^{\mathcal{F}}_{m_1}(\boldsymbol{x}) + e^{\mathcal{F}}_{m_2}(\boldsymbol{x})\right)^\intercal \\
&\quad - \frac{1}{4}\left(e^{\mathcal{F}}_{m_1}(\boldsymbol{x}) - e^{\mathcal{F}}_{m_2}(\boldsymbol{x})\right)^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1}\left(e^{\mathcal{F}}_{m_1}(\boldsymbol{x}) - e^{\mathcal{F}}_{m_2}(\boldsymbol{x})\right)^\intercal .
\end{aligned} \tag{5.152}$$

Thus

$$\mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x}) = \frac{1}{4Z_N(k)} \int_{\mathcal{X}^N} \left(\Psi^+(\boldsymbol{x}) - \Psi^-(\boldsymbol{x})\right) \otimes_{i\in[N]} \mathrm{d}\omega(x_i), \tag{5.153}$$

where

$$\Psi^+(\boldsymbol{x}) = \left(e^{\mathcal{F}}_{m_1}(\boldsymbol{x}) + e^{\mathcal{F}}_{m_2}(\boldsymbol{x})\right)^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1}\left(e^{\mathcal{F}}_{m_1}(\boldsymbol{x}) + e^{\mathcal{F}}_{m_2}(\boldsymbol{x})\right) \operatorname{Det} \boldsymbol{K}(\boldsymbol{x}), \tag{5.154}$$

and

$$\Psi^-(\boldsymbol{x}) = \left(e^{\mathcal{F}}_{m_1}(\boldsymbol{x}) - e^{\mathcal{F}}_{m_2}(\boldsymbol{x})\right)^\intercal \boldsymbol{K}(\boldsymbol{x})^{-1}\left(e^{\mathcal{F}}_{m_1}(\boldsymbol{x}) - e^{\mathcal{F}}_{m_2}(\boldsymbol{x})\right) \operatorname{Det} \boldsymbol{K}(\boldsymbol{x}). \tag{5.155}$$

We proceed as in Section 5.8.8 and we use Proposition 5.7 to prove that

$$\begin{aligned}
\Psi^+(\boldsymbol{x}) &= \partial_t \operatorname{Det}\left(\boldsymbol{K}(\boldsymbol{x}) + t\left(e^{\mathcal{F}}_{m_1}(\boldsymbol{x}) + e^{\mathcal{F}}_{m_2}(\boldsymbol{x})\right)\left(e^{\mathcal{F}}_{m_1}(\boldsymbol{x}) + e^{\mathcal{F}}_{m_2}(\boldsymbol{x})\right)^\intercal\right)|_{t=0^+} \\
&= \partial_t \operatorname{Det}\left(\boldsymbol{K}^+_t(\boldsymbol{x})\right)|_{t=0^+}.
\end{aligned} \tag{5.156}$$

and

$$\begin{aligned}
\Psi^-(\boldsymbol{x}) &= \partial_t \operatorname{Det}\left(\boldsymbol{K}(\boldsymbol{x}) + t\left(e^{\mathcal{F}}_{m_1}(\boldsymbol{x}) - e^{\mathcal{F}}_{m_2}(\boldsymbol{x})\right)\left(e^{\mathcal{F}}_{m_1}(\boldsymbol{x}) - e^{\mathcal{F}}_{m_2}(\boldsymbol{x})\right)^\intercal\right)|_{t=0^+} \\
&= \partial_t \operatorname{Det}\left(\boldsymbol{K}^-_t(\boldsymbol{x})\right)|_{t=0^+}.
\end{aligned} \tag{5.157}$$

In order to prove that $\phi^+_{m_1,m_2}$ and $\phi^-_{m_1,m_2}$ are right differentiable in zero along with the second equation in Proposition 5.5, one can follow the same steps as in the end of Section 5.8.8. In particular, the interchange of the derivation and the integration operations follows from the same arguments, upon noting that both $\int_{\mathcal{X}} k^+_t(x,x)\mathrm{d}\omega(x)$ and $\int_{\mathcal{X}} k^-_t(x,x)\mathrm{d}\omega(x)$ are finite.

5.8.9 *Proof of Proposition 5.4*

The proof is a straightforward computation now that we have Proposition 5.5 and Proposition 5.6.

The expected value of the m-th leverage score

Let $m \in \mathbb{N}^*$. We have by Proposition 5.5 and Proposition 5.1,

$$\mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x}) = \frac{1}{N! \sum\limits_{\substack{U \subset \mathbb{N}^* \\ |U|=N}} \prod\limits_{u \in U} \sigma_u} \frac{\partial \phi_m}{\partial t}|_{t=0^+}, \tag{5.158}$$

where

$$\phi_m(t) = \int_{\mathcal{X}^N} \mathrm{Det}\left(\boldsymbol{K}(\boldsymbol{x}) + t e_m^{\mathcal{F}}(\boldsymbol{x}) e_m^{\mathcal{F}}(\boldsymbol{x})^\intercal\right) \otimes_{i=1}^N \mathrm{d}\omega(x_i). \tag{5.159}$$

Now by Proposition 5.6 and Proposition 5.1,

$$\phi_m(t) = N! \sum_{U \in \mathcal{U}_N} \prod_{u \in U} \tilde{\sigma}_u(t), \tag{5.160}$$

where for $u \in \mathbb{N}^*$, $\tilde{\sigma}_u(t) = \sigma_u + t\delta_{m,u}\sigma_u$. Therefore,

$$\phi_m(t) = N! \sum_{\substack{U \in \mathcal{U}_N \\ m \in U}} \prod_{u \in U} \tilde{\sigma}_u(t) + N! \sum_{\substack{U \in \mathcal{U}_N \\ m \notin U}} \prod_{u \in U} \tilde{\sigma}_u(t) \tag{5.161}$$

$$= N!\, \sigma_m (t+1) \sum_{\substack{U \in \mathcal{U}_{N-1} \\ m \notin U}} \prod_{u \in U} \sigma_u + N! \sum_{\substack{U \in \mathcal{U}_N \\ m \notin U}} \prod_{u \in U} \sigma_u. \tag{5.162}$$

Thus,

$$\frac{\partial \phi_m}{\partial t}|_{t=0} = N!\, \sigma_m \sum_{\substack{U \in \mathcal{U}_{N-1} \\ m \notin U}} \prod_{u \in U} \sigma_u, \tag{5.163}$$

so that (5.158) becomes

$$\mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x}) = \left(\cancel{N!} \sum_{U \in \mathcal{U}_N} \prod_{u \in U} \sigma_u\right)^{-1} \cancel{N!}\, \sigma_m \sum_{\substack{U \in \mathcal{U}_{N-1} \\ m \notin U}} \prod_{u \in U} \sigma_u, \tag{5.164}$$

which concludes the proof.

The expected value of cross-leverage scores

Let $m_1, m_2 \in \mathbb{N}^*$ such that $m_1 \neq m_2$. We have by Proposition 5.5 and Proposition 5.1,

$$\mathbb{E}_{\mathrm{VS}}\, \tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x}) = \frac{1}{4N! \sum\limits_{U \in \mathcal{U}_N} \prod\limits_{u \in U} \sigma_u} \left(\frac{\partial \phi_{m_1,m_2}^+}{\partial t} - \frac{\partial \phi_{m_1,m_2}^-}{\partial t}\right)\Bigg|_{t=0^+}, \tag{5.165}$$

where

$$\phi_{m_1,m_2}^+(t) = \int_{\mathcal{X}^N} \mathrm{Det}\left(\boldsymbol{K}(\boldsymbol{x}) + t\left(e_{m_1}^{\mathcal{F}}(\boldsymbol{x}) + e_{m_2}^{\mathcal{F}}(\boldsymbol{x})\right)\left(e_{m_1}^{\mathcal{F}}(\boldsymbol{x}) + e_{m_2}^{\mathcal{F}}(\boldsymbol{x})\right)^\intercal\right) \otimes_{i=1}^N \mathrm{d}\omega(x_i), \tag{5.166}$$

and

$$\phi_{m_1,m_2}^-(t) = \int_{\mathcal{X}^N} \mathrm{Det}\left(\boldsymbol{K}(\boldsymbol{x}) + t\left(e_{m_1}^{\mathcal{F}}(\boldsymbol{x}) - e_{m_2}^{\mathcal{F}}(\boldsymbol{x})\right)\left(e_{m_1}^{\mathcal{F}}(\boldsymbol{x}) - e_{m_2}^{\mathcal{F}}(\boldsymbol{x})\right)^\intercal\right) \otimes_{i=1}^N \mathrm{d}\omega(x_i). \tag{5.167}$$

Now by Proposition 5.6, for $t \geq 0$,

$$\phi^+_{m_1,m_2}(t) = N! \sum_{U \in \mathcal{U}_N} \prod_{u \in U} \tilde{\sigma}^+_u(t) = N! \sum_{U \in \mathcal{U}_N} \prod_{u \in U} \tilde{\sigma}^-_u(t) = \phi^-_{m_1,m_2}(t). \tag{5.168}$$

Plugging this back into (5.165) yields $\mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x}) = 0$.

5.8.10 *Proof of Theorem 5.2*

A decomposition result for the error

We start with a lemma.

Lemma 5.3. *Let $\mu \in \mathcal{F}$ such that $\|\mu\|_{\mathcal{F}} \leq 1$. Under Assumption 5.4,*

$$\mathbb{E}_{\mathrm{VS}}\, \mathcal{E}(\mu; x)^2 \leq (1 + B) \sum_{m \in [N]} \frac{\sigma_N}{\sigma_m} \langle \mu, e^{\mathcal{F}}_m \rangle^2_{\mathcal{F}} + \sum_{m \geq N+1} \langle \mu, e^{\mathcal{F}}_m \rangle^2_{\mathcal{F}}. \tag{5.169}$$

Proof. Using the same arguments as in the proof of Lemma 5.1 in Section 5.8.4, it comes that, for $\boldsymbol{x} \in \mathcal{X}^N$ such that $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) > 0$,

$$\|\mu - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu\|^2_{\mathcal{F}} = \sum_{m \in \mathbb{N}^*} \langle \mu, e^{\mathcal{F}}_m \rangle^2_{\mathcal{F}} \left(1 - \tau^{\mathcal{F}}_m(\boldsymbol{x})\right) - \sum_{\substack{m_1, m_2 \in \mathbb{N}^* \\ m_1 \neq m_2}} \langle \mu, e^{\mathcal{F}}_{m_1} \rangle_{\mathcal{F}} \langle \mu, e^{\mathcal{F}}_{m_2} \rangle_{\mathcal{F}} \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x}). \tag{5.170}$$

We want to take expectations in both sides of (5.170). For the first term in the RHS, we prove, using the same arguments as for the proof of Theorem 5.1 in Section 5.8.5, that

$$\mathbb{E}_{\mathrm{VS}} \sum_{m \in \mathbb{N}^*} \langle \mu, e^{\mathcal{F}}_m \rangle^2_{\mathcal{F}} \left(1 - \tau^{\mathcal{F}}_m(\boldsymbol{x})\right) = \sum_{m \in \mathbb{N}^*} \langle \mu, e^{\mathcal{F}}_m \rangle^2_{\mathcal{F}} \left(1 - \mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_m(\boldsymbol{x})\right). \tag{5.171}$$

For the second term in the RHS of (5.170), we need to justify that

$$\begin{aligned} \mathbb{E}_{\mathrm{VS}} \sum_{\substack{m_1, m_2 \in \mathbb{N}^* \\ m_1 \neq m_2}} \langle \mu, e^{\mathcal{F}}_{m_1} \rangle_{\mathcal{F}} \langle \mu, e^{\mathcal{F}}_{m_2} \rangle_{\mathcal{F}}\, \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x}) &= \sum_{\substack{m_1, m_2 \in \mathbb{N}^* \\ m_1 \neq m_2}} \langle \mu, e^{\mathcal{F}}_{m_1} \rangle_{\mathcal{F}} \langle \mu, e^{\mathcal{F}}_{m_2} \rangle_{\mathcal{F}}\, \mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x}) \\ &= 0. \end{aligned} \tag{5.172}$$

This can be done using dominated convergence. Indeed, let $M \in \mathbb{N}^*$. We have

$$\begin{aligned} \mathbb{E}_{\mathrm{VS}} \sum_{\substack{m_1, m_2 \in [M] \\ m_1 \neq m_2}} \langle \mu, e^{\mathcal{F}}_{m_1} \rangle_{\mathcal{F}} \langle \mu, e^{\mathcal{F}}_{m_2} \rangle_{\mathcal{F}}\, \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x}) &= \sum_{\substack{m_1, m_2 \in [M] \\ m_1 \neq m_2}} \langle \mu, e^{\mathcal{F}}_{m_1} \rangle_{\mathcal{F}} \langle \mu, e^{\mathcal{F}}_{m_2} \rangle_{\mathcal{F}}\, \mathbb{E}_{\mathrm{VS}}\, \tau^{\mathcal{F}}_{m_1,m_2}(\boldsymbol{x}) \\ &= 0. \end{aligned} \tag{5.173}$$

Moreover,

$$\begin{aligned}
&\left| \sum_{\substack{m_1,m_2\in[M]\\ m_1\neq m_2}} \langle \mu, e_{m_1}^{\mathcal{F}}\rangle_{\mathcal{F}} \langle \mu, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}}\, \tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x}) \right| \\
&\qquad = \left| \sum_{m_1,m_2\in[M]} \langle \mu, e_{m_1}^{\mathcal{F}}\rangle_{\mathcal{F}} \langle \mu, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}}\, \tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x}) - \sum_{m\in[M]} \langle \mu, e_m^{\mathcal{F}}\rangle_{\mathcal{F}}^2\, \tau_m^{\mathcal{F}}(\boldsymbol{x}) \right| \\
&\qquad \leq \left| \sum_{m_1,m_2\in[M]} \langle \mu, e_{m_1}^{\mathcal{F}}\rangle_{\mathcal{F}} \langle \mu, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}}\, \tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x}) \right| + \left| \sum_{m\in[M]} \langle \mu, e_m^{\mathcal{F}}\rangle_{\mathcal{F}}^2\, \tau_m^{\mathcal{F}}(\boldsymbol{x}) \right| \\
&\qquad = \left\| \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} \sum_{m\in[M]} \langle \mu, e_m^{\mathcal{F}}\rangle_{\mathcal{F}} e_m^{\mathcal{F}} \right\|_{\mathcal{F}}^2 + \left| \sum_{m\in[M]} \langle \mu, e_m^{\mathcal{F}}\rangle_{\mathcal{F}}^2\, \tau_m^{\mathcal{F}}(\boldsymbol{x}) \right| \\
&\qquad \leq \left\| \sum_{m\in[M]} \langle \mu, e_m^{\mathcal{F}}\rangle_{\mathcal{F}} e_m^{\mathcal{F}} \right\|_{\mathcal{F}}^2 + \sum_{m\in[M]} \langle \mu, e_m^{\mathcal{F}}\rangle_{\mathcal{F}}^2 \\
&\qquad = 2\|\mu\|_{\mathcal{F}}^2 < +\infty. \qquad (5.174)
\end{aligned}$$

Combining (5.173) and (5.174), we deduce (5.172) by the dominated convergence theorem.

Finally, we combine (5.171) and (5.172) to get

$$\begin{aligned}
\mathbb{E}_{\mathrm{VS}} \|\mu - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu\|_{\mathcal{F}}^2 &= \sum_{m\in\mathbb{N}^*} \langle \mu, e_m^{\mathcal{F}}\rangle_{\mathcal{F}}^2 \left(1 - \mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x})\right) \\
&= \sum_{n\in[N]} \langle \mu, e_n^{\mathcal{F}}\rangle_{\mathcal{F}}^2 \left(1 - \mathbb{E}_{\mathrm{VS}}\, \tau_n^{\mathcal{F}}(\boldsymbol{x})\right) + \sum_{m\geq N+1} \langle \mu, e_m^{\mathcal{F}}\rangle_{\mathcal{F}}^2 \left(1 - \mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x})\right). \qquad (5.175)
\end{aligned}$$

On the one hand,

$$\forall m \geq N+1, \quad 1 - \mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x}) \leq 1, \qquad (5.176)$$

and on the other hand, remember that by Theorem 5.1, the sequence $\epsilon_m(N)$ is non-increasing, so that

$$\begin{aligned}
\forall n \in [N],\ \sigma_n(1 - \mathbb{E}_{\mathrm{VS}}\, \tau_n^{\mathcal{F}}(\boldsymbol{x})) &= \mathbb{E}_{\mathrm{VS}} \|\mu_{e_n} - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_{e_n}\|_{\mathcal{F}}^2 \qquad (5.177)\\
&= \epsilon_n(N) \qquad (5.178)\\
&\leq \epsilon_1(N), \qquad (5.179)
\end{aligned}$$

and by (5.25) in the same theorem one gets

$$\sigma_n(1 - \mathbb{E}_{\mathrm{VS}}\, \tau_n^{\mathcal{F}}(\boldsymbol{x})) \leq (1 + \beta_N)\sigma_N, \qquad (5.180)$$

so that

$$(1 - \mathbb{E}_{\mathrm{VS}}\, \tau_n^{\mathcal{F}}(\boldsymbol{x})) \leq (1 + \beta_N)\frac{\sigma_N}{\sigma_n}. \qquad (5.181)$$

Assumption 5.4 yields

$$\forall n \in [N],\ 1 - \mathbb{E}_{\mathrm{VS}}\, \tau_n^{\mathcal{F}}(\boldsymbol{x}) \leq (1 + B)\frac{\sigma_N}{\sigma_n}. \qquad (5.182)$$

This concludes the proof of the lemma. □

The expected value of the interpolation error

If there exists $r \in [0, 1/2]$ such that $\|\Sigma^{-r}\mu\|_{\mathcal{F}} < +\infty$, we have

$$\sum_{m \geq N+1} \langle \mu, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}^2 = \sum_{m \geq N+1} \sigma_m^{2r} \frac{\langle \mu, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}^2}{\sigma_m^{2r}} \tag{5.183}$$

$$\leq \sigma_{N+1}^{2r} \sum_{m \geq N+1} \frac{\langle \mu, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}^2}{\sigma_m^{2r}} \tag{5.184}$$

$$\leq \sigma_{N+1}^{2r} \|\Sigma^{-r}\mu\|_{\mathcal{F}}^2, \tag{5.185}$$

and

$$(1+B) \sum_{m \in [N]} \frac{\sigma_N}{\sigma_m} \langle \mu, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}^2 = (1+B) \sum_{m \in [N]} \frac{\sigma_N}{\sigma_m^{1-2r+2r}} \langle \mu, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}^2 \tag{5.186}$$

$$= (1+B) \sum_{m \in [N]} \frac{\sigma_N}{\sigma_m^{1-2r}} \frac{\langle \mu, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}^2}{\sigma_m^{2r}} \tag{5.187}$$

$$\leq (1+B) \sigma_N^{2r} \sum_{m \in [N]} \frac{\langle \mu, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}^2}{\sigma_m^{2r}} \tag{5.188}$$

$$= (1+B) \sigma_N^{2r} \|\Sigma^{-r}\mu\|_{\mathcal{F}}^2. \tag{5.189}$$

By Lemma 5.3, $\mathbb{E}_{\mathrm{VS}} \|\mu - \mathbf{\Pi}_{\mathcal{T}(x)}\mu\|_{\mathcal{F}}^2$ converges at the slow rate $\mathcal{O}(\sigma_N^{2r})$.

On the other hand, if there exists $r > 1/2$ such that $\|\Sigma^{-r}\mu\|_{\mathcal{F}} < +\infty$, we have

$$(1+B) \sum_{m \in [N]} \frac{\sigma_N}{\sigma_m} \langle \mu, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}^2 = (1+B) \sum_{m \in [N]} \frac{\sigma_N}{\sigma_m^{1-2r+2r}} \langle \mu, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}^2 \tag{5.190}$$

$$\leq (1+B) \sigma_N \sigma_1^{2r-1} \sum_{m \in [N]} \frac{\langle \mu, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}^2}{\sigma_m^{2r}} \tag{5.191}$$

$$\leq (1+B) \sigma_N \sigma_1^{2r-1} \|\Sigma^{-r}\mu\|_{\mathcal{F}}^2, \tag{5.192}$$

and

$$\sum_{m \geq N+1} \langle \mu, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}^2 = \sum_{m \geq N+1} \sigma_m^{2r} \frac{\langle \mu, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}^2}{\sigma_m^{2r}} \tag{5.193}$$

$$\leq \sigma_{N+1}^{2r} \sum_{m \geq N+1} \frac{\langle \mu, e_m^{\mathcal{F}} \rangle_{\mathcal{F}}^2}{\sigma_m^{2r}} \tag{5.194}$$

$$\leq \sigma_{N+1}^{2r} \|\Sigma^{-r}\mu\|_{\mathcal{F}}^2. \tag{5.195}$$

This time, the bound in Lemma 5.3 is dominated by its first term, so that $\mathbb{E}_{\mathrm{VS}} \|\mu - \mathbf{\Pi}_{\mathcal{T}(x)}\mu\|_{\mathcal{F}}^2$ converges at the faster rate $\mathcal{O}(\sigma_N)$.

5.8.11 *Proof of Theorem 5.3*

Proof of the bias identity

First, recall that, as f and g belong to $\mathbb{L}_2(\mathrm{d}\omega)$, we have

$$\int_{\mathcal{X}} f(x) g(x) \mathrm{d}\omega(x) = \sum_{m \in \mathbb{N}^*} \langle f, e_m \rangle_{\mathrm{d}\omega} \langle g, e_m \rangle_{\mathrm{d}\omega}, \tag{5.196}$$

thus, in order to prove the result, it is enough to prove that

$$\mathbb{E}_{\mathrm{VS}} \sum_{i\in[N]} \widehat{w}_i f(x_i) = \sum_{m\in\mathbb{N}^*} \langle f, e_m\rangle_{\mathrm{d}\omega} \langle g, e_m\rangle_{\mathrm{d}\omega}\, \mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x}). \tag{5.197}$$

Let $\boldsymbol{x} \in \mathcal{X}^N$ such that $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) > 0$. The optimal kernel quadrature weights satisfy

$$\widehat{\boldsymbol{w}} = \boldsymbol{K}(\boldsymbol{x})^{-1} \mu_g(\boldsymbol{x}), \tag{5.198}$$

so that

$$\begin{aligned}
\sum_{i\in N} \widehat{w}_i f(x_i) &= \widehat{\boldsymbol{w}}^{\intercal} f(\boldsymbol{x}) && (5.199)\\
&= \mu_g(\boldsymbol{x})^{\intercal} \boldsymbol{K}(\boldsymbol{x})^{-1} f(\boldsymbol{x}) && (5.200)\\
&= \sum_{m_1,m_2\in\mathbb{N}^*} \sigma_{m_1} \langle g, e_{m_1}\rangle_{\mathrm{d}\omega} \langle f, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}}\, e_{m_1}(\boldsymbol{x})^{\intercal} \boldsymbol{K}(\boldsymbol{x})^{-1} e_{m_2}^{\mathcal{F}}(\boldsymbol{x}) && (5.201)\\
&= \sum_{m_1,m_2\in\mathbb{N}^*} \sqrt{\sigma_{m_1}} \langle g, e_{m_1}\rangle_{\mathrm{d}\omega} \langle f, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}}\, e_{m_1}^{\mathcal{F}}(\boldsymbol{x})^{\intercal} \boldsymbol{K}(\boldsymbol{x})^{-1} e_{m_2}^{\mathcal{F}}(\boldsymbol{x}). && (5.202)
\end{aligned}$$

We want to use the dominated convergence theorem to take expectations in (5.202). Let $M \in \mathbb{N}^*$. By Lemma 5.2 and by the fact that $\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}$ is an $\langle .,. \rangle_{\mathcal{F}}$-orthogonal projection, it comes

$$\begin{aligned}
&\left| \sum_{m_1,m_2\in[M]} \sqrt{\sigma_{m_1}} \langle g, e_{m_1}\rangle_{\mathrm{d}\omega} \langle f, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}}\, e_{m_1}^{\mathcal{F}}(\boldsymbol{x})^{\intercal} \boldsymbol{K}(\boldsymbol{x})^{-1} e_{m_2}^{\mathcal{F}}(\boldsymbol{x}) \right| && (5.203)\\
&\quad = \left| \sum_{m_1,m_2\in[M]} \sqrt{\sigma_{m_1}} \langle g, e_{m_1}\rangle_{\mathrm{d}\omega} \langle f, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}}\, \langle \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_{m_1}^{\mathcal{F}}, \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}} \right| && (5.204)\\
&\quad = \left| \left\langle \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} \sum_{m_1\in[M]} \sqrt{\sigma_{m_1}} \langle g, e_{m_1}\rangle_{\mathrm{d}\omega} e_{m_1},\ \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} \sum_{m_2\in[M]} \langle f, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}} e_{m_2}^{\mathcal{F}} \right\rangle_{\mathcal{F}} \right| && (5.205)\\
&\quad \leq \left| \left\langle \sum_{m_1\in[M]} \sqrt{\sigma_{m_1}} \langle g, e_{m_1}\rangle_{\mathrm{d}\omega} e_{m_1},\ \sum_{m_2\in[M]} \langle f, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}} e_{m_2}^{\mathcal{F}} \right\rangle_{\mathcal{F}} \right| && (5.206)\\
&\quad \leq \left\| \sum_{m_1\in[M]} \sqrt{\sigma_{m_1}} \langle g, e_{m_1}\rangle_{\mathrm{d}\omega} e_{m_1} \right\|_{\mathcal{F}} \left\| \sum_{m_2\in[M]} \langle f, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}} e_{m_2}^{\mathcal{F}} \right\|_{\mathcal{F}}. && (5.207)
\end{aligned}$$

Now,

$$\begin{aligned}
&\left\| \sum_{m_1\in[M]} \sqrt{\sigma_{m_1}} \langle g, e_{m_1}\rangle_{\mathrm{d}\omega} e_{m_1} \right\|_{\mathcal{F}} \left\| \sum_{m_2\in[M]} \langle f, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}} e_{m_2}^{\mathcal{F}} \right\|_{\mathcal{F}} && (5.208)\\
&\qquad = \left\| \sum_{m_1\in[M]} \langle g, e_{m_1}\rangle_{\mathrm{d}\omega} e_{m_1}^{\mathcal{F}} \right\|_{\mathcal{F}} \left\| \sum_{m_2\in[M]} \langle f, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}} e_{m_2}^{\mathcal{F}} \right\|_{\mathcal{F}} && (5.209)\\
&\qquad = \left\| \sum_{m_1\in[M]} \langle g, e_{m_1}\rangle_{\mathrm{d}\omega} e_{m_1} \right\|_{\mathrm{d}\omega} \left\| \sum_{m_2\in[M]} \langle f, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}} e_{m_2}^{\mathcal{F}} \right\|_{\mathcal{F}} && (5.210)\\
&\qquad \leq \left\| \sum_{m_1\in\mathbb{N}^*} \langle g, e_{m_1}\rangle_{\mathrm{d}\omega} e_{m_1} \right\|_{\mathrm{d}\omega} \left\| \sum_{m_2\in\mathbb{N}^*} \langle f, e_{m_2}^{\mathcal{F}}\rangle_{\mathcal{F}} e_{m_2}^{\mathcal{F}} \right\|_{\mathcal{F}} && (5.211)\\
&\qquad < +\infty, && (5.212)
\end{aligned}$$

since $\sum_{m\in\mathbb{N}^*}\sqrt{\sigma_m}\langle g,e_m\rangle_{\mathrm{d}\omega}e_m\in\mathcal{F}$. Dominated convergenve thus yields

$$\mathbb{E}_{\mathrm{VS}}\sum_{m_1,m_2\in\mathbb{N}^*}\sqrt{\sigma_{m_1}}\langle g,e_{m_1}\rangle_{\mathrm{d}\omega}\langle f,e^{\mathcal{F}}_{m_2}\rangle_{\mathcal{F}}\,e^{\mathcal{F}}_{m_1}(\boldsymbol{x})^\intercal\boldsymbol{K}(\boldsymbol{x})^{-1}e^{\mathcal{F}}_{m_2}(\boldsymbol{x}) \tag{5.213}$$

$$=\sum_{m_1,m_2\in\mathbb{N}^*}\sqrt{\sigma_{m_1}}\langle g,e_{m_1}\rangle_{\mathrm{d}\omega}\langle f,e^{\mathcal{F}}_{m_2}\rangle_{\mathcal{F}}\;\mathbb{E}_{\mathrm{VS}}\,e^{\mathcal{F}}_{m_1}(\boldsymbol{x})^\intercal\boldsymbol{K}(\boldsymbol{x})^{-1}e^{\mathcal{F}}_{m_2}(\boldsymbol{x}). \tag{5.214}$$

Using Proposition 5.4, we continue our derivation as

$$\mathbb{E}_{\mathrm{VS}}\sum_{m_1,m_2\in\mathbb{N}^*}\sqrt{\sigma_{m_1}}\langle g,e_{m_1}\rangle_{\mathrm{d}\omega}\langle f,e^{\mathcal{F}}_{m_2}\rangle_{\mathcal{F}}\,e^{\mathcal{F}}_{m_1}(\boldsymbol{x})^\intercal\boldsymbol{K}(\boldsymbol{x})^{-1}e^{\mathcal{F}}_{m_2}(\boldsymbol{x}) \tag{5.215}$$

$$=\sum_{m\in\mathbb{N}^*}\sqrt{\sigma_m}\langle g,e_m\rangle_{\mathrm{d}\omega}\langle f,e^{\mathcal{F}}_m\rangle_{\mathcal{F}}\;\mathbb{E}_{\mathrm{VS}}\,e^{\mathcal{F}}_m(\boldsymbol{x})^\intercal\boldsymbol{K}(\boldsymbol{x})^{-1}e^{\mathcal{F}}_m(\boldsymbol{x}) \tag{5.216}$$

$$=\sum_{m\in\mathbb{N}^*}\langle g,e_m\rangle_{\mathrm{d}\omega}\sqrt{\sigma_m}\langle f,e^{\mathcal{F}}_m\rangle_{\mathcal{F}}\;\mathbb{E}_{\mathrm{VS}}\,\tau^{\mathcal{F}}_m(\boldsymbol{x}). \tag{5.217}$$

Finally, (5.197) is obtained upon noting that

$$\forall m\in\mathbb{N}^*,\ \langle f,e_m\rangle_{\mathrm{d}\omega}=\sqrt{\sigma_m}\langle f,e^{\mathcal{F}}_m\rangle_{\mathcal{F}}. \tag{5.218}$$

Proof of the asymptotic unbiasedness of the quadrature

The expected value of the bias writes

$$\mathbb{E}_{\mathrm{VS}}\left(\int_{\mathcal{X}}f(x)g(x)\mathrm{d}\omega(x)-\sum_{i\in[N]}\widehat{w}_i f(x_i)\right)=\sum_{m\in\mathbb{N}^*}\langle f,e_m\rangle_{\mathrm{d}\omega}\langle g,e_m\rangle_{\mathrm{d}\omega}\left(1-\mathbb{E}_{\mathrm{VS}}\,\tau^{\mathcal{F}}_m(\boldsymbol{x})\right). \tag{5.219}$$

Now, by Theorem 5.1, for $m\in\mathbb{N}^*$,

$$\mathbb{E}_{\mathrm{VS}}\,\|\mu_{e_m}-\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_{e_m}\|^2_{\mathcal{F}}\leq\epsilon_1(N)\leq\sigma_N(1+\beta_N)\leq\sigma_N+\sum_{n\geq N}\sigma_n. \tag{5.220}$$

Thus

$$0\leq 1-\mathbb{E}_{\mathrm{VS}}\,\tau^{\mathcal{F}}_m(\boldsymbol{x})=\sigma_m{}^{-1}\,\mathbb{E}_{\mathrm{VS}}\,\|\mu_{e_m}-\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_{e_m}\|^2_{\mathcal{F}}\leq\sigma_m{}^{-1}\sigma_N+\sum_{n\geq N}\sigma_n, \tag{5.221}$$

so that

$$\lim_{N\to\infty}\langle f,e_m\rangle_{\mathrm{d}\omega}\langle g,e_m\rangle_{\mathrm{d}\omega}\left(1-\mathbb{E}_{\mathrm{VS}}\,\tau^{\mathcal{F}}_m(\boldsymbol{x})\right)=\langle f,e_m\rangle_{\mathrm{d}\omega}\langle g,e_m\rangle_{\mathrm{d}\omega}(1-\lim_{N\to\infty}\mathbb{E}_{\mathrm{VS}}\,\tau^{\mathcal{F}}_m(\boldsymbol{x}))=0. \tag{5.222}$$

To conclude, it is thus enough to apply the dominated convergence theorem to (5.219). By Lemma 5.2, $\tau^{\mathcal{F}}_m(\boldsymbol{x})\in[0,1]$, so that $1-\mathbb{E}_{\mathrm{VS}}\,\tau^{\mathcal{F}}_m(\boldsymbol{x})\in[0,1]$. In particular, for all $N\in\mathbb{N}^*$,

$$|\langle f,e_m\rangle_{\mathrm{d}\omega}\langle g,e_m\rangle_{\mathrm{d}\omega}\left(1-\mathbb{E}_{\mathrm{VS}}\,\tau^{\mathcal{F}}_m(\boldsymbol{x})\right)|\leq|\langle f,e_m\rangle_{\mathrm{d}\omega}\langle g,e_m\rangle_{\mathrm{d}\omega}| \tag{5.223}$$

$$\leq\frac{1}{2}\left(\langle f,e_m\rangle^2_{\mathrm{d}\omega}+\langle g,e_m\rangle^2_{\mathrm{d}\omega}\right), \tag{5.224}$$

which is the generic term of a convergent series as $f,g\in\mathbb{L}_2(\mathrm{d}\omega)$. This concludes the proof.

6

CONCLUSION AND PERSPECTIVES

6.1 CONCLUSION

We started this manuscript by reviewing the existing work in sub-sampling for problems with a linear structure. These problems appear in many fields such as data analysis, signal processing, machine learning, or statistics; with applications that include continuous signal discretization, numerical integration, dimension reduction, learning on budget, or preconditioning. In particular, we reviewed existing work on randomized sub-sampling based on first-order information: leverage scores, ridge leverage scores... A natural question that arose was to investigate sub-sampling using high-order information such as volumes. In particular, we were questioning whether DPPs would be beneficial for some sub-sampling problems.

We started by the problem of column subset selection: for a given matrix $\boldsymbol{X} \in \mathbb{R}^{N \times d}$, select the set $S \subset [d]$ that defines the most representative columns of $\boldsymbol{X}$. The idea was to investigate volume sampling through the lens of DPPs, and to challenge its optimality for this approximation task. Indeed, we observed that volume sampling can be seen as a mixture of projection DPPs that depends on the spectral decomposition of the matrix $\boldsymbol{X}^\intercal \boldsymbol{X}$. From this observation, we can expect that there may exist a better choice of DPP rather than this seemingly arbitrary mixture. This motivated the work done in Chapter 3, where we proposed to study the column subset selection task using the projection DPP with the largest weight in the mixture of volume sampling. This choice was motivated by a new geometric interpretation of projection DPPs: they naturally define a random space that "hovers" around a reference subspace. We formalized this idea using the notion of principal angles between subspaces. This new parametrization enabled the theoretical analysis of the projection DPP as a column sampler and led to improved theoretical guarantees compared to volume sampling under some conditions on the matrix $\boldsymbol{X}$.

The intuitions and techniques acquired in Chapter 3 were used to tackle another approximation problem that can be informally called the *continuous column subset selection problem*. This time the starting point was the work of Bach, 2017 on quadrature problems for functions living in RKHSs: for a given kernel k defined on some metric space $\mathcal{X}$, select the most representative set of nodes $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ which can then be used for approximating integrals of functions that belong to the RKHS associated to k. The quadrature proposed by Bach, 2017 relies on independent random sampling from the continuous ridge leverage score distribution, and the weights are calculated through a regularized quadratic optimization problem. However, the proposed distribution is intractable in general. As an alternative, we proposed in Chapter 4, to use optimal kernel quadrature based on nodes that follow the distribution of the projection DPP associated to the first eigenfunctions of the integration operator. Using again this new interpretation of a projection DPP (a projection DPP defines a random subspace), we succeeded in deriving the theoretical bound for optimal kernel quadrature when

the nodes follow the distribution of this projection DPP. In particular, the rates of convergence of this class of quadratures depends on the eigenvalues of the integration operator: the smoother the kernel, the faster the convergence of the quadrature. However, empirical investigations showed that these theoretical rates were pessimistic.

This observation motivated the work of Chapter 5, where we investigated an extension of volume sampling to continuous domains. Once again, we used the accumulated intuitions from Chapter 3 and Chapter 4 to study kernel quadrature and kernel interpolation under the distribution of continuous volume sampling. In particular, we proved tractable formulas for the expected value of the interpolation error when the nodes follow this repulsive distribution. These formulas were useful to derive sharp upper bounds for kernel quadrature that scale as the existing lower bounds. Moreover, the continuous volume sampling distribution has the advantage to be approximated using a fully kernelized MCMC algorithm i.e. an MCMC that can be implemented using the knowledge of the evaluation of the kernel k but without requiring the spectral decomposition of the integration operator. This property places continuous volume sampling as a natural alternative to the intractable continuous ridge leverage score distribution.

We conclude with a thought on the scientific impact (or the scientific leverage) of this thesis. One way to measure the impact of an algorithm is to estimate how hard it is to replace it. From the lengthy bibliographic work, which was conducted across the chapters of this thesis, we can conclude that the scientific impact of DPPs and their variants depends on the problem to study. Indeed, we have three settings with three different levels of impact.

It seems that the impact of DPPs and their variants may be limited for sub-sampling in discrete sets. Indeed, as it was mentioned in the introduction, DPPs and their variants replace leverage score sampling in the minimalist sampling regime, i.e. when the sub-sampling budget is equal to the dimensionality of the problem. This can be useful for applications such as feature selection where we need to keep the smallest possible number of columns. However, for other sub-sampling problems, this stringency is not necessary and DPPs may easily be replaced by leverage scores sampling. In other words, in discrete sets, sub-sampling using first-order information is sufficient for the majority of problems and there is no need to use high-order information! Nevertheless, the work done in Chapter 3 was crucial to develop the necessary techniques and intuitions to study DPPs in settings where they may have higher impact.

These settings correspond to sub-sampling problems on continuous domains. Indeed, the projection DPP, we proposed in Chapter 4, is an implementable alternative to continuous ridge leverage scores distribution. Moreover, when it is is possible to sample from the latter, like in the periodic Sobolev spaces, the projection DPP has better empirical performance. In other words, in continuous domains, high-order information is worthwhile for sub-sampling.

Still, even in this setting we have two different levels of impact that correspond to two different situations. In the first situation, the RKHS is "classical" and there exists some configuration with good approximation guarantees for kernel quadrature or kernel interpolation. This is the case of the periodic Sobolev space where the uniform grid is known to achieve the optimal rates. In these cases, we know from the theoretical and the empirical results that the projection DPP and continuous volume sampling

would be as competitive as the deterministic sequence but with the supplementary cost of sampling.

In the second situation, the RKHS is not classical and there is no simpler alternative algorithm that can compete with the optimality of the projection DPP or continuous volume sampling. In these cases DPPs and their variants cannot be easily replaced.

6.2 PERSPECTIVES

6.2.1 *Exploring non classical kernels*

The empirical investigations in Chapter 4 and Chapter 5 were restricted to classical RKHSs such as Sobolev spaces and Gaussian spaces. Nevertheless, our theoretical results are applicable to a broader class of less-known kernels.

In particular, our results are valid for RKHSs defined by *localization operators*. For example, the RKHS defined by band-limited functions restricted to a compact interval of $\mathbb{R}$ was the topic of intense research in signal processing (Slepian and Pollak, 1961; Landau and Pollak, 1961). This RKHS represents band-limited signals measured in a limited span of time. The introduction of this RKHS was motivated by the limitation of the Whittaker-Shannon-Kotel'nikov reconstruction theory. The latter requires to evaluate a band-limited function along the whole real line, which may be impossible in practice. For this RKHS, we may use the projection DPP associated to the eigenfunctions of the corresponding integration operator, also known as the *Prolate spheroidal wavefunctions* (Xiao et al., 2001). The use of localization operators was extended to time-frequency analysis in (Daubechies, 1988; Daubechies and Paul, 1988) and they are connected to *Bargmann* and *Bergman* spaces (Seip, 1991).

Interpolation and quadrature on the hyperspheres are another field of application of our results. Indeed, many kernels are known to diagonalize in the spherical harmonics basis (Smola et al., 2001); and localization operators may be defined on the sphere (Miranian, 2004).

Walsh spaces constitute another class of RKHS that may be explored in the future. These functional spaces were investigated in the QMC community for their connections with high-order digital nets (Dick, 2008). The corresponding eigenfunctions are Walsh functions, which are piecewise constant, and the corresponding projection DPP may be sampled efficiently using the HKPV algorithm as it was illustrated in Example 2.6 of Section 2.2.5.

RKHSs defined on manifolds (Gao, Kovalsky, and Daubechies, 2019)(Ehler, Gräf, and Oates, 2019) would be another field of application of our results. In particular, in these domains, the spectral decomposition of the integration operator is usually intractable, and we may use continuous volume sampling to construct a set of nodes with good interpolation properties.

6.2.2 *Exploring new approximation tasks*

In this thesis, we have dealt with two classes of approximation problems. The first class gathers problems for which the squared approximation error is tractable under the determinantal distribution. This is the case of kernel quadrature under continuous

volume sampling in Chapter 5. This tractability seems to be recurrent in other approximation problems: CSSP using volume sampling, linear regression on budget using dual volume sampling (Derezinski and Warmuth, 2017), ridge linear regression under regularized volume sampling(Dereziński and Warmuth, 2017), A-optimal design under proportional volume sampling (Nikolov et al., 2019)...

The second class gathers problems for which the squared approximation error is not tractable under the determinantal distribution. This is the case of CSSP under projection DPP of Chapter 3, or kernel quadrature under projection DPP of Chapter 4.

It seems that this thesis offers the first analysis of an approximation task under a determinantal distribution, when the error term is not tractable. This opens the door to study other approximation problems where the tractability is not satisfied. Examples of such problems include, regularized interpolation as in (Bach, 2017), kernel quadrature in misspecified settings (Kanagawa et al., 2016), or experimental design problems for Gaussian process approximations (Wynne et al., 2020).

6.2.3 *DPP-based quadratures that achieve the Bakhvalov rate in RKHSs?*

One of the main motivations behind the investigations that we conducted in Chapter 4 and Chapter 5 was to connect the previous work on numerical integration using DPPs (Bardenet and Hardy, 2020) and the work of (Bach, 2017) on kernel quadrature. In particular, we have elucidated the importance of the spectral decomposition of the integration operator $\mathbf{\Sigma}$ in the construction of quadratures with fast rates using DPPs. Still, our analysis does not allow an understanding of one intriguing grey area on this topic that we believe should be lightened.

Indeed, previous work on DPP-based quadratures deals with functional spaces $\mathcal{F}$ that satisfy

$$\forall f \in \mathcal{F}, \sum_{m \in \mathbb{N}^*} \frac{\langle f, e_m \rangle^2_{\mathrm{d}\omega}}{\sigma_m} < +\infty, \tag{6.1}$$

where $(e_n)_{n \in \mathbb{N}^*}$ is an orthonormal basis of $\mathbb{L}_2(\mathrm{d}\omega)$, and $(\sigma_m)_{m \in \mathbb{N}^*}$ is a sequence of positive numbers satisfying $\sum_{m \in \mathbb{N}^*} \sigma_m = +\infty$. However, the rate of convergence of these quadratures scales as $\mathcal{O}(N^{-1}\sigma_{N+1})$. For example, the asymptotic rate (4.50) in Theorem 4.4 of Bardenet and Hardy, 2020 scales as $\mathcal{O}(N^{-1-1/d}) = \mathcal{O}(N^{-1}\sigma_{N+1})$ with $\sigma_m = m^{-2s/d}$ and $s = 1/2$.

Compare that to the rate of optimal kernel quadrature that scales as $\mathcal{O}(\sigma_{N+1})$ when $\sum_{n \in \mathbb{N}^*} \sigma_n < +\infty$. The previous work on DPPs based quadrature does not fit in the framework of kernel quadrature. Indeed, the considered functional spaces correspond to Sobolev spaces of smoothness degree $s \leq d/2$, therefore they cannot be represented by RKHSs; see (Berlinet and Thomas-Agnan, 2011)[Theorem 132].

An interesting line of research would be to investigate a universal random quadrature based on DPPs or their variants, that achieve the rate $\mathcal{O}(N^{-1}\sigma_{N+1})$ in the context of kernel quadrature. The rate of such a quadrature would fulfil truly the "faster than Monte Carlo" promise: the $1/N$ corresponds to the Monte Carlo rate, and σ_{N+1} corresponds to the smoothness expressed by the functional space $\mathcal{F}$. This corresponds to the *Bakhvalov rate* (Bakhvalov, 1959): the rate that could be achieved by a randomized quadrature for approximating one integral; see (Novak, 2016) and (Oates et al., 2014) for details. Note that this improvement in the convergence rate does not contradict

the lower bound in Section 5.2.3. The latter is a lower bound for approximating all the integrals $\int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x)$ uniformly on g.

Figure 6.1 illustrates the rate $\mathcal{O}(N^{-1}\sigma_{N+1})$ compared to the optimal kernel quadrature rate $\mathcal{O}(\sigma_{N+1})$ in the case of the Korobov space $\mathcal{K}_{2,1}$: the Bakhvalov rate "breaks" the plateau of eigenvalues and makes numerical integration tractable even in high dimensional domains.

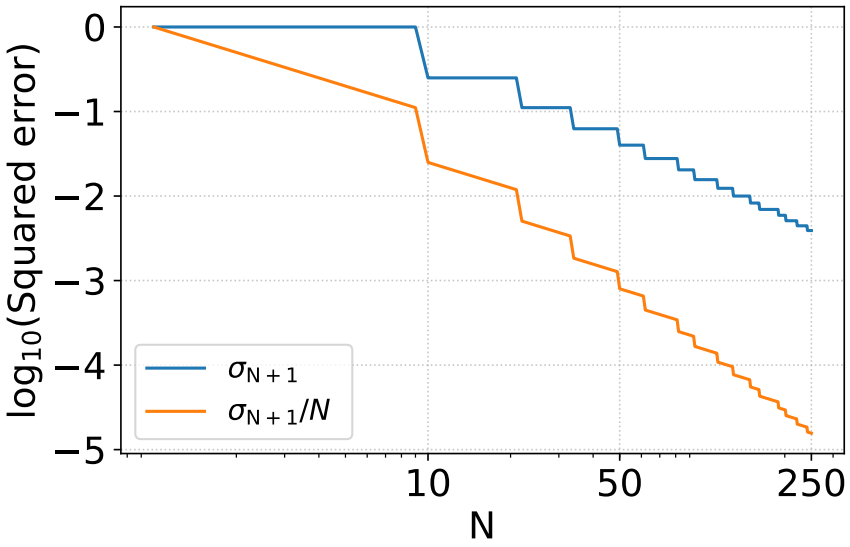


Figure 6.1 – The spectral rate σ_{N+1} compared to the Bakhvalov rate σ_{N+1}/N in the case of the Korobov space of order $s = 1$ and $d = 2$.

6.2.4 *A Grassmannian trigonometry problem*

We conclude with a technical open question that concerns the control of principal angles under the distribution of projection DPPs.

As we have seen in Chapter 3 and Chapter 4, the control of the largest principal angle between subspaces under the distribution of a well-chosen projection DPP lead to prove a theoretical guarantee for the column subset selection problem and for kernel quadrature.

However, the analysis we proposed relies on some loose majorizations. For example, in the case of kernel quadrature we used the symmetrizations

$$\frac{1}{\cos^2\theta_N(\mathcal{T}(\boldsymbol{x}),\mathcal{E}_N^{\mathcal{F}})} - 1 \leq \prod_{n\in[N]} \frac{1}{\cos^2\theta_n(\mathcal{T}(\boldsymbol{x}),\mathcal{E}_N^{\mathcal{F}})} - 1, \tag{6.2}$$

or

$$\frac{1}{\cos^2\theta_N(\mathcal{T}(\boldsymbol{x}),\mathcal{E}_N^{\mathcal{F}})} - 1 \leq \sum_{n\in[N]} \frac{1}{\cos^2\theta_n(\mathcal{T}(\boldsymbol{x}),\mathcal{E}_N^{\mathcal{F}})} - N. \tag{6.3}$$

The only reason to introduce such symmetrizations was tractability: the expected value of the r.h.s. of (6.2) and (6.3) are tractable under the distribution of the projection DPP, while the l.h.s is not. Improving the theoretical rates of optimal kernel quadrature under the projection DPP would require to avoid such symmetrizations.

One way to do so is treat the $\cos^2 \theta_n(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}})$ as the eigenvalues of the matrix

$$\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})\boldsymbol{K}(\boldsymbol{x})^{-1}\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal}, \tag{6.4}$$

and to use the tools of random matrix theory (Tao, 2012). In particular, investigating the expectation of the Stieltjes transform of the atomic measure associated to the random matrix (6.4), defined by

$$\mathfrak{s}(z) = \mathbb{E}_{\mathrm{DPP}} \operatorname{Tr}\left(\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})\boldsymbol{K}(\boldsymbol{x})^{-1}\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal} - z\mathbb{I}_N\right)^{-1}, \tag{6.5}$$

would give a better understanding of the fluctuations of its eigenvalues. For example, by Proposition 4.7 we have

$$\mathfrak{s}(0) = \mathbb{E}_{\mathrm{DPP}} \operatorname{Tr}\left(\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal^{-1}}\boldsymbol{K}(\boldsymbol{x})\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{-1}\right) = N + \sum_{v\in[N]} \frac{1}{\sigma_v} \sum_{w\in\mathbb{N}^*\setminus[N]} \sigma_w. \tag{6.6}$$

Unfortunately, $\mathfrak{s}(z)$ has no tractable formula for other values of z. However, we can study the properties of the expected characteristic polynomial, that we have introduced in the proof of Proposition 4.7

$$\mathfrak{p}(t) = \mathbb{E}_{\mathrm{DPP}} \operatorname{Det}\left(\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{\intercal^{-1}}\boldsymbol{K}(\boldsymbol{x})\boldsymbol{E}^{\mathcal{F}}(\boldsymbol{x})^{-1} - t\mathbb{I}_N\right). \tag{6.7}$$

Indeed, we showed that $\mathfrak{p}$ has the following expression

$$\mathfrak{p}(t) = \frac{1}{\prod_{n\in[N]} \sigma_n} \sum_{\ell=0}^{N} (1-t)^{\ell} \sum_{\substack{V\subset[N]\\|V|=\ell}} \prod_{v\in V} \sigma_v \sum_{\substack{W\subset\mathbb{N}^*\setminus[N]\\|W|=N-|V|}} \prod_{w\in W} \sigma_w. \tag{6.8}$$

In particular, this polynomial is equal to $(1-t)^N$ when $\mathcal{F}$ is a finite dimensional RKHS of dimension N: the roots of this polynomial are all equal to 1. This may suggests that in this case

$$\mathbb{E}_{\mathrm{DPP}} \frac{1}{\cos^2 \theta_N(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}})} = 1. \tag{6.9}$$

Indeed, this is true regarding our discussion after Theorem 4.8. In general, deriving sharper bounds for $\mathbb{E}_{\mathrm{DPP}} 1/\cos^2 \theta_N(\mathcal{T}(\boldsymbol{x}), \mathcal{E}_N^{\mathcal{F}})$ using the polynomial $\mathfrak{p}$ may be challenging but not impossible; see some recent advances in matrix sparsification that led to the resolution of a long-standing conjecture known as Kadison-Singer (Marcus, Spielman, and Srivastava, 2015).

After matrix concentration inequalities and the Stieltjes transform, the expected characteristic polynomial is an alternative way to study the eigenvalues of random matrices. Moreover it seems that this tool is more adequate to use in our case.

RÉSUMÉ EN FRANÇAIS

Le sous-échantillonnage est une tâche récurrente en mathématiques appliquées. Ce paradigme a des applications en traitement du signal, en apprentissage automatique, en analyse des données ou bien en statistiques: la discrétisation des signaux analogiques, le calcul approché des intégrales, la réduction de dimension, la réduction du budget d'étiquetage des algorithmes d'apprentissage... Alors qu'ils paraissent différents, ces problèmes peuvent être abordés avec la même stratégie: chercher les éléments les plus représentatifs d'un ensemble. Un bon sous-ensemble de représentants doit éviter de contenir des informations redondantes. Pour certains problèmes à structure linéaire, l'ensemble peut être plongé dans un espace vectoriel et la redondance d'un sous-ensemble peut se mesurer à l'aide du volume du polytope engendré par ce sous-ensemble. Il se trouve qu'il existe une famille de modèles probabilistes qui définissent des sous-ensembles aléatoires avec une propriété de répulsion: d'une façon informelle, la probabilité d'apparition d'un sous-ensemble est proportionnelle au volume qu'il engendre dans cet espace vectoriel. Ces modèles sont connus sous le nom des processus ponctuels déterminantaux et ils ont été étudiés dans plusieurs domaines: les matrices aléatoires, l'optique quantique, les statistiques spatiales, le traitement des images, l'apprentissage automatique et récemment l'intégration numérique.

Cette thèse est consacrée à l'étude de la pertinence des DPPs pour certaines tâches de sous-échantillonnage. Dans un premier temps, nous avons considéré le problème de sélection d'attributs: pour une matrice qui représente des données exprimées sur un système d'attributs, on cherche à sélectionner les attributs les plus représentatifs. En particulier, nous avons étudié l'échantillonnage volumique, un algorithme bien connu dans la littérature, à travers la théorie des DPPs. Nous avons proposé un algorithme impliquant un DPP avec de meilleures garanties théoriques et de meilleures performances empiriques. Le choix de ce DPP était motivé par une nouvelle interprétation géométrique que nous avons mise en évidence: un DPP définit naturellement un sous-espace vectoriel aléatoire qui "flotte" autour d'un sous-espace de référence.

A l'aide de cette nouvelle interprétation, nous avons réussi à étudier un autre problème d'approximation: l'approximation d'intégrales de fonctions qui vivent dans un espace à noyau, aussi appelé le problème de quadrature à noyau.

Pour ce problème, nous avons proposé une nouvelle classe de quadratures: les quadratures à noyau optimisées et basées sur des nœuds qui suivent la distribution d'un DPP. La définition de ce DPP est basée sur les fonctions propres de l'opérateur d'intégration correspondant. Nous avons montré que les taux de convergence de cette classe de quadratures dépendent des valeurs propres de cet opérateur: plus le noyau est régulier, meilleure est la convergence de la quadrature. Néanmoins, les expériences numériques montrent que ces taux de convergence sont pessimistes pour certains espaces fonctionnels.

Cette observation a motivé l'extension de l'échantillonnage volumique au domaine continu. Nous avons étudié le problème de quadrature à noyau ainsi que le problème d'interpolation à noyau pour des nœuds qui suivent cette nouvelle distribution. En

particulier, nous avons démontré des formules closes de l'espérance de l'erreur sous cette distribution répulsive. Ces formules ont permis de démontrer l'optimalité de l'échantillonnage volumique pour cette classe de problèmes d'approximation. De plus, cette nouvelle distribution peut être approchée par un algorithme MCMC qui peut être implémenté sans le recours à la décomposition spectrale de l'opérateur d'intégration.

Dans ce qui suit, on présente brièvement le contenu des chapitres de cette thèse ainsi que les publications associées.

DETERMINANTAL POINT PROCESSES

Ce chapitre est dédié à donner une définition rigoureuse et universelle des processus ponctuels déterminantaux. Cette définition peut être utilisée en domaine discret ou continu. En plus, on a revu quelques propriétés fondamentales des DPPs qui seront utilisées tout au long de ce manuscrit. Finalement, on a rappelé l'état de l'art de la simulation numérique des DPPs.

COLUMN SUBSET SELECTION USING PROJECTION DPPS

Dans le Chapitre 3, on a proposé et analysé un algorithme de sélection de colonnes, à base d'un DPP de projection, ayant comme but la factorisation de faible rang des matrices.

En particulier, on a comparé le DPP de projection, associé aux premiers vecteurs propres d'une matrice donnée, contre l'échantillonnage volumique, connu en anglais sous le nom de *volume sampling*, et on a prouvé que ce DPP peut avoir des meilleures garanties théoriques ainsi que des meilleures performances empiriques sous l'hypothèse dite *la parcimonie des k-leviers*. En plus, on a démontré des garanties théoriques sous une condition plus réaliste dite *la parcimonie relâchée des k-leviers*. Cette dernière a été constatée sur plusieurs jeux de données réelles.

Une contribution importante de ce chapitre est la mise en lumière de l'importance des angles principaux entre les sous-espaces dans l'étude des DPPs de projection. Plus précisément, ces paramètres géométriques ont permis d'obtenir des formules closes d'une borne supérieure de l'espérance de l'erreur d'approximation sous la distribution du DPP de projection.

Le contenu de ce chapitre est basé sur l'article suivant.

- A. Belhadji, R. Bardenet, and P. Chainais (2020a). "A determinantal point process for column subset selection". In: *Journal of Machine Learning Research* 21.197, pp. 1–62.

PROJECTION DPPS FOR KERNEL QUADRATURE

Dans le Chapitre 4, on a étudié la quadrature à noyau à base d'un DPP de projection. Une telle quadrature est une approximation

$$\int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x) \approx \sum_{i\in[N]} w_i f(x_i), \tag{6.10}$$

avec g est une fonction de carré intégrable et f appartient à un RKHS associé à un noyau k. En particulier, on s'intéresse à l'erreur d'intégration du pire cas, définit par

$$\sup_{\|f\|_{\mathcal{F}}\leq 1} \left| \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x) - \sum_{i\in[N]} w_i f(x_i) \right|^2. \tag{6.11}$$

Ce terme s'écrit également comme

$$\|\mu_g - \sum_{i\in[N]} w_i k(x_i,.)\|^2_{\mathcal{F}}, \tag{6.12}$$

avec $\mu_g = \int_{\mathcal{X}} k(x,.)g(x)\mathrm{d}\omega(x)$ le plongement de g dans l'RKHS $\mathcal{F}$.

Les nœuds de la quadrature qu'on a proposée, forment un ensemble aléatoire $\boldsymbol{x} = \{x_1, \dots, x_N\}$ qui suit la distribution d'un DPP de projection qui dépend uniquement des premières fonctions propres de l'opérateur d'intégration $\boldsymbol{\Sigma}$. Alors que le vecteur des poids $\boldsymbol{w}$ résout le problème d'optimisation

$$\min_{\boldsymbol{w}\in\mathbb{R}^N} \|\mu_g - \sum_{i\in[N]} w_i k(x_i,.)\|^2_{\mathcal{F}}. \tag{6.13}$$

La solution de (6.13) s'écrit $\hat{\boldsymbol{w}} = \boldsymbol{K}(\boldsymbol{x})^{-1}\mu_g(\boldsymbol{x})$, et la mixture optimale $\sum_{n\in[N]} \hat{w}_n k(x_n,.)$ est la projection de μ_g sur le sous-espace $\mathcal{T}(\boldsymbol{x}) = \mathrm{Span}(k(x_n,.)_{n\in[N]})$ qu'on dénote $\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g$, et la valeur optimale de (6.13) est appelée l'erreur d'interpolation de μ_g.

On donne une analyse théorique de cette quadrature et on démontre que le taux de convergence dépend des valeurs propres $(\sigma_n)_{n\in\mathbb{N}^*}$ de l'opérateur d'intégration comme $\mathcal{O}(N\sum_{n\geq N+1}\sigma_n)$ avec N est le nombre des nœuds de la quadrature. Les simulations numériques suggèrent que le taux de convergence est plus rapide et se comporte comme $\mathcal{O}(\sigma_N)$.

D'un point de vue technique, on a exploité les intuitions géométriques du Chapitre 3 pour développer l'analyse théorique de ces quadratures. En particulier, on a utilisé les angles principaux entre les sous-espaces pour borner l'espérance de l'erreur d'interpolation au carré

$$\mathbb{E}_{\mathrm{DPP}} \|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|^2_{\mathcal{F}}, \tag{6.14}$$

qui n'a pas de formule close connue.

Le contenu de ce chapitre est basé sur l'article suivant.

- A. Belhadji, R. Bardenet, and P. Chainais (2019a). "Kernel quadrature with DPPs". In: *Advances in Neural Information Processing Systems* 32, pp. 12907–12917.

CONTINUOUS VOLUME SAMPLING FOR KERNEL INTERPOLATION

Dans Chapitre 5, on continue sur la ligne de recherche initiée dans Chapitre 3 et Chapitre 4, et on étudie la quadrature à noyau ainsi que l'interpolation à noyau sous la distribution de l'échantillonnage volumique continu (Continuous Volume Sampling en anglais).

Contrairement au cas du DPP de projection étudié dans Chapitre 4, on montre que l'espérance du carré de l'erreur d'interpolation admet une formule close sous la distribution CVS

$$\mathbb{E}_{\mathrm{CVS}} \|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|^2_{\mathcal{F}} = \sum_{m\in\mathbb{N}^*} \langle g, e_m\rangle^2_{\mathrm{d}\omega} \epsilon_m(N), \tag{6.15}$$

avec

$$\epsilon_m(N) = \sum_{\substack{|T|=N \\ m \notin T}} \prod_{t \in T} \sigma_t \Big/ \sum_{|T|=N} \prod_{t \in T} \sigma_t. \tag{6.16}$$

En plus, on montre que $\epsilon_m(N) = \mathcal{O}(\sigma_{N+1})$ pour plusieurs RKHSs, et on donne la garantie de convergence pour l'interpolation des fonctions au-delà du cadre de la quadrature à noyau. Finalement, on démontre que la quadrature est asymptotiquement non-biaisé, et on donne une interprétation à ce résultat.

L'avantage derrière cette distribution est qu'elle peut être approximée par un schéma MCMC qui ne nécessite pas la décomposition spectrale de $\mathbf{\Sigma}$ contrairement au DPP de projection de Chapitre 4 et la distribution des leviers régularisés de (Bach, 2017).

Le contenu de ce chapitre est basé sur l'article suivant.

- A. Belhadji, R. Bardenet, and P. Chainais (2020b). "Kernel interpolation with continuous volume sampling". In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 725–735.

CONCLUSION AND PERSPECTIVES

On conclut le manuscrit par une discussion des contributions de la thèse et on soulève des questions ouvertes concernant les approximations à base de DPPs.

NOTATIONS

The symbols with the indication(∗∗) appear in Chapter 3 and they have a different meaning in Chapter 4 and Chapter 5.

FUNCTIONAL ANALYSIS

$\mathcal{X}$	Metric space
$x_1, \dots, x_N$	Elements of $\mathcal{X}$ (even if $\mathcal{X}$ is a high dimensional domain)
$\boldsymbol{x}$	Discrete subset of $\mathcal{X}$: $\boldsymbol{x} = \{x_1, \dots, x_N\} \subset \mathcal{X}$. It may be seen as an element of $\mathcal{X}^N$. In this case we write $\boldsymbol{x} = (x_1, \dots, x_N)$.
ω, $\mathrm{d}\omega$	Measure on $\mathcal{X}$. We use the two interchangeably
$\omega^{\otimes L}$, $\mathrm{d}\omega^{\otimes L}$	Tensor product of the measure ω defined on $\mathcal{X}^L$
$w(t)$	Density of an absolutely continuous measure with respect to the Lebesgue measure: $\mathrm{d}\omega(t) = w(t)\mathrm{d}t$
N (**)	Cardinality of $\boldsymbol{x}$
k (**)	Kernel defined on $\mathcal{X}$
$\boldsymbol{K}(\boldsymbol{x})$	Kernel matrix $(k(x_i, x_{i'}))_{(i,i') \in [N] \times [N]}$
$\mathcal{F}$	RKHS associated to k
$\mathbb{L}_2(\mathrm{d}\omega)$	Set of square-integrable functions with respect to $\mathrm{d}\omega$ (or ω)
$\mathbb{L}_2([0,1]^d)$	Set of square-integrable functions with respect to the Lebesgue measure on $[0,1]^d$.
$\mathbb{I}_{\mathcal{H}}$	Identity operator of a Hilbert space $\mathcal{H}$
$\mathbb{1}_S$	Indicator function of the set S
$\boldsymbol{\Pi}_{\mathcal{P}}$	Orthogonal projection onto the subspace $\mathcal{P}$
$\|.\|_{\mathrm{d}\omega}$, $\langle .,. \rangle_{\mathrm{d}\omega}$	Norm and bilinear form defined by $\mathrm{d}\omega$ (or ω)
$\|.\|_{\mathcal{F}}$, $\langle .,. \rangle_{\mathcal{F}}$	Norm and bilinear form of $\mathcal{F}$
$\boldsymbol{\Sigma}$ (**)	Integration operator associated to k and $\mathrm{d}\omega$
σ_m	m-th eigenvalue of $\boldsymbol{\Sigma}$
e_m	m-th eigenfunction of $\boldsymbol{\Sigma}$ with $\|e_m\|_{\mathrm{d}\omega} = 1$
$e_m^{\mathcal{F}}$	Scaled m-th eigenfunction of $\boldsymbol{\Sigma}$ with $\|e_m^{\mathcal{F}}\|_{\mathcal{F}} = 1$
μ_g	Embedding of g equivalently the mean-element of the measure $g\mathrm{d}\omega$

$\mathbb{I}_\delta$	Identity matrix of dimension δ
$\mathbf{\Pi}_{\mathcal{P}}$	Orthogonal projection onto the subspace $\mathcal{P}$
$\boldsymbol{X}$	Real matrix
$\boldsymbol{X}_{:,S}$	Sub-matrix of $\boldsymbol{X}$ containing the columns S
$\boldsymbol{X}_{S,:}$	Sub-matrix of $\boldsymbol{X}$ containing the rows S
$\boldsymbol{X}_{S,S'}$	Sub-matrix of $\boldsymbol{X}$ defined by the rows S and the columns S'
$\boldsymbol{U}$	Left eigenvectors of $\boldsymbol{X}$
$\mathbf{\Sigma}$ (**)	Matrix of singular values of $\boldsymbol{X}$
$\boldsymbol{V}$	Right eigenvectors of $\boldsymbol{X}$
N (**)	Number of the rows of $\boldsymbol{X}$
d	Number of the columns of $\boldsymbol{X}$
k (**)	Spectral cut-off
$\boldsymbol{X}^+$	Moore-Penrose pseudo-inverse of $\boldsymbol{X}$
$\|.\|_2$	Spectral norm
$\|.\|_{\text{Fr}}$	Frobenius norm
$\mathbf{0}_{n,m}$	Matrix of zeros in $\mathbb{R}^{n\times m}$

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

Alaoui, A. and M. W. Mahoney (2015). "Fast randomized kernel ridge regression with statistical guarantees". In: *Advances in Neural Information Processing Systems*, pp. 775–783.

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine (1999). "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays". In: *Proceedings of the National Academy of Sciences* 96.12, pp. 6745–6750.

Anari, N., S. O. Gharan, and A. Rezaei (2016). "Monte Carlo Markov chain algorithms for sampling strongly Rayleigh distributions and determinantal point processes". In: *Conference on Learning Theory*, pp. 103–115.

Baccelli, F., B. Błaszczyszyn, and M. Karray (2020). *Random Measures, Point Processes, and Stochastic Geometry*.

Bach, F. (2013). "Sharp analysis of low-rank kernel matrix approximations". In: *Conference on Learning Theory*, pp. 185–209.

Bach, F. (2017). "On the equivalence between kernel quadrature rules and random feature expansions". In: *The Journal of Machine Learning Research* 18.1, pp. 714–751.

Bakhvalov, NS (1959). "On approximate computation of integrals, Vestnik MGU, Ser". In: *Math. Mech. Astron. Phys. Chem* 4.3, p. 18.

Bardenet, R. and A. Hardy (2020). "Monte Carlo with determinantal point processes". In: *The Annals of Applied Probability* 30.1, pp. 368–417.

Baryshnikov, Y. (2001). "GUEs and queues". In: *Probability Theory and Related Fields* 119.2, pp. 256–274.

Beck, A. and M. Teboulle (2003). "Mirror descent and nonlinear projected subgradient methods for convex optimization". In: *Operations Research Letters* 31.3, pp. 167–175.

Belhadji, A., R. Bardenet, and P. Chainais (2019a). "Kernel quadrature with DPPs". In: *Advances in Neural Information Processing Systems 32*, pp. 12907–12917.

Belhadji, A., R. Bardenet, and P. Chainais (2019b). "Un processus ponctuel déterminantal pour la sélection d'attributs". In: *GRETSI 2019*.

Belhadji, A., R. Bardenet, and P. Chainais (2020a). "A determinantal point process for column subset selection". In: *Journal of Machine Learning Research* 21.197, pp. 1–62.

Belhadji, A., R. Bardenet, and P. Chainais (2020b). "Kernel interpolation with continuous volume sampling". In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 725–735.

Ben-Israel, A. (1992). "A volume associated with m x n matrices". In: *Linear Algebra and its Applications* 167, pp. 87 –111.

Ben-Tal, A. and M. Teboulle (1990). "A geometric property of the least squares solution of linear equations". In: *Linear algebra and its applications* 139, pp. 165–170.

Berlinet, A. and Ch. Thomas-Agnan (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Bilyk, D., V. N. Temlyakov, and R. Yu (2012). "The L_2 discrepancy of two-dimensional lattices". In: *Recent Advances in Harmonic Analysis and Applications*. Springer, pp. 63–77.

Björck, A. and G. H. Golub (1973). "Numerical methods for computing angles between linear subspaces". In: *Mathematics of computation* 27.123, pp. 579–594.

Bojanov, B.D. (1981). "Uniqueness of the optimal nodes of quadrature formulae". In: *Mathematics of computation* 36.154, pp. 525–546.

Bos, L. P. and S. De Marchi (2011). "On optimal points for interpolation by univariate exponential functions". In: *Dolomites Research Notes on Approximation* 4.1.

Bos, L. P. and U. Maier (2002). "On the asymptotics of Fekete-type points for univariate radial basis interpolation". In: *Journal of Approximation Theory* 119.2, pp. 252–270.

Boutsidis, C., M. W. Mahoney, and P. Drineas (2009). "An Improved Approximation Algorithm for the Column Subset Selection Problem". In: *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '09. New York, New York: Society for Industrial and Applied Mathematics, pp. 968–977.

Boutsidis, C., P. Drineas, and M. Magdon-Ismail (2011). "Near Optimal Column-Based Matrix Reconstruction". In: *Proceedings of the 2011 IEEE 52Nd Annual Symposium on Foundations of Computer Science*. FOCS '11. Washington, DC, USA: IEEE Computer Society, pp. 305–314.

Brezis, H. (2010). *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media.

Briol, F. X., C. J. Oates, M. Girolami, M. A. Osborne, D. Sejdinovic, et al. (2019). "Probabilistic integration: A role in statistical computation?" In: *Statistical Science* 34.1, pp. 1–22.

Briol, F.X., C. Oates, M. Girolami, and M.A. Osborne (2015). "Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees". In: *Advances in Neural Information Processing Systems*, pp. 1162–1170.

Brunel, V. E. (2018). "Learning signed determinantal point processes through the principal minor assignment problem". In: *Advances in Neural Information Processing Systems*, pp. 7365–7374.

Burden, R.L. and J.D. Faires (1997). *Numerical Analysis*. Mathematics Series. Brooks/Cole Publishing Company.

Caillol, J. M. (1981). "Exact results for a two-dimensional one-component plasma on a sphere". In: *Journal de Physique Lettres* 42.12, pp. 245–247.

Calandriello, D. (2017). "Efficient Sequential Learning in Structured and Constrained Environments". PhD thesis.

Calandriello, D., A. Lazaric, and M. Valko (2017). "Distributed adaptive sampling for kernel matrix approximation". In: *Artificial Intelligence and Statistics*, pp. 1421–1429.

Chen, Y., M. Welling, and A. Smola (2010). "Super-samples from Kernel Herding". In: *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. UAI'10. Catalina Island, CA: AUAI Press, pp. 109–116.

Chikuse, Y. (2012). *Statistics on special manifolds*. Vol. 174. Springer Science & Business Media.

Christoffel, E. B. (1877). "Sur une classe particuliere de fonctions entieres et de fractions continues". In: *Annali di Matematica Pura ed Applicata (1867-1897)* 8.1, pp. 1–10.

Coeurjolly, J. F., A. Mazoyer, and P. O. Amblard (2020). "Monte Carlo integration of non-differentiable functions on $[0,1]^\iota$, $\iota = 1,\dots,d$, using a single determinantal point pattern defined on $[0,1]^d$". In: *arXiv preprint arXiv:2003.10323*.

Cohen, M. B., Y. T. Lee, C. Musco, Ch. Musco, R. Peng, and A. Sidford (2015). "Uniform sampling for matrix approximation". In: *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ACM, pp. 181–190.

Cucker, F. and D.X. Zhou (2007). *Learning theory: an approximation theory viewpoint*. Vol. 24. Cambridge University Press.

Daley, D.J. and D. Vere-Jones (2007). *An introduction to the theory of point processes: Volume II: general theory and structure*. Springer Science & Business Media.

Daubechies, I. (1988). "Time-frequency localization operators: a geometric phase space approach". In: *IEEE Transactions on Information Theory* 34.4, pp. 605–612.

Daubechies, I. and T. Paul (1988). "Time-frequency localisation operators-a geometric phase space approach: II. The use of dilations". In: *Inverse problems* 4.3, p. 661.

Davis, Ph. J. and Ph. Rabinowitz (1984). "Methods of numerical integration. 1984". In: *Comput. Sci. Appl. Math*.

De Marchi, S. (2003). "On optimal center locations for radial basis function interpolation: computational aspects". In: *Rend. Splines Radial Basis Functions and Applications* 61.3, pp. 343–358.

De Marchi, S., R. Schaback, and H. Wendland (2005). "Near-optimal data-independent point locations for radial basis function interpolation". In: *Advances in Computational Mathematics* 23.3, pp. 317–330.

Dereziński, M. and M. K. Warmuth (2017). "Subsampling for ridge regression via regularized volume sampling". In: *arXiv preprint arXiv:1710.05110*.

Derezinski, M. and M. K. Warmuth (2017). "Unbiased estimates for linear regression via volume sampling". In: *Advances in Neural Information Processing Systems*, pp. 3084–3093.

Derezinski, M. and M. K. Warmuth (2018). "Reverse iterative volume sampling for linear regression". In: *The Journal of Machine Learning Research* 19.1, pp. 853–891.

Derezinski, M., M. K. Warmuth, and D. J. Hsu (2018). "Leveraged volume sampling for linear regression". In: *Advances in Neural Information Processing Systems*, pp. 2505–2514.

Deshpande, A. and L. Rademacher (2010). "Efficient volume sampling for row/column subset selection". In: *CoRR* abs/1004.4057.

Deshpande, A. and S. Vempala (2006). "Adaptive Sampling and Fast Low-rank Matrix Approximation". In: *Proceedings of the 9th International Conference on Approximation Algorithms for Combinatorial Optimization Problems, and 10th International Conference on Randomization and Computation*. APPROX'06/RANDOM'06. Barcelona, Spain: Springer-Verlag, pp. 292–303.

Deshpande, A., L. Rademacher, S. Vempala, and G. Wang (2006). "Matrix Approximation and Projective Clustering via Volume Sampling". In: *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*. SODA '06. Miami, Florida: Society for Industrial and Applied Mathematics, pp. 1117–1126.

Deutsch, F. (1995). "The angle between subspaces of a Hilbert space". In: *Approximation theory, wavelets and applications*. Springer, pp. 107–130.

Dhillon, I., R. Heath, M. Sustik, and J. Tropp (2005). "Generalized Finite Algorithms for Constructing Hermitian Matrices with Prescribed Diagonal and Spectrum". In: *SIAM Journal on Matrix Analysis and Applications* 27.1, pp. 61–71.

Diaconis, P. (1988). "Bayesian numerical analysis". In: *Statistical decision theory and related topics IV* 1, pp. 163–175.

Diaconis, P. and M. Shahshahani (1994). "On the eigenvalues of random matrices". In: *Journal of Applied Probability* 31.A, pp. 49–62.

Dick, J. (2008). "Walsh spaces containing smooth functions and quasi–Monte Carlo rules of arbitrary high order". In: *SIAM Journal on Numerical Analysis* 46.3, pp. 1519–1553.

Dick, J. and F. Pillichshammer (2010). *Digital nets and sequences: discrepancy theory and quasi–Monte Carlo integration*. Cambridge University Press.

Dick, J. and F. Pillichshammer (2014). "Discrepancy theory and quasi-Monte Carlo integration". In: *A Panorama of Discrepancy Theory*. Springer, pp. 539–619.

Drineas, P., A. Frieze, R. Kannan, S. Vempala, and V. Vinay (June 2004). "Clustering Large Graphs via the Singular Value Decomposition". In: *Mach. Learn.* 56.1-3, pp. 9–33.

Drineas, P., M. W. Mahoney, and S. Muthukrishnan (2006). "Sampling algorithms for l 2 regression and applications". In: *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*. Society for Industrial and Applied Mathematics, pp. 1127–1136.

Drineas, P., M. W. Mahoney, and S. Muthukrishnan (2007). *Relative-Error CUR Matrix Decompositions*.

Drineas, P., M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff (2012). "Fast approximation of matrix coherence and statistical leverage". In: *Journal of Machine Learning Research* 13.Dec, pp. 3475–3506.

Dumitriu, I. and A. Edelman (2002). "Matrix models for beta ensembles". In: *Journal of Mathematical Physics* 43.11, pp. 5830–5847.

Dumitriu, I. and A. Edelman (2005). "Eigenvalues of Hermite and Laguerre ensembles: large beta asymptotics". In: *Annales de l'IHP Probabilités et statistiques*. Vol. 41. 6, pp. 1083–1099.

Dyson, F. J. (1962). "Statistical theory of the energy levels of complex systems. I". In: *Journal of Mathematical Physics* 3.1, pp. 140–156.

Ehler, M., M. Gräf, and C. J. Oates (2019). "Optimal Monte Carlo integration on closed manifolds". In: *Statistics and Computing* 29.6, pp. 1203–1214.

Fickus, M., D. G. Mixon, and M. J. Poteet (2011). "Frame completions for optimally robust reconstruction". In:

Fickus, M., D. G. Mixon, M. J. Poteet, and N. Strawn (2013). "Constructing all self-adjoint matrices with prescribed spectrum and diagonal". In: *Advances in Computational Mathematics* 39.3-4, pp. 585–609.

Forrester, P.J, B. Jancovici, and J. Madore (1992). "The two-dimensional Coulomb gas on a sphere: exact results". In: *Journal of statistical physics* 69.1-2, pp. 179–192.

Gao, T., S. Z. Kovalsky, and I. Daubechies (2019). "Gaussian process landmarking on manifolds". In: *SIAM Journal on Mathematics of Data Science* 1.1, pp. 208–236.

Gauss, C. F. (1815). *Methodus nova integralium valores per approximationem inveniendi*. apvd Henricvm Dieterich.

Gautier, G., R. Bardenet, and M. Valko (2017). "Zonotope hit-and-run for efficient sampling from projection DPPs". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1223–1232.

Gautier, G., R. Bardenet, and M. Valko (2019). "DPPy: Sampling Determinantal Point Processes with Python". In: *Journal of Machine Learning Research – Machine Learning Open Source Software*.

Gautier, G., R. Bardenet, and M. Valko (2020). *Fast sampling from β-ensembles*.

Ginibre, J. (1965). "Statistical ensembles of complex, quaternion, and real matrices". In: *Journal of Mathematical Physics* 6.3, pp. 440–449.

Glasserman, P. (2013). *Monte Carlo methods in financial engineering*. Vol. 53. Springer Science & Business Media.

Gnewuch, M., A. Srivastav, and C. Winzen (2009). "Finding optimal volume subintervals with k points and calculating the star discrepancy are NP-hard problems". In: *Journal of Complexity* 25.2, pp. 115–127.

Golub, G. (June 1965). "Numerical Methods for Solving Linear Least Squares Problems". In: *Numer. Math.* 7.3, pp. 206–216.

Golub, G. H. and G. Meurant (2009). *Matrices, moments and quadrature with applications*. Vol. 30. Princeton University Press.

Golub, G. H. and C. F. Van Loan (1996). *Matrix Computations (3rd Ed.)* Baltimore, MD, USA: Johns Hopkins University Press.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al. (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring". In: *science* 286.5439, pp. 531–537.

Gu, M. and S. C. Eisenstat (July 1996). "Efficient Algorithms for Computing a Strong Rank-revealing QR Factorization". In: *SIAM J. Sci. Comput.* 17.4, pp. 848–869.

Guruswami, V. and A. K. Sinop (2012). "Optimal column-based low-rank matrix reconstruction". In: *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, pp. 1207–1214.

Halton, J.H. (1964). "Algorithm 247: Radical-inverse quasi-random point sequence". In: *Communications of the ACM* 7.12, pp. 701–702.

Heath, T. L. (2003). *A manual of Greek mathematics*. Courier Corporation.

Hickernell, F. J. (1996). "Quadrature error bounds with applications to lattice rules". In: *SIAM Journal on Numerical Analysis* 33.5, pp. 1995–2016.

Hickernell, F. J. (1998). "A generalized discrepancy and quadrature error bound". In: *Mathematics of computation* 67.221, pp. 299–322.

Hinrichs, A. and J. Oettershagen (2016). "Optimal point sets for quasi-Monte Carlo integration of bivariate periodic functions with bounded mixed derivatives". In: *Monte Carlo and Quasi-Monte Carlo Methods*. Springer, pp. 385–405.

Hlawka, E. (1961). "Funktionen von beschränkter variatiou in der theorie der gleichverteilung". In: *Annali di Matematica Pura ed Applicata* 54.1, pp. 325–333.

Holtz, M. (2008). *Sparse grid quadrature in high dimensions with applications in finance and insurance*. PhD Thesis, University of Bonn.

Horn, A. (1954). "Doubly stochastic matrices and the diagonal of a rotation matrix". In: *American Journal of Mathematics* 76.3, pp. 620–630.

Hough, J. B., M. Krishnapur, Y. Peres, and B. Virág (2006). "Determinantal processes and independence". In: *Probability surveys* 3, pp. 206–229.

Hough, J. B., M. Krishnapur, Y. Peres, and B. Virág (2009). *Zeros of Gaussian analytic functions and determinantal point processes*. Vol. 51. American Mathematical Soc.

Huszár, Ferenc and David Duvenaud (2012). "Optimally-weighted Herding is Bayesian Quadrature". In: *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*. UAI'12. AUAI Press, pp. 377–386.

Ipsen, I.C.F. and T. Wentworth (2014). "The effect of coherence on sampling from matrices with orthonormal columns, and preconditioned least squares problems". In: *SIAM Journal on Matrix Analysis and Applications* 35.4, pp. 1490–1520.

Jacobi, C. G. J. (1826). "Ueber Gauß neue Methode, die Werthe der Integrale näherungsweise zu finden." In: *Journal für die reine und angewandte Mathematik* 1826.1, pp. 301–308.

Johansson, K. (1997). "On random matrices from the compact classical groups". In: *Annals of mathematics*, pp. 519–545.

Johansson, K. (Oct. 2005). "Random matrices and determinantal processes". In: *ArXiv Mathematical Physics e-prints*.

Kanagawa, M., B. K. Sriperumbudur, and K. Fukumizu (2016). "Convergence guarantees for kernel-based quadrature rules in misspecified settings". In: *Advances in Neural Information Processing Systems*, pp. 3288–3296.

Karvonen, T. and S. Särkkä (2019). "Gaussian kernel quadrature at scaled Gauss-Hermite nodes". In: *BIT Numerical Mathematics*, pp. 1–26.

Karvonen, T., S. Särkkä, and K. Tanaka (2019). "Kernel-based interpolation at approximate Fekete points". In: *arXiv preprint arXiv:1912.07316*.

Khinchin, A. Ya. (1997). *Continued Fractions*. Dover books on mathematics. Dover Publications.

Killip, Rowan and Irina Nenciu (2004). "Matrix models for circular ensembles". In: *International Mathematics Research Notices* 2004.50, p. 2665.

Koksma, J.F. (1942). "Een algemeene stelling uit de theorie der gelijkmatige verdeeling modulo 1". In: *Mathematica B (Zutphen)* 11.7-11, p. 43.

Kress, Rainer (1999). *Linear Integral Equations*. Springer New York.

Krishnapur, M. et al. (2009). "From random matrices to random analytic functions". In: *The Annals of Probability* 37.1, pp. 314–346.

Kulesza, A. and B. Taskar (2012). "Determinantal point processes for machine learning". In:

Lacoste-Julien, S., F. Lindsten, and F. Bach (2015). "Sequential kernel herding: Frank-Wolfe optimization for particle filtering". In: *arXiv preprint arXiv:1501.02056*.

Landau, H. J. and H. O. Pollak (1961). "Prolate spheroidal wave functions, Fourier analysis and uncertainty—II". In: *Bell System Technical Journal* 40.1, pp. 65–84.

Larkin, FM (1972). "Gaussian measure in Hilbert space and applications in numerical analysis". In: *The Rocky Mountain Journal of Mathematics*, pp. 379–421.

Launay, C., A. Desolneux, and B. Galerne (2020a). "Determinantal point processes for image processing". In: *to appear in SIAM Journal on Imaging Sciences*.

Launay, C., B. Galerne, and A. Desolneux (2020b). "Exact Sampling of Determinantal Point Processes without Eigendecomposition". In: *to appear in the Journal of Applied Probability, JAP 57.4* Dec.

Lavancier, F., J. Møller, and E. Rubak (2015). "Determinantal point process models and statistical inference". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.4, pp. 853–877.

Lax, P.D. (2002). *Functional analysis*. Pure and applied mathematics. Wiley.

Li, C., S. Jegelka, and S. Sra (2017a). "Polynomial time algorithms for dual volume sampling". In: *Advances in Neural Information Processing Systems*, pp. 5038–5047.

Li, J., K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu (2017b). "Feature selection: A data perspective". In: *ACM Computing Surveys (CSUR)* 50.6, p. 94.

Lorch, L. (1983). "Alternative proof of a sharpened form of Bernstein's inequality for Legendre polynomials". In: *Applicable Analysis* 14.3, pp. 237–240.

Ma, P., M. W. Mahoney, and B. Yu (2015). "A statistical perspective on algorithmic leveraging". In: *The Journal of Machine Learning Research* 16.1, pp. 861–911.

Macchi, O (Mar. 1975). "The coincidence approach to stochastic point processes". In: 7, pp. 83–122.

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

Marcus, A. W., D. A. Spielman, and N. Srivastava (2015). "Interlacing families II: Mixed characteristic polynomials and the Kadison—Singer problem". In: *Annals of Mathematics*, pp. 327–350.

Marshall, A. W., I. Olkin, and B. C. Arnold (2011). *Inequalities: Theory of Majorization and its Applications*. Second. Vol. 143. Springer.

Mercer, J. (1909). "Functions of positive and negative type, and their connection with the theory of integral equations". In: *Philosophical Transactions of the Royal Society, London* 209, pp. 415–446.

Metropolis, N. and S. Ulam (1949). "The Monte Carlo method". In: *Journal of the American statistical association* 44.247, pp. 335–341.

Miao, J. and A. Ben-Israel (1992). "On principal angles between subspaces in $\mathbb{R}^n$". In: *Linear Algebra and its Applications* 171, pp. 81 –98.

Minsker, S. (2017). "On some extensions of bernstein's inequality for self-adjoint operators". In: *Statistics & Probability Letters* 127, pp. 111–119.

Miranian, L. (2004). "Slepian functions on the sphere, generalized Gaussian quadrature rule". In: *Inverse Problems* 20.3, p. 877.

Mor-Yosef, L. and H. Avron (2019). "Sketching for principal component regression". In: *SIAM Journal on Matrix Analysis and Applications* 40.2, pp. 454–485.

Muandet, K., K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. (2017). "Kernel mean embedding of distributions: A review and beyond". In: *Foundations and Trends® in Machine Learning* 10.1-2, pp. 1–141.

Nashed, M. Z. and G. G. Walter (1991). "General sampling theorems for functions in reproducing kernel Hilbert spaces". In: *Mathematics of Control, Signals and Systems* 4.4, p. 363.

Nikolov, A., M. Singh, and U. T. Ta. (2019). "Proportional volume sampling and approximation algorithms for A-optimal design". In: *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, pp. 1369–1386.

Novak, E. (2016). "Some results on the complexity of numerical integration". In: *Monte Carlo and Quasi-Monte Carlo Methods*. Springer, pp. 161–183.

Novak, E., M. Ullrich, and H. Woźniakowski (2015). "Complexity of oscillatory integration for univariate Sobolev spaces". In: *Journal of Complexity* 31.1, pp. 15–41.

Oates, C. J., M. Girolami, and N. Chopin (2014). "Control functionals for Monte Carlo integration". In: *arXiv preprint arXiv:1410.2392*.

Oettershagen, J. (2017). *Construction of optimal cubature algorithms with applications to econometrics and uncertainty quantification*. PhD Thesis, University of Bonn.

Papailiopoulos, D., A. Kyrillidis, and C. Boutsidis (2014). "Provable Deterministic Leverage Score Sampling". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, New York, USA: ACM, pp. 997–1006.

Pinkus, A. (2012). *N-widths in Approximation Theory*. Vol. 7. Springer Science & Business Media.

Puy, G., N. Tremblay, R. Gribonval, and P. Vandergheynst (2018). "Random sampling of bandlimited signals on graphs". In: *Applied and Computational Harmonic Analysis* 44.2, pp. 446–475.

Raskutti, G. and M. W. Mahoney (2016). "A statistical perspective on randomized sketching for ordinary least-squares". In: *The Journal of Machine Learning Research* 17.1, pp. 7508–7538.

Rasmussen, Carl Edward (2003). "Gaussian processes in machine learning". In: *Summer School on Machine Learning*. Springer, pp. 63–71.

Rezaei, A. and S. O. Gharan (2019). "A Polynomial Time MCMC Method for Sampling from Continuous Determinantal Point Processes". In: *International Conference on Machine Learning*, pp. 5438–5447.

Robert, C. P. and G. Casella (2004). *Monte Carlo statistical methods*. Springer.

Robins, G. and C. Shute (1987). "The Rhind Mathematical Papyrus. An Ancient Egyptian Text." In:

Rudelson, M. and R. Vershynin (2007). "Sampling from large matrices: An approach through geometric functional analysis". In: *Journal of the ACM (JACM)* 54.4, p. 21.

Rudin, W. (1991). *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill.

Santin, G. and B. Haasdonk (2017). "Convergence rate of the data-independent P-greedy algorithm in kernel-based approximation". In: *Dolomites Research Notes on Approximation* 10.Special_Issue.

Schaback, R. (2005). "Multivariate interpolation by polynomials and radial basis functions". In: *Constructive Approximation* 21.3, pp. 293–317.

Schaback, R. and H. Wendland (2006). "Kernel techniques: from machine learning to meshless methods". In: *Acta numerica* 15, pp. 543–639.

Schneider, R. and W. Weil (2008). *Stochastic and integral geometry*. Springer Science & Business Media.

Schölkopf, B. and A.J. Smola (2018). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series.

Seip, K. (1991). "Reproducing formulas and double orthogonality in Bargmann and Bergman spaces". In: *SIAM journal on mathematical analysis* 22.3, pp. 856–876.

Shadrin, A. (2004). "Twelve proofs of the Markov inequality". In: *Approximation theory: a volume dedicated to Borislav Bojanov*, pp. 233–298.

Shannon, C. E. (1948). "A mathematical theory of communication". In: *Bell system technical journal* 27.3, pp. 379–423.

Shawe-Taylor, J. and N. Cristianini (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Shohat, J. A. (1929). "On a certain formula of mechanical quadratures with non-equidistant ordinates". In: *Transactions of the American Mathematical Society*, pp. 448–463.

Simon, B. (2010). *Szegő's theorem and its descendants: spectral theory for L2 perturbations of orthogonal polynomials*. Vol. 6. Princeton university press.

Slawski, M. (2018). "On principal components regression, random projections, and column subsampling". In: *Electronic Journal of Statistics* 12.2, pp. 3673–3712.

Slepian, D. and H. O. Pollak (1961). "Prolate spheroidal wave functions, Fourier analysis and uncertainty—I". In: *Bell System Technical Journal* 40.1, pp. 43–63.

Smola, A., A. Gretton, L. Song, and B. Schölkopf (2007). "A Hilbert space embedding for distributions". In: *International Conference on Algorithmic Learning Theory*. Springer, pp. 13–31.

Smola, A. J. and B. Schölkopf (2000). "Sparse greedy matrix approximation for machine learning". In:

Smola, A. J., Z. L. Ovari, and R. C. Williamson (2001). "Regularization with dot-product kernels". In: *Advances in neural information processing systems*, pp. 308–314.

Smolyak, S.A. (1963). "Quadrature and interpolation formulas for tensor products of certain classes of functions". In: *Doklady Akademii Nauk*. Vol. 148. 5. Russian Academy of Sciences, pp. 1042–1045.

Soshnikov, A. (Oct. 2000). "Determinantal random point fields". In: *Russian Mathematical Surveys* 55, pp. 923–975.

Spielman, D. A. and N. Srivastava (2011). "Graph sparsification by effective resistances". In: *SIAM Journal on Computing* 40.6, pp. 1913–1926.

Spielman, D. A. and S. Teng (2004). "Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems". In: *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pp. 81–90.

Steele, J.M. (2004). *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. New York, NY, USA: Cambridge University Press.

Steinwart, I. and A. Christmann (2008). *Support Vector Machines*. 1st. Springer Publishing Company, Incorporated.

Steinwart, I. and C. Scovel (2012). "Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs". In: *Constructive Approximation* 35.3, pp. 363–417.

Sun, H. (2005). "Mercer theorem for RKHS on noncompact sets". In: *Journal of Complexity* 21.3, pp. 337–349.

Szegő, G. (1939). *Orthogonal polynomials*. Vol. 23. American Mathematical Soc.

Tanaka, K. (2019). "Generation of point sets by convex optimization for interpolation in reproducing kernel Hilbert spaces". In: *Numerical Algorithms*, pp. 1–31.

Tao, T. (2012). *Topics in random matrix theory*. Vol. 132. American Mathematical Soc.

Tremblay, N., P.O. Amblard, and S. Barthelmé (2017). "Graph sampling with determinantal processes". In: *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1674–1678.

Tremblay, N., S. Barthelmé, and P.-O. Amblard (2018). "Optimized Algorithms to Sample Determinantal Point Processes". In: *arXiv preprint arXiv:1802.08471*.

Tropp, J. A. (2015). "An introduction to matrix concentration inequalities". In: *arXiv preprint arXiv:1501.01571*.

Wahba, Grace (1990). *Spline Models for Observational Data*. Vol. 59. SIAM.

Warnock, T. T. (1972). "Computational investigations of low-discrepancy point sets". In: *Applications of number theory to numerical analysis*. Elsevier, pp. 319–343.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Wendland, H. (2004). *Scattered Data Approximation*. Cambridge University Press.

Weyl, H. (1946). *The Classical Groups: Their Invariants and Representations*. Vol. 1. Princeton University Press.

Widom, H. (1963). "Asymptotic behavior of the eigenvalues of certain integral equations. I". In: *Transactions of the American Mathematical Society* 109.2, pp. 278–295.

Widom, H. (1964). "Asymptotic behavior of the eigenvalues of certain integral equations. II". In: *Archive for Rational Mechanics and Analysis* 17.3, pp. 215–229.

Williams, C.K.I. and M. Seeger (2001). "Using the Nyström method to speed up kernel machines". In: *Advances in neural information processing systems*, pp. 682–688.

Wynne, G., F. X. Briol, and M. Girolami (2020). "Convergence guarantees for Gaussian process approximations under several observation models". In: *arXiv preprint arXiv:2001.10818*.

Xiao, H., V. Rokhlin, and N. Yarvin (2001). "Prolate spheroidal wavefunctions, quadrature and interpolation". In: *Inverse problems* 17.4, p. 805.

Xu, Y. (1994a). "Block Jacobi matrices and zeros of multivariate orthogonal polynomials". In: *Transactions of the American Mathematical Society* 342.2, pp. 855–866.

Xu, Y. (1994b). *Common Zeros of Polynominals in Several Variables and Higher Dimensional Quadrature*. Vol. 312. CRC Press.

Yao, K. (1967). "Applications of reproducing kernel Hilbert spaces-bandlimited signal models". In: *Information and Control* 11.4, pp. 429–444.

Zhu, P. and A. V. Knyazev (2013). "Angles between subspaces and their tangents". In: *Journal of Numerical Mathematics* 21.4, pp. 325–340.

Zouzias, Anastasios. "Euclidean Embeddings that Preserve Volumes". MA thesis.

Échantillonnage des sous-espaces à l'aide des processus ponctuels déterminantaux

Les processus ponctuels déterminantaux (DPP) sont des modèles probabilistes impliquant une répulsion entre les points. Ces modèles ont été étudiés dans différents domaines allant de l'étude des matrices aléatoires à l'optique quantique, en passant par le traitement d'image, l'apprentissage automatique et plus récemment les quadratures. Dans cette thèse, on étudie l'échantillonnage de sous-espaces à l'aide des DPP. Ce problème se trouve à l'intersection de trois branches de la théorie de l'approximation : la sous-sélection dans les ensembles discrets, la quadrature à noyau et l'interpolation à noyau. On étudie ces questions classiques à travers une nouvelle interprétation des DPP : un DPP est une façon naturelle de définir un sous-espace aléatoire ayant de bonnes propriétés. En plus de donner une analyse unifiée de l'intégration et de l'interpolation sous les DPPs, cette nouvelle approche permet de démontrer les garanties théoriques de plusieurs algorithmes impliquant des DPPs.
Mots-clés : Processus ponctuels déterminantaux, méthodes Monte Carlo, interpolation, quadrature, méthodes à noyau

Subspace sampling using determinantal point processes

Determinantal point processes are probabilistic models of repulsion. These models were studied in various fields: random matrices, quantum optics, spatial statistics, image processing, machine learning, and recently numerical integration. In this thesis, we study subspace sampling using determinantal point processes. This problem takes place within the intersection of three sub-domains of approximation theory: subset selection, kernel quadrature, and kernel interpolation. We study these classical topics, through a new interpretation of these probabilistic models: a determinantal point process is a natural way to define a random subspace. Besides giving a unified analysis to numerical integration and interpolation under determinantal point processes, this new perspective allows to work out the theoretical guarantees of several approximation algorithms and to prove their optimality in some settings.
Keywords: Determinantal point processes, Monte Carlo methods, interpolation, quadrature, kernel methods